NEW HORIZON
COLLEGE OF ENGINEERING
New Horizon Knowledge Park, Ring Road, Marathalli
Autonomous College Permanently Affiliated to VTU, Approved by AICTE & UGC
Accredited by NAAC with 'A' Grade, Accredited by NBA

A

MINI PROJECT REPORT

ON

"DIABETES PREDICTION SYSTEM"

Submitted in the partial fulfillment of the requirements in the V semester of

BACHELOR OF ENGINEERING

IN

INFORMATION SCIENCE AND ENGINEERING

BY

AKHILA S - [1NH17IS008]

COURSE NAME: MINI PROJECT
COURSE CODE: ISE57

Under the guidance of

Dr. S Mohan Kumar
Professor
Dept. of ISE, NHCE

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

NEW HORIZON COLLEGE OF ENGINEERING
(Autonomous College Permanently Affiliated to VTU, Approved by AICTE, Accredited by NBA & NAAC with 'A' Grade)
New Horizon Knowledge Park, Ring Road, Bellandur Post, Near Marathalli,
Bangalore-560103, INDIA

NEW HORIZON
COLLEGE OF ENGINEERING
New Horizon Knowledge Park, Ring Road, Marathalli
Autonomous College Permanently Affiliated to VTU, Approved by AICTE & UGC
Accredited by NAAC with 'A' Grade, Accredited by NBA

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

CERTIFICATE

I hereby certify that, the report entitled **"Diabetes Prediction System"** as a part of Mini Project Component in partial fulfillment of the requirements during 5th semester Bachelor of Engineering in Information Science and Engineering during the year 2019-20(Aug 2019-Nov 2019) is an authentic record of my own work carried out by **Akhila S (1NH17IS008)**, a bonafied student of NEW HORIZON COLLEGE OF ENGINEERING.

Name & Signature of Student

(Ms. Akhila S)

Name & Signature of Guide

(Dr. S Mohan Kumar)

Name & signature of HOD

(Dr. R J Anandhi)

ABSTRACT

Diabetes is one of the deadliest diseases in the world. The general process is that patients need to visit a diagnostic center, consult their doctor, and have to wait to get their reports. And, every time they want to get their diagnosis report, they have to waste their money in vain. Using Machine Learning we can develop an algorithm which will be able to predict if a patient has diabetes or not. Predicting the disease early leads to treating the patients before it becomes critical. The prediction is done based on the given dataset. Also, it will help individuals to get acquainted about their health status and future possible diabetic condition so that they can get chance to adopt better lifestyle to prevent the disease.

One of the supervised machine learning algorithms such as logistic regression can be used to predict if a patient is diabetic or not. Logistic Regression is a supervised binary classification algorithm which gives us answers in the form of yes or no, true or false and so on. In this project we apply this algorithm on our dataset to perform prediction. The dataset is divided into train, test and check sets where the model is first trained using the train model and prediction is done on the check set.

Upon applying logistic regression algorithm on the dataset, we are able to predict if the person has diabetes or not. Diabetes prediction using logistic regression in this project gives an accuracy of 78%. There are many other algorithms which can be used to improve the accuracy.

ACKNOWLEDGEMENT

Any achievement, be it scholastic or otherwise does not depend solely on the individual efforts but on the guidance, encouragement and cooperation of intellectuals, elders and friends. A number of personalities, in their own capacities have helped me in carrying out this mini project. I would like to take this opportunity to thank them all.

I thank the management, **Dr. Mohan Manghnani,** Chairman, New Horizon Educational Institutions for providing necessary infrastructure and creating conducive environment for effective learning.

I also record here the constant encouragement, support and facilities extended to us by **Dr. Manjunatha,** Principal, New Horizon College of Engineering, Bengaluru.

I extent sincere gratitude for constant encouragement and facilities provided to us by **Dr. R.J Anandhi**, Professor and Head of the Department, Department of Information Science and Engineering, New Horizon College of Engineering, Bengaluru.

I sincerely acknowledge the encouragement, timely help and guidance to me by **Dr. S Mohan Kumar,** Professor, Department of Information Science and Engineering, New Horizon College of Engineering, Bengaluru, to complete the mini project within stipulated time successfully.

Finally, a note of thanks to the teaching and non-teaching staff of Information Science and Engineering Department for their cooperation extended to us and our friends, who helped me directly or indirectly in the successful completion of this mini project.

