

SUPERVISED LEARNING METHOD OF DIABETES PREDICTION

Submitted in partial fulfillment of the requirements
for the award of
Bachelor of Engineering degree in Computer Science and Engineering

by

Vemireddy Praneeth Kumar Reddy (38110421)
Yalamarthi Sivva (38110659)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade "A" by NAAC

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI - 600 119**

MARCH – 2022



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600119

www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Vemireddy Praneeth Kumar Reddy(38110421)** and **Yalamarathi Sivva(38110659)** who carried out the project entitled "**SUPERVISED LEARNING METHOD OF DIABETES PREDICTION**" under my supervision from November 2021 to March 2022.

Internal Guide

M.S.Roobini, M.E., Ph. D.,

Head of the Department

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I **Vemireddy Praneeth Kumar Reddy(38110421)** and **Yalamarthi Sivva(38110659)** hereby declare that the Project Report entitled “**SUPERVISED LEARNING METHOD OF DIABETES PREDICTION**” done by us under the guidance of **M.S.Roobini, M.E., Ph. D** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in 2018-2022.

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E.,Ph.D., Dean**, School of Computing **Dr.S.Vigneshwari M.E., Ph.D.** and **Dr.L.Lakshmanan M.E., Ph.D.** , Heads of the Department of Computer Science and Engineering for providing us necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and a deep sense of gratitude to my Project Guide **M.S. Roobini, M.E., Ph. D.**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

- Nowadays, e-Health service has become a booming area, which refers to computer-based health care and information delivery to improve health service locally, regionally and worldwide. An effective disease risk prediction model by analyzing electronic health data benefits not only to care a patient but also to provide services through the corresponding data-driven e-Health systems. In this paper, we particularly focus on predicting and analyzing diabetes mellitus, an increasingly prevalent chronic disease that refers to a group of metabolic disorders characterized by a high blood sugar level over a prolonged period of time. K-Nearest Neighbor (KNN) is one of the most popular and simplest machine learning techniques to build such a disease risk prediction model utilizing relevant health data. In order to achieve our goal, we present an optimal K-Nearest Neighbor (OPT-KNN) learning based prediction model based on patient's habitual attributes in various dimensions. This approach determines the optimal number of neighbors with low error rate for providing better prediction outcome in the resultant model. The effectiveness of this machine learning e-Health model is examined by conducting experiments on the real-world diabetes mellitus data collected from medical hospitals.

TABLE OF CONTENTS

Chapter No.	TITLE	Page No.
	ABSTRACT	v
	LIST OF FIGURES	vii
1	INTRODUCTION	1
	1.1. OUTLINE OF THE PROJECT	1
	1.2 MACHINE LEARNING	2
	1.3 SCOPE AND OBJECTIVE	2
2	LITERATURE SURVEY	3
	2.1. RELATED WORK	3
3	METHODOLOGY	6
	3.1. EXISTING SYSTEM	6
	3.2. PROPOSED SYSTEM	6
	3.3 ALGORITHMS USED	6
	3.3.1. K-Nearest Neighbor(KNN)	6
	3.4 SYSTEM ARCHITECTURE	8
	3.5 SYSTEM REQUIREMENTS	8
	3.6 MODULES	9
	3.7 UML DIAGRAMS	10
	3.8 LANGUAGES USED	12
	3.9 REQUIREMENT ANALYSIS	20
4	RESULTS AND DISCUSSION	27
	4.1. WORKING	27
5	CONCLUSION	29
	5.1. CONCLUSION	29
	5.2. REFERENCES	29
	5.3. APPENDIX	30
	A. SCREENSHOTS	30
	B. PLAGIARISM REPORT	33
	C. JOURNAL PAPER	34

LIST OF FIGURES

Figure No.	Figure Name	Page No.
3.4	System Architecture	8
3.7.1	Data Flow Diagram	11
B.1	Input Page	30
B.2	Output Page	31
B.3	Output Page	31
B.4	Accuracy Graph	32

CHAPTER – 1

INTRODUCTION

1.1 OUTLINE OF THE PROJECT

Diabetes may be a set of metabolic problems known by high blood sugar levels over a protracted period of our time. Diabetes is outlined as a bunch of metabolic disorders in the main caused by abnormal insulin secretion and/or action. Symptoms of high aldohexose incorporate excessive voiding, continually feeling thirsty and enlarged hunger. If not treated on time, diabetes will cause serious health problems in a person like diabetic acidosis, hyperosmolar hyperglycemic state, or maybe result in death. This could result in time period complications as well as vas upset, brain stroke, failure, ulcers within the foot, and eye complications etc. Diabetes is caused once the duct gland within the body is unable to come up with insulin in enough amounts or once the cells and tissues within the body fail to utilize the insulin created. Diabetes exists in 3 forms: Diabetes Mellitus Type-1 is characterized by duct gland generating insulin but what's needed by the body, a condition conjointly referred to as "insulin-subordinate diabetes mellitus"(IDDM). Folks littered with type-1 DM need external insulin indefinite quantity to form up for the less insulin created by the duct gland. Diabetes Mellitus Type-2 is marked by the body resisting insulin because the body cells react otherwise to insulin than they traditional would. This could ultimately result in no insulin within the body. This can be otherwise referred to as "non-insulin subordinate diabetes mellitus"(NIDDM) or "adult beginning diabetes". This sort of diabetes is often found in folks with high BMI or people who lead associate degree inactive manner. Gestational diabetes is that the third principle structure that's ascertained throughout physiological state. Generally, for a traditional person, aldohexose levels vary from seventy to ninety-nine milligrams per deciliter. An individual is taken into account diabetic providing the fast aldohexose level is found to be over 126 mg/dL. Within the practice, an individual having an aldohexose concentration of a hundred to one hundred twenty-five mg/dL is taken into account as pre-diabetic.

1.2 MACHINE LEARNING

Machine learning could be a subfield of computer science (AI). The goal of machine learning typically is to know the structure information of knowledge of information and match that data into models which will be understood and used by folks. Although machine learning could be a field inside technology, it differs from ancient process approaches.

In ancient computing, algorithms are sets of expressly programmed directions employed by computers to calculate or downside solve. Machine learning algorithms instead give computers to coach on knowledge inputs and use applied math analysis so as to output values that fall inside a particular vary. thanks to this, machine learning facilitates computers in building models from sample knowledge so as to modify decision-making processes supported knowledge inputs.

1.3 SCOPE AND OBJECTIVE

The system should be useful in many e-commercial websites for maintaining the security and reliability of customers and people online

The system should be useful in preventing online frauds leading to leakage of important and private user data

The scope of using Machine Language over other Traditional Detecting Methods
Objectives:

Diabetes mellitus (DM) is one of the most prevalent chronic non-communicable diseases (NCD) around the world; about 90% of the patients who have diabetes suffer from Type 2 DM (T2DM)

The risk of developing T2DM is strongly associated with many predispositions, behavioral, and environmental risk factors and also genetic factors

About 90% of patients who have diabetes suffer from Type 2 DM (T2DM)

Many studies suggest using the significant role of Incrnas to improve the diagnosis of T2DM.

Machine learning is the techniques are tools that can improve the analysis and interpretation or extraction of knowledge from the data

CHAPTER – 2

LITERATURE SURVEY

2.1 RELATED WORK

prediction of Diabetes using Classification Algorithms

Diabetes isn't a hereditary disorder however heterogeneous group of disorder which could ultimately result in an boom of glucose within the blood and lack of glucose inside the urine. Diabetes is typically resulting from genetics, way of life and surroundings. Eating an dangerous weight loss plan, being overweight play role in developing the diabetes. High blood sugar tiers can also result in kidney diseases, coronary heart illnesses. The excess of sugar in the blood can harm the tiny blood vessels in your frame. Signs of diabetes are blurry imaginative and prescient , extreme hunger, unusual weight reduction, common urination and thirsty. In this paper, parameters used within the facts set to locate the diabetes are Glucose, Blood pressure, pores and skin thickness, Insulin, Age. Huge volumes of statistics units are generated by health care industries. Those facts sets is a collection of patient information about the diabetes from the hospitals. Big records analytics is the processing which it examines the information units and exhibits the hidden information. Pima Indians Diabetes Database (PIDD), this dataset is taken from the national Institute of Diabetes and Digestive diseases. The objective of the dataset is to predict whether or not the patient has diabetes or not, primarily based on diagnostic measurements in the dataset. Several constraints were taken from the massive database.

A Novel Technique To Predict Diabetic Disease Using Data Mining Classification Techniques

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. The software of data mining is an analytical tool for analyzing data. Data mining has become a main strategy in

many industries to improve outputs and decrease costs. Now days in healthcare management this field will become very useful. Data mining techniques has became great potential for the healthcare industry to predict health deceases by using systematic data and analytics to identify inefficiencies and best practices that improve care and reduce costs. These techniques are fast in nature and take less time for the prediction system to improve the diabetic decease with more accuracy. In this paper we are applying the various classification techniques over diabetic mellitus decease dataset for the prediction of decease and non decease person. The diabetic database is preprocessed to make the mining process more efficient. The preprocessed data is used to predict using classification algorithms like Discriminant analysis, KNN, Naïve Bayes and Support vector machine. These classifiers can be efficiently used in bioinformatics problem. We are analyzing the various classification techniques like Discriminant analysis, KNN, Naïve Bayes and Support vector machine with linear and RBF kernel function and showing their accuracy.

