Faculty of Engineering and Technology

Department of Electronics and Communication Engineering

Jain Global Campus, Kanakapura Taluk - 562112
Ramanagara District, Karnataka, India

2017-2021

A Project Report on

"EARLY DETECTION OF DIABETES USING MACHINE LEARNING"

Submitted in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

Submitted by

PAMULURU YASWANTH KUMAR REDDY
16BT6EC101

GOKULA BHAVYA TEJ
17BTREC035

M S MOHAMED ASHFAQ
17BTREC056

MALLELA VISHNU
17BTREC057

Under the guidance of

CHETHANA GS
Assistant Professor
Department of Electronics and Communication Engineering
Faculty of Engineering & Technology
JAIN(DEEMED-TO-BE UNIVERSITY)

Faculty of Engineering and Technology

Department of Electronics and Communication Engineering

Jain Global Campus, Kanakapura Taluk - 562112
Ramanagara District, Karnataka, India

2017-2021

A Project Report on

"EARLY DETECTION OF DIABETES USING MACHINE LEARNING"

Submitted in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

Submitted by

PAMULURU YASWANTH KUMAR REDDY
16BT6EC101

GOKULA BHAVYA TEJ
17BTREC035

M S MOHAMED ASHFAQ
17BTREC056

MALLELA VISHNU
17BTREC057

Under the guidance of

CHETHANA GS
Assistant Professor
Department of Electronics and Communication Engineering
Faculty of Engineering & Technology
JAIN(DEEMED-TO-BE UNIVERSITY)

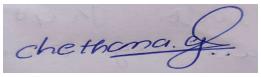
Faculty of Engineering & Technology

Department of Electronics & Communication Engineering

Jain Global campus
Kanakapura Taluk - 562112
Ramanagara District
Karnataka, India

CERTIFICATE

This is to certify that the project worktitled **"EARLY DETECTION OF DIABETES USING MACHINE LEARNING"**is carried out by **Pamuluru Yaswanth Kumar Reddy (16BT6EC101), Gokula Bhavya Tej (17BTREC035), M S Mohamed Ashfaq (17BTREC056), Mallela Vishnu (17BTREC057),** are bonafide students of Bachelor of Technology at the Faculty of Engineering & Technology, JAIN DEEMED-TO-BE UNIVERSITY, Bengaluru in partial fulfillment for the award of degree in Bachelor of Technology in Electronics and Communication Engineering, during the year **2020-2021**.



Chethana GS
Assistant Professor
Dept. of ECE,
Faculty of Engineering & Technology,
JAIN DEEMED-TO-BE UNIVERSITY
Date:

Dr. R. Sukumar
Head of the Department,
Electronics and Communication,
Faculty of Engineering & Technology,
JAIN DEEMED-TO-BE UNIVERSITY
Date:

Dr. Hariprasad S.A
Director,
Faculty of Engineering & Technology,
JAIN DEEMED-TO-BE UNIVERSITY

Date:

Name of the Examiner Signature of Examiner

1.

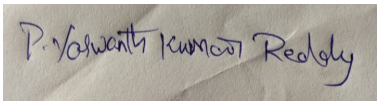
2.

DECLARATION

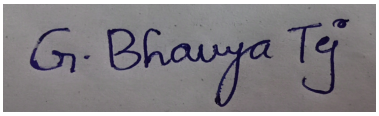
We, **Pamuluru Yaswanth Kumar Reddy (16BT6EC101), Gokula Bhavya Tej (17BTREC035), M S Mohamed Ashfaq (17BTREC056), Mallela Vishnu (17TREC057),**are student's of eighth semester B.Tech in **Electronics and Communication Engineering**, at Faculty of Engineering & Technology,**JAIN DEEMED-TO-BE UNIVERSITY**, hereby declare that the project titled **"Early Detection Of Diabetes Using Machine Learning"** has been carried out by us and submitted in partial fulfilment for the award of degree in**Bachelor of Technology in Electronics and Communication Engineering**during the academic year **2020-2021**. Further, the matter presented in the project has not been submitted previously by anybody for the award of any degree or any diploma to any other University, to the best of our knowledge and faith.

Signature

Pamuluru Yaswanth Kumar Reddy
16BT6EC101

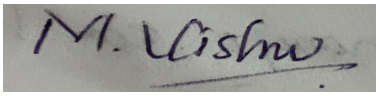


Gokula Bhavya Tej
17BTREC035



M S Mohamed Ashfaq
17BTREC056

Mallela Vishnu
17BTREC057



Place : Bengaluru
Date :

ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and support of a large number of individuals who have been responsible for the successful completion of this project work.

First, we take this opportunity to express our sincere gratitude to **Faculty of Engineering & Technology,JAIN DEEMED-TO-BE UNIVERSITY** *for providing us with a great opportunity to pursue ourBachelor's Degree in this institution.*

In particular we would like to thank ***Dr. Hariprasad S.A****,* ***Director****,* **Faculty of Engineering & Technology**,**JAIN DEEMED-TO-BE UNIVERSITY***for his constant encouragement and expert advice.*

It is a matter of immense pleasure to express our sincere thanks to ***Dr. R. Sukumar****,***Head of the department**, **Electronics and communication Engineering**, **JAIN DEEMED-TO-BE UNIVERSITY,** *for providing right academic guidance that made ourtask possible.*

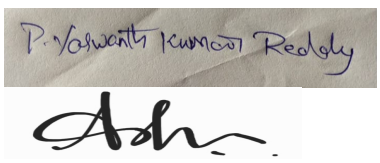
We would like to thank our guide ***Mr.Chethana GS****,* ***Assistant Professor****,* **Dept. of Electronics and Communication Engineering**, **JAIN DEEMED-TO-BE UNIVERSITY***, for sparing his valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project.*

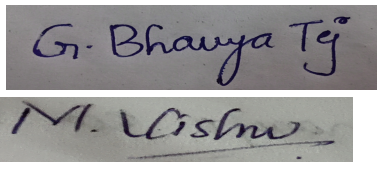
*We would like to thank our Project Coordinator****Mr. Sunil M P****and all the staff members ofElectronics and Communicationfor their support.*

We are also grateful to ourfamily and friends who provided us with everyrequirement throughout the course.

We would like to thank one and all who directly or indirectly helped us in completing the Project work successfully.

Signature of Students





ABSTRACT

The explosive population growth and health maintenance is an extremely crucial matter worldwide. Many lethal diseases are causing threats at a high peak in recent years.Introducing machine learning technologies into healthcare for early prognosis and diagnosis need to be more accurate based on the parameters and frames selected from the available clinical databases.

Healthcare domain is a very prominent research field with rapid technological advancement and increasing data day by day.

The diabetes is one of lethal diseases in the world. It is additional a inventor of various varieties of disorders foe example: coronary failure, blindness, urinary organ diseases etc. In such case the patient is required to visit a diagnostic center, to get their reports after consultation. But with the growth of Machine Learning methods we have got the flexibility to search out an answer to the current issue, we have got advanced system mistreatment information processing that has the ability to forecast whether the patient has polygenic illness or not. The exploration inferred that more variables and hybrid disciplines should be considered for an accuracy of result which can overcome the existing limitations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

1. INTRODUCTION

Diabetes mellitus (DM) is one of the most lethal noncommunicable diseases in the world. Earlier medical records show that the prediction and prevention of diabetes have become a major challenge. As the number of diagnosed patients is increasing, the medications are still not enough to control the disease. So a better predictive analysis is required to treat diabetes at an early stage which can help solve fewer issues that can help to treat the patient with fewer medications and affordability. In general, diabetes occurs when the use or production of insulin in a person's body is not balanced well i.e. the body is not able to produce or respond to the insulin hormone but the genetic and environmental factors also play a crucial role into this. Many cases are undiagnosed because of no early check-up or awareness about mainly the blood glucose level among different age groups which increases the risk of other diseases.