Akhila S

(1NH17IS008)

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

Diabetes is a very serious disease which poses a great threat to human health. Diabetes is caused when the blood glucose is higher than normal level. Diabetes can cause dysfunction of various organs in our body like loss of vision, kidney neuropathy, liver problems, and heart problems. With the current living standards diabetes is increasingly common in people's life. This situation must be diagnosed before it causes major health issues and prediction is the first and important step in the process. This prediction can be done using machine learning. Machine learning has the ability to learn and improve from experience. By learning it means the system is able to understand the input data and make decisions and predictions based on it. In this project diabetes is predicted using one of the machine learning algorithms known as logistic regression.

1.1 Motivation of Project

The World Health Associations annual health report states that the number of people experiencing diabetes is 422 million the year. The number keeps increasing every year. Hence it must be diagnosed at initial stages itself else it may cause various health issues. The normal way to identify if a person has diabetes or not is by visiting a diagnostic center, consult the doctors and wait for their results. Moreover every time they want to get their diagnosis report, they have to spend their money. But with machine learning approach we are developing a system using data mining which will predict if a person has diabetes or not. And predicting the disease early leads to treating the patient before the problem becomes critical.

1.2 Methodology

Machine learning can be described as automating and improving the process of learning of computers based on experience. The process starts by giving inputs to the system, then training the system by building models. It can simply be described as “ability to learn”. Machine learning algorithms can be used for online fraud detection, product recommendation, spam filtering and in many more day to day applications. The algorithms can be classified into 3 types:

a) Supervised Learning: The dataset provided is labeled. The algorithms generally used are classification and regression algorithms.
b) Unsupervised Learning: The dataset provided is not labeled or classified. The algorithms generally used are clustering and grouping algorithms.
c) Reinforcement Learning: They are related to dynamic programming algorithms frequently used to solve optimization problems.

The common machine learning algorithms used are Linear Regression, Logistic Regression, Naïve Bayes, KNN, K-means, Decision Tree and SVM. In this project we make use of Logistic Regression algorithm to predict diabetes.

Logistic Regression is regression model where the dependent variable is categorical, namely binary dependent variable that can take only 2 values that is 0 and 1 which can mean yes/no, true/false, pass/fail, win/lose. In this project we make use of dataset provided by Pima Indian Diabetes Database. It discusses the Indian population’s medical record regarding diabetes. In this dataset there are total 768 instances which can be classified into two classes: diabetic or non-diabetic along with eight factors: Pregnancy (no of times pregnant), Plasma Glucose, Diastolic Blood Pressure, Skin Thickness, Insulin, Body Mass Index, Diabetes Pedigree Function and Age.

	A	B	C	D	E	F	G	H	I
1	Pregnanci	Glucose	BloodPres	SkinThickı	Insulin	BMI	DiabetesF	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1
22	3	126	88	41	235	39.3	0.704	27	0
23	8	99	84	0	0	35.4	0.388	50	0
24	7	196	90	0	0	39.8	0.451	41	1
25	9	119	80	35	0	29	0.263	29	1
26	11	143	94	33	146	36.6	0.254	51	1
27	10	125	70	26	115	31.1	0.205	41	1

diabetes

Fig 1.2.1 First few instances of diabetes dataset

1.3 Problem Statement

To design and implement an application which will predict if a person has diabetes or not using one of the machine learning algorithms known as logistic regression.

Chapter 2

SYSTEM REQUIREMENT & LANGUAGE USED

Purpose: To design and implement an application which will predict if a person has diabetes or not using one of the machine learning algorithms known as logistic regression.

2.1 Hardware System Configuration:

Processor	- Intel core i5
Speed	- 1.8 GHz
RAM	- 256 MB (min)
Hard Disk	- 10 GB

2.2 Software System Configuration:

Operating System	-	Windows 8.1
Programming Language	-	Python
Compiler	-	Anaconda

2.2 About the language used

Anaconda is an open source distribution of Python and R programming which is used for machine learning, data science, and predictive analysis. Anaconda navigator is a desktop GUI which consists of these applications: Jupyter Notebook, Spyder, Rstudio, Orange, and Visual Studio. In this project we make use of the Jupyter Notebook application to write our code in Python. Python is a high level programming language which makes use of object oriented programming concepts.

Features of Python language are:

1. It is user friendly and easy to understand.
2. It is readable.
3. It is platform independent; it can work on any operating system.
4. It is open source.
5. It is an interpreted language where debugging happens line by line.