Review on Prediction of Diabetes using Data Mining Technique

Diabetes mellitus is one of the world's major diseases. Millions of people are affected by the disease. The risk of diabetes is increasing day by day and is found mostly in women than men. The diagnosis of diabetes is a tedious process. So with improvement in science and technology it is made easy to predict the disease. The purpose is to diagnose whether the person is affected by diabetes or not using K Nearest Neighbor classification technique. The diabetes dataset is a taken as the training data and the details of the patient are taken as testing data. The training data are classified by using the KNN classifier and secondly the target data is predicted. KNN algorithm used here would be more efficient for both classification and prediction. The results are analyzed with different values for the parameter k.

A Prediction Technique in Data Mining for Diabetes Mellitus

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. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health. In this study, we used decision tree, random forest and neural network to predict diabetes mellitus. The dataset is the hospital physical examination data in Luzhou, China. It contains 14 attributes. In this study, five-fold cross validation was used to examine the models. In order to verify the universal applicability of the methods, we chose some methods that have the better performance to conduct independent test experiments. We randomly selected 68994 healthy people and diabetic patients' data, respectively as training set. Due to the data unbalance, we randomly extracted 5 times data. And the result is the average of these five experiments. In this study, we used principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) to reduce the dimensionality. The results showed that prediction with random forest could reach the highest accuracy ($ACC = 0.8084$) when all the attributes were used.

CHAPTER 3

METHODOLOGY

3.1 EXISTING SYSTEM

The existing system was taking in order to meets the demands of this system and solve the problems of the existing system by implementing the naïve bayes classifier.

DISADVANTAGES OF EXISTING SYSTEM

- The system is not fully automated, it needs data from user for full diagnosis

3.2 PROPOSED SYSTEM

The proposed diabetes prediction system has two main stages that work together to achieve the desired results. The first stage of the proposed system is the data preparation, and the second one is the classification. However, the input into the system is the dataset and the output will be one class which represent the healthy or the diabetic. The implementation of this new system will help to reduce the stressful process, the result of the experiment shows that the proposed system has a better prediction in terms of accuracy. We have applied K- Nearest neighbor algorithm on the training and test sample data and obtained results for different values of K which is number of nearest neighbors.

ADVANTAGES OF PROPOSED SYSTEM

- User can diagnose their diabetes and get instant result.
- K- Nearest neighbor Algorithm is a fast, highly scalable algorithm.

3.3 ALGORITHMS USED

3.3.1 DECISION TREE CLASSIFIER

The k-nearest neighbor's is a ML algorithm is the non-parametric method proposed by Thomas Cover used for Regression and Classification. This

algorithm is mainly used for the classification of problems in the industry. KNN algorithm is a type of instance-based learning method. This algorithm relies on the distance for objects classification, training data normalizing to the improve its accuracy dramatically. The neighbors are derived from the set of things for which classes or object property values are known. It can be thought of as a training set for the algorithm, although no explicit training steps are required (remove)Quite possibly the most impressive and famous calculation. The choice tree calculation is beneath the control level of the calculation. It works in two ways, similar to yield. Input:

STEP 1: BEGIN

STEP 2: Input: $D = \{(x_1, c_1), \dots, (x_N, c_N)\}$.

STEP 3: $x = (x_1 \dots x_n)$ new instance to be classified

STEP 4 FOR each labelled instance (x_i, c_i) calculate $d(x_i, x)$

STEP 5: Order $d(x_i, x)$ from lowest to highest, $(i = 1 \dots N)$

STEP 6: Select the K nearest instances to x : D_{Kx} .

STEP 7: Assign to x the most frequent class in D_{Kx}

STEP 8: END

3.4 SYSTEM ARCHITECTURE

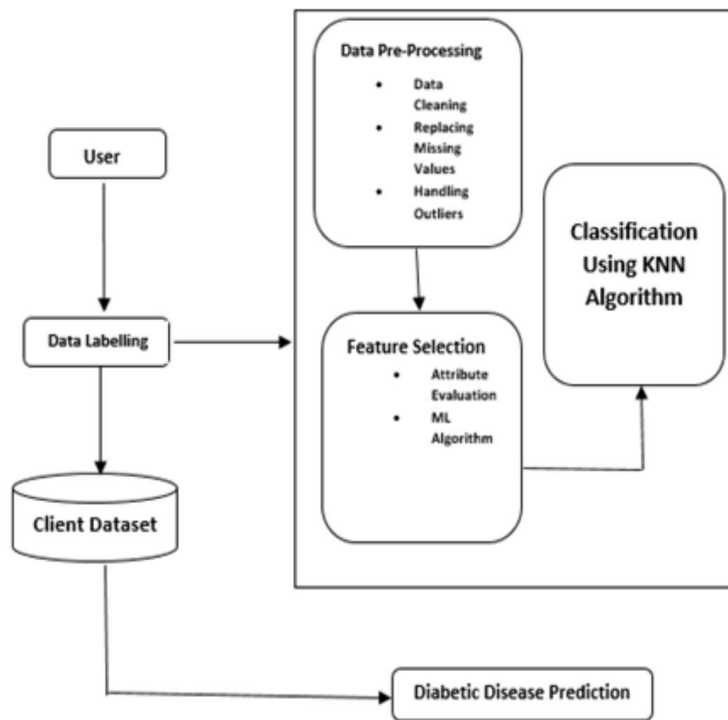


Fig:3.4 :System Architecture

3.5 SYSTEM REQUIREMENTS

Hardware Requirements:

System	: Pentium Dual Core.
Monitor	: 15'' LED
Input Devices	: Keyboard, Mouse.
Ram	: 4.00 GB
Hard Disk	: 120 GB

Software Requirements:

Operating system	: Windows 7,
Tool	: python 3.6.1
Language	: Python 3

3.6 MODULES

- Data Constraints
- Train Dataset and Test Dataset
- Pre-processing of data
- Feature Extraction

Data Constraints

Data is a collection global dataset. IN this system use Pima Indian data set is used for training a model.

Train Dataset and Test Dataset

The training data is a initial set of data which is used to understand the program. This is the one in which we have to train the model first because to set the feature and this data is available on system. This data is used to teach the machine for do different actions. It is the data in which model can learn with algorithm to teach the model and doing work automatic. Testing data is the input given to a software. It shows the data affects when the execution of the module that specifying and this is basically used for testing.

Pre-processing of data

Data preprocessing is a process in which that is actual use for converting the basic data into the clean data set. It is the step in which the data transform or an encode to the state that the machine can be easily parse. The major task of data preprocessing in learning process is to remove the unwanted data and filling the missed value. So that it help to machine can be trained easily.

Feature Extraction

Feature Extraction is the method in which it used for alter the key data for features of outcomes. This, trait square is used to compute the characteristics of designs given that facilitate in different amid the class of key pattern details. This method

involving to decrease the counts of resource required to describe the huge set of data. Feature extraction is an attribute reduction process. This is also used to increasing the speed and effectiveness of supervised learning.

3.7 UML DIAGRAMS

UML is simply another graphical representation of a common semantic model. UML provides a comprehensive notation for the full lifecycle of object-oriented development.

ADVANTAGES

- To represent complete systems (instead of only the software portion) using object oriented concepts
- To establish an explicit coupling between concepts and executable code
- To take into account the scaling factors that are inherent to complex and critical systems
- To creating a modeling language usable by both humans and machines
- UML defines several models for representing systems
- The class model captures the static structure
- The state model expresses the dynamic behavior of objects
- The use case model describes the requirements of the user
- The interaction model represents the scenarios and messages flows
- The implementation model shows the work units
- The deployment model provides details that pertain to process allocation

➤ DATA FLOW DIAGRAM

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.

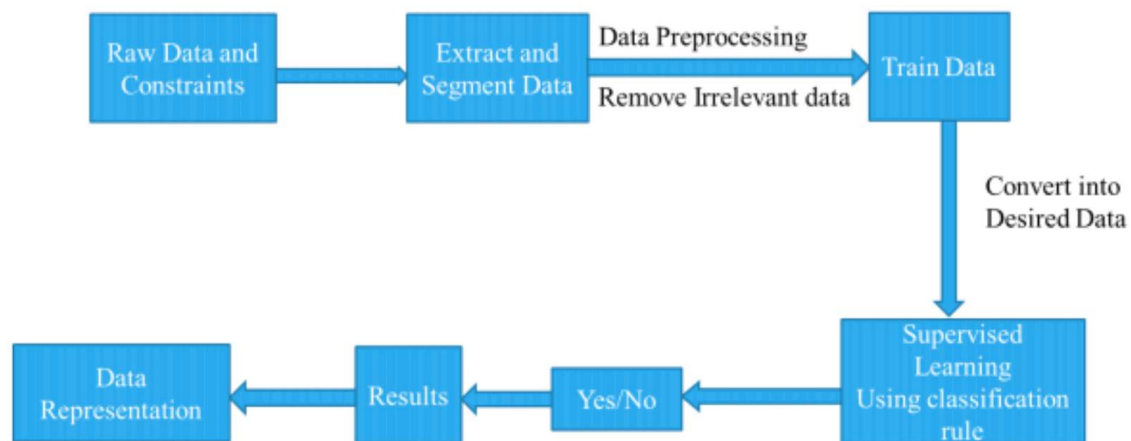
2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

3. DFD shows how the information moves through the system and how it

is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.