Diabetes can be classified into following types i.e.

Type-1 diabetes- It's a cause of insulin deficiency because the pancreas is unable to produce enough insulin. It happens due to the beta cells(insulin-making cells) being destroyed by the immune system by mistaken thinking them as invaders which leads to rising in blood sugar levels as not enough insulin is produced to use the glucose by cells. The exact reason behind this is still not known but mostly the genes are attributed to be playing a role in it. It usually occurs in both kids and youngsters. Although about 5% to 10% of people suffer from type-1 diabetes the severity can be high if not treated early.

Type-2 diabetes- This form of diabetes is a very common condition that occurs when the insulin is produced enough in the body but the cells do not use it

efficiently which leads the glucose level rise in the blood itself. Over time, the pancreas may slowdown in producing insulin. It accounts for about 90% of all cases(mostly youngsters and old aged people) with a risk of developing many medical complications like diabetic retinopathy (damages blood vessels in eyes),diabetic neuropathy, kidneys and cardiovascular diseases.It's an ongoing disease with no cure yet.

Gestational diabetes- It's a short-term condition occurs during pregnancy which is identified as a risk of type-2 diabetes in the future for both mother and child. Though in most cases, these conditions go away after the delivery but still the chances of developing the disease in later life are present

1.1 Literature Survey

Various computing techniques are applied in Healthcare domain. The focus of literature survey here is on the use of Big Data Analytics and Machine Learning in healthcare domain . In order to make a smart learning Healthcare system there are unresolved analytical data challenges. Through Big data analytics the relationship between data patterns are understood and additional value from the huge healthcare data is uncovered. There are various research trends and challenges throughout the data life cycle while implementing Big Data Analytics and is well described in. Healthcare domain challenges are in improving research phases.The reality mining is a new approach i.e. using big data to study the patient's behavior through mobile phone sensors that helps to improve the healthcare quality. Extracting useful information from Electronic Health Record is surveyed in. An intelligent design using Big Data is proposed in which is a web

based application that provides efficient platform for simplification of complex assessment of health along with monitoring procedure.

In this Analysis and Prediction of Diabetes Using Machine Learning Survey. (International Journal of EmergingTechnology and Innovative Engineering) :

Compared different classifiers and accuracy for better prediction using Pima Indian dataset by building a predictive model. WEKA software is used to implement the Decision Tree, Naive Bayes, and KNN algorithms. The bootstrapping technique is used to enhance the accuracy rates. The proposed ensemble method obtained an accuracy of 94.44%.

In this Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare Survey:

Analyzed the Pima dataset to suggest the optimal algorithm based on their experimental results using the WEKA tool. Some of the previous works with their outcomes and limitations. The confusion matrix is used to examine the performance of each algorithm. SVM, Naïve Bayes, Random Forest, and simple CART algorithms were used for the experiment in which the SVM achieved better accuracy with 79.13%.

In this Prediction of Diabetes Using Data Mining Techniques Survey:

The Pima dataset To classify diabetic and non-diabetic persons using SVM, Naïve Bayes,J48 and Backpropagation algorithms. PIMA dataset is evaluated using the Rstudio tool and the Chi-square test was used for feature selection of the dataset.

1.2 Limitations of the Current Work

In the early detection of diabetes,we use different types of datasets for analyzing the diabetes.Using the local Clinical dataset is some complicated.We can't stick only one type of dataset for analyzing diabetes.For early detection of diabetes we can more than one methods and technologies.

Diabetes is lethal non communicable disease,which is leads to kill internal organs.By the early dection of diabetes we can give the proper medication to control the diabetes.Early detection and treatment of diabetes is an impotant step towards keeping people with diabetes healthy.It can help to reduce the risk of serious complications such as premature heart disease and stroke,blindness,limb amputations, and kidney failure.

1.3. Problem Definition

The explosive population growth and health maintenance is an extremely crucial matter worldwide. Many lethal diseases are causing threats at a high peak in recent years. Introducing machine learning technologies into healthcare for carly prognosis and diagnosis need to be more accurate based on the parameters and frames selected from the available clinical databases.Toover come this we use early detection of diabetes in early stage.The early detection of diabetes using the relation between Blood Glucose Level and Plantar Pressure,we are looking into this idea.

1.4 .Objectives

The objective of this paper is to analyze, explore various research outcomes of machine learning methodologies used in diabetes mellitus and how the efficiencies obtained could be helpful in future perspective of a predictive diabetes model designing. To provide the early awareness and information regarding this disease so that the individual would take action to prevent it as quickly as possible.

Now we are using the Random Forest Classifier algorithm for the Early Detection Of Diabetes Using Machine Learning.

1.5.Methodology

As per the problems mentioned in the introduction section, this is the existing classification model is used with machine learning techniques to enhance the prediction accuracy of predict diabetes. Both classification and clustering methods are applied to build the predictive model. Algorithms used are SVM, KNN, K-means, Random Forest, Regression,Decision Tree, Outlier Detection, Rule mining, etc....

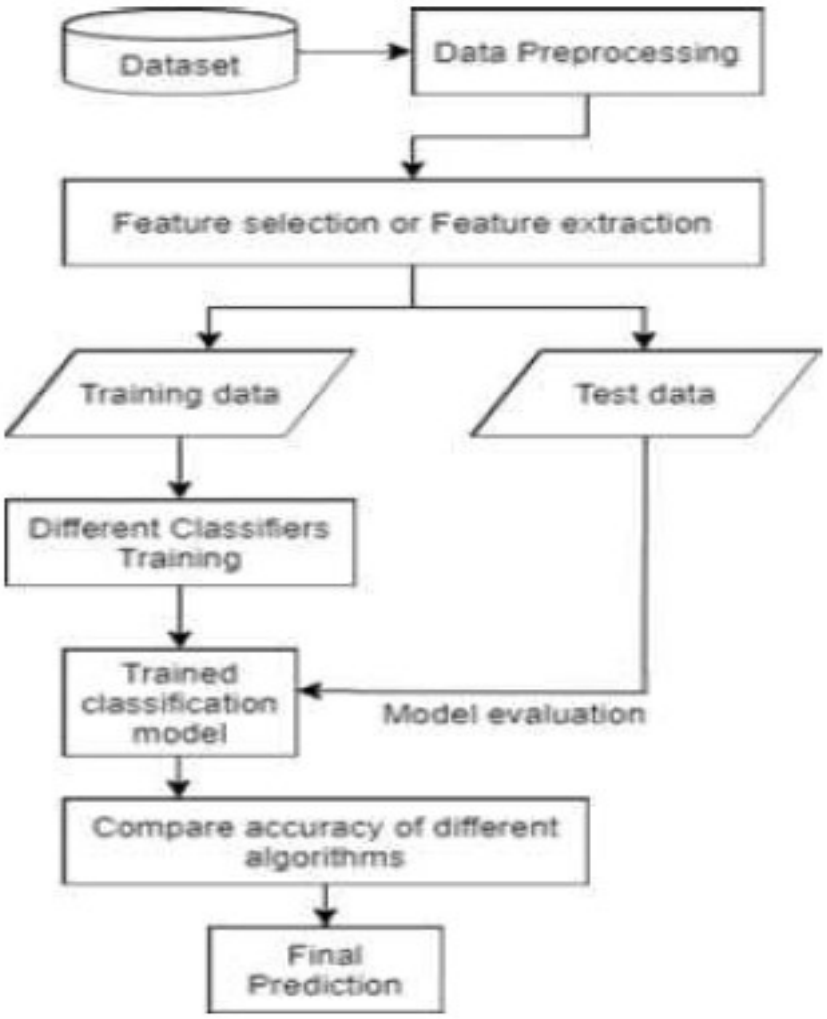


Fig 1 . Framework for Prediction Analysis

For this study, the PIMA Indian dataset is collected from the UCI Machine Learning Repository.The following are the parameters of the dataset.