Chapter 3

METHODOLOGY

3.1 Architecture

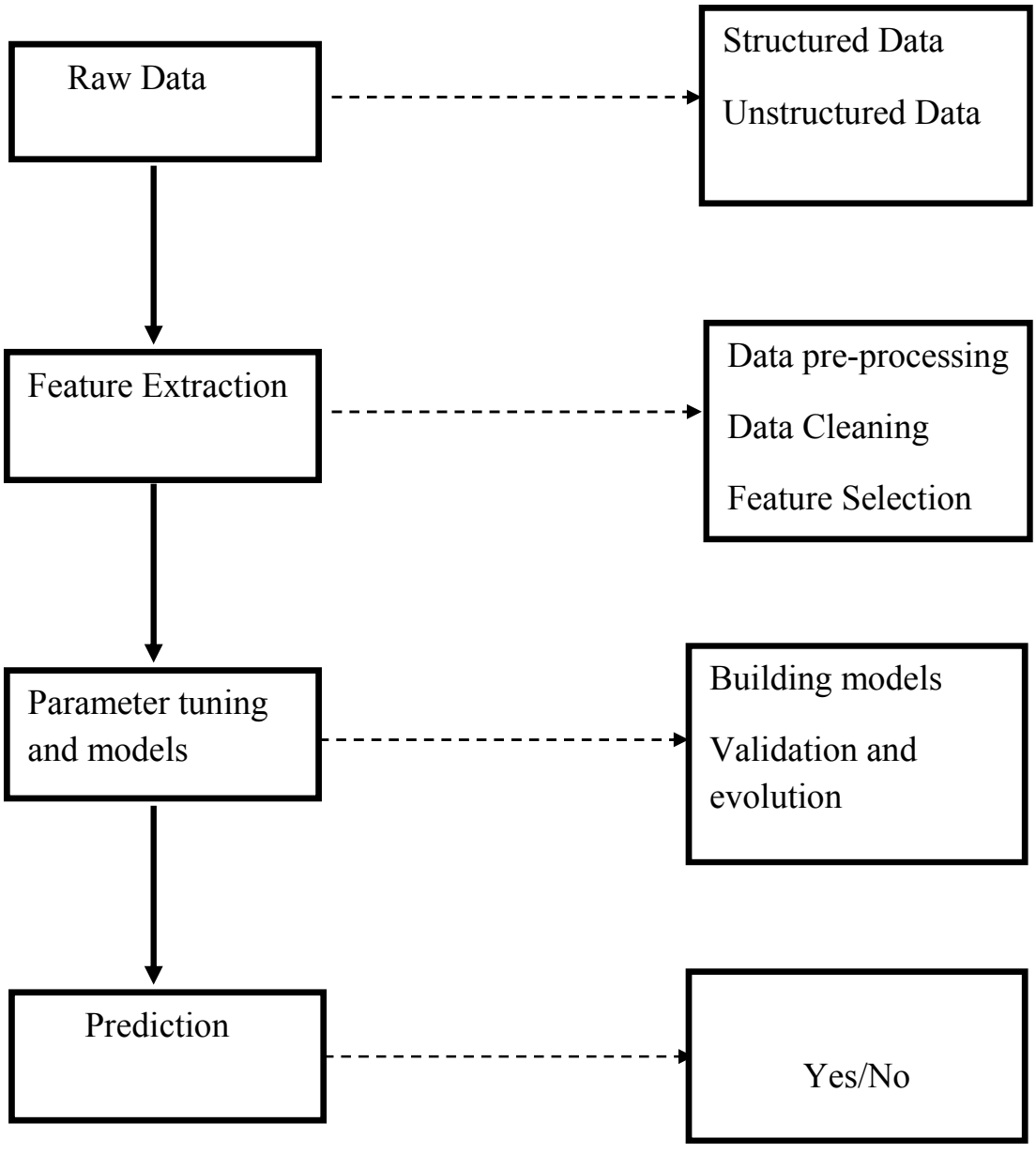


Fig 3.1.1 Architecture

3.2 Algorithm

Step 1: Start

Step 2: Import the necessary library packages such as pandas, numpy, seaborn, matplotlib and sklearn to use their functions.

Step 3: Read the dataset “diabetes.csv” in your program.

Step 4: Print the first five instances of the dataset with all the factors.

Step 5: Check if there are any null values in the imported dataset.

Step 6: Describing the dataset.

Step 7: Find the correlation for all the factors.

Step 8: Plotting the correlation as a heatmap.

Step 9: Counting the number of diabetic and non-diabetic patients using bar plot.