4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

LEVEL-0



DFD-LEVEL-1



DFD-LEVEL-2

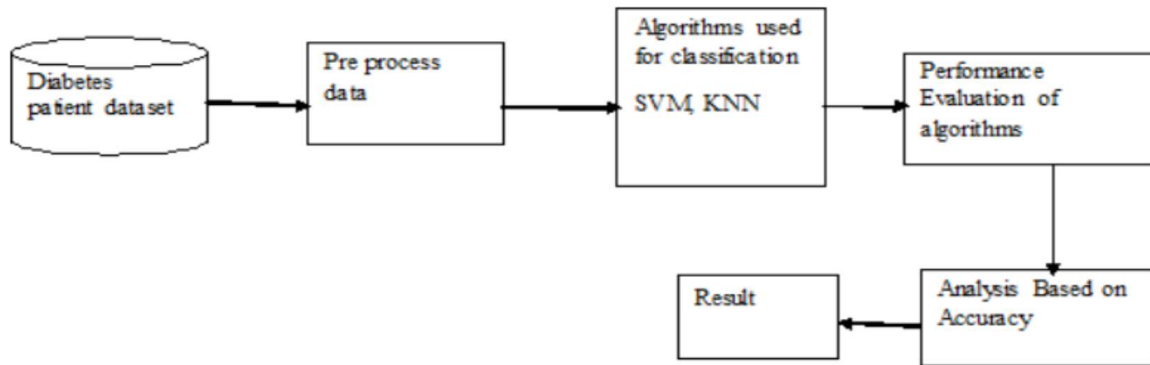


Fig:3.7.1 :Data Flow diagram

3.8 LANGUAGES USED

Python:

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

- **Python is Interpreted** – Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
- **Python is Interactive** – You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
- **Python is Object-Oriented** – Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
- **Python is a Beginner's Language** – Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

History of Python

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands.

Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk, and Unix shell and other scripting languages.

Python is copyrighted. Like Perl, Python source code is now available under the GNU General Public License (GPL).

Python is now maintained by a core development team at the institute, although Guido van Rossum still holds a vital role in directing its progress.

Python Features

Python's features include –

- **Easy-to-learn** – Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- **Easy-to-read** – Python code is more clearly defined and visible to the eyes.
- **Easy-to-maintain** – Python's source code is fairly easy-to-maintain.
- **A broad standard library** – Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
- **Interactive Mode** – Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- **Portable** – Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
- **Extendable** – You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- **Databases** – Python provides interfaces to all major commercial databases.
- **GUI Programming** – Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- **Scalable** – Python provides a better structure and support for large programs than shell scripting.

Apart from the above-mentioned features, Python has a big list of good features, few are listed below –

- It supports functional and structured programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.
- It supports automatic garbage collection.
- It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

Python is available on a wide variety of platforms including Linux and Mac OS X. Let's understand how to set up our Python environment.

Getting Python

The most up-to-date and current source code, binaries, documentation, news, etc., is available on the official website of Python <https://www.python.org>.

Windows Installation

Here are the steps to install Python on Windows machine.

- Open a Web browser and go to <https://www.python.org/downloads/>.
- Follow the link for the Windows installer python-XYZ.msifile where XYZ is the version you need to install.
- To use this installer python-XYZ.msi, the Windows system must support Microsoft Installer 2.0. Save the installer file to your local machine and then run it to find out if your machine supports MSI.
- Run the downloaded file. This brings up the Python install wizard, which is really easy to use. Just accept the default settings, wait until the install is finished, and you are done.

The Python language has many similarities to Perl, C, and Java. However, there are some definite differences between the languages.

First Python Program

Let us execute programs in different modes of programming.

Interactive Mode Programming

Invoking the interpreter without passing a script file as a parameter brings up the following prompt –

```
$ python

Python2.4.3(#1,Nov112010,13:34:43)

[GCC 4.1.220080704(RedHat4.1.2-48)] on linux2

Type"help","copyright","credits"or"license"for more information.

>>>
```

Type the following text at the Python prompt and press the Enter –

```
>>>print"Hello, Python!"
```

If you are running new version of Python, then you would need to use print statement with parenthesis as in **print ('Hello, Python!')**;. However in Python version 2.4.3, this produces the following result –

```
Hello, Python!
```

Script Mode Programming

Invoking the interpreter with a script parameter begins execution of the script and continues until the script is finished. When the script is finished, the interpreter is no longer active.

Let us write a simple Python program in a script. Python files have extension **.py**. Type the following source code in a test.py file –

```
print"Hello, Python!"
```

We assume that you have Python interpreter set in PATH variable. Now, try to

run this program as follows –

```
$ python test.py
```

This produces the following result –

```
Hello, Python!
```

Flask Framework:

Flask is a web application framework written in Python. Armin Ronacher, who leads an international group of Python enthusiasts named Pocco, develops it. Flask is based on Werkzeug WSGI toolkit and Jinja2 template engine. Both are Pocco projects.

Http protocol is the foundation of data communication in world wide web. Different methods of data retrieval from specified URL are defined in this protocol.

The following table summarizes different http methods –

Sr.No	Methods & Description
1	GET Sends data in unencrypted form to the server. Most common method.
2	HEAD Same as GET, but without response body
3	POST Used to send HTML form data to server. Data received by POST method is not cached by server.
4	PUT Replaces all current representations of the target resource with the uploaded content.

5

DELETE

Removes all current representations of the target resource given by a URL

By default, the Flask route responds to the **GET** requests. However, this preference can be altered by providing methods argument to **route()** decorator.

In order to demonstrate the use of **POST** method in URL routing, first let us create an HTML form and use the **POST** method to send form data to a URL.

Save the following script as login.html

```
<html>

<body>

<formaction="http://localhost:5000/login"method="post">

<p>Enter Name:</p>

<p><inputtype="text"name="nm"/></p>

<p><inputtype="submit"value="submit"/></p>

</form>

</body>

</html>
```

Now enter the following script in Python shell.

```
from flask import Flask, redirect, url_for, request

app=Flask(__name__)
```

```
@app.route('/success/<name>')

def success(name):

    return'welcome %s'% name

@app.route('/login',methods=['POST','GET'])

def login():

    ifrequest.method=='POST':

        user=request.form['nm']

        return redirect(url_for('success',name= user))

    else:

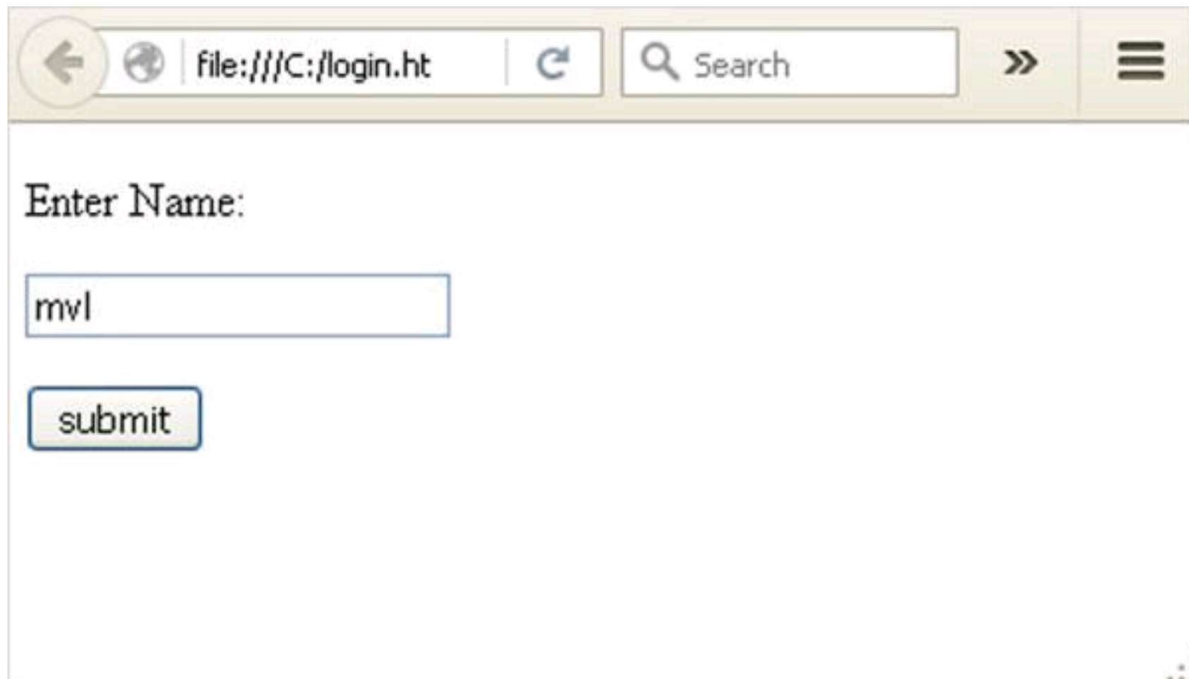
        user=request.args.get('nm')

        return redirect(url_for('success',name= user))

if __name__ == '__main__':

    app.run(debug =True)
```

After the development server starts running, open **login.html** in the browser, enter name in the text field and click **Submit**.



