MAULANA AZAD

NATIONAL INSTITUTE OF TECHNOLOGY

BHOPAL, INDIA 462003

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Diabetes Prediction (DIABTELLER)

Minor Project Report

Semester 6

Submitted by:

1.	Gumutch Mishra	191112009
2.	Kunal Thite	191112015
3.	Harshit Prajapati	191112020
4.	Tanish Rangnekar	191112058

Under the Guidance of

Dr. Jaytrilok Choudhary

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Session: 2021-22

MAULANA AZAD

NATIONAL INSTITUTE OF TECHNOLOGY

BHOPAL ,INDIA, 462003

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the project report carried out on "**Diabetes Prediction**" by the 3rd year students:

1.	Gumutch Mishra	191112009
2.	Kunal Thite	191112015
3.	Harshit Prajapati	191112020
4.	Tanish Rangnekar	191112058

Have successfully completed their project in partial fulfillment of their Degree in Bachelor of Technology in Computer Science and Engineering.

Dr. Jaytrilok Choudhary

(Minor Project Mentor)

DECLARATION

We, hereby declare that the following report which is being presented in the Minor Project Documentation Entitled as **"Diabetes Prediction"** is an authentic documentation of our own original work and to the best of our knowledge. The following project and its report, in part or whole, has not been presented or submitted by us for any purpose in any other institute or organization. Any contribution made to the research by others, with whom we have worked at Maulana Azad National Institute of Technology, Bhopal or elsewhere, is explicitly acknowledged in the report.

1.	Gumutch Mishra	191112009
2.	Kunal Thite	191112015
3.	Harshit Prajapati	191112020
4.	Tanish Rangnekar	191112058

ACKNOWLEDGEMENT

With due respect, we express our deep sense of gratitude to our respected guide and coordinator Dr. Jaytrilok Choudhary, for his valuable help and guidance. We are thankful for the encouragement that he has given us in completing this project successfully.

It is imperative for us to mention the fact that the report of a minor project could not have been accomplished without the periodic suggestions and advice of our project guide Dr. Jaytrilok Choudhary and project coordinators Dr. Namita Tiwari and Dr. Sanyam Shukla.

We are also grateful to our respected HoD, Dr. Nilay Khare and Head AI and central computing facility, Dr Meenu Chawla for permitting us to utilize all the necessary facilities of the college.

We are also thankful to all the other faculty, staff members and laboratory attendants of our department for their kind cooperation and help. Last but certainly not the least; we would like to express our deep appreciation towards our family members and batch mates for providing the much needed support and encouragement.

ABSTRACT

Diabetes mellitus commonly known as just diabetes, is an interminable disease that is said to have affected over 246 million people worldwide. Diabetes has been named the 5th deadliest disease in the US with no imminent cure in sight.

However, with the rise of IT and its continued advancement into the medical and healthcare sector, the cases of diabetes as well as their symptoms are well documented.

The proposed project provides a method to help future diabetic patients by using collected data from various medical and research clinics, analyzing it and using it to develop a prediction model in order to attempt to find quicker and more efficient techniques of diagnosing the disease, leading to timely treatment of the patients. Current data models have not had great accuracy regarding the same and this project attempts to fix those gaps and give more concrete and consistent results.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Diabetes mellitus (DM) is a metabolic disease, involving inappropriately elevated blood glucose levels and is a significant public health problem which is increasing at an unprecedented rate. By 2025, the prevalence of diabetes is projected to be 6.3%, which is a 24% increase as compared with 2003. Nearly 2-5 million patients every year are said to lose their lives due to diabetes. It is estimated that by the year 2045 this will rise to 629 million. Predicting the disease at the early stage of life can save valuable human resources. Age, hereditary diabetes, obesity, lack of exercise, high blood pressure etc are all possible causes of diabetes mellitus.

Diabetes mellitus is generally classified into 3 major types.The first is Type-1 also known as Insulin-Dependent Diabetes Mellitus (IDDM). The lack of ability of the human body to generate sufficient insulin is the reason behind this type of DM and therefore it is needed to inject insulin to a patient. This type of diabetes can generally influence any age and is known to affect the younger population. Type-2 known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM) also is the most well-known. This is usually observed when one's body cells are not able to properly utilize insulin. There also exists Gestational Diabetes, a unique type of diabetes that occurs during a pregnancy, where there is an increase in blood sugar level. DM is known to have serious long term complications and problems associated with it.

Big Data analytics play an important role in the medical and healthcare industries, which have large volumes of databases. Analyzing such data may help in finding hidden patterns and information. In this project we are hoping to analyze data from previous diabetes patients in order to develop a prediction model for better classification of possible diabetes causing factors which include insulin, glucose levels, BMI, age, number of pregnancies etc. We are also utilizing predictive analysis, which incorporates a variety of machine learning algorithms,statistical methods, and data mining techniques that uses current and past history and data in order to help predict future events. Existing methods for diabetes detection use time consuming lab tests such as fasting blood glucose and oral glucose tolerance which is not ideal. This project is aiming to minimize the time which is needed in order to diagnose the possibility of diabetes in the future with best possible accuracy . We are hoping to accomplish this by building a predictive model using machine learning algorithms and data mining techniques for diabetes prediction.

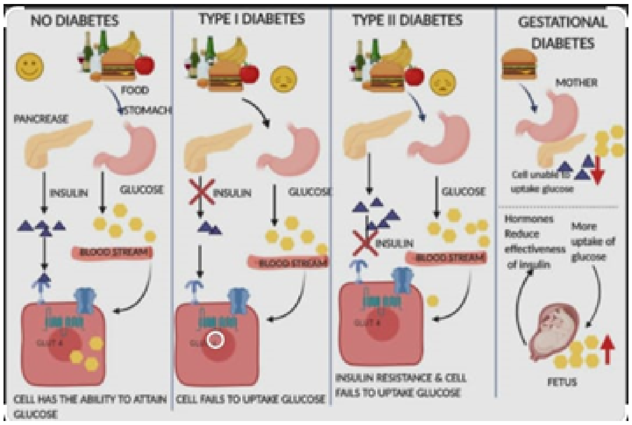


Figure 1: Intro To Diabetes

2. Literature Review and Survey

Machine learning is used in numerous different fields such as medical fields , banking or even in the agriculture department. It is often used for prediction, classification and clustering. For diabetes research, analysis of related work gives results, which show creation of various prediction models which have been developed and implemented by numerous researchers from around the world with the usage of data mining techniques, machine learning algorithms or also combination of these techniques.

