**A PROJECT REPORT**

**on**

#### “DIABETES PREDICTION SYSYTEM”

**Submitted to**

#### KIIT Deemed to be University

**In Partial Fulfillment of the Requirement for the Award of BACHELOR’S DEGREE IN**

**INFORMATION TECHNOLOGY**

**BY**

**GEETANSH VERMA 1729201**

**SAURABH KUMAR 1729218**

**SINDHUSUTA MOHANTY 1729222**

**ANSUMAN SAHOO 1729242**

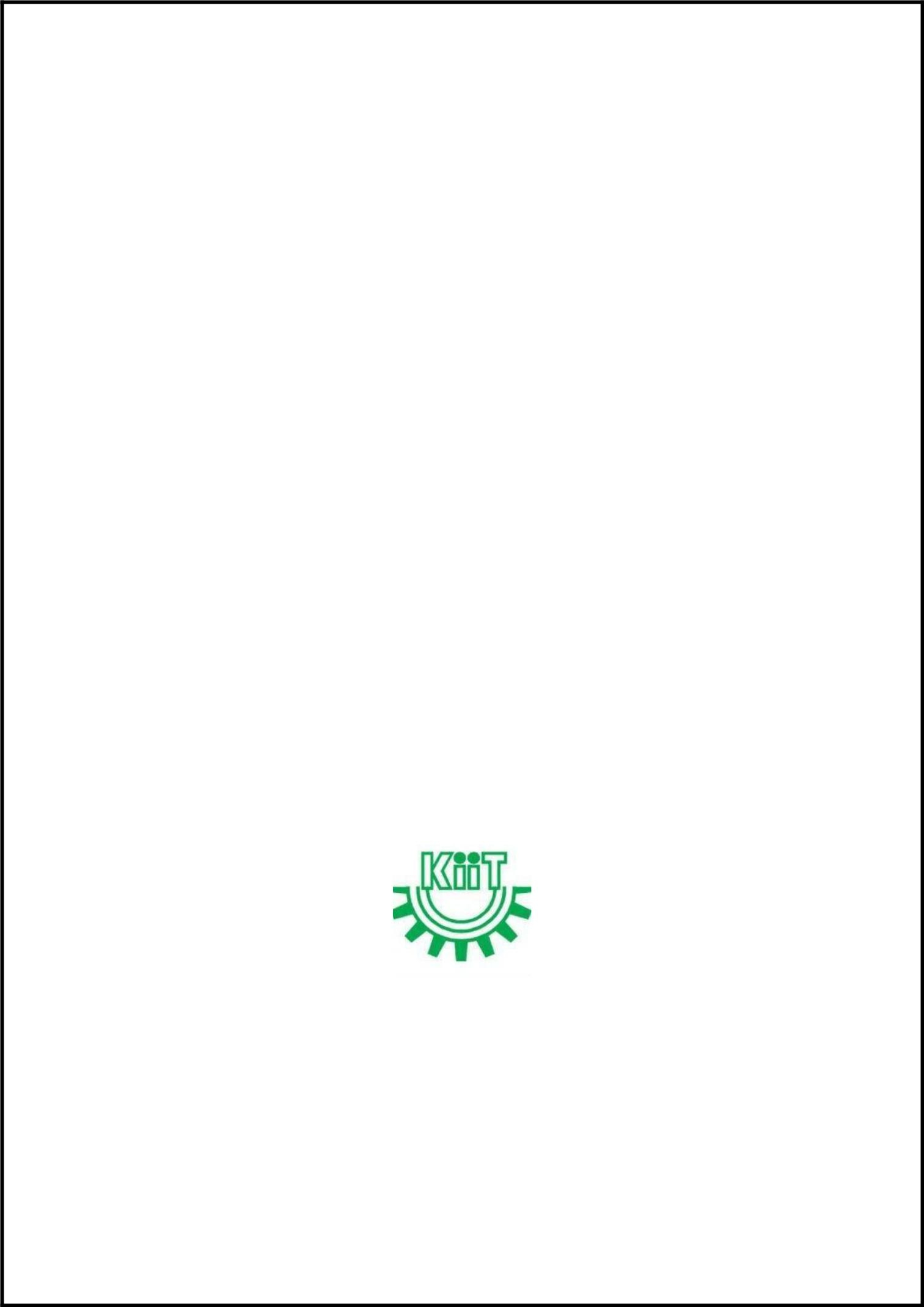
**UNDER THE GUIDANCE OF PROF. Dr. TANMOY MAITRA**

**SCHOOL OF COMPUTER ENGINEERING**

**KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY**

**BHUBANESWAR, ODISHA - 751024**

**May 2020**

**A PROJECT REPORT**

on

“**DIABETES PREDICTION SYSYTEM**”

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR’S DEGREE IN INFORMATION TECHNOLOGY

BY

|  |  |
| --- | --- |
| **GEETANSH VERMA** | **1729201** |
| **SAURABH KUMAR** | **1729218** |
| **SINDHUSUTA MOHANTY** | **1729222** |
| **ANSUMAN SAHOO** | **1729242** |

UNDER THE GUIDANCE OF

**PROF. Dr. TANMOY MAITRA**

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAE, ODISHA -751024

May 2020

KIIT Deemed to be University

School of Computer Engineering Bhubaneswar, ODISHA 751024



## CERTIFICATE

This is certify that the project entitled

“**DIABETES PREDICTION SYSYTEM**“

submitted by

|  |  |
| --- | --- |
| **GEETANSH VERMA** | **1729201** |
| **SAURABH KUMAR** | **1729218** |
| **SINDHUSUTA MOHANTY** | **1729222** |
| **ANSUMAN SAHOO** | **1729242** |

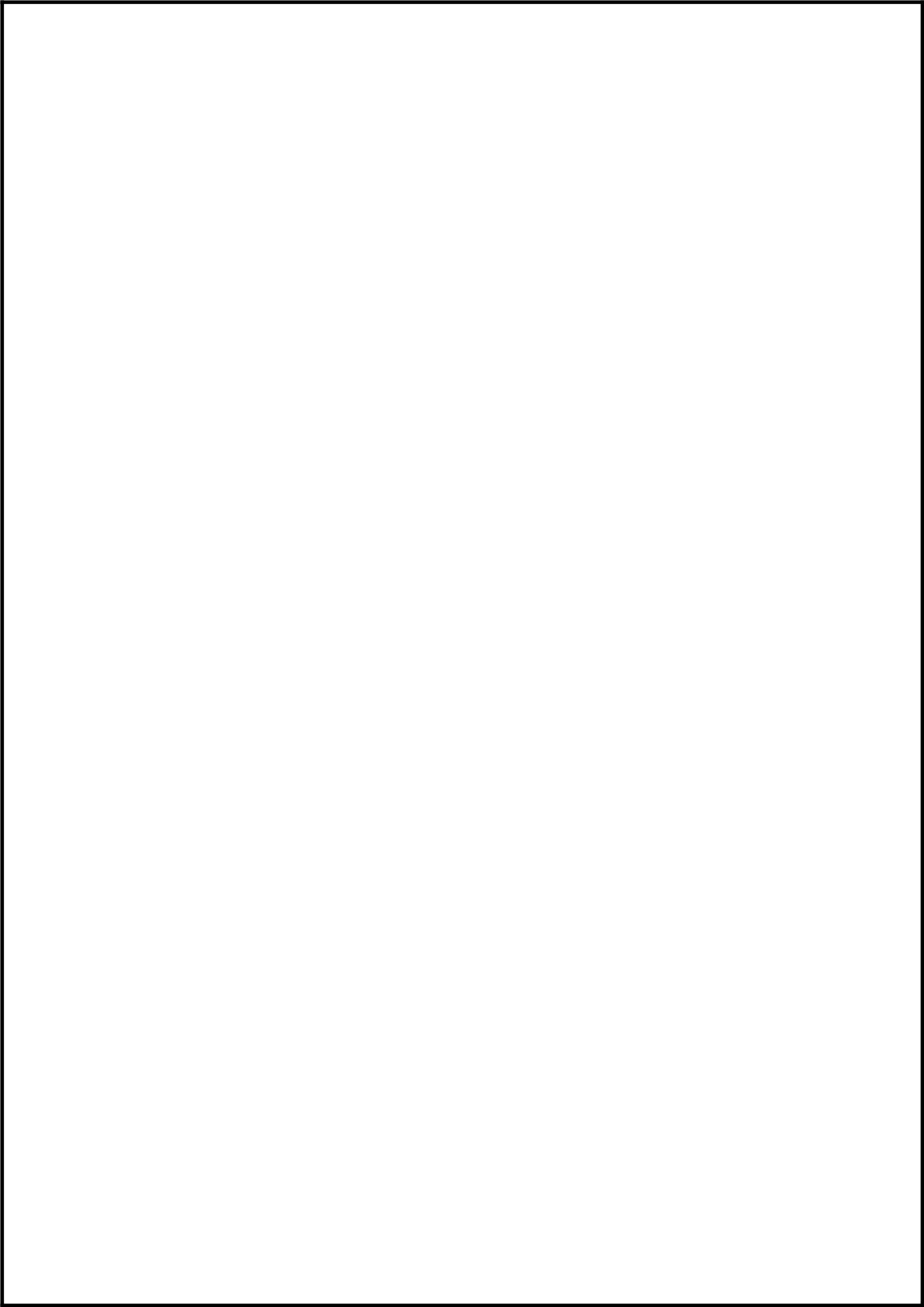
is been recorded as a bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be university,

Bhubaneswar. This work is done during year 2019-2020, under our guidance.