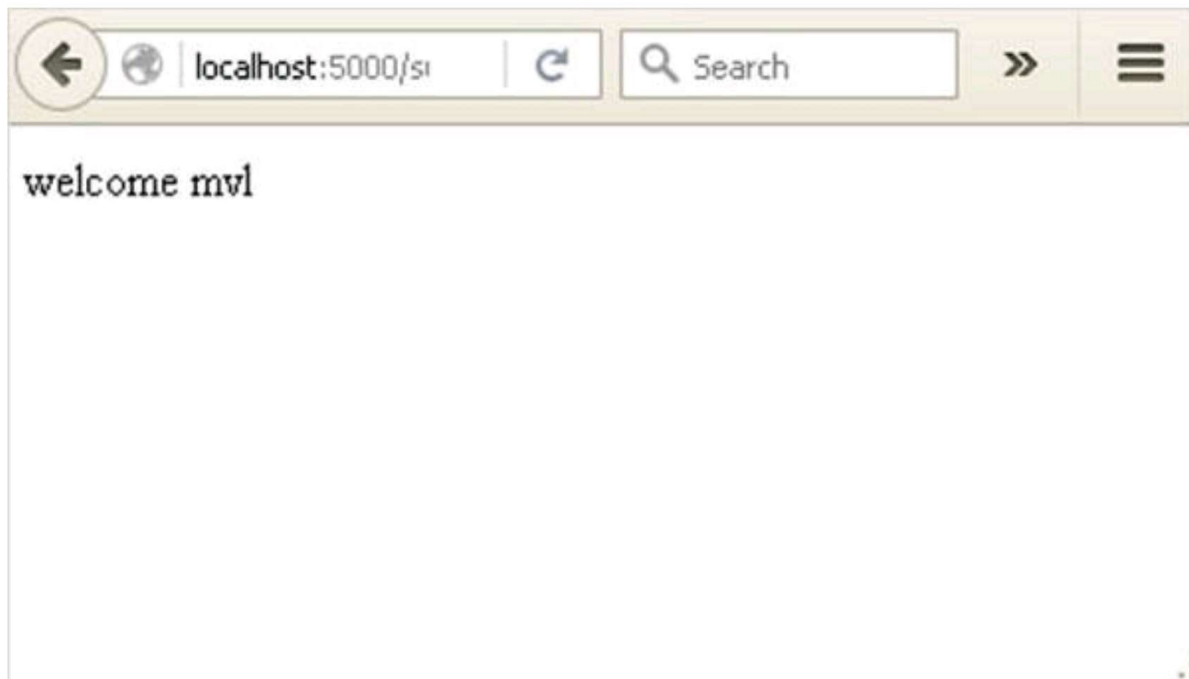
Enter Name:

Form data is POSTed to the URL in action clause of form tag.

`http://localhost/login` is mapped to the `login()` function. Since the server has received data by POST method, value of 'nm' parameter obtained from the form data is obtained by –

```
user = request.form['nm']
```

It is passed to '/success' URL as variable part. The browser displays a welcome message in the window.



welcome mvl

Change the method parameter to '**GET**' in **login.html** and open it again in the browser. The data received on server is by the **GET** method. The value of 'nm'

parameter is now obtained by –

```
User = request.args.get('nm')
```

Here, **args** is dictionary object containing a list of pairs of form parameter and its corresponding value. The value corresponding to 'nm' parameter is passed on to '/success' URL as before.

3.9 REQUIREMENT ANALYSIS

Requirement analysis, also called requirement engineering, is the process of determining user expectations for a new modified product. It encompasses the tasks that determine the need for analyzing, documenting, validating and managing software or system requirements. The requirements should be documentable, actionable, measurable, testable and traceable related to identified business needs or opportunities and define to a level of detail, sufficient for system design.

FUNCTIONAL REQUIREMENTS

It is a technical specification requirement for the software products. It is the first step in the requirement analysis process which lists the requirements of particular software systems including functional, performance and security requirements. The function of the system depends mainly on the quality hardware used to run the software with given functionality.

Usability

It specifies how easy the system must be use. It is easy to ask queries in any format which is short or long, porter stemming algorithm stimulates the desired response for user.

Robustness

It refers to a program that performs well not only under ordinary conditions but also under unusual conditions. It is the ability of the user to cope with errors for irrelevant queries during execution.

Security

The state of providing protected access to resource is security. The system provides good security and unauthorized users cannot access the system there by providing high security.

Reliability

It is the probability of how often the software fails. The measurement is often expressed in MTBF (Mean Time Between Failures). The requirement is needed in order to ensure that the processes work correctly and completely without being aborted. It can handle any load and survive and survive and even capable of working around any failure.

Compatibility

It is supported by version above all web browsers. Using any web servers like localhost makes the system real-time experience.

Flexibility

The flexibility of the project is provided in such a way that is has the ability to run on different environments being executed by different users.

Safety

Safety is a measure taken to prevent trouble. Every query is processed in a secured manner without letting others to know one's personal information.

NON- FUNCTIONAL REQUIREMENTS

Portability

It is the usability of the same software in different environments. The project can be run in any operating system.

Performance

These requirements determine the resources required, time interval, throughput and everything that deals with the performance of the system.

Accuracy

The result of the requesting query is very accurate and high speed of retrieving information. The degree of security provided by the system is high and effective.

Maintainability

Project is simple as further updates can be easily done without affecting its stability. Maintainability basically defines that how easy it is to maintain the system. It means that how easy it is to maintain the system, analyse, change and test the application. Maintainability of this project is simple as further updates can be easily done without affecting its stability.

SYSTEM DESIGN AND TESTING PLAN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's

relationship to help user decision-making.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

The feasibility study investigates the problem and the information needs of the stakeholders. It seeks to determine the resources required to provide an information systems solution, the cost and benefits of such a solution, and the feasibility of such a solution.

The goal of the feasibility study is to consider alternative information systems solutions, evaluate their feasibility, and propose the alternative most suitable to the organization. The feasibility of a proposed solution is evaluated in terms of its components.

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system is well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is

the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

- Valid Input : identified classes of valid input must be accepted.
- Invalid Input : identified classes of invalid input must be rejected.
- Functions : identified functions must be exercised.
- Output : identified classes of application outputs must be exercised.
- Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining

to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

CHAPTER 4

RESULTS AND DISCUSSION

WORKING

4.1 Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

4.2 Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software

applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

4.3 Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

CHAPTER 5

CONCLUSION

5.1 CONCLUSION

There is a concern among physicians how to detect diabetes at its infancy stage. This study had made an effort in designing the system in predicting the diabetes. The experimental work implemented ML algorithms and the evaluation was done on various measures. The experiment was carried on the Diabetes data set and the results confirmed the designed system had an accuracy of 79.17% using the supplying regression classification formula. The designed system using this ML algorithm can also be customized in predicting other alternative diseases. The research can be further enhanced in implementing other ML algorithm in improving the prediction of diabetes.

5.2 REFERENCES

- [1] Deepti Sisodiaa, Dilip Singh Sisodiab, 2018, Prediction of Diabetes using Classification Algorithms, International Conference on Computational Intelligence and Data Science - ICCIDS 2018 ,Science Direct Procedia Computer Science 132 (2018) 1578–1585.
- [2] G. Krishnaveni*, T. Sudha,” A Novel Technique To Predict Diabetic Disease Using Data Mining Classification Techniques” in International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT2017), vol. 3, Issue 1, pp. 5-11, 2017.
- [3] Vrushali B., and Rakhi W., “Review on Prediction of Diabetes using Data Mining Technique”, International Journal of Research and Scientific Innovation (IJRSI), Volume IV, Issue IA, pp. 43-46, January 2017.
- [4] Harleen and Dr. Pankaj B.,”A Prediction Technique in Data Mining for Diabetes Mellitus,” Journal of Management Sciences and Technology, vol. 4, Issue 1, pp. 1-12, 2016.
- [5] Thirumal P., and Nagarajan N.,” Utilization of Data Mining Techniques for Diagnosis of Diabetes Mellitus - A Case Study”, ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 1, pp. 8-13, January 2015.
- [6] Iyer A., Jeyalatha S., Sumbaly R., “Diagnosis of diabetes using classification mining techniques,” International Journal of Data Mining & Knowledge Management Process

(IJDKP), vol. 5, no. 1, 2015.