Step 10: Splitting the imported dataset into the following 3 categories

a) 650 instances for training
b) 100 instances for testing
c) 18 instances for checking

Step 11: Separate the label and features and convert them to numpy array for training and test dataset by dropping the “outcome” feature.

Step 12: Normalize the inputs to understand importance of each feature.

Normalize in such a way that each variable has mean is 0, standard deviation 1.

Step 13: Build a logistic regression model.

Step 14: Train the model.

Step 15: Use test data to find accuracy of model.

Step 16: Save the model

Step 17: The test data is used to check the accuracy of the saved model

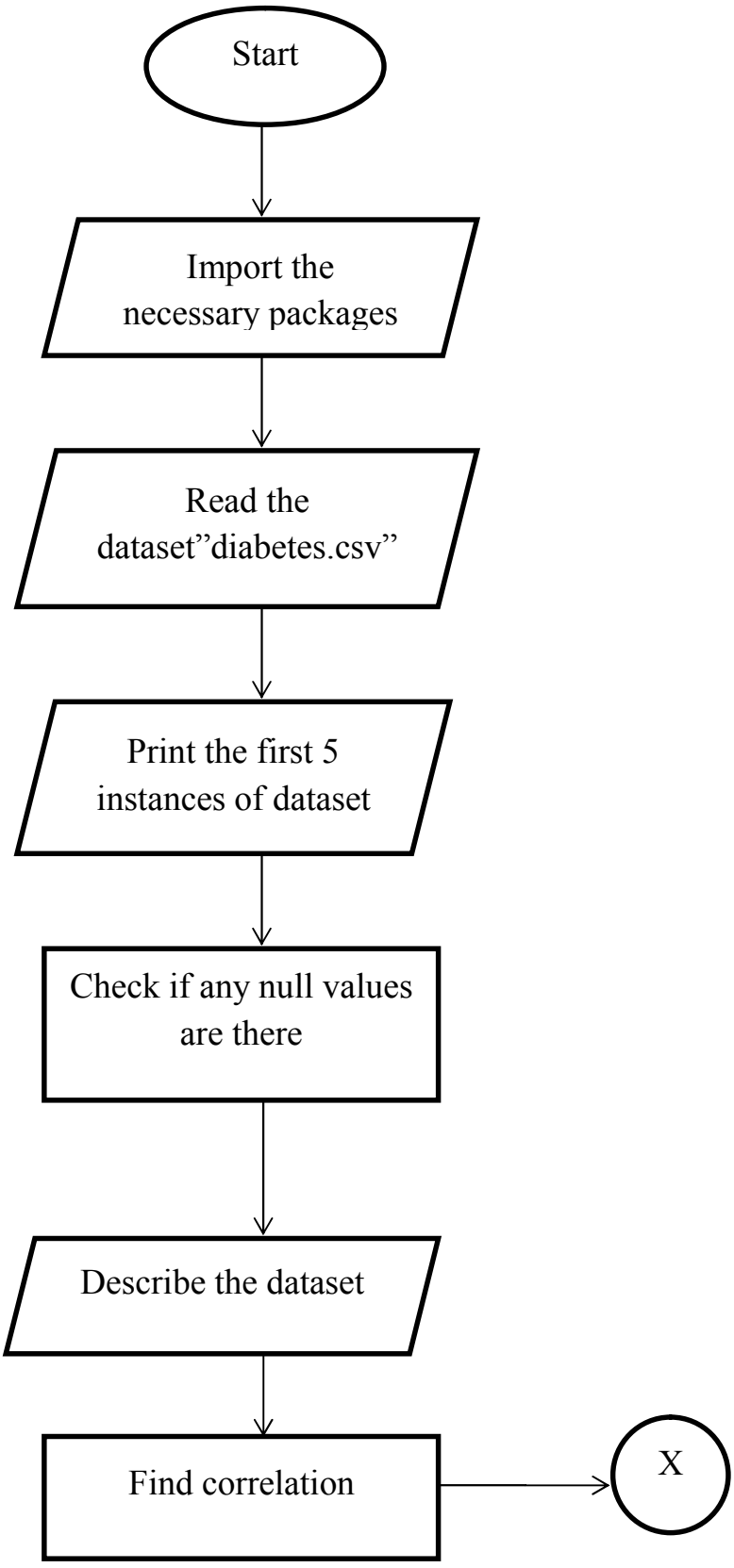
There will be no change in accuracy if the model is saved properly.