Date: 28/11/2020

(**Prof. TANMOY MAITRA**)

Project Guide



#### Acknowledgements

We are extremely grateful to Prof. Tanmoy Maitra for his insightful guidance and continual encouragement throughout to ensure that this project maintains its target from its inception to its completion. His constant advice helped us to complete the project within the required time frame. Last but not the least, we would like to thank our project colleagues for helping us solve all the problems we faced and for providing us with a pleasant experience.

GEETANSH VERMA SAURABH KUMAR SINDHUSUTA MOHANTY ANSUMAN SAHOO





#### ABSTRACT

Diabetes is considered one of the most lethal and chronic diseases that causes an increase in blood glucose. There are numerous complications arises if diabetes is not treated and not identified in early ages. The tedious identification process translates into a patient's visit to a diagnostic center and consultation with a physician. But the increase in the automatic learning approach solves this critical problem.

The purpose of this study is to develop a model that can predict the probability of diabetes in patients with the greatest precision.. Consequently six machine learning classification algorithms namely Decision Tree,

KNN ,Logistic regression ,Naive Bayes,Random Forest,SVM are used in the experiment to detect early-stage diabetes. Experiments are carried out on the Pima Indians Diabetes Database (PIDD), which is produced by the UCI Machine Learning Repository.

**Keywords: Diabetes, Decision Tree, KNN, Logistic regrassion, Naive Bayes Random forest, SVM**

## Contents

1. Introduction
2. Literature Survey 2
3. Software Requirements Specification 3
   1. [System Requirement……… 3](#_bookmark0)
   2. Technical Requirement…… 3
   3. [Libraries Requirement…….. 4](#_bookmark1)
4. Requirement Analysis 5
   1. [Logical Dataset…………… 5](#_bookmark2)
   2. [Parameter…………… 5](#_bookmark3)
5. System Design

5.1.MODEL DIAGRAM 6

5 .2System Testing 7

1. [Test Cases and Test Results………… 8](#_bookmark4)
   1. Project Planning 9
2. [Project Planning Included……… 9](#_bookmark5)
3. Implementation 10
   1. [Collect the Dataset…………. 10](#_bookmark6)
   2. [Load the dataset…………… 10](#_bookmark7)
   3. [Describe the Dataset……………… 10](#_bookmark8)
   4. Analysis of the dataset for cleaning ………… 10
   5. [Split the dataset for training and testing……. 11](#_bookmark9)
   6. Train the dataset using algorithms ……..
   7. Compare the accuracy of all algorithms

11-14

* 1. Create GUI for the model……….. 14-18

1. Hyper parameter Optimization 19
   1. Hyper parameter Optimization of KNN 20
   2. Hyper parameter Optimization of Decision tree 20
   3. Hyper parameter Optimization of Logistic Regression 21
   4. Hyper parameter Optimization of Naive Bayes 22
   5. Hyper parameter Optimization of Random Forest 22
   6. Hyper parameter Optimization of SVM 23
2. Screen shots of Project 23
   1. Data Entry page…….. 23
   2. After clicking on KNN button…. 23
   3. After clicking on Logistic regression button…. 24
   4. After clicking on Naive Bayes button…. 24
   5. After clicking on Decision tree button…. 25
   6. After clicking on compare button…. 25
   7. After entering data and clicking on predict

button…. 26

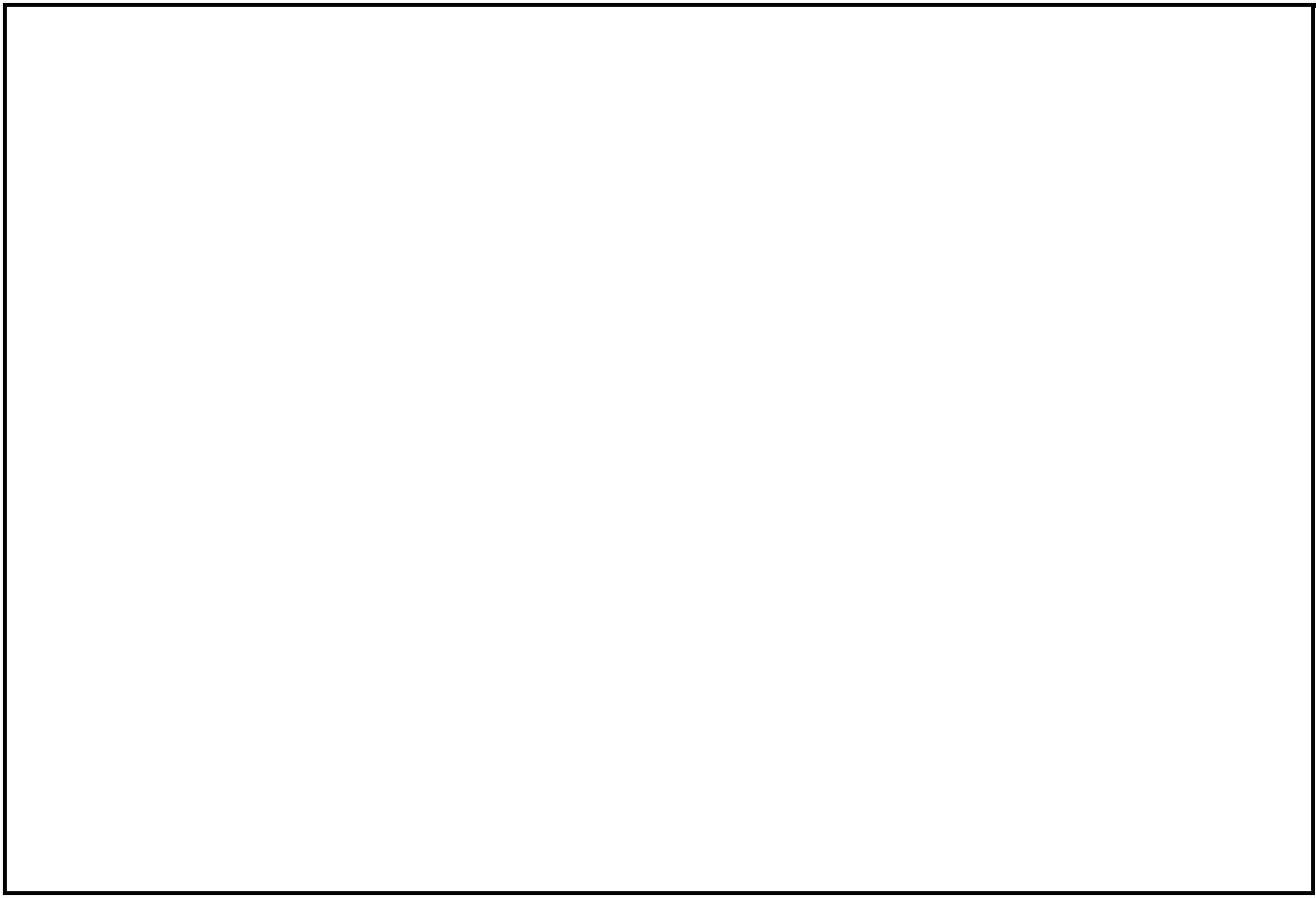
* 1. After clicking on Reset button…. 26

1. Conclusion and Future Scope 27

11.1Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27

11.2 Future Scope . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27

11.3 References 28



List of Figures

1.1 MODEL DIAGRAM 06

## Chapter 1

### Introduction

Now days from health care industries large amount of data is generating. This data contains lots of useful information that can be used to take efficient medical decisions. Hence it is necessary to store and process such data to extract knowledge from it and by using that take efficient decisions. In current situations the numbers of people suffering from Diabetes are increasing per day. Diabetic Mellitus (DM) belongs to the family of Non Communicable Diseases (NCD). It is one of the dangerous diseases as it has long term impacts on patient’s body. Diabetic Mellitus (DM) is categorized into three types. First is Type 1 DM which is called as Insulin - Dependent Diabetes Mellitus (IDDM). Type 1 DM is caused when patient’s body is unable to produce insulin and requires injecting insulin externally to the patient.The second type of DM is Type 2 DM which is called as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of diabetes is caused when patient’s body cell are not able to use insulin properly. Third type of DM is gestational diabetes which iscaused in pregnant women due to the development of high blood sugar without the knowledge of Pre-diagnosis of Diabetes. This type of DM is very dangerous and can lead to Type 2 DM. As numbers of patients of this serious disease are increasing day by day, the size of diabetic data set is also increasing. With the utilization ofmachine learning algorithm, it can be possible to analyze the diabetes data and accordingly provide early diagnosis and better treatment. Machine learning is the set of different methods that can be used to find patterns from the dataset and then use those patterns to predict future conditions or to make efficient decisions under some conditions.Machine learning can make decisions independently at itsown.