[7] Perveen, S., Shahbaz, M.,Guergachi,A.,Keshavjee,K.,2016.PerformanceAnalysis of Data MiningClassificationTechniques to Predict Diabetes. Procedia Computer Science 82, 115–121. doi:10.1016/j.procs.2016.04.016.

[8] Orabi, K.M.,Kamal, Y.M.,Rabah,T.M.,2016.EarlyPredictive SystemforDiabetes MellitusDisease,in:IndustrialConference on Data Mining, Springer. Springer. pp. 420–427.

5.3 APPENDIX

A. SCREENSHOTS

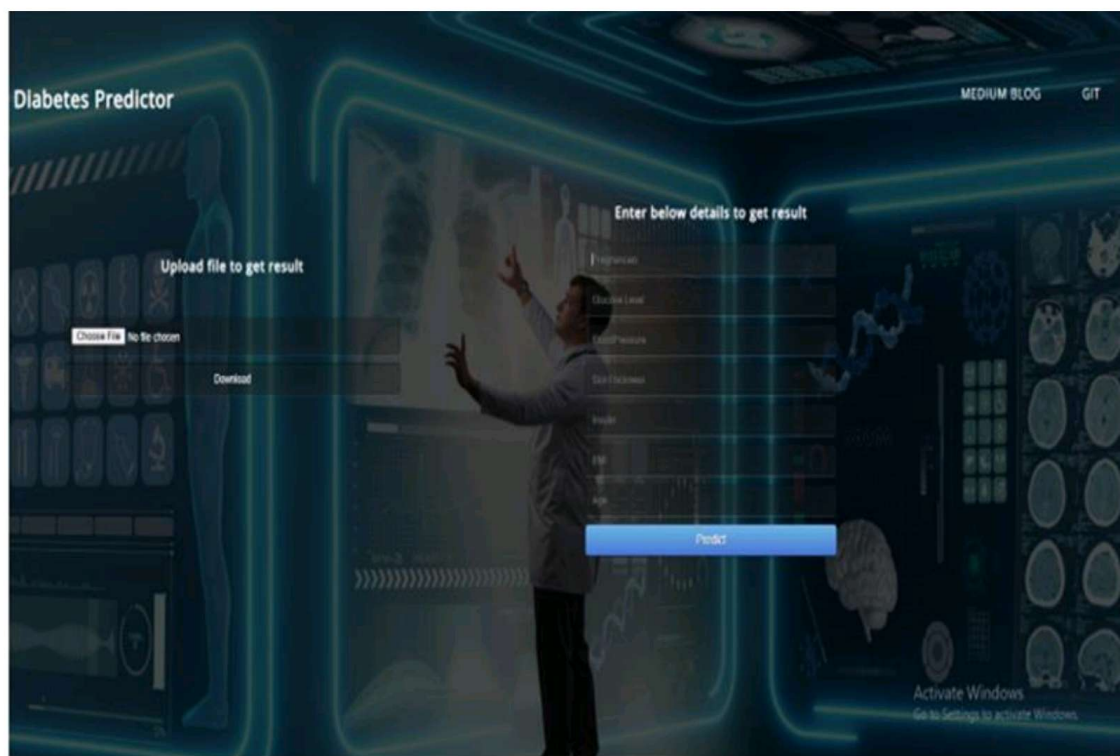


Fig:B.1 : Input page

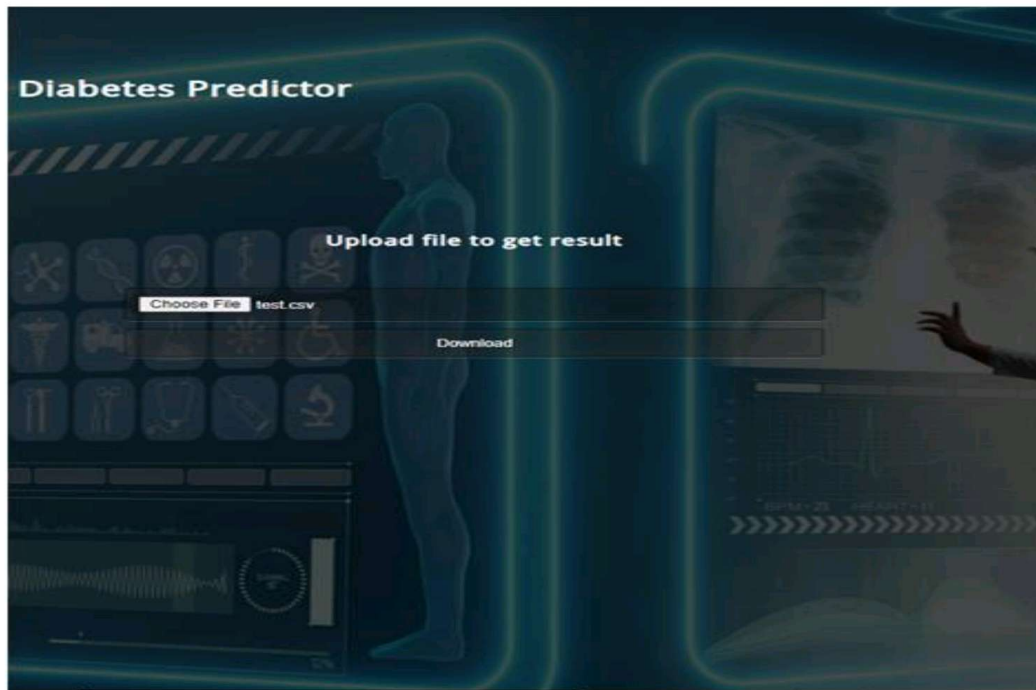


Fig:B.2 : Output page

1		Pregnanci	Glucose	BloodPres	SkinThickr	Insulin	BMI	Age	pred
2	0	6	148	72	35	0	33.6	50	1
3	1	1	85	66	29	0	26.6	31	0
4	2	8	183	64	0	0	23.3	32	1
5	3	1	89	66	23	94	28.1	21	0
6	4	0	137	40	35	168	43.1	33	1
7	5	5	116	74	0	0	25.6	30	0
8	6	3	78	50	32	88	31	26	0
9	7	10	115	0	0	0	35.3	29	1
10	8	2	197	70	45	543	30.5	53	1
11	9	8	125	96	0	0	0	54	0
12	10	4	110	92	0	0	37.6	30	0
13	11	10	168	74	0	0	38	34	1
14	12	10	139	80	0	0	27.1	57	1
15	13	1	189	60	23	846	30.1	59	1
16	14	5	166	72	19	175	25.8	51	1
17	15	7	100	0	0	0	30	32	0
18	16	0	118	84	47	230	45.8	31	0
19	17	7	107	74	0	0	29.6	31	0
20	18	1	103	30	38	83	43.3	33	0