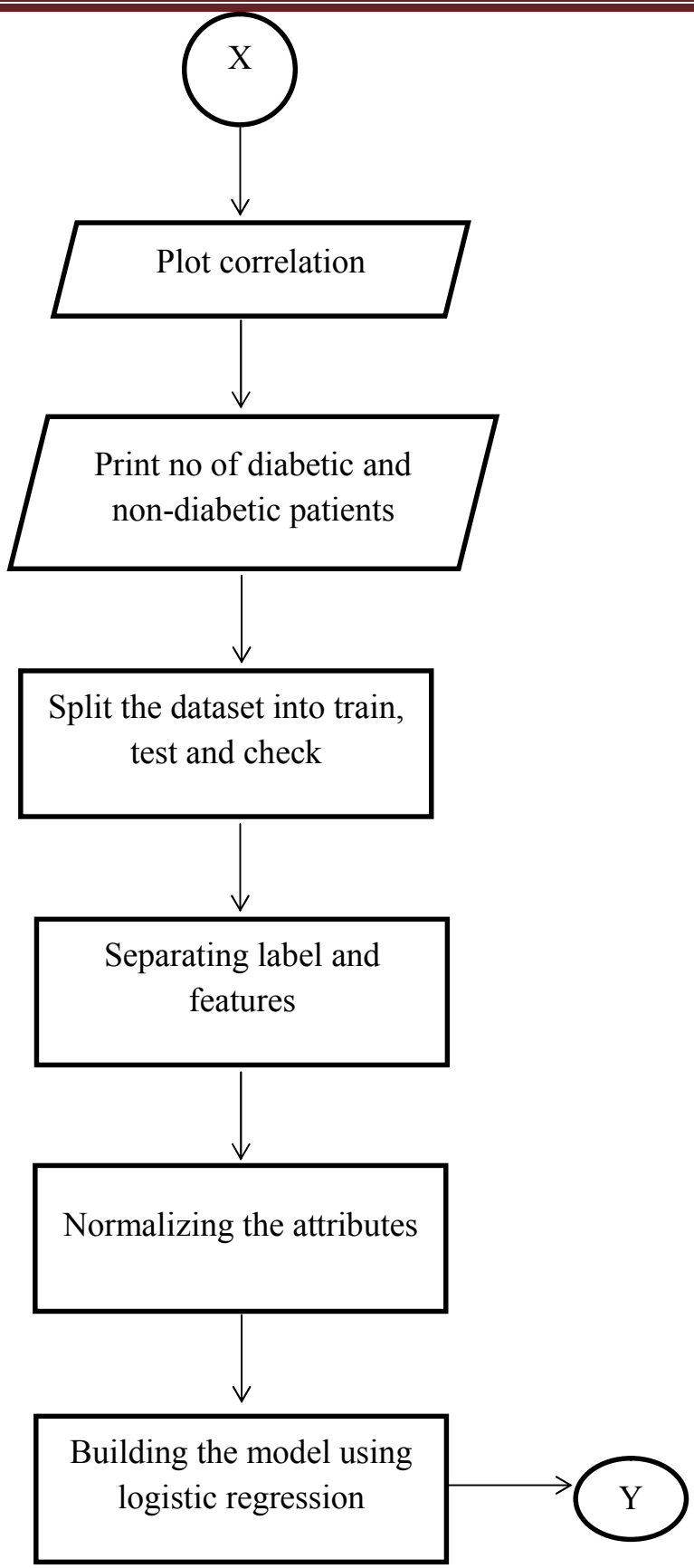
Step 18: Now we can do the prediction using the data from checking dataset.