Machine learning is categorized into to two types, supervised learning and unsupervised learning.

1

Chapter 2

### Literature Survey

Orabi et al. in designed a system for diabetes prediction, whose main aim is the prediction ofdiabetes a candidateissufferingataparticularage. Theproposedsystemisdesignedbasedon the concept of machine learning, by applying decision tree. Obtained results were satisfactory as the designed system works well in predicting the diabetes incidents at a particular age, with

higher accuracy using Decision tree,

Pradhan etal in usedGenetic programming (GP) for the training and testing of the database for prediction of diabetes by employing Diabetes data set which is sourced from UCI repository. Results achieved using Genetic Programming , gives optimal accuracy as compared to other implemented techniques. There can be significant improve in accuracy by taking less time for classifier generation. It proves to be useful for diabetes prediction at low cost.

Rashid et al. in designed a prediction model with two sub-modules to predict diabetes-chronic disease.ANN(ArtificialNeuralNetwork)isusedinthefirstmoduleandFBS(FastingB lood Sugar) is used in the second module. Decision Tree is used to detect the symptoms of diabetes on patients health.

Nongyao et al. in applied an algorithm which classifies the risk of diabetes mellitus. To fulfill the objective au- thor has employed four following renowned machine learning classification methods namely Decision Tree, Artificial Neural Networks, Logistic Regression and Naive Bayes. For improving the robustness of designed model Bagging and Boosting techniques are

used. [3]

2



## Chapter 3

**Software Requirements Specification**

* 1. System Requirement

Operating system:- Linux or Windows 10 RAM requirement:- 8 GB.

ROM:-256 GB

System Type:- 64-bit Operating System

x-64 based processor

Processor:-Intel® Core™ i5-7200 CPU 2.50GHz 2.71GHz

To obtain efficient results, more computational power is favourable .

* 1. Technical Requirements

**Python3** : Python is an interpreted high-level programming language for general-purpose programming. Created by GuidovanRossum andfirst released in 1991, It has a design philosophy that emphasizes code readability, notably using significant white space. It provides constructs that enable clear programming on both small and large scales.

**PyCharm :-** It is developed by Jet brains,which is an Integrated Development Environment used in computer programming and specifically for the Python language.

**Jupyter Notebook :-**TheJupyter Notebook is an open-source web application that allows us to create and also share document that contains live code, equations, visualizations and narrative text.

**Spyder: -** Spyder is an open source cross-platform integrated development environment for scientific programming in the Python language. Spyder integrates NumPy, SciPy, Matplotlib and IPython, as well as other open source software.

3

* 1. Libraries Requirement:

Pandas: - *It is a software library written for the Python programming language fordata manipulation and analysis.It also offers datastructures and operations for manipulating numerical tables and*

*time series.*

Matplotlib: - *It is a Python 2D plotting library which produces publication quality figures in a variety of hard copy formats and interactive environments across platforms.*

NumPy: - It *is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.*

**[Scikit-learn](http://scikit-learn.org/stable/) *: It is a library in Python that provides many unsupervised and supervised learning algorithms.***

**The functionality that scikit-learn provides include:**

***Regression, including Linear and Logistic Regression* Classification**, ***including K-Nearest Neighbors* Clustering**, ***including K-Means and K-Means++***

**Model selection**

**Preprocessing**, ***including Min-Max Normalization***

Tkinter:- Tkinter is the standard GUI library for Python. Python when combined with Tkinter provides a fast and easy way to create GUI applications. Tkinter provides a powerful object-oriented interface to the Tk GUI toolkit.

Creating a GUI application using Tkinter is an easy task. All we need to do is perform the following steps .

1. We should Import the *Tkinter* module.
2. Then after we should create the GUI application main window.
3. Then we should add one or more of the above-mentioned widgets to the GUI application.
4. Finally,we sholud Enter the main event loop to take action against each event triggered by the user.

**4**



## Chapter 4 Requirement Analysis

* 1. Logical Dataset:
* Experiments are performed on *Pima Indians Diabetes Database (PIDD*) which is sourced from

UCI machine learning repository.

* *PIDD:*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. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The dataset contains 8 attributes and 1 class variable. This data set contains total 768 diabetic and non-diabetic women records whose age is above 21 years. All attributes in data

set are numeric. For this work the dataset is considered as complete data set having no missing value in it.

* 1. Parameter :

1. **Pregnancies:** No. of times pregnant
2. **Glucose:** Plasma Glucose Concentration a 2 hour in an oral glucose tolerance test (mg/dl) A 2- hour value between 140 and 200 mg/dL (7.8 and 11.1 mmol/L) is called impaired glucose tolerance. This is called "pre- diabetes." It means you are at increased risk of developing

diabetes over time. A glucose level of 200 mg/dL (11.1 mmol/L) or higher is used to diagnose diabetes.

1. **Blood Pressure:** Diastolic Blood Pressure(mmHg): If Diastolic B.P > 90 means High B.P(High

Probability of Diabetes) Diastolic B.P < 60 means low B.P (Less Probability of Diabetes**)**

1. **Skin Thickness:** Triceps Skin Fold Thickness (mm) – A value used to estimate body fat.

Normal Triceps SkinFold Thickness in women is 23mm. Higher thickness leads to obesity and chances of diabetes increases.

1. **Insulin:** 2-Hour Serum Insulin (mu U/ml) Normal Insulin Level 16-166 mIU/L Values above

this range can be alarming.

1. **BMI:** Body Mass Index (weight in kg/ height in m2) Body Mass Index of 18.5 to 25 is within the normal range BMI between 25 and 30 then it falls within the overweight range. A BMI of 30

or over falls within the obese range**.**

1. **Diabetes Pedigree Function:** It provides information about diabetes history in relatives and genetic relationship of those relatives with patients. Higher Pedigree Function means patientis

more likely to have diabetes.

1. **Age** (years)
2. **Outcome:** Class Variable (0 or 1) where ‘0’ denotes patient is not having diabetes and ‘1’ denotes patient having diabetes The dependent variable is whether the patient is having diabetes or not.More than 70% Pima Indian population is suffering from the diabetes.
3. ***4.3 Data Cleaning*** will take place as data has got lot of missing values. Handling missing values can be done either by replacing null values with mode or mean or replacing the

null value with a random variable.



## Chapter 5 -System Design

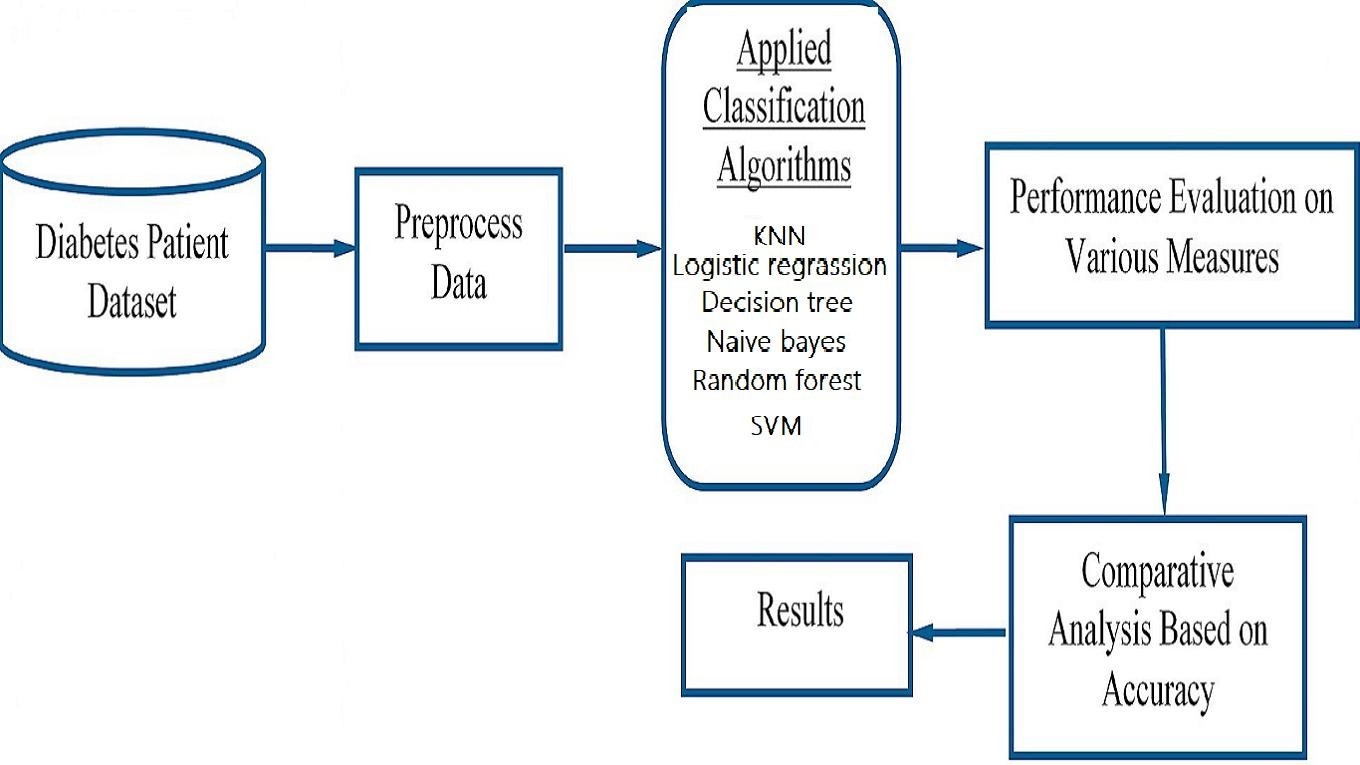


Fig :1.1(model diagram)