Fig:B.3 : Dataset

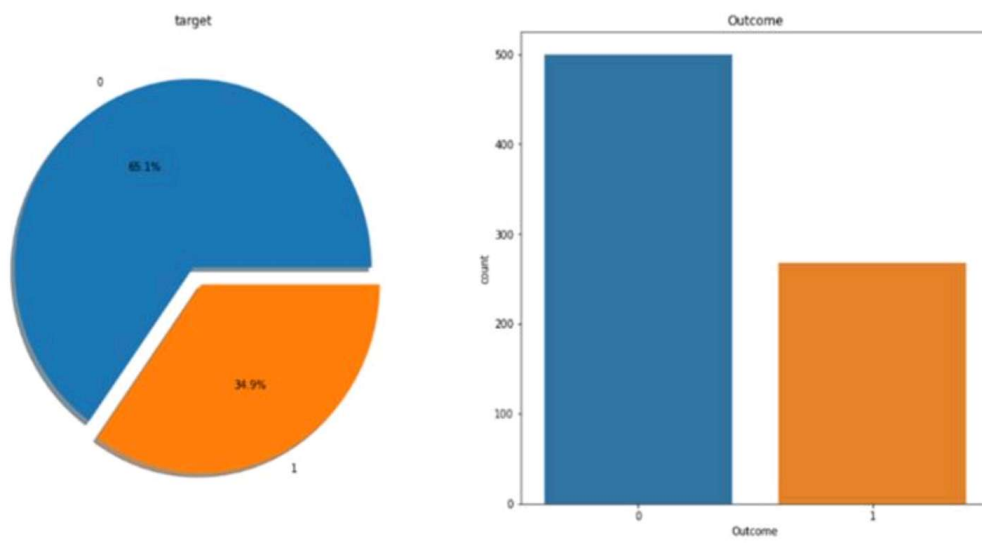
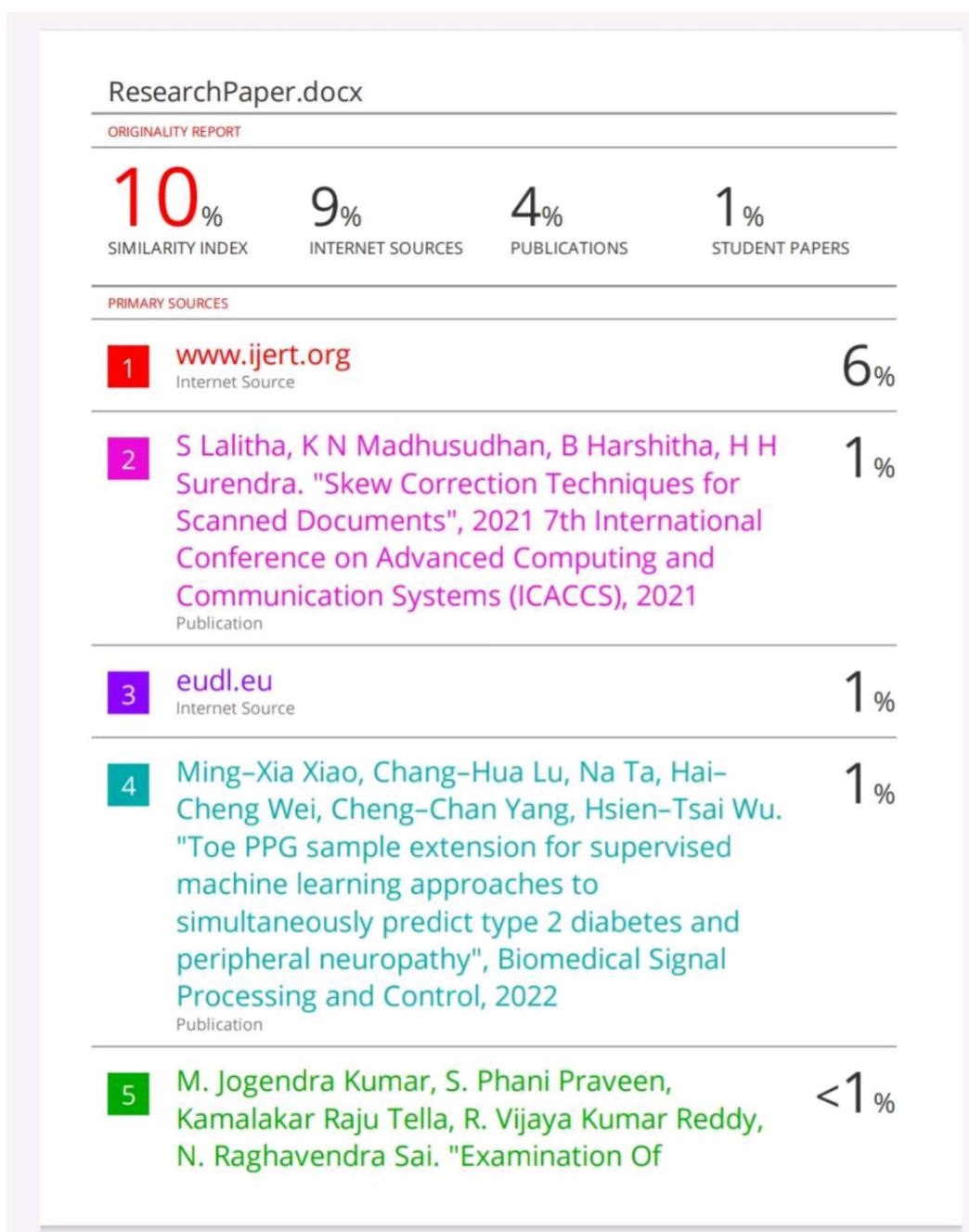


Fig:B.4 : Accuracy Graph



Fig:B.5 : Corelation Matrix

B. PLAGIARISM REPORT



Supervised Learning method of Diabetes Prediction

M.S. Roobini¹, S.R. Srividhya², Sugnaya³, Yalamarathi Sivva⁴, Vemireddy Praneeth Kumar Reddy⁵

^{1,2,3,4,5}Sathyabama Institute of Science and Technology, Chennai-600127, India.

roobinims@gmail.com, srividhyasr.cse@sathyabama.ac.in, suganiya.cse@sathyabama.ac.in,
yalamarthisivva18@gmail.com, vemireddy.praneeth25@gmail.com

Abstract— *In the sphere of software hospital care and data, healthcare services functions as it does now. in the context of expanding local, provincial, and global health administrations. A real-life example of anticipating the risk of sickness using electronic clinical data analysis aids in not only focusing on the patient, but also in providing various types of help through an electronic health-care system in this article, we'll discuss attempt to predict and analyzing diabetes, a dangerous sickness that is part of a group of metabolic illnesses in which glucose levels are consistently elevated throughout time. K-Nearest Neighbor is well-known as well as simple to grasp AI approaches as a result of this a thing as predicting sickness risk based on health details. To achieve this goal, we provide a reasonable a speculation model based on analysis on ideal K-Nearest Neighbor (KNN) with light on patient impressions in diverse regions. This method selects the greatest number of neighbors-blunder in order to achieve the optimal end result given in the upshot. Continuous data regarding diabetes collected in clinical facilities is used to test the presentation of that kind of e-wellbeing model.*

Keywords— *Data Constraints, Diabetes, Feature Extraction, Pre-processing.*

I. INTRODUCTION

Type 2 mellitus (DM) is the most common noncommunicable disease (NCD) on the globe; type 2 diabetes affects up to 90% of individuals (T2DM) The risk of T2DM infection is linked to a variety of mental, behavioural, and natural dangers and health issues. Diabetes is a collection of metabolic problems that result in elevated glucose levels over a lengthy amount of time. Diabetes is characterised by a variety of metabolic disorders caused by abnormal insulin release and/or work. Unreasonable cravings, constant thirst, and general desire are among symptoms of aldohexose. If diabetes is not treated, it can lead to serious medical problems including as diabetic, hyperosmolar hyperglycemic corrosive, and even death. Nervousness, brain paralyze, slowness, foot ulcers, and eye anguish might all occur at the same time. Diabetes develops when the body's organs are unable to provide enough insulin, or if the body's basic elements as well as the insulin they create are not utilised. Insulin-incited diabetics is characterised along the a tubular function that manufactures insulin but does not meet the body's ments (IDDM). Insulin blockage describes type 2 diabetes, which happens when the

body's defensive system responds to insulin uniquely in contrast to anticipated. At the point when there isn't sufficient insulin in the body, this can happen. This is known as "non-insulin diabetes mellitus" "adults with diabetes mellitusDiabetics are more likely among those who have a higher Metabolic rate or are overweight. poor academic performance. Physiologically, diabetes has a third standard. In general, the amount of aldohexose in a normal person ranges from 70 to 99 milligrammes per deciliter. Pre-diabetic patients have aldohexosis levels ranging from one to 25 mg/dL. Individuals with high cholesterol but low long-term hypotension of 40 mg/dL were shown to have a variety of ovarian growths.

II. LITERATURE SURVEY

G. Krishnaveni et al. use a variety of ways to score the diabetic extortion informative index in order to anticipate deception and deceit in this article. The Diabetes Warehouse aspires to improve mining efficiency. The processed data is used to produce forecasts using old-style computations such as discriminant analysis, KNN, Nave Bayes, other vector machines. In the discipline of bioinformatics, similar orders can be put to good use. We look into many characterization techniques, such as insightful segregation, K - nearest neighbors, Naive Bayes, support vector preparations, RBF lines, and approval. [1].

In this peer-reviewed study, Talha Mahboob Alam and al discuss diabetes in terms of important characteristics, as well as the relationship between different civilizations. To choose practises, assess, predict, and develop affiliation criteria, many methods are used. In the inspection of important components, key items were determined. The findings of the study demonstrate that diabetes is strongly linked to body mass (BMI) and, of course, glucose levels. The downside of studying is that it takes time. information acquired is arranged; however, this method is used in numerous areas of medicine to investigate different types of cancerParkinson's disease, psoriasis, and.. infection. Diabetes testing should also take into account characteristics such as real occupation, family heritage, and smoking habits [2]. The goal of Deepti Sisodia et al's is to create a version that clearly demonstrated how a sufferer might cultivate diabetes. As a result, three AI computations, such as Decision Graph, SVM, and Naive Bayes, were used in this study to distinguish early diabetes. The review is being undertaken by the UCI Machine Training Foundation's (PIDD). One of the most appealing