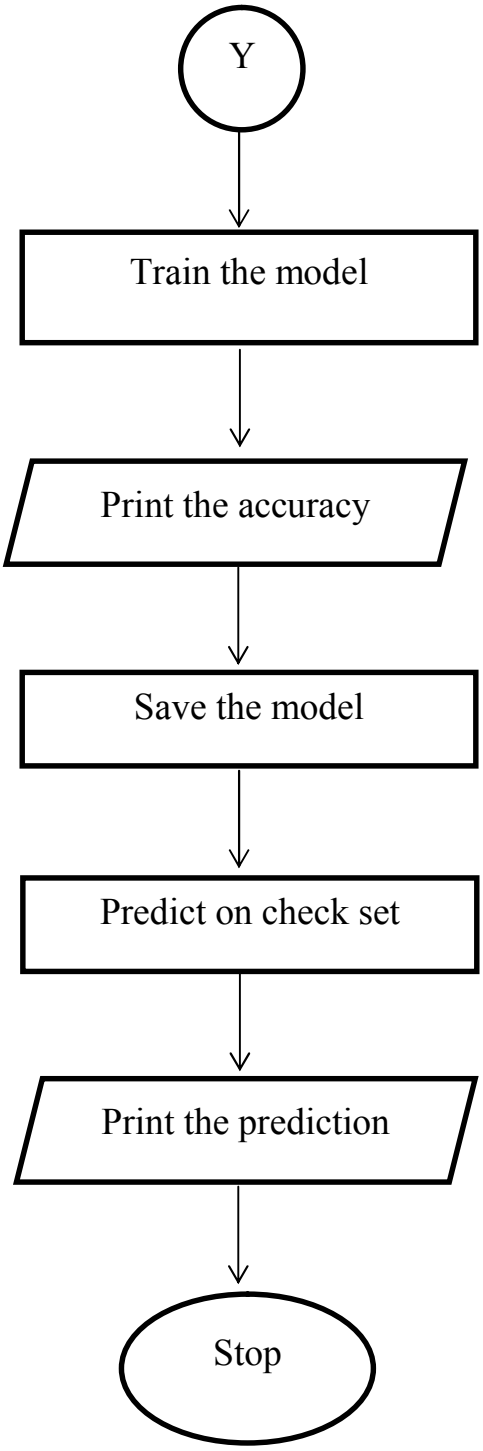
Step 19: We will use the first record in the checking dataset to make the prediction.

Step 20: Stop

3.3 Flowchart




X
Plot correlation
Print no of diabetic and non-diabetic patients
Split the dataset into train, test and check
Separating label and features
Normalizing the attributes
Building the model using logistic regression
Y


Y
Train the model
Print the accuracy
Save the model
Predict on check set
Print the prediction
Stop

3.4 Code and Implementation

```
#importing the library packages

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

% matplotlib inline

from sklearn.linear_model import LogisticRegression

from sklearn.externals import joblib


#reading the dataset

dataframe = pd.read_csv("Documents/diabetes.csv")


#print the first 5 instances of the dataset

dataframe.head()


#checking if there are any null values in the dataset

print (dataframe.isnull().values.any())


#finding the correlation

dataframe.corr()
```

```
#plotting the correlation as heatmap
sns.heatmap(corr,
        xticklabels=corr.columns,
        yticklabels=corr.columns)
#counting the number of diabetic and non-diabetic patients
sns.countplot(y=df['Outcome'],palette='Set1')


#splitting the dataset into train, test and check
dfTrain = dataframe[:650]
dfTest = dataframe[650:750]
dfCheck = dataframe[750:]


#separating the label and features
trainLabel = np.asarray(dfTrain['Outcome'])
trainData = np.asarray(dfTrain.drop('Outcome',1))
testLabel = np.asarray(dfTest['Outcome'])
testData = np.asarray(dfTest.drop('Outcome',1))


#normalizing the attributes
means = np.mean(trainData, axis=0)
```

```
stds = np.std(trainData, axis=0)

trainData = (trainData - means)/stds

testData = (testData - means)/stds


#building a logistic regression model and training it

diabetesCheck = LogisticRegression()

diabetesCheck.fit(trainData, trainLabel)


#finding accuracy

accuracy = diabetesCheck.score(testData, testLabel)

print("accuracy = ", accuracy * 100, "%")


#saving the model

joblib.dump([diabetesCheck, means, stds], 'diabeteseModel.pkl')


#test data used to check accuracy of saved model

diabetesLoadedModel, means, stds = joblib.load('diabeteseModel.pkl')

accuracyModel = diabetesLoadedModel.score(testData, testLabel)

print("accuracy = ",accuracyModel * 100,"%")


print(dfCheck.head())
```

```
sampleData = dfCheck[:1]

# prepare sample

sampleDataFeatures = np.asarray(sampleData.drop('Outcome',1))


sampleDataFeatures = (sampleDataFeatures - means)/stds

# predict

predictionProbability = diabetesLoadedModel.predict_proba(sampleDataFeatures)

prediction = diabetesLoadedModel.predict(sampleDataFeatures)

print('Probability:', predictionProbability)

print('prediction:', prediction)
```

Chapter 4

RESULTS AND DISCUSSION

4.1 Summary of result obtained

The output obtained in this project will indicate if a person has diabetes or not. The output will be displayed in the following format.