To perform the experimentation Pima Indians Diabetes Dataset has been used.The dataset contains 8 attributes and 1 class variable .This data set contains total 768 diabetic and non-diabetic women records whose age is above 21 years.Class variable (0 or 1).0-It indicates False Diabetic Test and1-It indicates True Diabetic Test .First we load the dataset as a CSV file and identify each column then process the dataset by finding a mean value and perform some cleaning operation.Divided the whole data into 8:2 for train and test respectively.

Train and test of dataset done over the all four algorithm i.e DESCISION TREE,KNN,LOGISTIC REGRASSION ,NAIVE BAYES ,Random Forest and Support Vector Machine. From the testing part we analyze the performance of each and every algorithm in percent .Most of the case we got SVM gives us the highest accuracy as compared to other so finally we pick it.

Finally it provides a interface to the user ,where a user has to provide the value of the parameter that she has and our machine will take that value and start the process from the algorithm comparison and use that which has maximum accuracy . Input data will take to test and provides the best possible output.

We are also trying to test some other algorithm in future so that our machine can perform more accurately and achieve the goal of system. [6] 6

Diabetes prediction system

7

## Chapter 6 System Testing

The validity of program was checked by inserting various input parameters and based upon the input the system does the classification into: person will have diabetes or will not have diabetes in future.

###### 6.1 Test Cases and Test Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test ID | Test Case Title | Test Condition | System Behavior | Expected Result |
|  | Pregnant=2 | After  compariso n SVM is found to have highest  accuracy.  Therefore here SVM has being  used |  | You have no diabetes  (0) |
|  | Glucose=200 |  |
|  | Blood pressure=140 | Knn-78.0% |
|  | Skin=19 | Decision tree-71.0% |
|  | Insulin=140  Body mass index=23 | Naive bayes-79.0%  Logistic regression-80% Random forest - 79.0% |
|  | Pedigree=1 | SVM - 82.0% |
| T01 | Age=34 |  |
|  | Pregnant=1 | After  compariso n SVM is found to have highest  accuracy.  Therefore here SVM has being  used |  | You have no diabetes  (0) |
|  | Glucose=140 |  |
|  | Blood pressure=65 | Knn-78.0% |
|  | Skin=21 | Decision tree-71.0% |
|  | Insulin=110  Body mass index=19 | Naive bayes-79.0%  Logistic regression-80% Random forest - 79.0% |
|  | Pedigree=0 | SVM - 82.0% |
| T02 | Age=29 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| T03 | Pregnant=4 Glucose=187  Blood pressure=110  Skin=17  Insulin=47  Body mass index=18  Pedigree=2 Age=25 | After compariso n SVM is found to have highest  accuracy.  Therefore here SVM has being  used | Knn-78.0%  Decision tree-71.0% Naive bayes-79.0%  Logistic regression-80% Random forest - 79.0%  SVM - 82.0% | You have  Diabetes or may get soon  (1) |

*School of Computer Engineering, KIIT, BBSR*

8



## Chapter 7

**Project Planning**

7.1 PROJECT PLANNING INCLUDED

-Project purpose

-Business and project goals and objectives

-Scope and expectations

-Assumptions and constraints

-Project management approach

-Ground rules for the project

-Project budget

-Project timeline

-The conceptual design of new technology

9

*NAME OF PROJECT*

## Chapter 8

**Implementation**

DESCRIPTION OF EACH SEGMENT OF CODE IMPLIMENTATION

* 1. Collect the dataset:

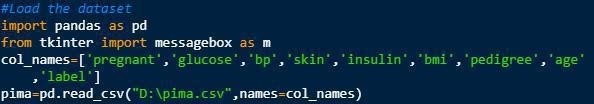
At first the dataset was collected from the UCI Repository and was saved as csv (coma

separated value) file.

* 1. Load the dataset:

Then the dataset was loaded into the system through the code using read\_csv()

present in ***pandas*** package.



*to load the dataset*

* 1. Describe the dataset:

Then the dataset was described (i.e. the names were given to the columns for

easy access) through the code.



*to describe the dataset*

* 1. Analysis of the dataset for cleaning purpose:

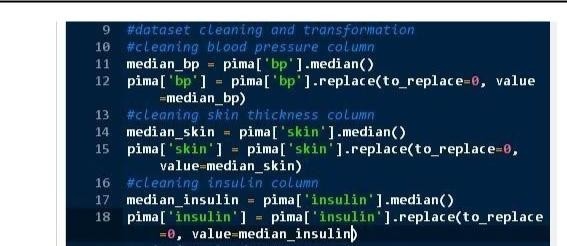
Then the analysis of the dataset was done. Some anomalies were found during this process which was mandatory to correct to achieve an accurate result. Thus, the mandatory columns (like blood pressure, skin thickness, insulin) which should not have 0 attribute were cleaned and replaced with the median value of

the respective whole column.

10







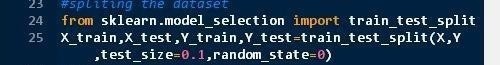
*to clean the dataset*

* 1. **Split the dataset for training and testing:**

Then the dataset was split in ratio 8:2 for training and testing purpose respectively.

The dataset was split through the code using train\_test\_split() which was present in

model\_selection module of ***sklearn*** package.



*to split the dataset*

* 1. **Train the dataset using algorithms**

Confusion matrix

To check the adequacy of classifiers performance measures need to be taken into account. True positives (TP) refers to the positive cases that were correctly labeled by the classifier, while true negatives (TN) are the negative cases that were correctly labeled by the classifier. False positives (FP) are the negative cases that were incorrectly labeled, while false negatives (FN) are the positive cases that were incorrectly labeled . Evaluation measures of classifiers such as accuracy was calculated. The accuracy of the classifier on a given test data set is the percentage of test cases that are correctly classified by classifier. [7]

Accuracy=(TN)/(TP+TN+FP+FN)

**Using KNN Algorithm**

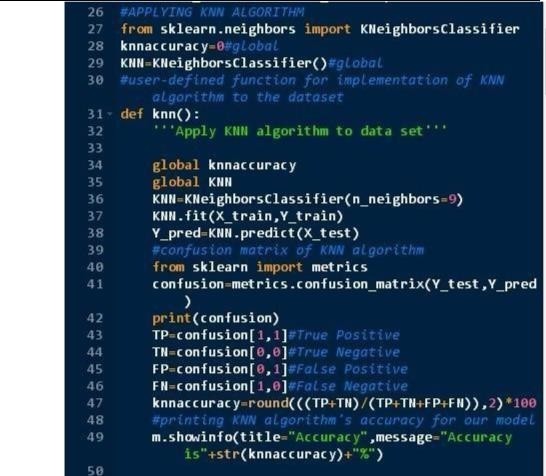
The system was thus trained using KNN Algorithm.

11





*NAME OF PROJECT*



*Training the model using KNN Algorithm*

**Using Logistic Regression Algorithm**

The system was then trained using Logistic Regression Algorithm.



*Training the model using Logistic Regression*

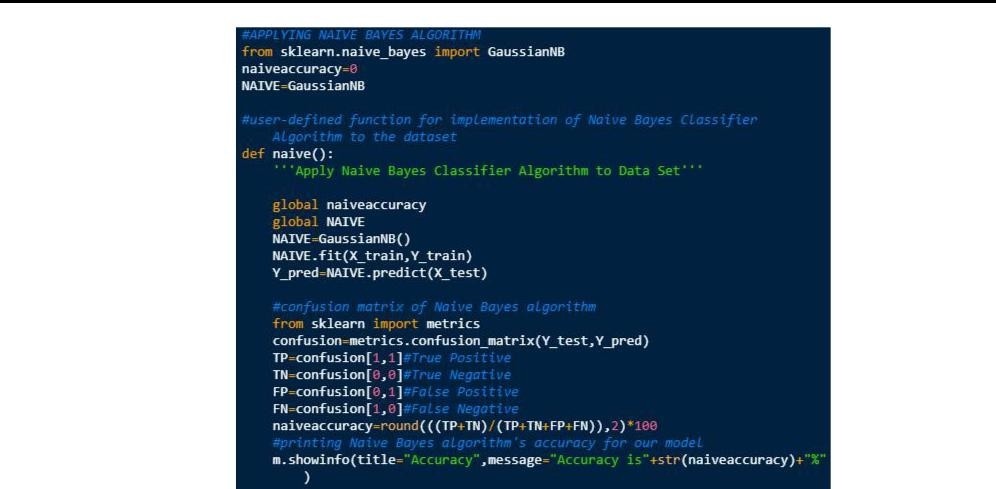
**Using Naive Bayes Algorithm**

The system was then trained using Gaussian Naïve Bayes Algorithm.