2.1. Eurekalert, "Insufficient sleep may be linked to increased diabetes risk," July 11, 2010

Statistics given by the Centre for Disease Control (CDC) states that 26.9% of the population affected by diabetes are people whose age is greater than 65, 11.8% of all men aged 20 years or older are affected by diabetes and 10.8% of all women aged 20 years or older are affected by diabetes. The dataset which is utilized for analysis and modeling has 50,000+ records with 37 variables.

2.2. Bhatt K., Dalal P., Panwar A., "A Cluster Centres Initialization Method for Clustering Categorical Data Using Genetic Algorithm" International Journal of Digital Application & Contemporary research, 2013, Volume-2 Issue-1.

According to research work performed here, the Pima Indian diabetic database (PIDD) at the UCI Machine Learning Lab has been thoroughly tested using different data mining algorithms to predict their accuracy in diabetic status from a collection of individuals. Out of the total of 392 complete cases, they ended up with a 65.1% accuracy when trying to predict the non diabetic individuals. With the usage of ROSETTA software , a data mining predictive tool had been applied from the rough sets to PIDD. The accuracy of predicting diabetic status on the PIDD was an impressive 82.6% on the initial random sample, which exceeds the previously used machine learning algorithms that ranged from 66-81%.

2.3. Huang, Zhexue. "A Fast-Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining." DMKD. 1997.

This research study concluded that the women who had diabetes could be tracked by utilizing attribute-oriented induction techniques as well as clustering . This dataset was collected from the National Institute of Diabetes, Digestive and Kidney Diseases. The results were evaluated in five different clusters denoting concentrations of the various attributes and the percentage of women suffering from diabetes and they show that 23% of the women suffering from diabetes fall in cluster-0, 5% fall in cluster-1, 23% fall in cluster-2, 8% in cluster-3 and 25% in cluster-3

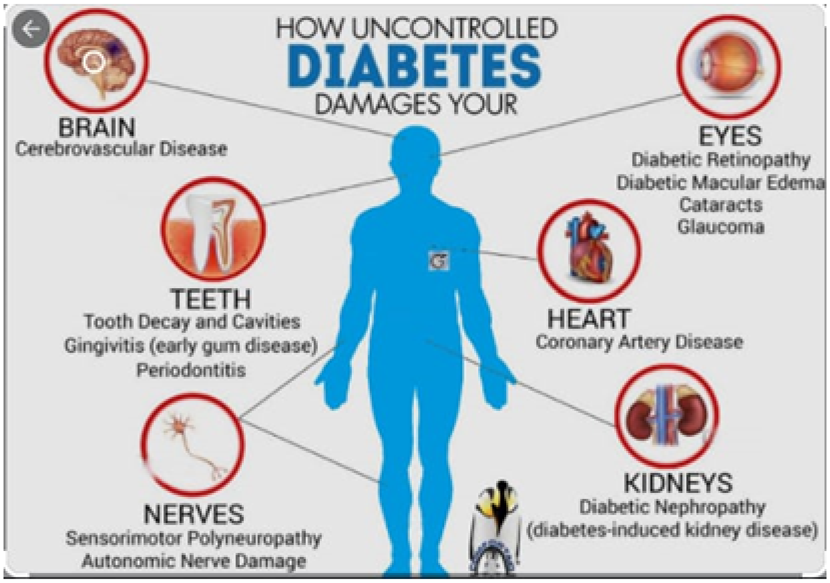


Figure 2: Effect of Diabetes

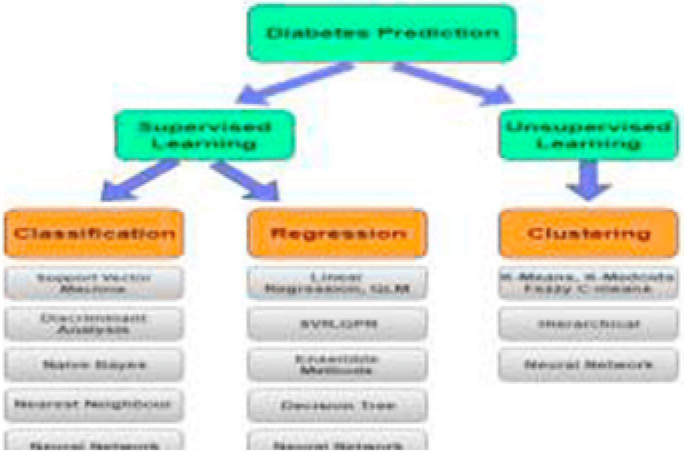


Figure 3: Taxonomy of Machine Learning Algorithms For Diabetes Prediction

A.The Supervised Learning/Predictive Models

Supervised learning algorithms/predictive models are used to construct predictive models. It predicts missing values using other values present in the dataset. It has a set of input data and also a set of output, and builds a model to make realistic predictions for the response to the new dataset. Supervised learning includes Bayesian Method,Decision Tree, Artificial Neural Network, Ensemble Method,Instance based learning. These are the most used techniques in Machine learning.

B. Unsupervised Learning / Descriptive Models

Descriptive models are developed using unsupervised learning methods. In this model we have a known set of inputs but output is unknown. Descriptive Models are mostly used on transactional data. Descriptive Models include clustering algorithms like k-Means clustering and k-Medians clustering.

C. Semi-supervised Learning

Semi Supervised learning method uses labeled and unlabeled data both on training dataset. Regression,Classification techniques come under Semi Supervised Learning. Linear Regression, Logistic Regression etc. are examples of regression techniques.

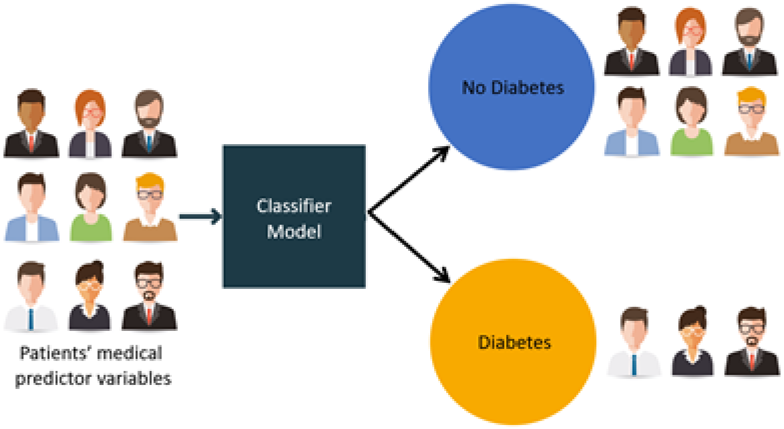


Figure 4: General Idea of Model

3. Gaps Identified

As we all know the drastic increase in the rate of people suffering from diabetes in the previous 15 years. Present trending human lifestyle is the main reason behind growth in diabetes which we can also see in various recent research. In current medical diagnosis methods, the gaps we have identified are:

1. In various causes,one of the causes is false-negative type in which a patient in reality is already a diabetic patient but test results are showing that the person is not having diabetes.