1) The first 5 instances of dataset are printed.
2) The attributes of the dataset are described.
3) The visualization (heatmap) of correlation of attributes is printed.
4) Using the bar plot, the count of diabetic and non-diabetic patients is printed.
5) The logistic model is built.
6) The accuracy of the model is found using test dataset.
7) After saving the model, prediction is made on the check set.

4.2 Outputs (Snapshots)

```
In [3]: dataframe.head()
```

Out[3]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
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 4.2.1: To print the first 5 instances

```
In [5]: print (dataframe.isnull().values.any())

        False
```

Fig 4.2.2: To check if there are any null values in the dataset

```
In [4]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
Pregnancies                 768 non-null int64
Glucose                     768 non-null int64
BloodPressure               768 non-null int64
SkinThickness               768 non-null int64
Insulin                     768 non-null int64
BMI                         768 non-null float64
DiabetesPedigreeFunction    768 non-null float64
Age                         768 non-null int64
Outcome                     768 non-null int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Fig 4.2.3: To describe the dataset

```
In [26]: dataframe.corr()
```

Out[26]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

Fig 4.2.4: To find the correlation

```
In [30]: sns.heatmap(corr,
                     xticklabels=corr.columns,
                     yticklabels=corr.columns)

Out[30]: <matplotlib.axes._subplots.AxesSubplot at 0xa820f7d828>
```

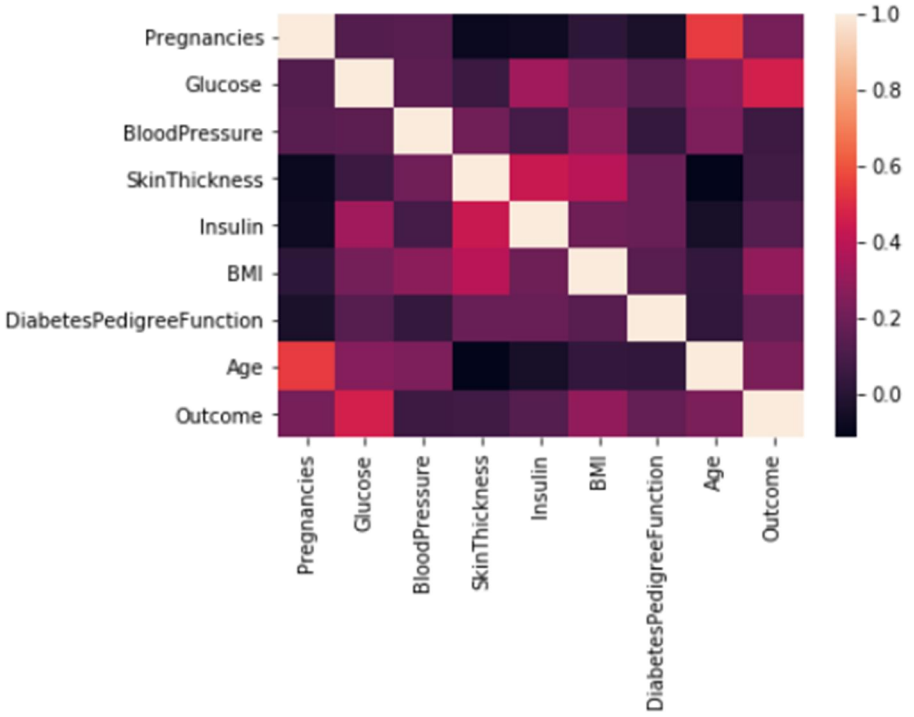


Fig 4.2.5: To plot the correlation as heatmap

```
In [9]: sns.countplot(y=df['Outcome'],palette='Set1')

Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x41ac7ef780>
```

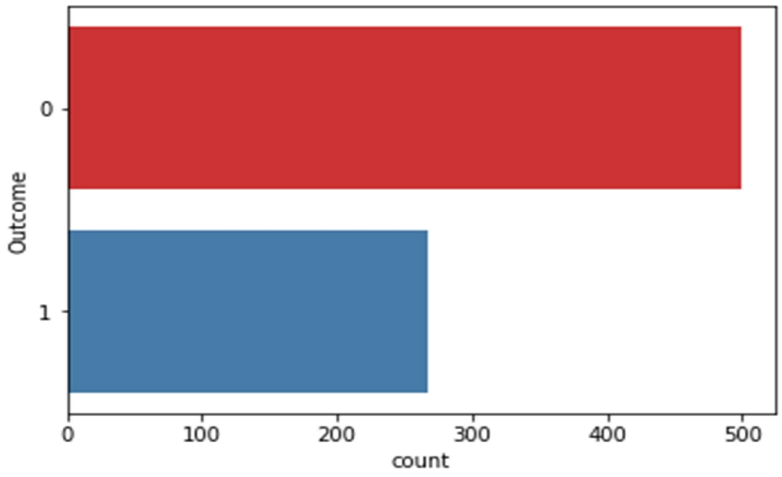


Fig 4.2.6: To count the number of diabetic and non-diabetic patients