12

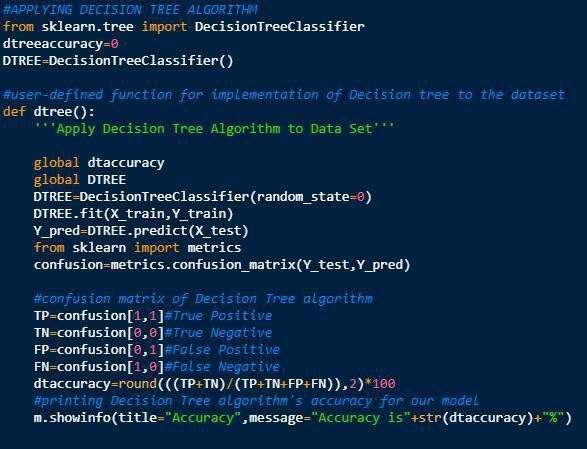






*Training the model using Gaussian Naïve Bayes*

**Using Decision Tree Algorithm**

The system was then trained using Decision Tree Algorithm.

*Training the model using Decision Tree Algorithm*

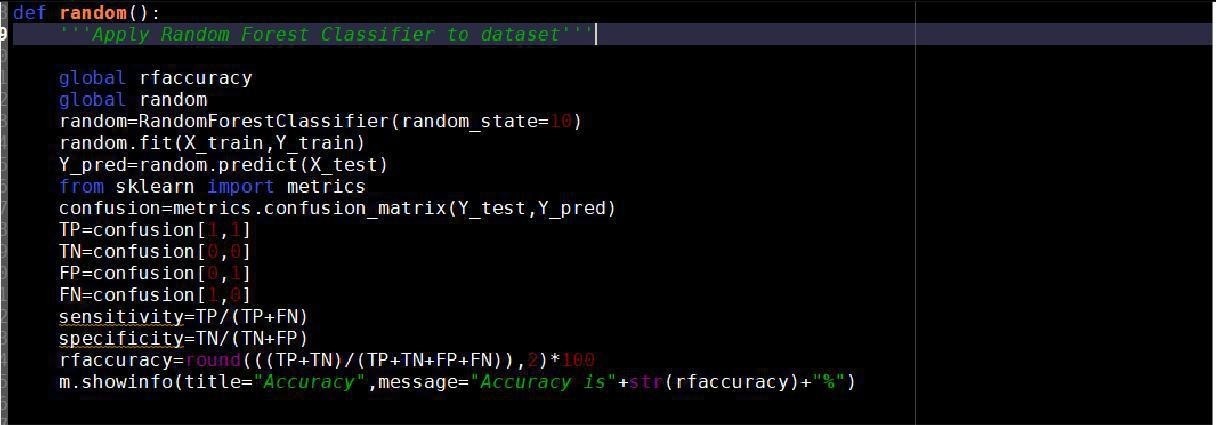
*13*

**Using Random forest algorithm**

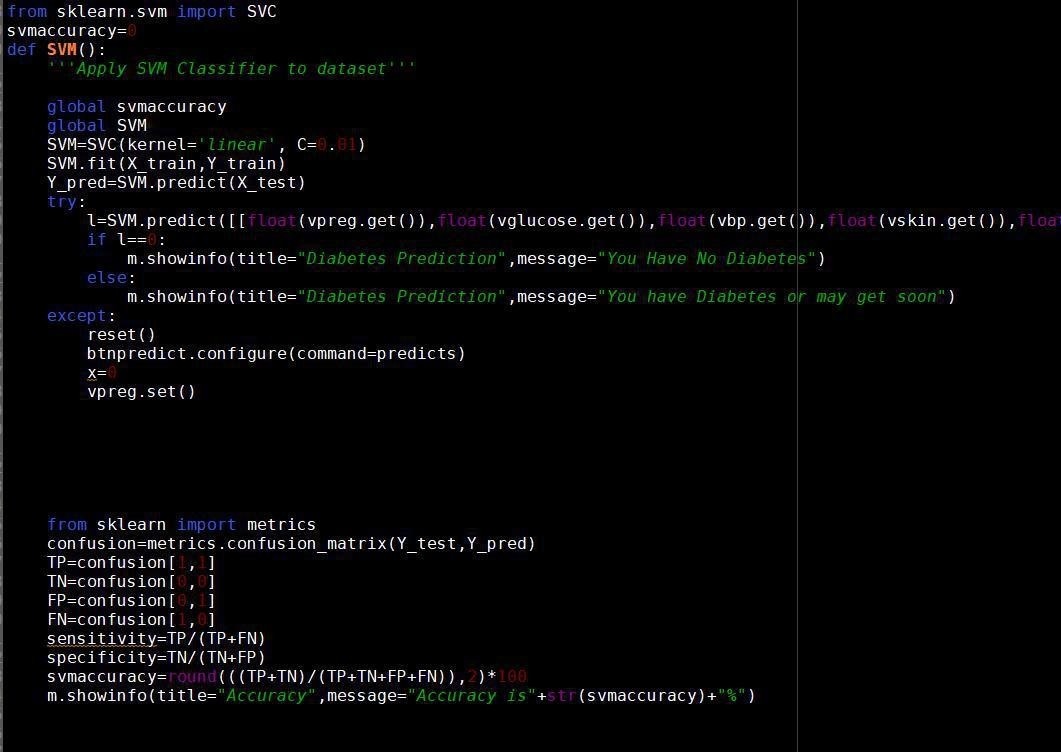




*NAME OF PROJECT*



Using Support vector machine algorithm

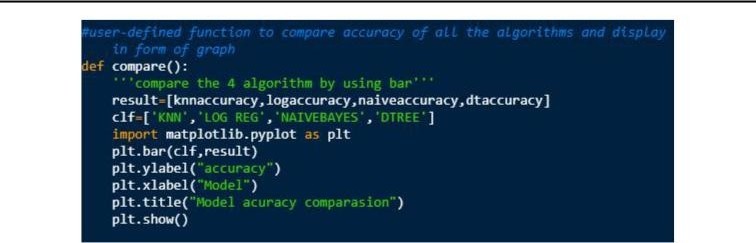


* 1. **Compare the accuracy of all the Algorithm**

After finding out the accuracy of both the algorithms, they were compared so as to find the optimized algorithm. And accuracy of both the algorithms were displayed in graphical representation.

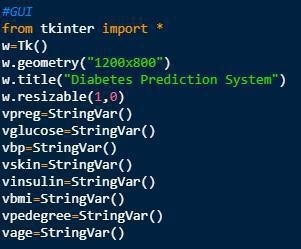






*To compare all the algorithm and display it as a graph*

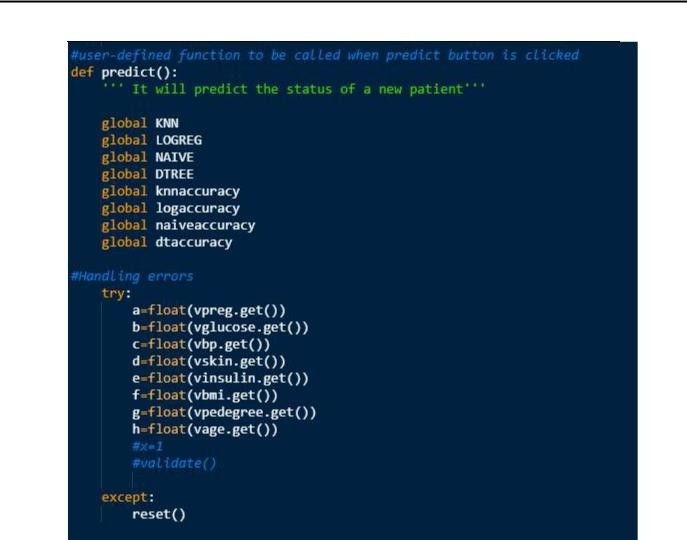
* 1. Create GUI for the model

Then the Graphical User Interfaces for the whole model were model were created.