Sr No	Attribute Used	Attribute Type	Attribute Description
1	Pregnancies	Numeric	No. of times pregnant
2	Glucose level	Numeric	Glucose concentration a 2 hours in an oral glucose tolerance Test
3	Blood Pressure (mmHg)	Numeric	Diastolic blood pressure (mm Hg)
4	BMI (Body Mass Index)	Numeric	weight in kg / height in square m
5	Skinfold thickness (mm)	Numeric	Triceps skin fold thickness (mm)
6	Insulin value in 2 hrs. (mu U/ml)	Numeric	2-Hour serum insulin (mu U/ml)
7	Diabetes Pedigree function	Numeric	Diabetes pedigree function
8	Age	Numeric	Age (years)
9	Outcome	Nominal	Tested positive or tested negative

Table 1 . PIMA Dataset Parameters

As well as with Pima dataset we constructed our own data set which consists of Seven parameters .

Our constructed dataset consist of :

1 BMI (Body Mass Index)
2 Habites
3 Walking
4 Yoga
5 Stress
6 Sleeping

7 Outcomes

1.6.Hardware and Software tools used

Project Jupyter exists to develop open-source software, openstandards,andservices for interactive computing across dozens of programming languages.

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Jupyter supports over 40 programming languages, including Python, R, Julia, and Scala.

Jupyter Notebooks are an open document format based on JSON. They contain a complete record of the user's sessions and include code, narrative text, equations and rich output.

Many data mining tools Anaconda, WEKA, Rstudio, MatLab,Orange, Knime, Apache, etc... are already available for carrying out this experiment. Anaconda distribution is an open-source tool and available for different OS platforms.Data visualization is very important as it needs to be clear for presenting the gaps of previous results with the experimental results so a clear analysis can be done for further researches.

.Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity

and diversity (it can be used for both classification and regression tasks). The random forest is classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

Chapter 2

2. BASIC Theory

The explosive population growth and health maintenance is an extremely crucial matter worldwide. Many lethal diseases are causing threats at a high peak in recent years.Introducing machine learning technologies into healthcare for early prognosis and diagnosis need to be more accurate based on the parameters and frames selected from the available clinical databases.

Diabetes mellitus (DM) is one of the most lethal noncommunicable diseases in the world. Earlier medical records show that the prediction and prevention of diabetes have become a major challenge. As the number of diagnosed patients is increasing, the medications are still not enough to control the disease. So a better predictive analysis is required to treat diabetes at an early stage which can help solve fewer issues that can help to treat the patient with fewer medications and affordability. In general, diabetes occurs when the use or
production of insulin in a person's body is not balanced well i.e. the body is not able to produce or respond to the insulin hormone but the genetic and environmental factors also play a crucial role into this.

Diabetes can be classified into following types i.e.

Type-1 diabetes- It's a cause of insulin deficiency because the pancreas is unable to produce enough insulin. It happens due to the beta cells(insulin-making cells) being destroyed by the immune system by mistaken thinking them as invaders which leads to rising in blood sugar levels as not enough insulin is produced to use the glucose by cells.

Type-2 diabetes- This form of diabetes is a very common condition that occurs when the insulin is produced enough in the body but the cells do not use it efficiently which leads the glucose level rise in the blood itself. Over time, the pancreas may slowdown in producing insulin.

Gestational diabetes- It's a short-term condition occurs during pregnancy which is identified as a risk of type-2 diabetes in the future for both mother and child.

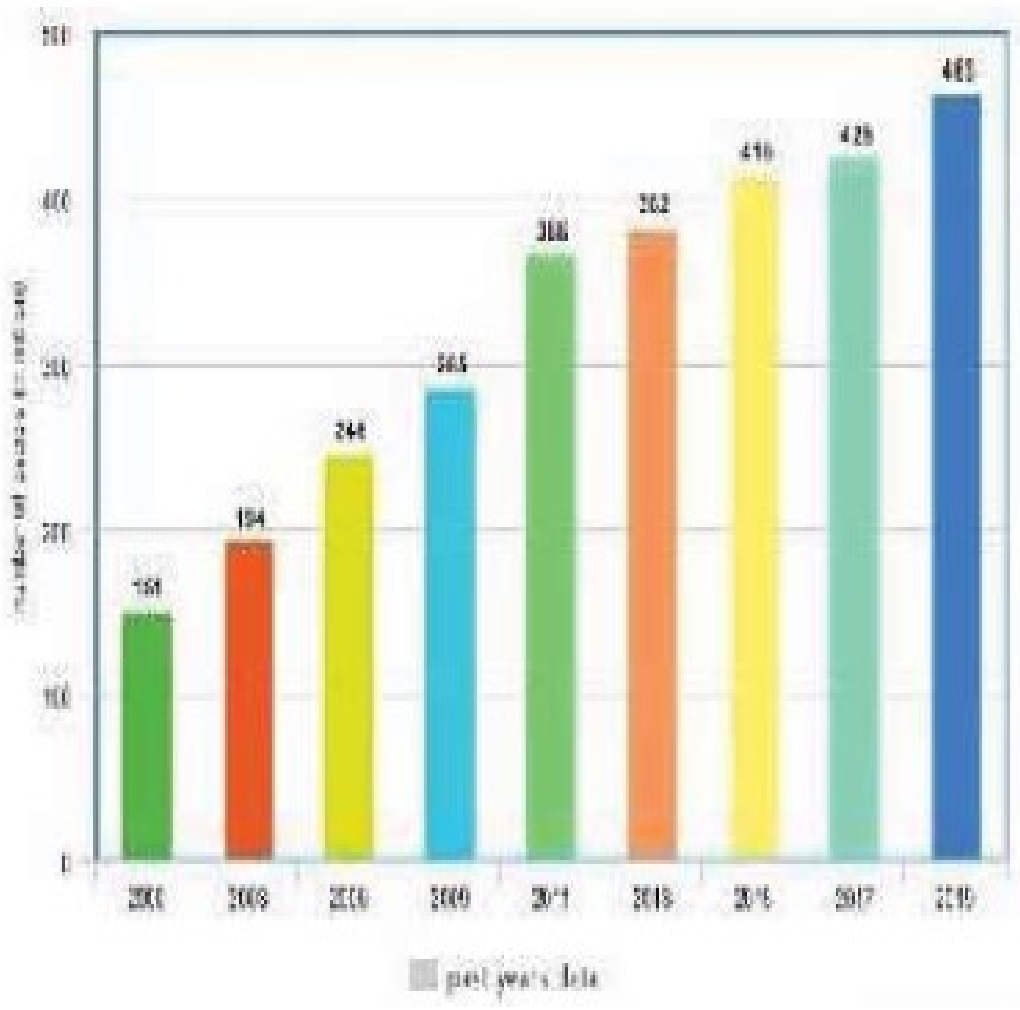


Fig 2 . Estimated number of people with diabetes in world by 2019

2.1 Machine learning in Diabetes

Using machine learning applications in healthcare is contributing a benefit to clinicians and patients for better decision making in less time and cost. The goal of using these applications is to provide quality services within the affordability of the patients. Although healthcare systems are complicated to be evaluated with emerging techniques it has become easier than before. Electronic Health Records(EHR) is widely used now for its security features and also information can be accessed quickly under authorized permissions. Various algorithms are being used combined to make the accuracies of experiments a way better than the existing ones. Well and precisely defined problems must be considered before building a prognosis or diagnosis model with higher accuracy and efficiency.