2. Second cause is false-positive type.Which is opposite of false-negative type.

3. Unclassifiable type is that type of cause in which a system cannot diagnose a given case. That happens due to insufficient knowledge of extraction from past data, a given patient may get predicted in an unclassified type. The patient must be able to determine whether to be in diabetic category or Non-diabteic category. Due to those errors in diagnosis may lead to unnecessary treatments or no treatments at all when required.

4. Current models which are there, are dependent on only a few methods and we know each method has its own advantages and disadvantages, so in that case the chances of incorrect prediction is Higher.

5. Currently there is no simple application in which a user can just input some data and get the results, current applications are not as such simply to use by people who don't know much about technology.

6. In our application we will be showing the user different probabilities of user having a diabetes according to the different algorithms and at last mean of all values for the final result also, so that if user will know how much is he or she having a risk at later stages of their life also, current systems lack such things.

7. In our model after analyzing the result and comparing with previous dataset any person from the medical field will be able to tell the main cause for the user diabetes like age,weight,insulin level etc. Current models as they exist are unable to do this.

8. From the above point if the cause is due to weight then research's show that the person has inherited the diabetes from their parents. Currently there are cases in which a person has diabetes but none of his or her parents have diabetes, and all also from other parameters it is not clear why the person has it? This is due to the limitation of the current system as it may cause diabetes to eventually be passed on to the next generation through inheritance.

4. PROPOSED WORK AND METHODOLOGY

In this section we are talking about our proposed work and also going to discuss methods to implement our model and also discuss all the tools and technologies we are using.

4.1. **Proposed Work:**

In this project, we will be providing a simple *GUI Application* which will allow users to enter their medical data which we on analysis will classify and label them as Diabetic or Non-Diabetic according to the dataset we will Take.

Our model, which we are supposed to make, will calculate the *probabilities* by using different Machine Learning Algorithms and taking the mean of all the values because each algorithm has its own set of Advantages and Disadvantages. So it is more definitive to take the mean of all such values so that chances of error will be reduced.

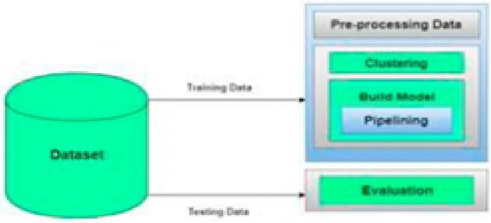


Figure 5: Diabetes Prediction Model

Fig 5, represents an architecture diagram for diabetes prediction model. This model has five unique modules, which include :

1. Dataset Collection
2. Data Pre-processing
3. Clustering
4. Build Model
5. Evaluation

After evaluating through each module, whose detailing will be given in the *Methodology* Section, we will figure out the *Result and Conclusion*, such that a person if diagnosed with diabetes can start treatment as soon as possible.

Our main aim is to build an accurate model.By which people can look after their health by taking care of some important steps, which can also reduce the patients of this deadly disease onword.

Figure 6: Urge to Stop Diabetes

4.2. Methodology:

In this section we will be describing each and every module, each and every algorithm in detail that is needed for our project.

4.2.1. Dataset Collection:

In this module it is seen to include data collection and to understand the data to study the specific trends and unique patterns which helps in prediction and evaluating the results. Dataset description is given below.

Table 1: Dataset Information

This table gives basic descriptions about attributes present in the pima dataset.

Sr.	Attribute Name	Attribute Description	Mean ± S.D
1	Pregnancies	Number of times a woman got pregnant	3.8±3.3
2	Glucose(mg/dl)	Glucose concentration in oral glucose tolerance test for 120 min	120.8±31.9
3	Blood Pressure (mmHg)	Diastolic Blood Pressure	69.1±19.3
4	Skin Thickness (mm)	Fold Thickness of Skin	20.3±15.9
5	Insulin(mu U/mL)	Serum for 2h	79.7±115.2
6	BMI(kg/m2)	Body Mass Index(weight/(height)^2)	31.9±7.8
7	Diabetes Pedigree Function	Diabetes Pedigree Function	0.4±0.3
8	Age	Age(Years)	33.2±11.7
9	Outcome	Class Variable	

4.2.2. Data Pre-processing:

This part of the model handles inconsistent data so as to be able to get more accurate and precise results. This dataset contains missing values. So we imputed missing values for a few selected attributes like Skin Thickness, Glucose level, Blood Pressure, BMI and Age because these attributes cannot have value as zero or null. Then we have scalize our given dataset in order to standardize all values.

We will be also categorizing the people according to different attributes and labeling them such that, this categorization will be helpful for further studies and research.

Table 2: Binning of age

Binning of age attribute is done as per given table.

Age(Years)	Age Bins
<=30	Youngest
31-40	Younger
41-50	Middle aged
51-60	Older
>=61	Oldest

Table 3: Binning of Glucose

Binning of Glucose attribute is done as per given table.

Glucose	Glucose Bins
<=60	Very Low
61-80	Low
81-140	Normal
141-180	Early Diabetes
>=181	Diabetes

Table 4: Binning of Blood Pressure

Binning of Blood Pressure attribute is done as per given table.

Blood Pressure	Diastolic Blood Pressure Bins
<61	Very Low
61-75	Low
75-90	Normal
91-100	Early Diabetes
>100	Diabetes

Table 5: Binning of BMI

Binning of BMI attribute is done as per given table.

BMI	BMI Bins
<19	Starvation
19-24	Normal
25-30	Overweight
31-40	Obese
>40	Very Obese

4.2.3 Model Building:

This phase is one of utter importance as it includes model building for prediction of diabetes. In this we make use of multiple different types of machine learning algorithms.These algorithms include,Linear Discriminant Analysis algorithm, Random Forest Classifier,, Extra Tree

Classifier, AdaBoost algorithm, Logistic Regression, K-Nearest Neighbor, Gaussian Naïve Bayes, Bagging algorithm, Gradient Boost Classifier, Perceptron.

4.2.4 MACHINE LEARNING ALGORITHMS TESTED:

There are in total 10 Machine Learning Algorithms that are tested, from each algorithm we calculate its accuracy and take that algorithm which has highest accuracy for our model and then pass the user data through it , to get final Probability.

The Machine Learning Algorithms are imported from the python in built Libraries

```
from sklearn.neighbors import KNeighborsClassifier as knn
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.naive_bayes import GaussianNB as gnb
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
from sklearn.ensemble import AdaBoostClassifier as ada
from sklearn.ensemble import RandomForestClassifier as rfc
from sklearn.linear_model import Perceptron
from sklearn.ensemble import ExtraTreesClassifier as etc
from sklearn.ensemble import BaggingClassifier as bc
from sklearn.linear_model import LogisticRegression as lr
from sklearn.ensemble import GradientBoostingClassifier as gbc
```

a) K Nearest Neighbor Algorithm (KNN):

K-Nearest Neighbor is the simplest Machine Learning algorithm. It is based on the Supervised Learning technique.