14







*NAME OF PROJECT*

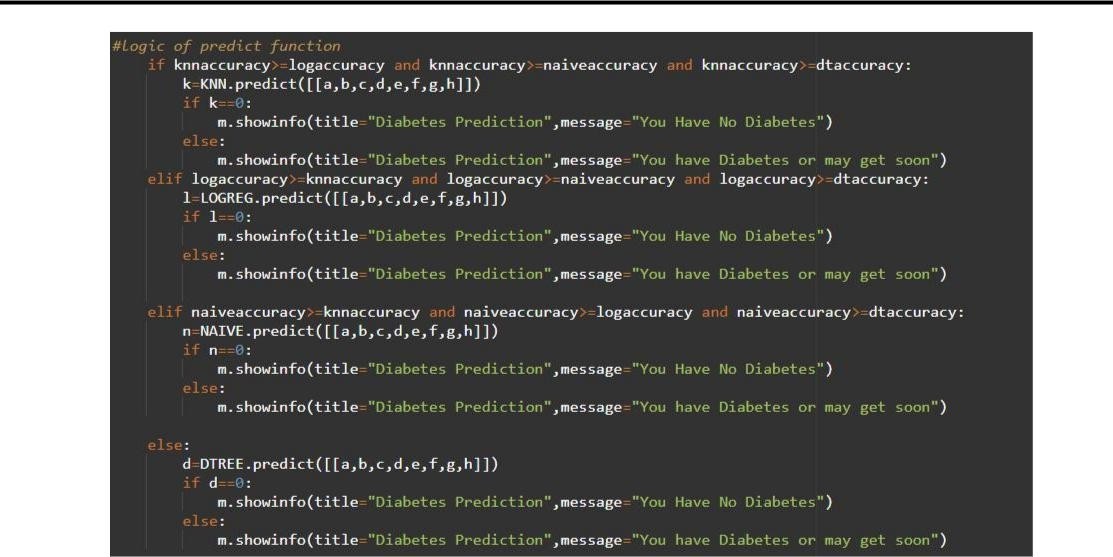
*To create space for the GUI*

*Function of Predict button*

*15*







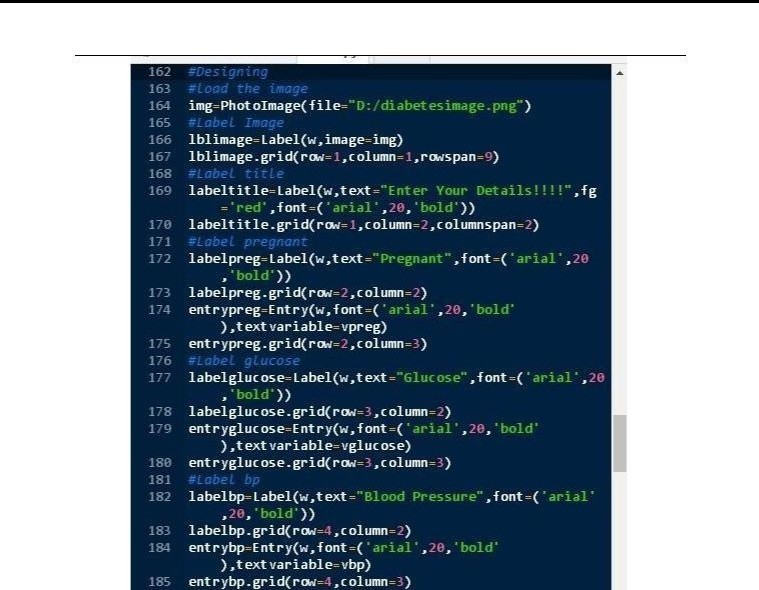
*Logic of Predict button*



*Function of Reset button 16*







*NAME OF PROJECT*

*Designing the page*



*Designing the page cont..*

17







*Designing the page cont..*

18





## Chapter 9:

*NAME OF PROJECT*

***Hyperparameter optimization of algorithms:***

1. **Hyper parameter optimization of KNN:**



# Hyper parameter optimization of Decision tree:

**19**

# Hyper parameter optimization of Logistic Regression:

**20**

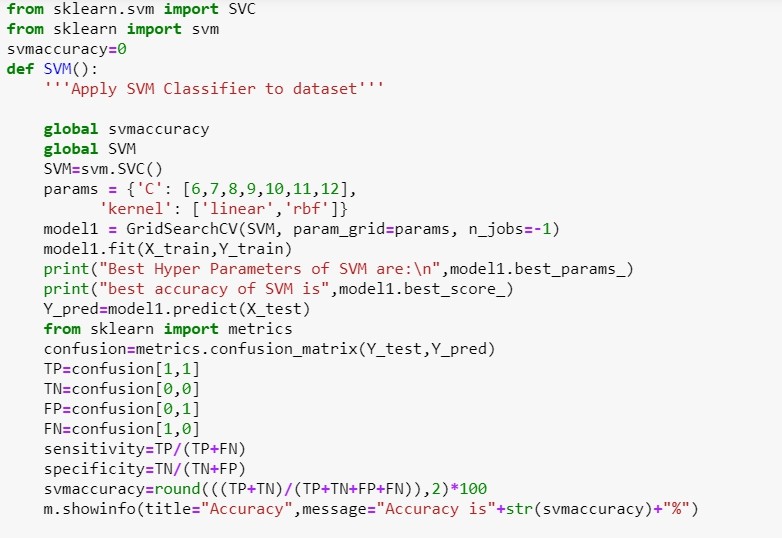
# Hyper parameter Optimization of Naive Bayes:

21

# Hyper parameter Optimization of Random Forest:

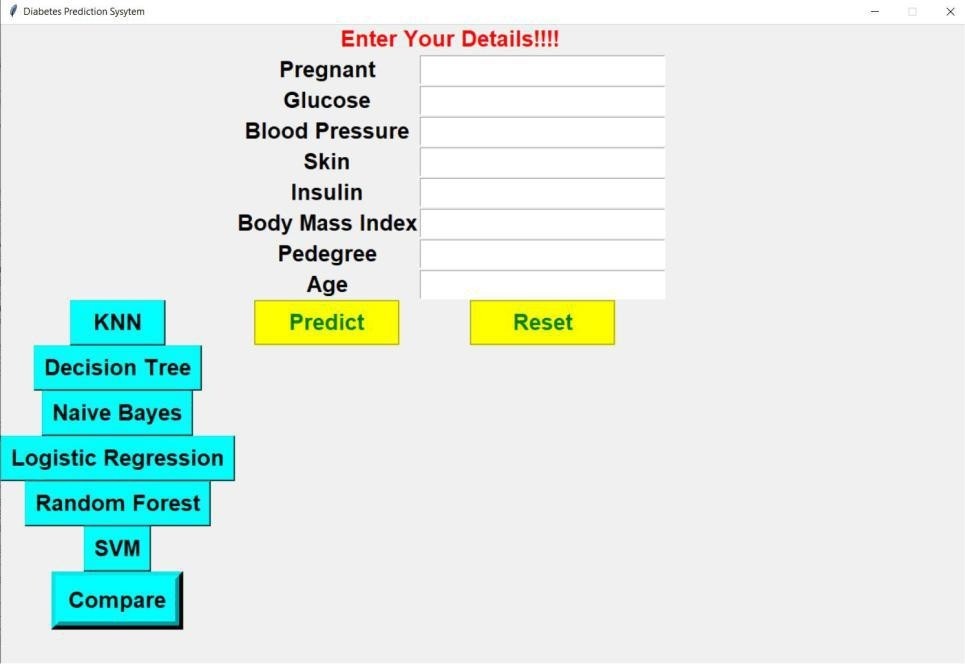


1. **Hyper parameter Optimization of SVM:**



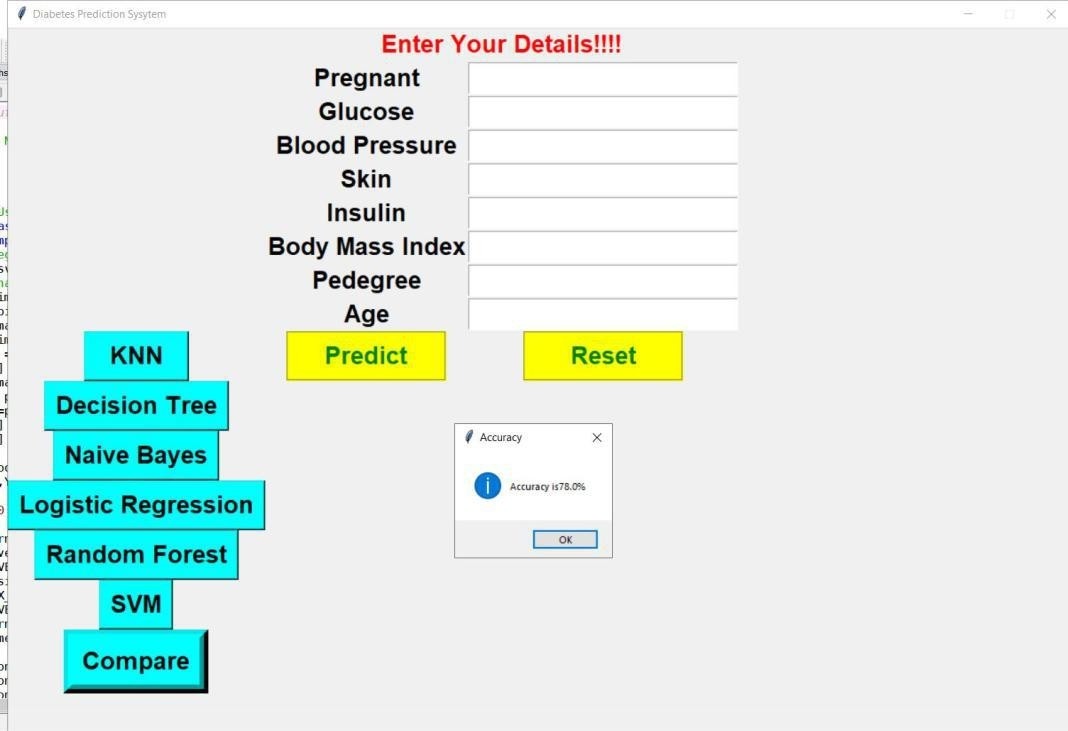
## Chapter 9

**Screen shots of Project**

* 1. Data Entry page

**22**

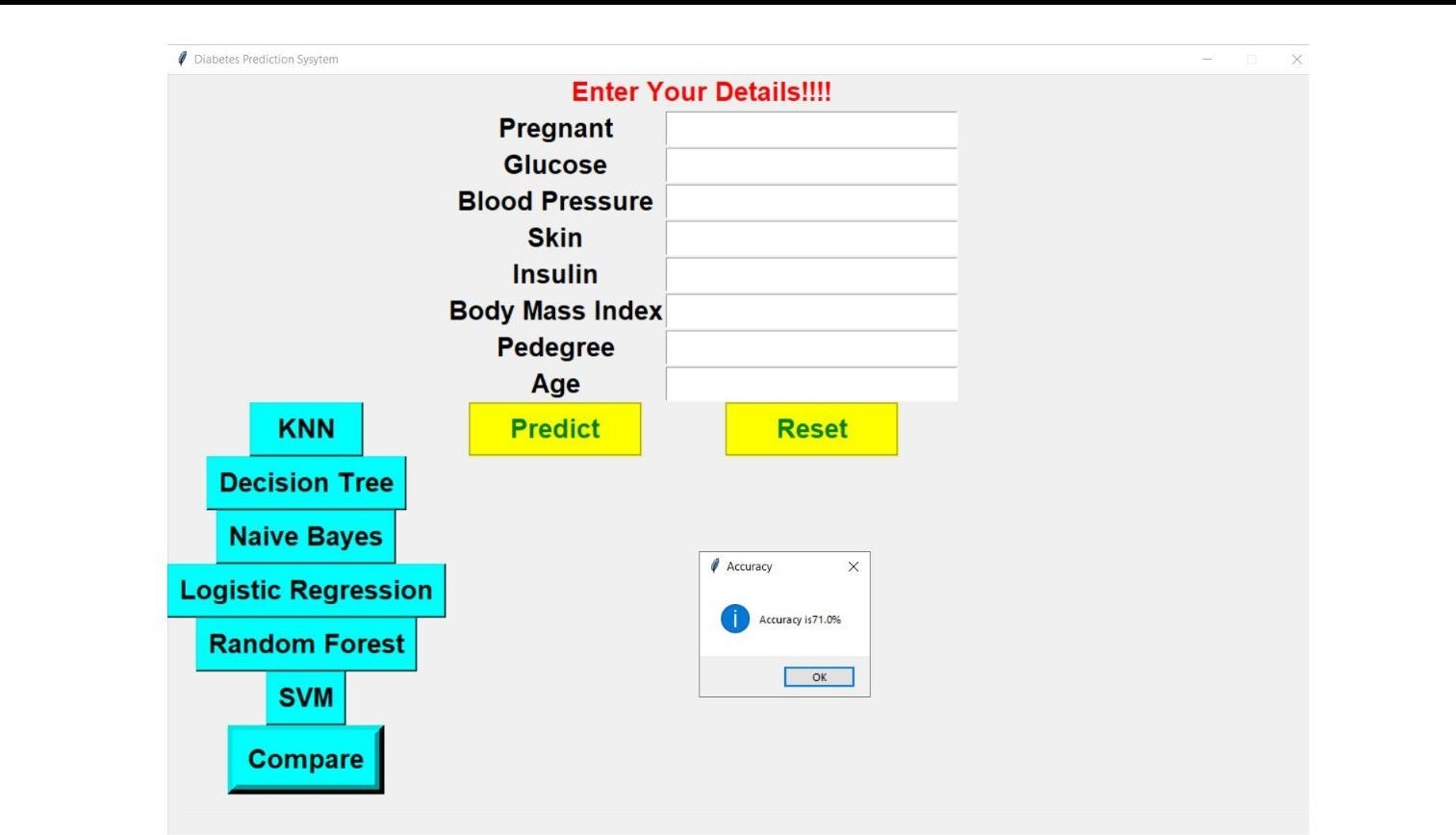
* 1. After clicking on the KNN button



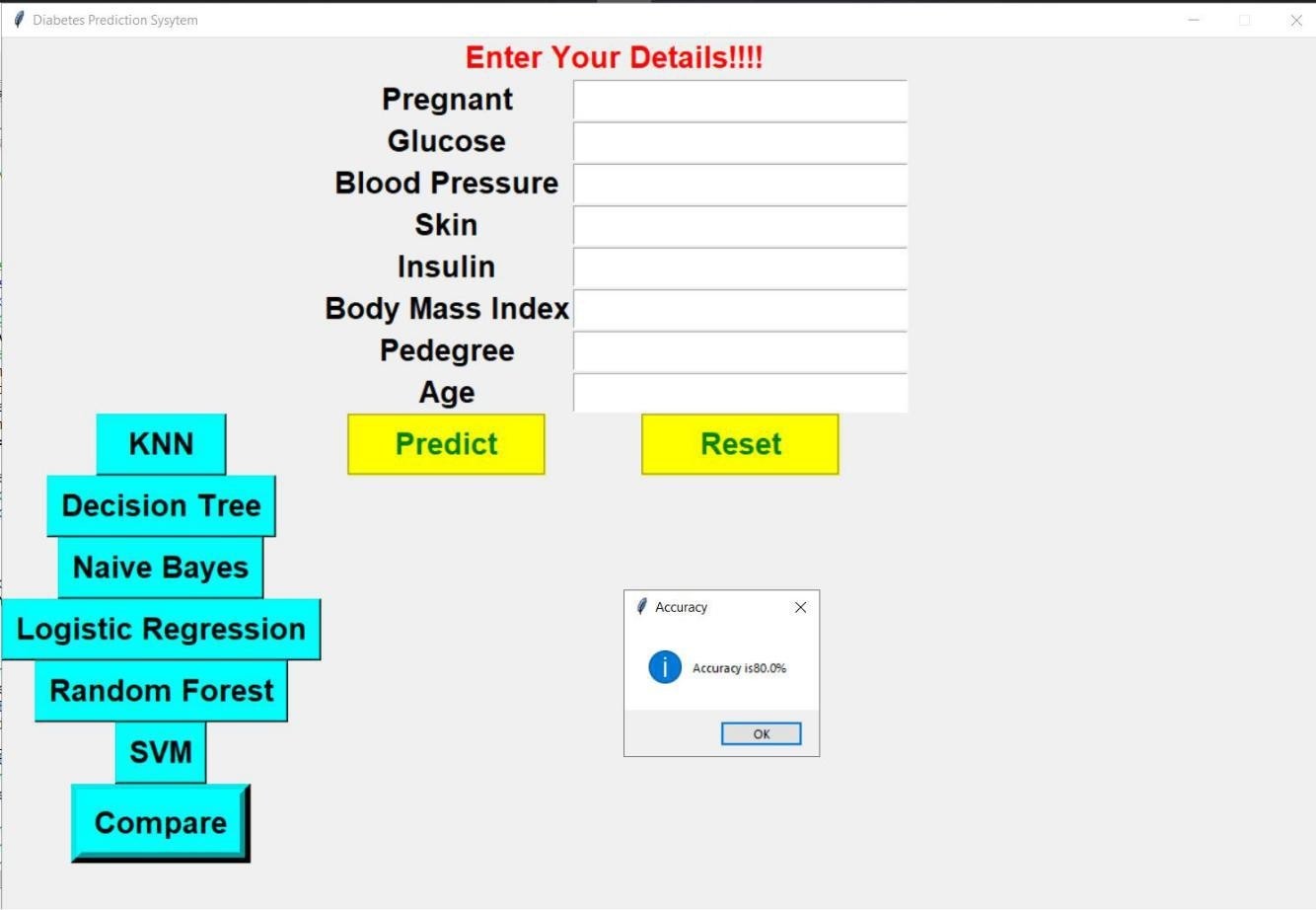
**23**







**9.3 After clicking on the Logistic Regression button**

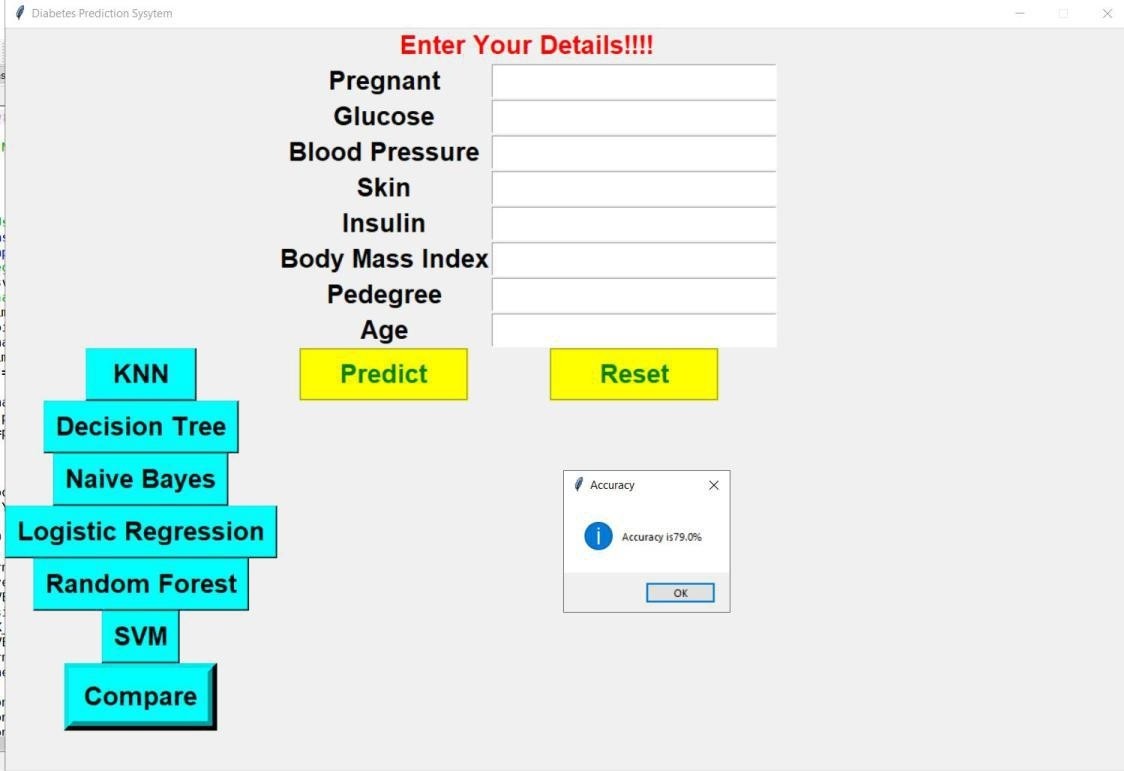
* 1. After clicking on the Naive Bayes button

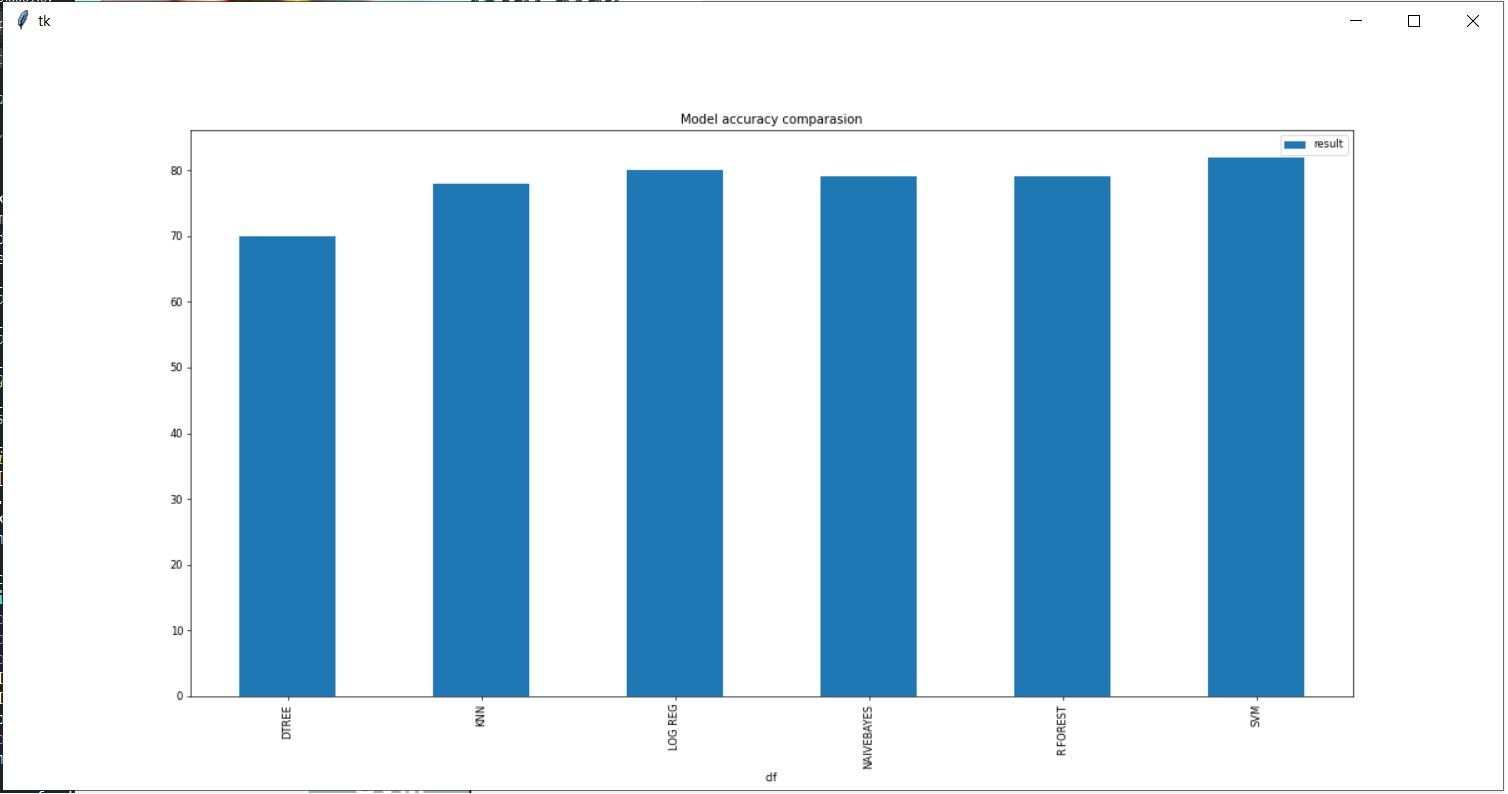
**23**





*NAME OF PROJECT*

* 1. **After clicking on the Decision Tree button**
  2. **After clicking on the Compare button**



*School of Computer Engineering, KIIT, BBSR* 24







* 1. **After entering the data and clicking on the Predict button**



* 1. **Then after clicking on the Reset button**



25





*NAME OF PROJECT*



**Chapter 11**

**Conclusion and Future Scope**

* 1. FUTURE SCOPE

We tested and optimized each algorithm and found out that SVM algorithm best suited for more application.But this is just not the end. We can also implement other classification algorithms to obtain more accurate and optimized result.

##### CONCLUSION

Machine learning algorithms in the medical field retrieve different hidden models of medical data. They can be used to analyse important clinical parameters, predicting various diseases, medical task forecasting, medical knowledge retrieval, therapeutic planning support, and