A better predictive analysis is required at an early stage to determine the severity of a disease. Machine learning algorithms are used as a hybrid model approach that compromises different results of various learning techniques that are used to compare and analyze patterns and accuracy by implementing using various tools which makes it easier for decision making in a medical field.

The machine learning algorithms are classified into two categories as follows-

• **Supervised learning**- with a collection of many training and test datasets that contains both input and desired output to make the algorithm learn for any new cases that are fed into the predictive model to produce a correct output. Classifications, regression, etc... techniques are used here.

• **Unsupervised learning**- used with unlabelled data for inferring the hidden patterns or layers using methods like clustering, feature learning, anomaly detection, dimensionality reduction, etc.

2.2 PIMA Dataset:

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database.

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1)

Body mass index (BMI)

Body mass index (BMI) is a value derived from the mass (weight) and height of a person. The BMI is defined as the body mass divided by the square of the body height, and is expressed in units of kg/m^2, resulting from mass in kilograms and height in metres.

Formula to find BMI is:BMI = weight (kg) / height ^{2}

BMI	Weight Status
Below 18.5	Underweight
18.5–24.9	Healthy
25.0–29.9	Over weight
30.0 and above	Obese

Table 2 : BMI Level Indicator

Glucose Level

The blood glucose level is the amount of glucose in the blood. Glucose is a sugar that comes from the foods we eat, and it's also formed and stored inside the body. It's the main source of energy for the cells of our body, and it's carried to each cell through the bloodstream.

Blood Pressure

Blood pressure is the force of your blood pushing against the walls of your arteries. Each time your heart beats, it pumps blood into the arteries. Your blood pressure is highest when your heart beats, pumping the blood. This is called systolic pressure. When your heart is at rest, between beats, your blood pressure falls. This is called diastolic pressure.

Your blood pressure reading uses these two numbers. Usually the systolic number comes before or above the diastolic number. For example, 120/80 means a systolic of 120 and a diastolic of 80.

Skinfold Thickness

An anthropometric measurement used to evaluate nutritional status by estimating the amount of subcutaneous fat. Calibrated calipers are used to measure the thickness of a fold of skin at defined body sites that include upper arm or triceps, subscapular region, and upper abdomen.

Fig 3: Calibrated Caliper

Diabetes Pedigree Function

Diabetes pedigree function (a function which scores likelihood of diabetes based on family history) Age: Age (years) Outcome: Class variable (0 if non-diabetic, 1 if diabetic).

- In Our Own dataset we had taken binary values to our dataset.

  For Example If "YES" means represented as "1"

  If "NO" means represented as "0"

Parameter	Represented as
BMI	Numeric
Habites(Smoking/Drinking)	Binary(0/1)
Walking	Binary(0/1)
Yoga	Binary(0/1)
Stress	Binary(0/1)
Sleeping	Binary(0/1)
Outcomes	Nominal

Table 2: Own Dataset

2.3 Random Forest Algorithm

- Random forests are a scheme proposed by Leo Breiman in the 2000's for building a predictor ensemble with a set of decision trees that grow in randomly selected subspaces of data.
- As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."
- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique.
- It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process

of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

- Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
- The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.
- Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction

2.3.1 Assumptions for Random Forest

- Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier.
- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

2.3.2 Use Random Forest

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

2.3.3 Working of the Random Forest algorithm:

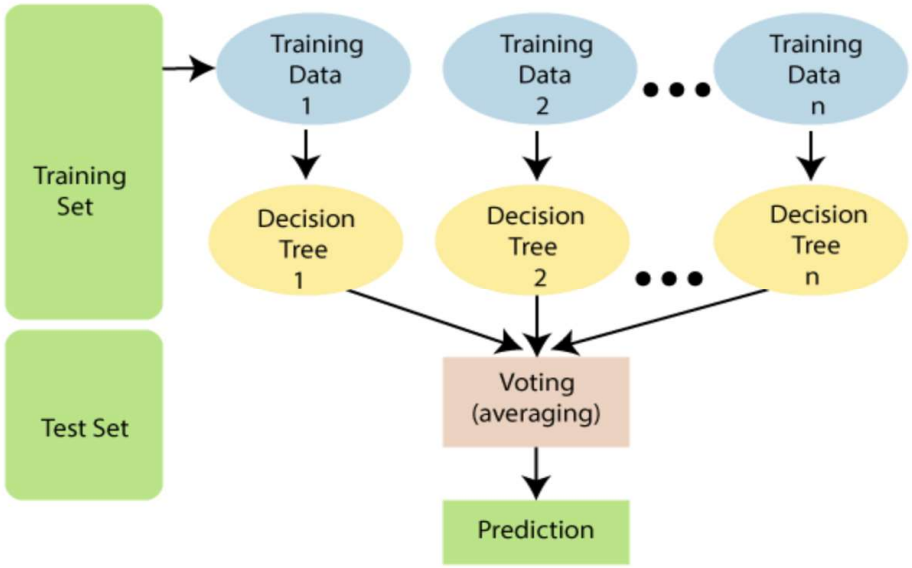


Fig 4: Working Of Random Forest Algorithm

It works in four steps:

- Select random samples from a given dataset.
- Construct a decisiontree for each sample and get a prediction result fromeach decision tree.
- Perform a vote for each predicted result.

- Select the prediction result with the most votes as the final prediction.

2.3.4 Applications of Random Forest

There are mainly four sectors where Random forest mostly used:

- **Banking:** Banking sector mostly uses this algorithm for the identification of loan risk.
- **Medicine:** With the help of this algorithm, disease trends and risks of the disease can be identified.
- **Land Use:** We can identify the areas of similar land use by this algorithm.
- **Marketing:** Marketing trends can be identified using this algorithm.

2.3.5 Advantages of Random Forest

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.
- It works well with both categorical and continuous values.
- It automates missing values present in the data.

2.3.6 Disadvantages of Random Forest

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.
- The main limitation of random forest is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions

From Fig 1 Flow chat explanation

- **Data set**-A dataset can be viewed as a collection of data objects, which are often also called as a records, points, vectors, patterns, events, cases, samples, observations, or entities. Data objects are described by a number of features, that capture the basic characteristics of an object, such as the mass of a physical object or the time at which an event occurred, etc.. Features are often called as variables, characteristics, fields,a ttributes, or dimensions.
- **Data preprocessing**-In any Machine Learning process, Data Preprocessing is that step in which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it. In other words, the features of the data can now be easily interpreted by the algorithm.
- **Feature extraction and feature selection** -Feature extraction and feature selection essentially reduce the dimensionality of the data, but feature extraction also makes the data more separable.
- **Selection-**Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.