The K-Nearest Neighbor algorithm is seen to adopt the similarity among the new data case and available cases and it places the new case into the most similar of the available categories.

The K-NN algorithm keeps all the available data. On the basis of similarity K-NN is able to classify a new data point. Hence, when new data is able to be quite easily sorted into a their most suitable category by using the specialized K- NN algorithm. This could be used for Regression as well as for Classification but we used for the Classification Problems i.e. either diabetes or not diabetes.

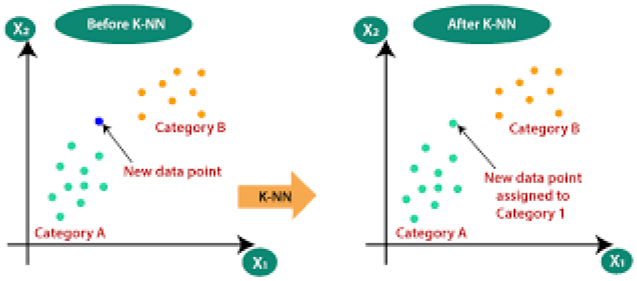


Figure 7: KNN Algorithm

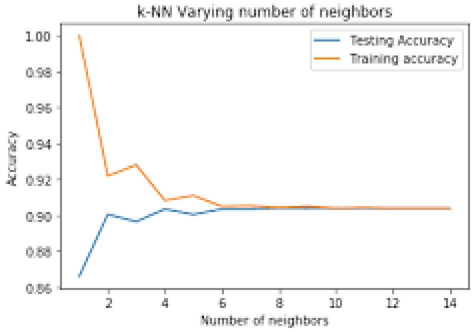


Figure 8: Accuracy of KNN

b) Gausssian Naïve Bayes (Gaussian NB):

When we are working with the continuous data (like blood pressure, skin thickness, sugar level), data which is not in an interval a consideration is always made that the continuous values associated with each of the classes are distributed or scattered according to a normal (or Gaussian) distribution. The likelihood of the features is assumed to be are as follows-

(Assuming that variance in some cases is to be considered)

- Should be independent of y (i.e., σi),
- Or Should be independent of x (i.e., σk)
- or both,sigma (i.e., σ)

GNB can support continuous valued features and models each as to obey a Gaussian distribution or what we also called Normal Distribution as well. Our aim is to create a very simple model, is to assume that the data is described by a Gaussian distribution with absolutely no or zero covariance which implies independent dimensions between the dimensions. This model can be fit anywhere by simply finding the mean(m) and standard deviation(sigma) of the points within each label, which is all we need to define a distribution.

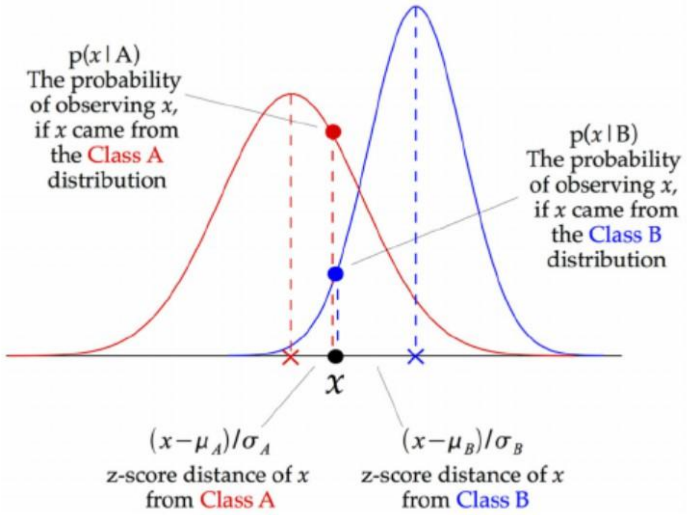

p(x | A)
The probability of observing x, if x came from the Class A distribution
p(x | B)
The probability of observing x, if x came from the Class B distribution
x
$(x-\mu_A)/\sigma_A$
z-score distance of x from Class A
$(x-\mu_B)/\sigma_B$
z-score distance of x from Class B

Figure 9: Working of GNB Classifier

c) Linear Discriminant Analysis (LDA):

LDA is an acronym of Linear Discriminant Analysis, which is a dimensionality reduction technique (like gender, patient name) that is commonly used for supervised Machine Learning classification problems.

LDA is used for. separating two or more classes. LDA is used to project the features in higher dimension space for e.g. 2D into a lower dimension space like 1D.

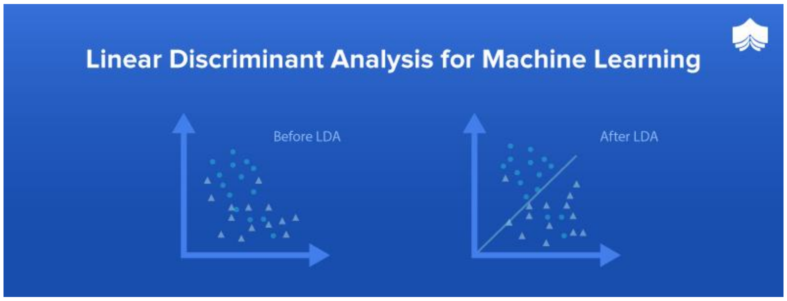


Figure 10: Plot of LDA

d) **Adaboost:**

AdaBoost algorithm, in which Ada stands for Adaptive ,It is a Boosting algorithm which is used as an Ensemble Method in *Machine Learning*.

AdaBoost is called Adaptive Boosting because the weights are re-assigned to each of the instances (in our project sugar level is a highly weighted instance), with higher weights assigned to incorrectly classified instances. Boosting helps to reduce bias as well as variance for supervised learning of Machine Learning Algorithms.

It works on the principle that instances are growing sequentially.Besides first, each subsequent instance is connected from previously connected instances. Simply, weak learners are converted into strong ones.