patient management. A number of algorithms were proposed for the prediction and diagnosis of diabetes. These algorithms provide more accuracy than the available traditional systems. We tried and optimize every algorithm and that we Acknowledged SVM algorithm best suitable for over application.

*School of Computer Engineering, KIIT, BBSR* 26







*NAME OF PROJECT*

## References

1. UCI Machine Learning Repository- Center for Machine Learningand Intelligent System, <http://archive.ics.uci.edu/ml/>
2. Gaganjot Kaur, Amit Chhabra, “Improved J48 Classification Algorithm for the Prediction of Diabetes”, International Journal of Computer

Applications (0975 – 8887) Volume 98 – No.22, July 2014.

1. Machine Learning tutorials and examples https://[www.toptal.com/machine-learning/machine-](http://www.toptal.com/machine-learning/machine-) learningtheory-

an-introductory-primer .

1. Sudip Mandal, Goutam Saha, Rajat K. Pal, “A Comparative Study on Disease Classification using Different Soft Computing Techniques”,The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 2, No. 3, May 2014.
2. Classification Techniques" IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014.
3. <http://www.diabetes.ca/about-diabetes/types-of-diabetes>
4. <http://www.diabetes.org/diabetes-basics/type>
5. G. F. Cooper. A simple constraint-based algorithm for eﬃciently mining observational databases for causal relationships. Machine Learning,

1(2):203–224, 1997

*School of Computer Engineering, KIIT, BBSR* 27