- **Extraction**-Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process

- **Training Data and Test Data**-The training data set in Machine Learning is used to train the model for carrying out abundant actions. Detailed features are fetched from the training set to train the model. These structures are therefore combined into the prototype. In sentiment analysis, single words or sequences of consecutive words are taken from the tweets. Therefore, if the training set is labelled correctly, then the model will be able to acquire.

Chapter 3

3. Tool Description

3.1 Python

Python, designed by Guido van Rossum and developed by the Python SoftwareFoundation is a renowned programming language. It was developed in 1991 and it has become one of the most popular high level programming languages.

- It follows simple syntax that is easily understandable even by those who are not from a programming background. It supports scalable and reliable development. The commands are written as sentence structures, allowing the user to get a brief of the idea just by a glance. In comparison with other languages python has the least redundancies.
- It is supported by all the major operating systems such as Windows, MacOS, Linux and UNIX. There is no step for compiling, hence it makes the edit-test-debug cycle astonishingly faster than other languages. If the program doesn't catch any exception, the interpreter prints a stack trace, making the task easy for the programmer.
- Python offers a variety of data structures to ease our work. It also supports various libraries that can be imported using the import command. It uses white spacing for structuring the code, making it easier to read and understand. Thus, Python is the most preferred programming language for Machine Learning.
- Python is Interpreted − The programmer does not have to compile the programbefore executing it. This is similar to PERL and PHP
- Python is Interactive − Python prompt is very helpful, allowing the programmer to interact and assists while typing the code.

- Python is Object-Oriented − It is an Object-oriented programming language that uses classes.
- Python is a Beginner's Language − Python is a considered the best language to begin with, in the journey of programming. Its easy-to-use features allows the beginners to understand the OOPS concepts thoroughly.

3.2 JUPYTER

The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter.

Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use.

Installation

If so, then you can use a handy tool that comes with Python called **pip** to install Jupyter Notebook like this:

```
$pip install jupyter
```

Starting the Jupyter Notebook Server

Now that you have Jupyter installed,Then just go to that location in your terminal and run the following command:

```
$jupyter notebook
```

Creating a Notebook

Now that you know how to start a Notebook server, you should now create an actual Notebook document.

All you need to do is click on the *New* button (upper right), and it will open up a list of choices. On my machine, I happen to have Python 3 installed, so I can create a Notebook that uses either of these. For simplicity's sake, let's choose Python 3.

Your web page should now look like this:

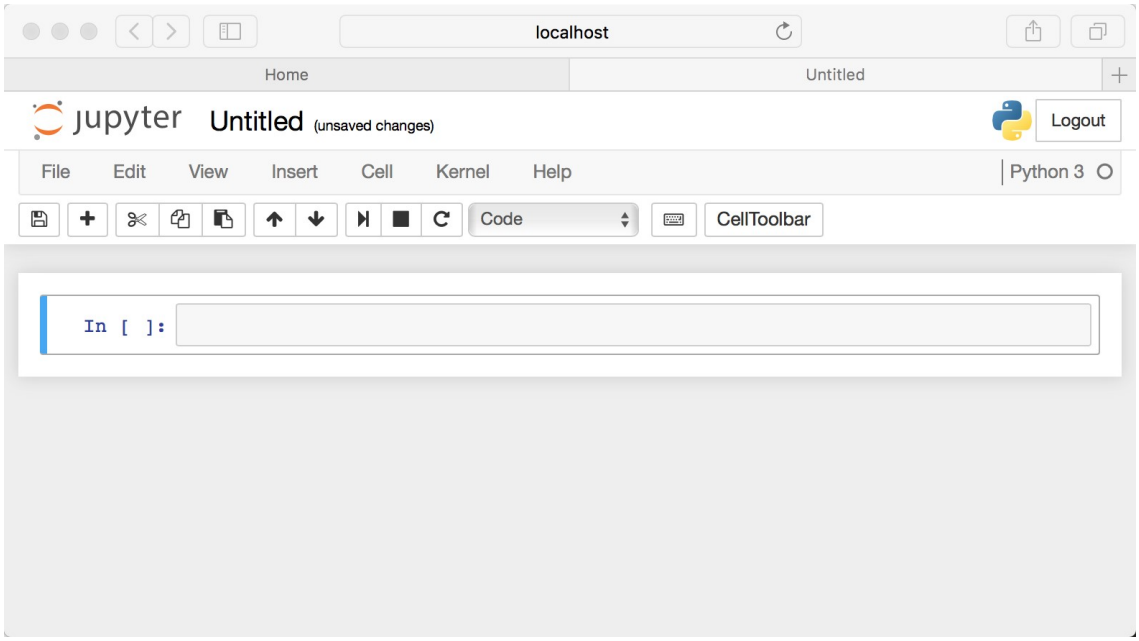


FIG 5: JUPYTER OVERVIEW

Running Cells

A Notebook's cell defaults to using code whenever you first create one, and that cell uses the kernel that you chose when you started your Notebook.

In this case, you started yours with Python 3 as your kernel, so that means you can write Python code in your code cells. Since your initial Notebook has only one empty cell in it, the Notebook can't really do anything.

Thus, to verify that everything is working as it should, you can add some Python code to the cell and try running its contents.

Let's try adding the following code to that cell:

```
print('Hello Jupyter!')
```

When I ran the code above, the output looked like this:

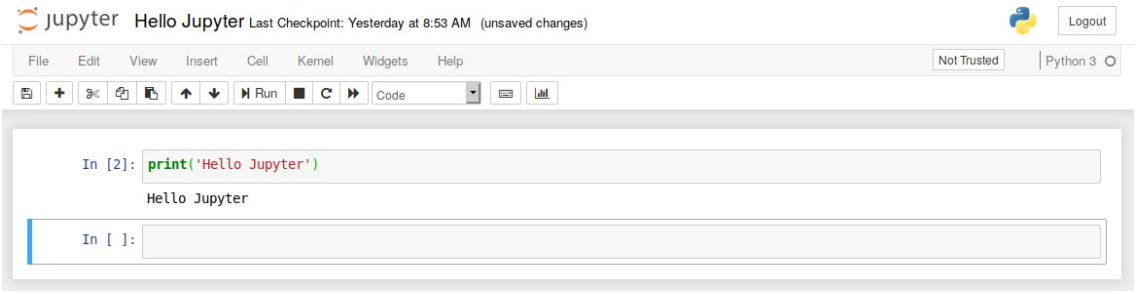


Fig 6: Hello Jupyter Output

The Menus

The Jupyter Notebook has several menus that you can use to interact with your Notebook. The menu runs along the top of the Notebook just like menus do in other applications. Here is a list of the current menus:

- *File*
- *Edit*
- *View*
- *Insert*
- *Cell*

- *Kernel*
- *Widgets*
- *Help*

Let's go over the menus one by one. This article won't go into detail for every single option in every menu, but it will focus on the items that are unique to the Notebook application.

The first menu is the **File** menu. In it, you can create a new Notebook or open a preexisting one. This is also where you would go to rename a Notebook. I think the most interesting menu item is the *Save and Checkpoint* option. This allows you to create checkpoints that you can roll back to if you need to.

Next is the ***Edit*** menu. Here you can cut, copy, and paste cells. This is also where you would go if you wanted to delete, split, or merge a cell. You can reorder cells here too.

The ***View*** menu is useful for toggling the visibility of the header and toolbar. You can also toggle *Line Numbers* within cells on or off. This is also where you would go if you want to mess about with the cell's toolbar.