aspects of using Dataset is that it may be planned. The most important The purpose of this study was to predict diabetic patients who will use the WEKA device and also the PIDD medical library. The limitations of this paradigm include the fact that certain information has not been taken into account while examining the diabetic class [3]. This method amounted to examining at all of the data right away, considerably reducing the number of live humans required to properly assess the patient. As a result, physicians may find it useful to contribute to a better understanding about AD as a people. In any event, therapy at this time will almost certainly be beneficial and advantageous in advancing the evaluation cycle. [4]. The main goal for this work, according to Minzhen Wang et al, is to provide a logical basis to a forecasting of expansion of mouth, hand, and foot disease by focusing on the consequences of environmental changes on the progression of hand, foot, and mouth disease there is a mouth. China. The disease and climatic variables are combined here, as well as The inquiry is ongoing. associated with the line and then how to modify it [5]. Zhang Y, Long JD, and colleagues proposed the Ghastly data, which is commonly used to assess the severity of various illnesses. In any event, such information has not been analysed in order to classify irresistible diseases. The study's purpose was to come up with a list of illnesses that might identify a class of leaf-borne diseases just at DS level [6]. Another quantifiable strategy has been devised for break its four voice of cardiovascular medicine, which is a means of revealing the health of the heart, as Hu Yuliang et al discuss more easily assess and treat cardiovascular illnesses. The study's subject is a common cardiac problem with symptoms as infection as a result of which it may be difficult to properly calculate and implement technique in suitable locations. [7]. In this research, Chengyu Liu1 et al discuss the importance of services and amenities portrayals on non-direct percentages of cardio vascular alterations in a group of patients with risk factors. Demonstrating speed, detection wavering inspection, and similar entropy were the three strategies used. Long-term episodes linked to pulse were measured using DFA. We will use several ways for acquiring diabetic data to discriminate passings and passings, according to G. Krishnaveni, Prof. T.Sudha/ 2018 [8]. Diabetes mellitus is designed to make the most prevalent method of getting rid of it easier. The KNN computation, Nave Bayes, and also the vector machine are used to predict using the pre-handled data. These classes have a lot of potential in bioinformatics. RBF, KNN, Nave Bayes, Support Vector machine (svm, and more algorithms are dissected and explained in detail alongside the atomic line. Talha Mahboob Alam, Muhammad AtifIqbal, Yasir Ali, and Abdul Wahab[9] are some of the most well-known people in Pakistan. Here In addition, a number of attributes are described. To choose important behaviors, diabetes, predict, and establish affiliation rules, several gadgets are used. In the research of vital sections, the key items were chosen. The study's findings demonstrate that diabetes is strongly linked to body mass index (BMI) and blood glucose measured in a deductive manner. The goal of this study, according to DeeptiSisodia,Dilip SinghSisodia [10], is to develop a model with a strong predictive incentive for predicting the

possibility of a diabetic patient. As a result, three AI computations,in this review to distinguish early diabetes: Decision Tree, SVM, and Naive Bayes.. This Pima Diabetes Data (PIDD) UCI Machines Training Foundation is the target of the audit.

III. EXISTING SYSTEM AND PROPOSED SYSTEM

The existing structure was handled by modifying the present situation structure to the frame requirements and using Naive Bayes to solve the problem. The framework isn't working properly, and data is being lost. supposed to provide an accurate assessment. To achieve the optimum outcomes, the suggested diabetic expectation framework contains two main steps. The information management phase is the most important part of the structure, followed by the interaction step. In any event, the framework's contribution is an informative index, with the end result being a stage in the recognition of wellness and diabetes. The implementation of the new frameworks will aid in the reduction of the influencing system, and the review's findings reveal that the framework has superior presumptions in terms of precision. Purchasers can get tested for diabetes and receive immediate results. The K-Neighbor Technique is a computation that is both speedy and large.

IV. SYSTEM ARCHITECTURE

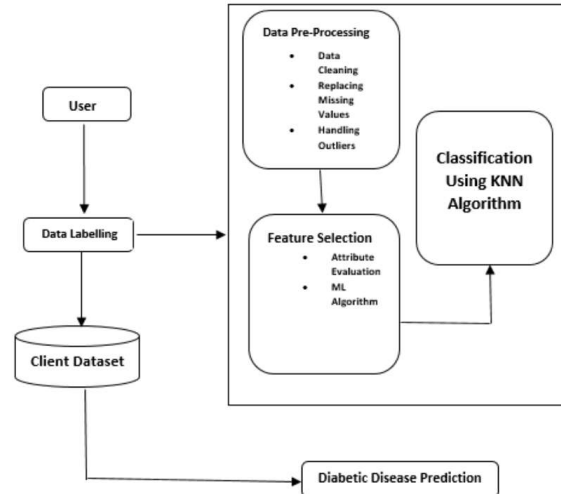


Fig.4.1. Architecture Diagram

- The system is about prediction of diabetic disease.
- The data is collected from Kaggle dataset and then pre-processing is performed to the data which is collected.
- After performing pre-processing to the data then feature extraction and selection is done then we will apply machine learning algorithms for detecting diabetic disease.

- By applying machine learning algorithms, it will predict diabetic disease.

V. REQUIREMENTS

A. Hardware Requirement

- System: Pentium Dual Core.
- HDD: 120 GB.
- Screen: 15'' LED
- Input Devices: Keyboard, Mouse
- Ram: 4 GB.

B. Software Requirement

- Operating system: Windows 7/10.
- Coding Language: Python
- Tool: python 3.6.1

VI. METHODOLOGY

A. Data Constraints

The term "data" refers to a wide range of information. As a result, Pima information from India is used to create the models. Preparation of data sets and data research

B. Train Dataset and Test Dataset

Data learning is the initial segment of data that is utilized to comprehend a program. To execute this usefulness, we should initially prepare the model, and this data is on the framework. This data is utilized in AI to play out an assortment of assignments. This is data that can be gained from a model utilizing calculations that show displaying and accomplish the work consequently. The news-casting test is a demonstration of programming. It shows what data means for the exhibition of a module and is generally utilized in estimation. Information handling ahead of time.

C. Pre-processing of data

Pre-handling information is a technique used to change fundamental data decidedly. This is a stage towards changing data or planning a machine in a manner that can be effectively investigated. The principle undertaking of handling data during preparing is to dispense with undesirable data and fill in lost data. This is not difficult to prepare the machine.

D. .Feature Extraction

Highlighting crucial data and the processes used to transform it into arrangements. It works out the graphs with a comparable quantity of characters to have an effect on the point by point classes of the fundamental structure. This method is anticipated to reduce the number of devices required to access large amounts of data. Unloading is a technique for lowering the number of highlights. It's also utilised to increase speed and the ability to find out how to regulate.

E. ML Algorithm: KNN

II. The ML calculation, a revolutionary approach introduced by Thomas Cover that uses backward and sequencing, is the closest neighbour to the k. This calculation is often utilised to solve current problems. The KNN computation is a learning-based model sort. This computation is based on a variety of elements, including the sequence in which data is prepared, as well as a significant amount of progress in general. Neighbors are part of a chain of events that define a class or entity. It's often thought of as a calculator's preparatory tool, yet it doesn't require any special preparation. K-Nearest Neighbour is a kind of example training, also called as slow learner, wherein the functionality is only estimated locally and all computation occurs after classification. One of the most fundamental machine learning algorithms is the k-NN approach. The neighbours are selected from a group of particles for which the classes (in k-NN classifications) or objects value (in quality k-NN regression) has been determined..

VII. OUTPUT SCREENSHOTS

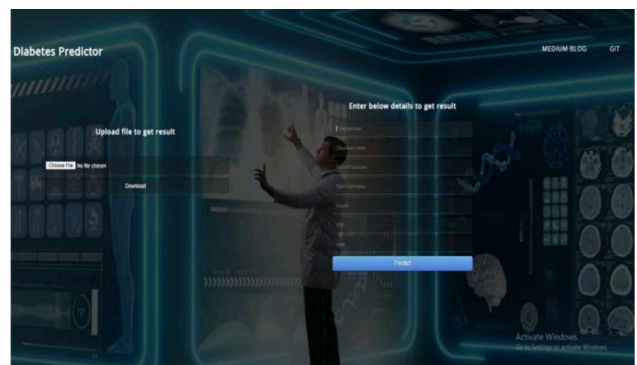


Fig.8.1. Home Page of Diabetes Predictor

To predict diabetes we have to enter some details of the person as shown in Fig 8.1. As shown in Fig 8.2 we have to upload dataset and download prediction of dataset. After downloading the prediction dataset we can see the diabetes prediction in the last column of Fig 8.3.