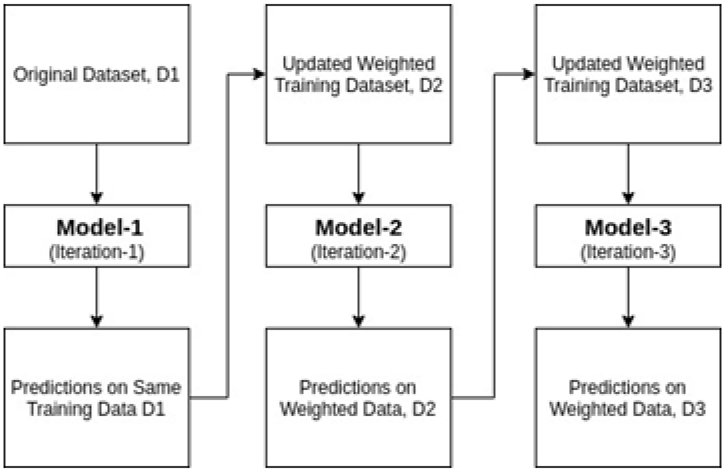

Original Dataset, D1
Updated Weighted Training Dataset, D2
Updated Weighted Training Dataset, D3
Model-1
(Iteration-1)
Model-2
(Iteration-2)
Model-3
(Iteration-3)
Predictions on Same Training Data D1
Predictions on Weighted Data, D2
Predictions on Weighted Data, D3

Figure 11: Working of Adaboost Classifier

e) Random Forest Classifier:

Random Forest classifier is a machine learning algorithm that belongs to the supervised learning technique. It can be used for classification and Regression problems both in ML (we used it as classifier). 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.

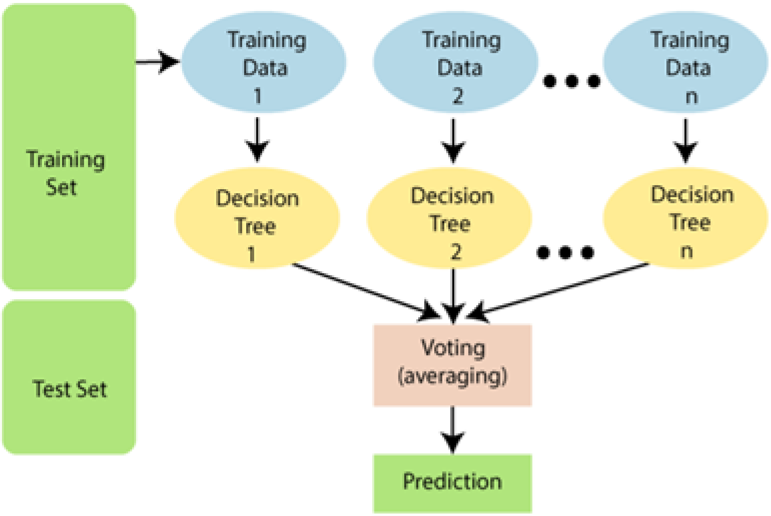


Figure 12: Working of Random Forest Classifier

f) **Perceptron:**

The Perceptron model is a supervised learning algorithm of binary classifiers (either diabetes or not). A one neuron, the perceptron model, detects whether any function is an input or not and classifies them in either of the classes.

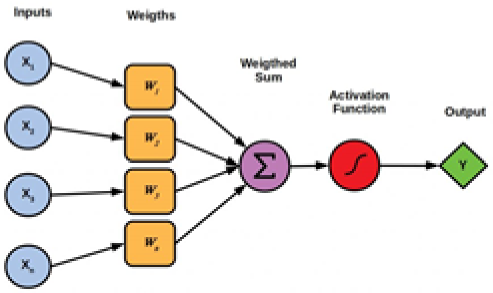
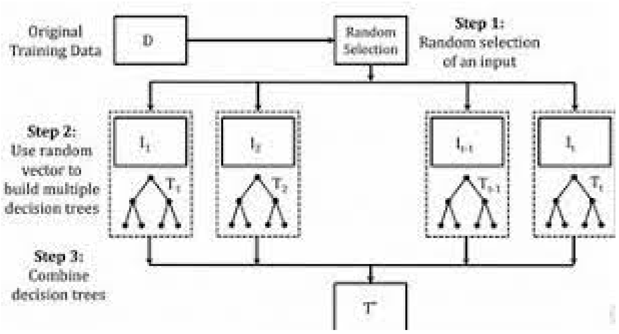


Figure 13: Working with Perceptron Model

g) <u>Extra Tree Classifier:</u>

This section involves a unique type of ensemble learning technique which is known as Extra Trees Classifier. This ETC works to combine the results of multiple, de-correlated decision trees which have been collected in a forest in order to output its classification result. From its theory it can be seen as being very similar to that of the Random Forest Classifier but its difference can be seen from the technique of the construction of the decision trees in the forest.

Prediction of diabetes at an early stage can lead to improved treatment, for this ETC can do better.


Original Training Data
D
Random Selection
Step 1: Random selection of an input
Step 2: Use random vector to build multiple decision trees
I_1
I_2
I_{t-1}
I_t
T_1
T_2
T_{t-1}
T_t
Step 3: Combine decision trees
T'

<u>Figure 14: Virtual representation of Extra Tree Classifier</u>

h) <u>Bagging:</u>

This ensemble learning method known as bagging, also called the bootstrap aggregation,is commonly used to reduce variance that is found inside of noisy datasets (noise due error during taking skin thickness).

In the process of bagging, 2 random data sets are chosen with one being the replacement. Which means that the one data point is able to be picked more than once.

Once multiple data samples have been generated , depending on the type of task whether it is regression or classification. The weak models are then trained in a non dependent manner.

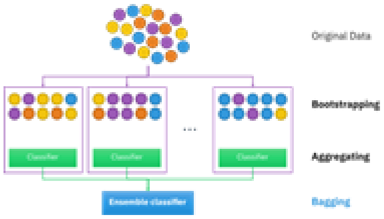


Figure 15: Working of Bagging

i) **Logistic regression:**

This is known as one of the most utilized and important ML algorithms, and is said to come under the Predictive Models. Logistic regression is utilized using a given number of independent variables in order to predict the categorical dependent variable.

Logistic regression is able to predict the categorical dependent variable. Therefore the outcome must be a discrete value. It can be either 0 or 1,Yes or No, True or False, etc. but instead of giving 0 and 1, it values in the range of (0,1). It will be helpful to determine relations between various instances, that help for better classification.

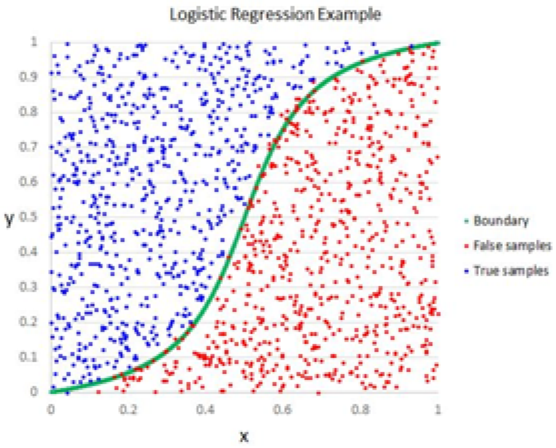


Figure 16: Logistic Regression

j) Gradient Boost Classifier:

We can say that Gradient boosting classifiers are a group of machine learning algorithms but not a single one that combines many weak learning models together to create a strong predictive model which can be used. That strong model is known as the Improved Model.