The ***Insert*** menu is just for inserting cells above or below the currently selected cell.

The ***Cell*** menu allows you to run one cell, a group of cells, or all the cells. You can also go here to change a cell's type, although I personally find the toolbar to be more intuitive for that.

The ***Kernel*** cell is for working with the kernel that is running in the background. Here you can restart the kernel, reconnect to it, shut it down, or even change which kernel your Notebook is using.

The ***Widgets*** menu is for saving and clearing widget state. Widgets are basically JavaScript widgets that you can add to your cells to make dynamic content using Python (or another Kernel).

Finally you have the ***Help*** menu, which is where you go to learn about the Notebook's keyboard shortcuts, a user interface tour, and lots of reference material.

Chapter 4

4. Implementation

4.1 Software Implementation:

For execution of Program written in Python Language using Machine Learning algorithm i.e. Random Forest Classifier is done in JupyterNotebook,which is a Open Source Software.

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Steps to open Jupyter Software

- Firstly in laptop/computer in search bar we have to type jypyter then we find a jupyter icon,by clicking on it we open Jupyter Notebook.

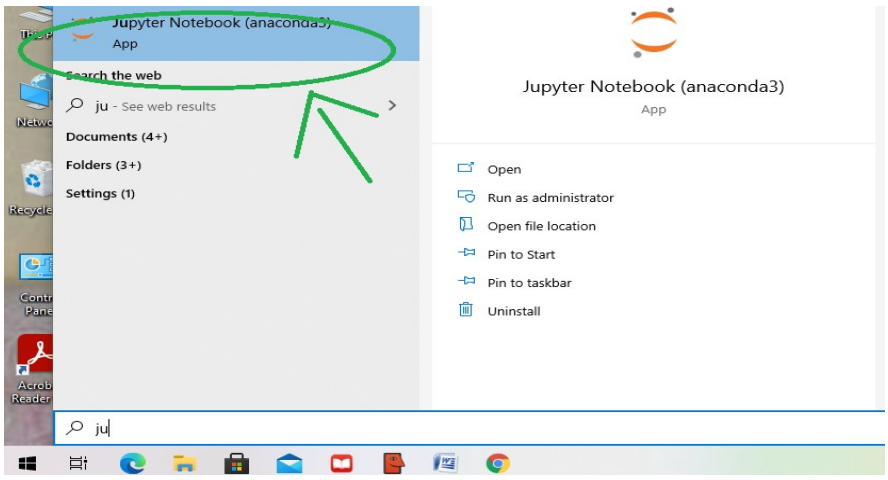


Fig 7: Searching Jupyter in Search Bar

- After clicking on the above we are navigate to jupyter page, which consists of all files that already exist in our local computer.

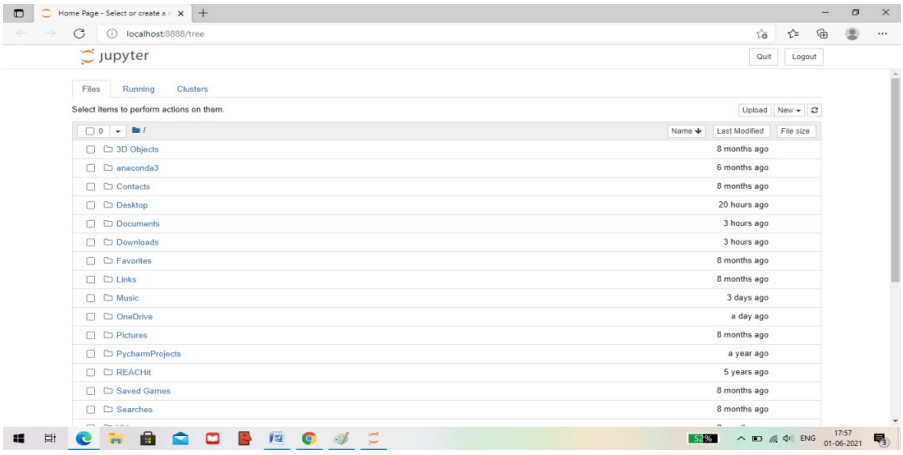


Fig 8: Overview of Jupyter Page

- From the above we picture we see new on right top of the screen.By clicking new we come to see like Python3,Textfile,Folder,Terminal.In that for programming the code we need to select Python3.

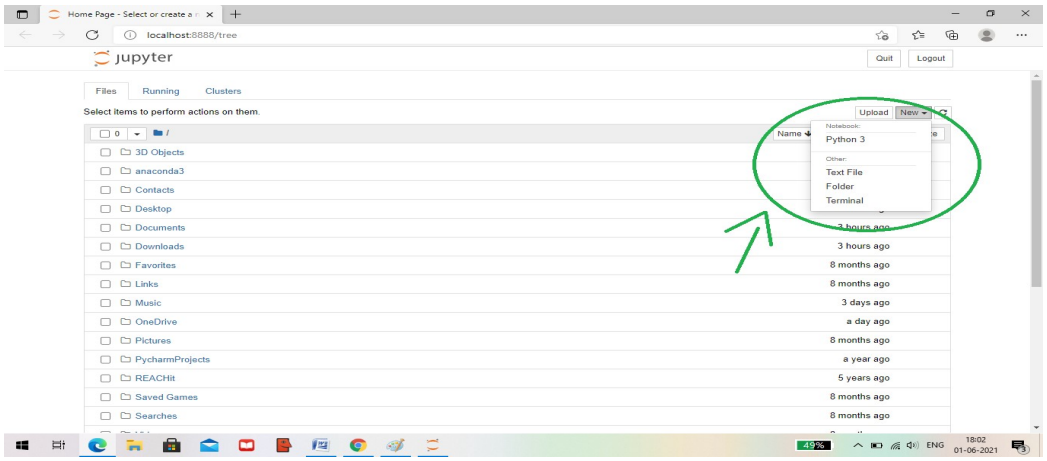


Fig 9: Overview of New options

- After selecting Python 3,we get a main page where we can write a code for programming.

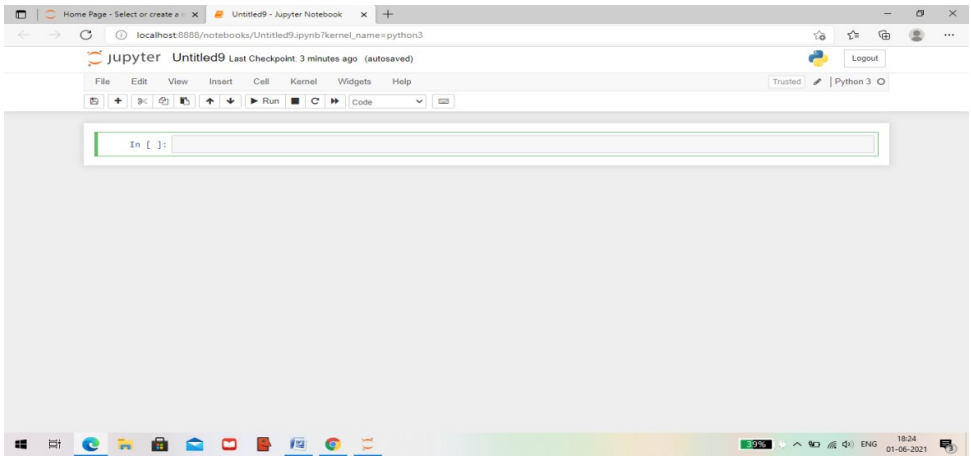


Fig 10: Overview of Jupyter

4.2 Uploading PIMA Dataset to Jupyter:

- Firstly,without uploading any dataset we can not use it in program, it is necessary to upload dataset to use in program.Now we have to open jupyter page in right to we can see upload option by clicking on it,we have to upload file i.e where the orginal pima dataset available in local computer folder by clicking on it we can download a any dataset to jupter notebook.