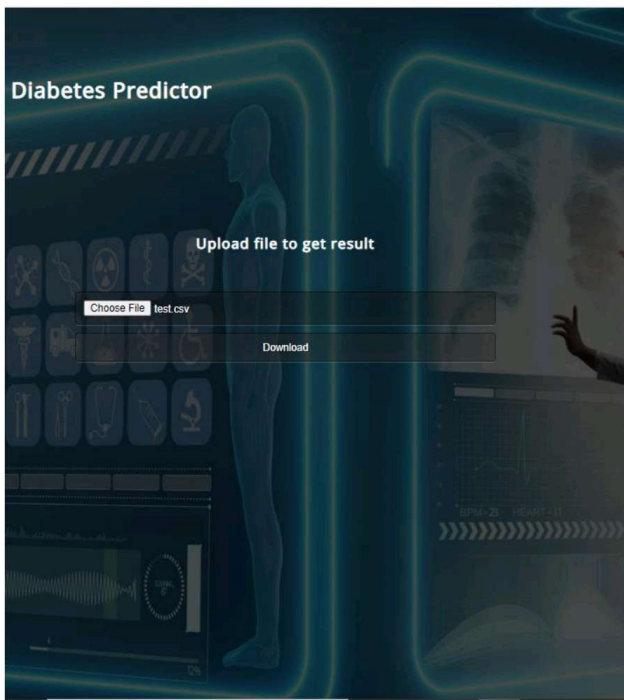


Fig.8.2. Data Set Upload Page

	Pregnanci	Glucose	BloodPres	SkinThickn	Insulin	BMI	Age	pred
1	0	6	148	72	35	0	33.6	50
2	1	1	85	66	29	0	26.6	31
3	2	8	183	64	0	0	23.3	32
4	3	1	89	66	23	94	28.1	21
5	4	0	137	40	35	168	43.1	33
6	5	5	116	74	0	0	25.6	30
7	6	3	78	50	32	88	31	26
8	7	10	115	0	0	0	35.3	29
9	8	2	197	70	45	543	30.5	53
10	9	8	125	96	0	0	0	54
11	10	4	110	92	0	0	37.6	30
12	11	10	168	74	0	0	38	34
13	12	10	139	80	0	0	27.1	57
14	13	1	189	60	23	846	30.1	59
15	14	5	166	72	19	175	25.8	51
16	15	7	100	0	0	0	30	32
17	16	0	118	84	47	230	45.8	31
18	17	7	107	74	0	0	29.6	31
19	18	1	103	30	38	83	43.3	33
20								

Fig .8.3. Data Set

	count	mean	std	min	10%	25%	50%	75%	90%	95%	99%	max
Pregnancies	768.0	3.045052	3.369578	0.000	0.000	1.00000	3.0000	6.00000	9.0000	10.00000	13.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	85.000	99.00000	117.000	140.25000	167.0000	181.00000	196.00000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	54.000	62.00000	72.0000	80.00000	88.0000	90.00000	106.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.000	0.00000	23.0000	32.00000	40.0000	44.00000	51.33000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.000	0.00000	30.5000	127.25000	210.0000	293.00000	519.90000	846.00
BMI	768.0	31.992578	7.804160	0.000	23.600	27.30000	32.0000	36.60000	41.5000	44.39500	50.75900	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.165	0.24375	0.3725	0.62625	0.8786	1.13285	1.69833	2.42
Age	768.0	33.240885	11.760232	21.000	22.000	24.00000	29.0000	41.00000	51.0000	58.00000	67.00000	81.00
Outcome	768.0	0.340958	0.476951	0.000	0.000	0.00000	0.0000	1.00000	1.0000	1.00000	1.00000	1.00

Fig.8.4. Exploratory Data Analysis

In the Fig 8.4 data preprocessing and exploratory data analysis is done for the dataset. Fig 8.5 is the graphical representation of outcome variable.

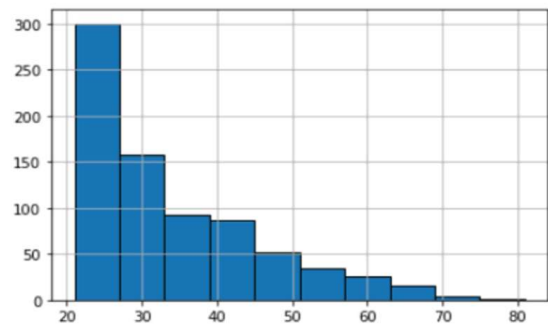


Fig.8.5. Outcome Variable

Fig 8.6 depicts the histogram and density graphs of variables in the dataset. The Distribution of the Outcome Variable in The Data was Examined and Visualized as shown in Fig 8.7. Fig8.8 represents Heat map plot diagram.

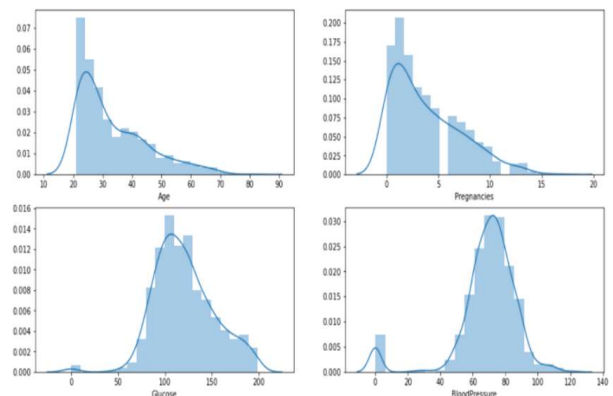


Fig.8.6. Histogram and Density Graphs of Variables

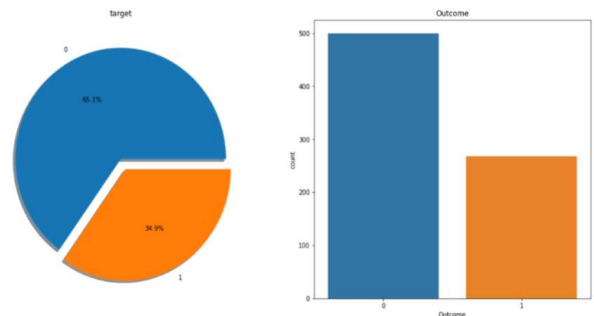


Fig. 8.7. The Distribution of the Outcome Variable in The Data was Examined and Visualized

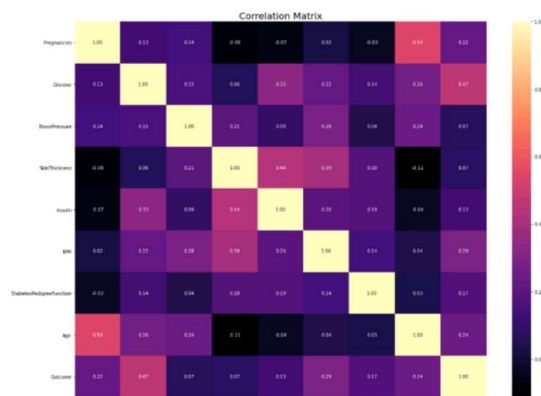


Fig.8.8. Correlation Matrix Graph of Data set

VIII. RESULT

Whenever this data is available in the frameworks, it will be able to split the figures using ML calculations and provide crucial results in promoting their discrepancies in order to choose the best diabetic therapy.

IX. CONCLUSION

In this essay, we'll look into diabetic counselling and e-health administrations with our nearest neighbours. In this example, diabetes data is used to survey a positive number of nearest neighbors in order to create a model for predicting illness. To create a reasonable predictive model, a positive value of K is combined with a small amount of errors. The diabetes information acquired at the medical emergency clinic is used to assess the model's exhibition. A variety of AI methodologies are being employed to examine the effects of theories by taking into account the various elements that might cause this sickness. The findings of the study will aid in making better therapeutic decisions for early detection and infection prevention, perhaps saving live

X. REFERENCES

- [1] P. Groves, B. Kayyali, D. Knott, S. Van Kuiken, The 'big data' revolution in healthcare, McKinsey Quarterly 2 (3) (2013).
- [2] National health and family planning commission of the people's republic of china, National status report on nutrition and chronic disease of residents in China, 2015.
- [3] N. Sneha, T. Gangil, Analysis of diabetes mellitus for early prediction using optimal features selection, Journal of Big Data 6 (1) (2019) 13.
- [4] M. Mashayekhi, F. Prescod, B. Shah, L. Dong, K. Keshav-gee, A. Guergachi, Evaluating the performance of the framingham diabetes risk scoring model in

canadian electronic medical records, Canadian journal of diabetes 39 (2) (2015) 152–156.

[5] M. F. Faruque, I. H. Sarker, et al., Performance analysis of machine learning techniques to predict diabetes mellitus, in: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, 2019, pp. 1–4.

[6] C. B. Giorda, V. Manicardi, J. Diago Cabezudo, The impact of diabetes mellitus on healthcare costs in Italy, Expert review of pharmacoeconomics & outcomes research 11 (6) (2011) 709–719.

[7] Q. Zou, K. Qu, Y. Ju, H. Tang, Y. Luo, D. Yin, Predicting diabetes mellitus with machine learning techniques, Frontiers in genetics 9 (2018) 515.

[8] A Novel Technique To Predict Diabetic Disease Using Data Mining Classification Techniques G. Krishnaveni, Prof. T. Sudha / 2018

[9] A model for early prediction of diabetes Talha Mahboob Alam, Muhammad Atif Iqbal, Yasir Ali, Abdul Wahab, Safdar Ijaz, Talha Imtiaz Baig, Ayaz Hussain, Muhammad Awais Malik, Muhammad Mehdi Raza, Salman Ibrar, Zunish Abbas / 2019

[10] Prediction of Diabetes using Classification Algorithms Deepti Sisodia, Dilip Singh Sisodia / 2018