When we are doing Gradient Boosting, generally Decision Tree is used to build the model.

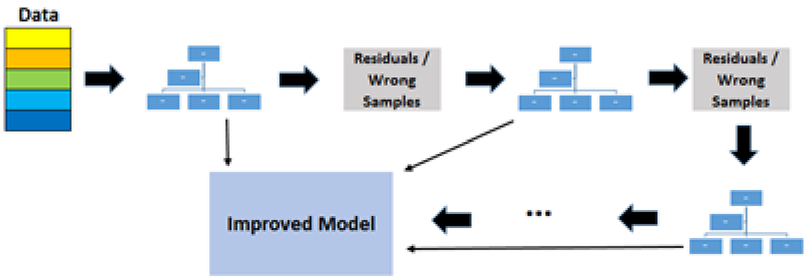


Figure 17: Gradient Boost Classifier working

4.2.5 Algorithm:

Diabetes Prediction using various machine learning algorithms

```
Generate training set and test set randomly.

Specify algorithms that are used in model

mn=[ KNN( ), GaussianNB( ), LDA(), AdaBoost(), RandomForestClassifier(), Perceptron(),
        ExtraTreeClassifier(), Bagging(), LogisticRegression(), GradientBoostClassifier()]

for(i=0; i<10; i++) do
  Model= mn[i];
  Model.fit();
  model.predict();
  print(Accuracy(i),confusion_matrix, classification_report);
End
```

4.2.6 Evaluation:

This was the last step of our prediction model. Now we evaluate the prediction results of our model using various parameters like classification accuracy, confusion matrix and f1-score

a) Classification Accuracy:

Accuray is defined as the ratio of the number of correct predictions to the total number of input samples given.

*A*ccuracy = No. of right Predictions /Total no. of predictions Made for our model

b) Confusion Matrix:

It will give us a 2-D Array(Matrix) as output and describes the Overall performance of the model.

Table 6: Confusion Matrix

This table gives a basic idea about the confusion matrix.

Positive (1)	Negative(0)
True Positive (1,1)	False Positive (1,0)
False Positive (0,1)	True Negative (0,0)

Accuracy for the matrix can be calculated by the following Formula:

Accuracy = ((*T*P+*F*N))/ *N* Where, N:Total number of samples

c) F1 Score:

-It is used to measure the test's accuracy.It is the Harmonic Mean between precision and recall.

The range for F1 Score is somewhat between 0 to 1. It tells us the precision of our model and also how robust our model is. *F*1 = (2 ∗ 1 /((1/Precision)+(1/recall)))

d) Precision:

Precision is defined as the ratio of the no of correct positive results to the no of positive results predicted by our model .

Precision= (*T*P/ (*T*P + *F*P))

e) Recall:

Recall is defined as the ratio of the no of correct positive results to the no of all the relevant samples.

Recall= (*T*P/ (*T*P + *F*N))

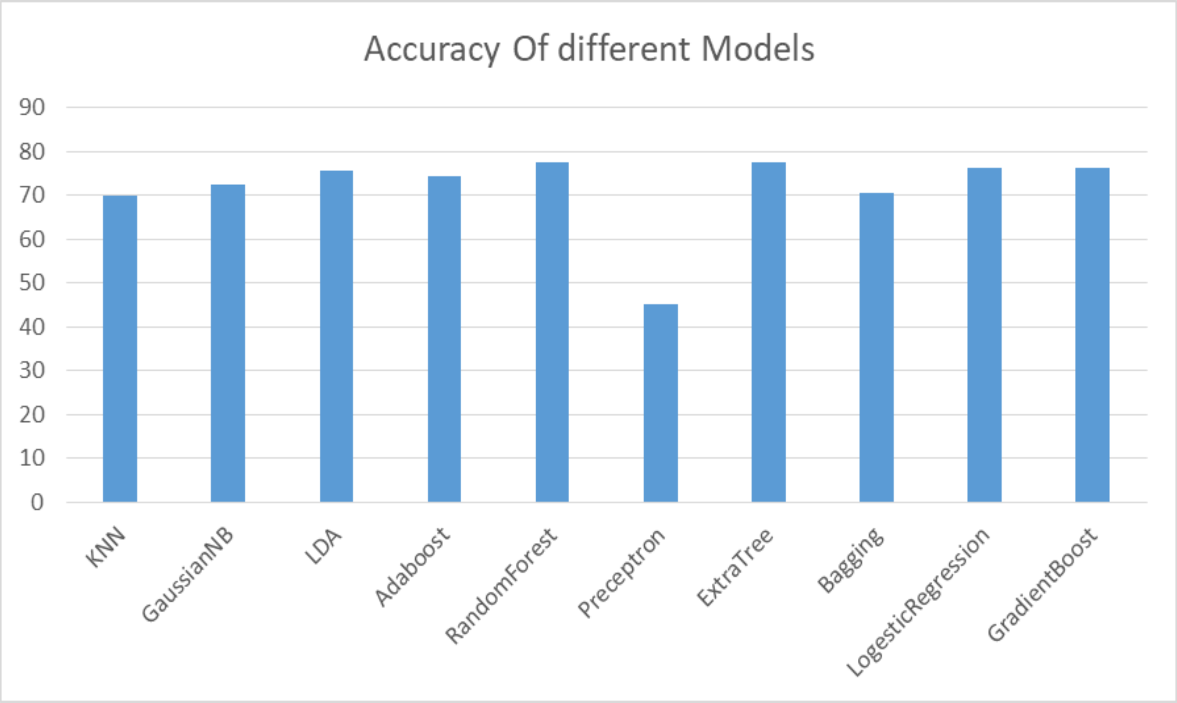


Figure 18: Comparison of Various ML Algorithms based on accuracy

4.3. TOOLS AND TECHNOLOGIES

4.3.1 Python3

The base Programming language that we are using to formulate the whole application

4.3.2 Flask

To develop User Interface we are using Python Framework Flask

4.3.3 Sklearn Library

We'll apply Machine Learning Algorithms for prediction of Diabetes by using this library

4.3.4 Jinja Templating

It contains variables and some programming logic ,when rendered into html are replaced with actual values.