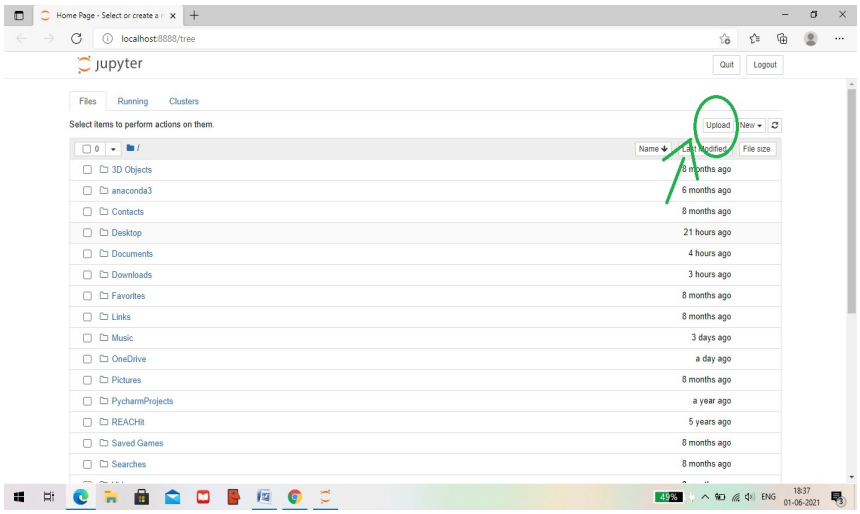


Fig 11: Upload Option Overview

- By clicking on it we get directed into local computer folders from there we have to download PIMA and Our Own Datasets.
- After uploading dataset that folder will be seen jupyter page.

Fig 12: Uploaded File Of PIMA Dataset

From the above 6 months indicate when we uploaded the dataset and 23.91kB indicate size of dataset.

- Similarly,We can upload our Own data Set.

4.3 Libraries Used in Python

- Import pandas as pd

  The above library is used to read the .csv files.

- Import matplotlib.pyplot as plt

The above library is used to plot the graphs.

- Import numpy as np

  The above library is used for Mathematical calculations.

```
In [1]: import pandas as pd
        import matplotlib.pyplot as plt
        import numpy as np

        %matplotlib inline
```

Fig 13: Libraries

- df.shape

  The above indicates the number of rows and columns of a dataset.

- df.head()

  Now that we have downloaded the data set from its source and converted that into a pandas Dataframe, let's display a few records from this dataframe. For this we will use the head() method.

```
In [3]: df.shape
Out[3]: (768, 9)

In [4]: df.head(5)
Out[4]:
```

	pregnancies	glucose	blood pressure	skin thickness	insulin	bmi	diabetes pedigree	age	outcomes
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Fig 14: Output of wanted number of data

- Import seaborn as sns

  The above library is used to draw heat map effectively.

  A heatmap is a plot of rectangular data as a color-encoded matrix. As parameter it takes a 2D dataset. That dataset can be coerced into an

ndarray. This is a great way to visualize data, because it can show the relation between variabels including time.

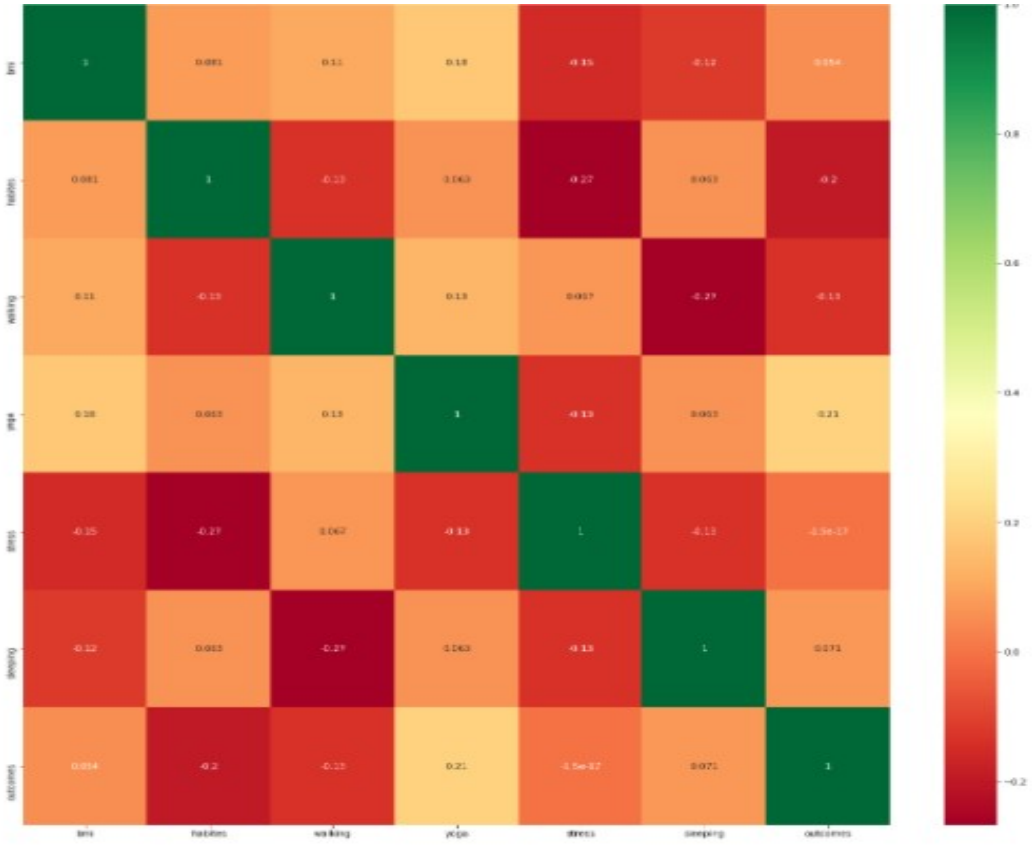


Fig 15: Heat map matrix

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees.

There needs to be some actual signal in our features so that models built using those features do better than random guessing.The predictions(and therefore the errors) made by individual trees need to have low correlations with each other.

4.4 sklearn algorithm in Python:

Sklearn-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

Scikit-learn is probably the most useful library for machine learning in **Python**. The **sklearn** library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

The following are used to import the Random Forest Classifier:

From sklearn ensemble import RandomForestClassifier

random_forest_model=RandomForestClassifier(random_state = 10)

random_forest_model fit(X_train,Y_train ravel())

```
In [11]: from sklearn.ensemble import RandomForestClassifier
         random_forest_model = RandomForestClassifier(random_state=10)

         random_forest_model.fit(X_train, y_train.ravel())
```

Fig 16: Importing Classifier

From the above ravel() is defined as the is used to create a contiguous flattenedarray.A 1-D array, containing the elements of the input, is returned

```
In [12]: predict_train_df = random_forest_model.predict(X_test)

         from sklearn import metrics

         print("Accuracy = {0:.3f}".format(metrics.accuracy_score(y_test, predict_train_df)))
```

Fig 17: Importing Metrics

From the above metrics is used to calculate the accuracy.