```
In [11]: diabetesCheck = LogisticRegression()
         diabetesCheck.fit(trainData, trainLabel)

Out[11]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
                   penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
                   verbose=0, warm_start=False)
```

Fig 4.2.7: To build a logistic regression model and train it

```
In [12]: accuracy = diabetesCheck.score(testData, testLabel)
         print("accuracy = ", accuracy * 100, "%")

         accuracy =  78.0 %
```

Fig 4.2.8: Finding accuracy of the model

```
In [14]: joblib.dump([diabetesCheck, means, stds], 'diabeteseModel.pkl')

Out[14]: ['diabeteseModel.pkl']
```

Fig 4.2.9: To save the train model

```
In [29]: sampleData = dfCheck[:1]
         # prepare sample
         sampleDataFeatures = np.asarray(sampleData.drop('Outcome',1))
         sampleDataFeatures = (sampleDataFeatures - means)/stds
         # predict
         predictionProbability = diabetesLoadedModel.predict_proba(sampleDataFeatures)
         prediction = diabetesLoadedModel.predict(sampleDataFeatures)
         print('Probability:', predictionProbability)
         print('prediction:', prediction)

         Probability: [[0.4385153 0.5614847]]
         prediction: [1]
```

Fig 4.2.10: To make a prediction using check set

Chapter 5:

CONCLUSION

There is no cure for diabetes but we can make an early detection to reduce the long term complications and reduce the costs on treatments. There are millions of people who are still unaware of the fact that they might have diabetes. The ability to predict diabetes early plays a vital role for a patient's treatment strategy. In this project we make use of one of the machine learning algorithms known as logistic regression which predicts if a person has diabetes or not. The accuracy obtained from building a logistic regression model in this project is 78%. This accuracy can be improved further by analyzing other attributes and making use of different combination of feature selection.

REFERENCES

1) Machine Learning

-Tom M Mitchell

2) Prediction Of Onset Diabetes using Machine Learning Techniques

- Md. Aminul Islam,Nusrat Jahan

3) Logistic Regression and SVM based Diabetes Prediction System

- Tejas N Joshi,Prof. Pramila M Chawan

4) An Analysis of Predicting Diabetes using Machine Learning

- Ujjwal Anand,Dr. Amit Sehgal,Shashank Tripathi

5) A Survey: Detection and Prediction of Diabetes using machine learning techniques

- Priyanka Indoria,Yogesh Kumar Rathore

6) Analysis of Diabetes Mellitus for early prediction using optimal features selection

- N Sneha, Tarun Gangil

7) Analysis and Prediction Of Diabetes using Machine Learning techniques

- S Subashree, S Saru

8) Predicting Diabetes in medical datasets using machine learning techniques

- Uswa Ali Zia, Dr. Naeem Khan

NEW HORIZON COLLEGE OF ENGINEERING

Autonomous College Permanently Affiliated to VTU, Approved by AICTE & UGC
Accredited by NAAC with 'A' Grade, Accredited by NBA

LIBRARY & INFORMATION CENTRE

CERTIFICATE ON PLAGIARISM CHECK

1	Name of the Student	AKHILA S
2	USN	1NH17IS008
3	Course	UG
4	Department	ISE
5	Mini Projects/Project Reports/ Ph.D Synopsis/Ph.D Thesis/ Paper Publication (N/I)	Mini Project
6	Title of the Document	Diabetes Prediction System
7	Similar Content(%)identified	5%
8	Acceptable maximum limit (%) of similarity	30%
9	Date of Verification	05.11.2019
10	Checked by (Name with signature)	Mr. Vinayak Kubihal
11	Specific remarks, if any :	1st Attempt (20 Pages only)

We have verified the contents as summarized above and certified that the statement made above are true to the best of our knowledge and belief.

Head-Library and Information Center

Head Library and Information Center
New Horizon College of Engineering
Ring Road, Kasubrisonahalli, Bellandur Post,
Near Marathalli, Bangalore-560 103