4.3.5 HTML,CSS and Bootstrap

To develop the user interface of the web app

4.3.6 Numpy

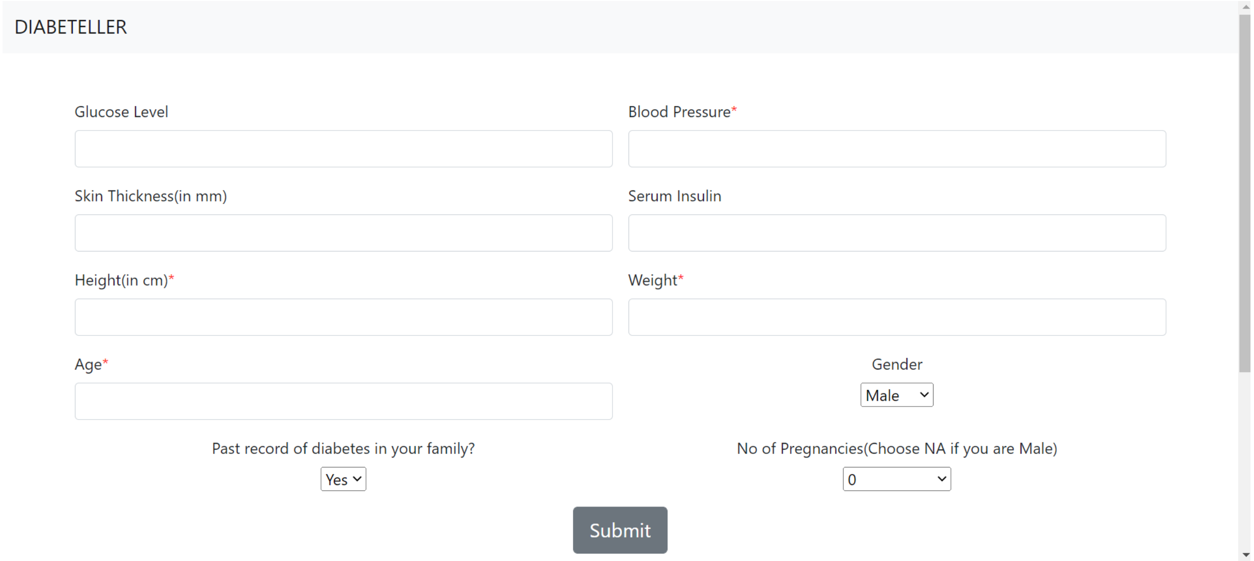
It is an open source Python library used to perform various mathematical and scientific tasks

4.3.7 Pandas

It is a python library that is used for data analysis and importing data from various file formats and manipulating it.

5) RESULTS

Firstly user has to enter the basic details some are optional and some are required



DIABETELLER

Glucose Level

Blood Pressure*

Skin Thickness(in mm)

Serum Insulin

Height(in cm)*

Weight*

Age*

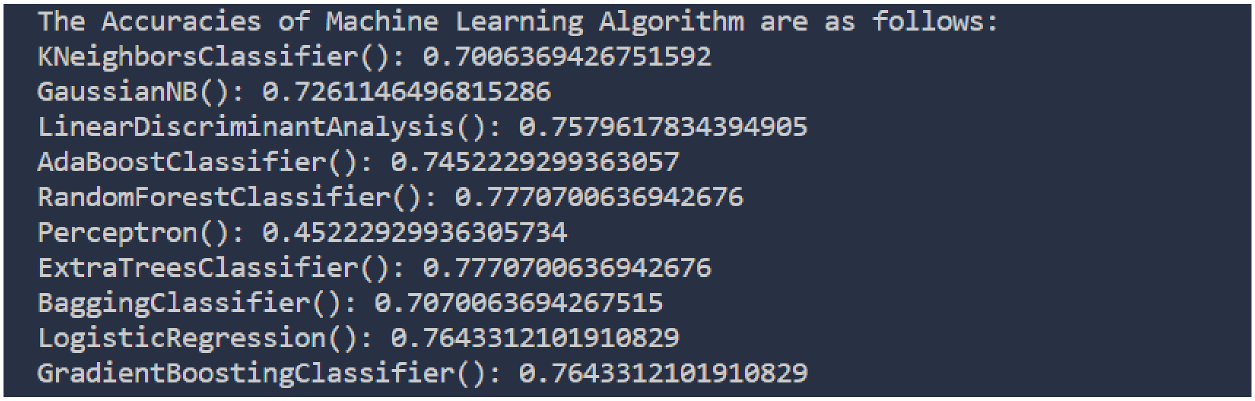
Gender: Male

Past record of diabetes in your family? Yes

No of Pregnancies(Choose NA if you are Male) 0

Figure 19: Basic User Interface of Model

After submitting the information firstly data cleaning of our dataset happens and then accuracy of all the Machine Learning Algorithms are calculated

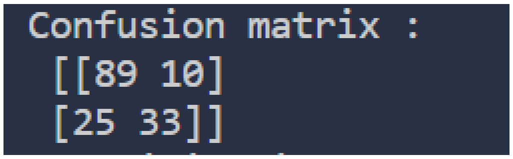


```
The Accuracies of Machine Learning Algorithm are as follows:
KNeighborsClassifier(): 0.7006369426751592
GaussianNB(): 0.7261146496815286
LinearDiscriminantAnalysis(): 0.7579617834394905
AdaBoostClassifier(): 0.7452229299363057
RandomForestClassifier(): 0.7770700636942676
Perceptron(): 0.45222929936305734
ExtraTreesClassifier(): 0.7770700636942676
BaggingClassifier(): 0.7070063694267515
LogisticRegression(): 0.7643312101910829
GradientBoostingClassifier(): 0.7643312101910829
```

Figure 20: Accurasies of all Machine Learning Algorithms

Now The maximum accuracy is 0.7770700636942676 which is of RandomForestClassifier() so it is selected for prediction.

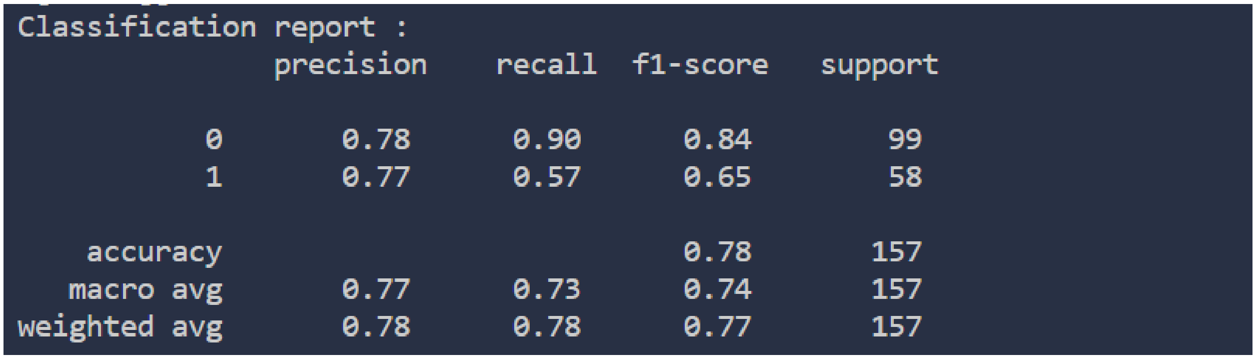
Confusion Matrix is Calculated from the data



```
Confusion matrix :
 [[89 10]
 [25 33]]
```