The Jupyter system supports over 100 programming languages (called "kernels" in the Jupyter ecosystem) including Python, Java, R, Julia, Matlab, Octave, Scheme, Processing, Scala, and many more.

In Programming we import many libraries such as, pandas as pd,matplotlib.pyplotasplt,numpyasnp,importtrain_test_split.

In this Jupyter we can have a line to line Error check and automatic save of the Programming written in any language.

We can have numericals,tables,pie chart in this implementation which is easy to understand .

Chapter 5

5. Results And Discussion

Minimum error is to be focused on a well-balanced training and testing data sets. After experimenting with the data set, if the outcome achieved is better than the existing works then other data sets can be considered for evaluations. The limitations to the previous works exist in many factors limiting to the dataset and methodologies used. Here the review shows that the Pima data set is widely used for the experiments but it is limited to only nine attributes. As diabetes is already a deadly disease in today's world, the parameters need a major focus to work on. The common symptoms cannot be avoided and limited to only a few attributes won't achieve better accuracy for real-world issues.

Till now we are working on PIMA dataset to obtain from the basic methodology in Machine Learning.The accuracy of the Random Forest Classifier algorithm using Machine Learning is 0.765.Now we are concentrating some parameters to identify the diabetes in early stage.

```
In [22]: predict_train_df = random_forest_model.predict(X_test)

         from sklearn import metrics

         print("Accuracy = {0:.3f}".format(metrics.accuracy_score(y_test, predict_train_df)))

         Accuracy = 0.745
```

Fig 18:Output of PIMA Dataset

In our own constructed dataset the accuracy we got by using Random Forest Classifier is 0.667.

```
In [17]: predict_train_df = random_forest_model.predict(X_test)

         from sklearn import metrics

         print("Accuracy = {0:.3f}".format(metrics.accuracy_score(y_test, predict_train_df)))

         Accuracy = 0.667
```

Fig 19: Output of Own Dataset

Random Forest have been used for experimentation on JUPYTER Notebook tool to predict Diabetes disease.

CONCLUSIONS AND FUTURE SCOPE

Exploring and reviewing various research findings infers that a single algorithm or method is not enough for a precise study. So, to achieve this goal many researchers are approaching new smart hybrid techniques to enhance their proposals with different techniques by evaluating and comparing their performance with the previous researches. The integration of classification algorithms and clustering with other technologies such as IoT, Cloud computing, etc... helps build intelligent systems and monitoring tools.

As we detect the accuracy above 60 percentage he/she may have changes exposed to Diabetes at early stage.If he/she have accuracy below 50 percentage have less chances to get exposed to diabetes.Early detection helps in taking care of health and avoiding to get exposed to diabetes.

In the future, we need more effective systems by taking the real-world data sets into account for analysis. There are still many complications and limitations which need more focus so the identified problems can be easily solved and can assist further research.

REFERENCES

Prediction of Diabetes Risk based on Machine Learning Techniques by Madhusmita Rout and AmandeepKaur,*Department of computer science and engineering,Lovely Professional University*

Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare byAyman Mir and Sudhir N. Dhage , *Department of Computer Engineering Sardar Patel Institute of Technology Mumbai, India.*

American Diabetes Association.(2019). 2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2019. Diabetes Care, 42(Supplement 1), S13-S28.

Rashid, T. A., & Abdullah, S. M. (2018).A Hybrid of Artificial Bee Colony, Genetic Algorithm, and Neural Network for Diabetic Mellitus Diagnosing. ARO-The Scientific Journal of Koya University, 6(1), 55-64.

Aada, M. T. S. A., &Tiwari, S. (2019). Predicting Diabetes in Medical Datasets Using Machine Learning Techniques

Shakeel, P. M., Baskar, S., Dhulipala, V. S., &Jaber, M. M. (2018). Cloud based framework for diagnosis of diabetes mellitus using Kmeans clustering. Health information science and systems, 6(1), 16.

APPENDIX – I

Python program for PIMA Dataset

```
import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

%matplotlib inline

df = pd.read_csv(r"pima-indians-diabetes.csv")

df.shape

df.head(5)

df.isnull().values.any()

import seaborn as sns

import matplotlib.pyplot as plt

corrmat = df.corr()

top_corr_features = corrmat.index

plt.figure(figsize=(20,20))

g=sns.heatmap(df[top_corr_features].corr(),annot=True,cmap="RdYlGn")

outcomes_true_count = len(df.loc[df['outcomes'] == True])

outcomes_false_count = len(df.loc[df['outcomes'] == False])

(outcomes_true_count,outcomes_false_count)

from sklearn.model_selection import train_test_split

feature_columns = ['pregnancies', 'glucose', 'blood pressure', 'insulin', 'bmi',
'diabetes pedigree', 'age', 'skin thickness']

predicted_class = ['outcomes']
```

```
X = df[feature_columns].values

y = df[predicted_class].values

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30,
random_state=10)

from sklearn.ensemble import RandomForestClassifier

random_forest_model = RandomForestClassifier(random_state=10)

random_forest_model.fit(X_train, y_train.ravel())

predict_train_df = random_forest_model.predict(X_test)

from sklearn import metrics

print("Accuracy = {0:.3f}".format(metrics.accuracy_score(y_test,
predict_train_df)))
```

APPENDIX – II

Python program for OWN Dataset

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
%matplotlib inline
df = pd.read_csv(r"project1.csv")
df.shape
df.head(3)
df.isnull().values.any()
import seaborn as sns
import matplotlib.pyplot as plt
corrmat = df.corr()
top_corr_features = corrmat.index
plt.figure(figsize=(20,20))
g=sns.heatmap(df[top_corr_features].corr(),annot=True,cmap="RdYlGn")
outcomes_true_count = len(df.loc[df['outcomes'] == True])
outcomes_false_count = len(df.loc[df['outcomes'] == False])
(outcomes_true_count,outcomes_false_count)
from sklearn.model_selection import train_test_split
feature_columns = ['bmi', 'habites', 'walking', 'yoga', 'stress', 'sleeping']
predicted_class = ['outcomes']
X = df[feature_columns].values
```

```
y = df[predicted_class].values

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30,
random_state=10)

from sklearn.ensemble import RandomForestClassifier

random_forest_model = RandomForestClassifier(random_state=10)

random_forest_model.fit(X_train, y_train.ravel())

predict_train_df = random_forest_model.predict(X_test)

from sklearn import metrics

print("Accuracy = {0:.3f}".format(metrics.accuracy_score(y_test,
predict_train_df)))
```

STUDENT NAME	EMAIL ID	ADDRESS	CELL NO	LAND LINE	PLACEMENT	PHOTO
Pamaluru Yaswanth Kumar Reddy	yaswanthkumar9171@gmail.com	Ayyavaripali Village, Vempalli(m) YSR Kadapa (DIST),AP 516350	8867635893		Not Placed	
Gokula Bhavya Tej	bhavyatejgokula@gmail.com	D/NO-3/60, Venkatadri Nagar, Proddatur, YSR Kadapa (DIST),AP 516360	9550404692		Not Placed	
M S Mohamed Ashfaq	shaikashwak2790@gmail.com	D/NO-6/118, Nehru Street,Piler, Chittoor (DIST),AP 517214	7032383220		Not Placed	
Mallela Vishnu	vishnutony66@gmail.com	D/NO-4/998, Pappampet, KLD Road, Anantapur (DIST),AP 515004	9493793109		Not Placed