Figure 21:Confusion Matrix

Classification report is also prepared in which tells us about F1-Score,Recall,Precision and other various Parameters



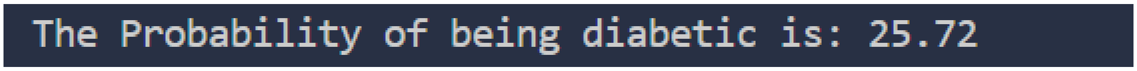
```
Classification report :
              precision    recall  f1-score   support

           0       0.78      0.90      0.84        99
           1       0.77      0.57      0.65        58

    accuracy                           0.78       157
   macro avg       0.77      0.73      0.74       157
weighted avg       0.78      0.78      0.77       157
```

Figure 22: Classification Report

Now The Probability is Calculated



```
The Probability of being diabetic is: 25.72
```

Figure 23: Probability of being diabetic

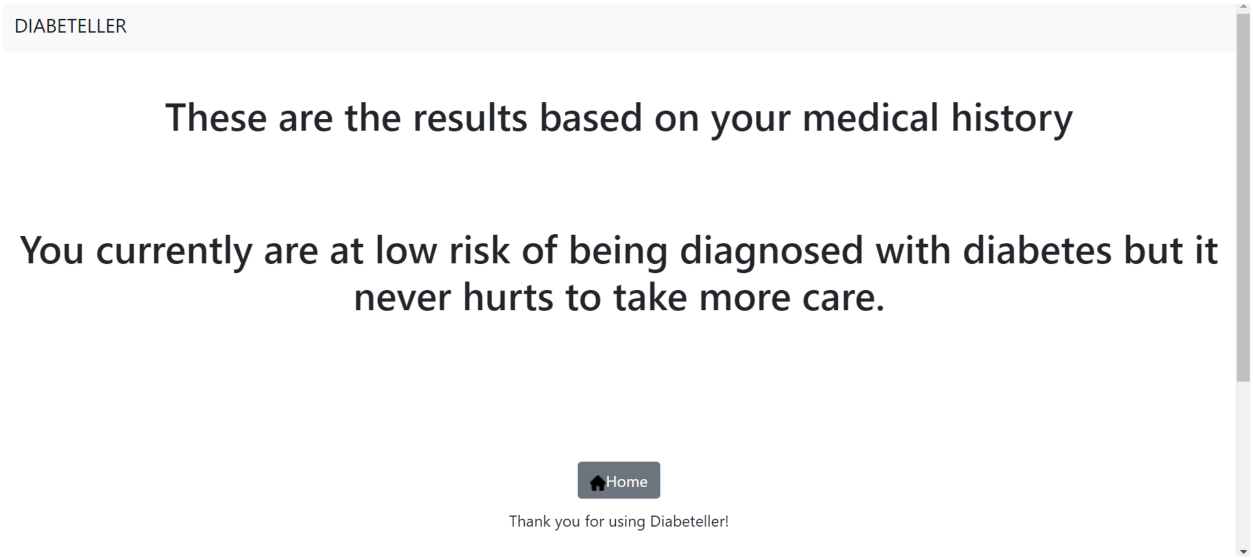
Based on the Probability user will can see one of four types of messages

5.1 If Probability greater than 80, user will get: Alert! You seem to be at a very high risk of getting diabetes.Please seek immediate medical attention

5.2 If Probability greater than 50 but less than 80, user will get: You are at a low risk of getting diabetes

5.3 If Probability greater than 20 but less than 50, user will get: You currently are at low risk of being diagnosed with diabetes but it never hurts to take more.

5.4 If Probability is less than 20, user will get: All's fine! You seem to have very low chances of getting diabetes.



These are the results based on your medical history

You currently are at low risk of being diagnosed with diabetes but it never hurts to take more care.

Thank you for using Diabeteller!

Figure 24: Final Result for User

6) CONCLUSION

In this project, we are implementing multiple different Machine learning algorithms and trying to fill the gaps present in the current model of detecting diabetes.

We are aiming to make an application which will enable people to check the probability of suffering from diabetes in the future. This is accomplished by entering some of the basic information such as Age, glucose, blood pressure, weight, height etc.

Our primary goal is to try to make people aware and minimize the time which is needed in order to diagnose the possibility of diabetes in the future.

7) REFERENCES

[1] Eurek alert, "Insufficient sleep may be reason to increased diabetes risk," July 11, 2010

[2] Bhatt K., Dalal P., Panwar A., “A Cluster Centres Initialization Method for Clustering Categorical Data Using Genetic Algorithm” Volume-2 Issue-1.,for clusterization.in International Journal of Digital Application & Contemporary research, 2013.

[3] Huang, Zhexue. "A Fast-Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining." DMKD. 1997.

[4] Dr. Vaidehi Vb Diabetes Prediction using Machine Learning Algorithms

[5] Ms. Shraddha Kumar Asst. Professor, Ms Mani Butwall and Sushila Devi . A Data Mining Method for the Diagnosis of Diabetes suggested using Random Forest Classifier. Bansal College of Technology, Indore (MP)

[6] Aiswarya Iyer, S. Jeyalatha and Ronak SumbalyDIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUES

[7] All The Images ,data,graphs and table are taken from Google

[8] Geeks for geeks machine Learning

[9] Idea of algorithms from wikipedia Page

[10] Overview of Data Analysis from Tutrorials Point

[11] Data Sciene Techniques From Wikipedia Page.

[12]https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes#:~:text=Diabetes+is+a+disease+that,to+be+used+for+energy

[13]https://drive.google.com/file/d/1IRlJBa4KxDzQr0u1_MuZ7h5Gz7-PCnOw/view?usp=sharing

[14]https://drive.google.com/file/d/1b0kcXhYsJfUuVqbdkhxyOhS8kUuuToUz/view?usp=sharing

[15]https://drive.google.com/file/d/1f0Te1STxo1UpLr_q_bEh6ob3_0E_XJiB/view?usp=sharing

[16]
https://drive.google.com/file/d/1X2YwiktLYV10zyuMGwB2L0-pOR9woik_/view?usp=sharing

[17] https://ieeexplore.ieee.org/document/9441935