SHRI VAISHNAV VIDYAPEETH VISHWAVIDAYALAYA

SHRI VAISHNAV INSTITUTE OF INFORMATION TECHNOLOGY

Department of Computer Science and Engineering

2020-2021

MINOR PROJECT

PREGNANT WOMEN DIABETES PREDICTION SYSTEM

GUIDED BY:

Prof. Shubham Kothari

SUBMITTED BY:

Harsh Agrawal
Saloni Patil
Vanshika Sawle

Contents

Contents of Tables

Diagrams

Pregnant Women Diabetes Prediction System

Introduction

The aim of this project is to develop a system that can predict the diabetic risk for a pregnant lady with higher accuracy. With the rise of machine learning we have developed a system using different significant attributes and the relationship among them.

This project is about predicting whether the pregnant woman will have diabetes or not. By using machine learning algorithms we developed this prediction system for pregnant women.

First of all we got the dataset of pregnant ladies, then we applied the different classification models like Logistic regression, Support Vector Machine (SVM), Decision tree, Random forest, K-Nearest Neighbours, Naive bayes, kernel SVM and then by deciding the significant attributes and the model with the greater accuracy we developed a system that predicts whether the woman will have diabetes during pregnancy or not which is known as Gestational Diabetes mellitus (GDM).

It is a web application based project in which first of all users has to register her and then after registering an account is created which can be used for login. After the user is logged into the system she has to fill the prediction form which contains some attributes required for predicting diabetes. After submitting the form the result will be displayed on the screen, if the woman will have diabetes in future it will suggest some precautions to follow and the list of doctors to whom she can contact.

Diabetes

Diabetes is a common, chronic disease which creates different types of diseases like heart attack, blindness, kidney disease etc. it is a disorder that occurs due to abnormal secretion of insulin from the pancreas which increases the glucose level in the body. The early identification is the only remedy to stay away from the complications.

It is a disease that affects how your body converts food into energy. There are three types of diabetes.

1. Type -1: The body does not produce insulin.
2. Type -2: The body produces insulin but does not use it well.
3. Gestational Diabetes: It is a type of diabetes that occurs during pregnancy.

Gestational diabetes mellitus (GDM) is defined as any degree of glucose intolerance with onset or first recognition during pregnancy. The definition applies whether insulin or only diet modification is used for treatment and whether or not the condition persists after pregnancy. It does not exclude the possibility that unrecognized glucose intolerance may have antedated or begun concomitantly with the pregnancy.

Machine Learning

Machine learning is the field of computer science in which machine learns from the experiences. The purpose of machine learning is the construction of system that can adapt and learn from their experiences. The machine learning algorithms are classified into three types that are supervised learning, unsupervised learning and semi-supervised learning or reinforcement learning. The supervised learning algorithms are classified into different types such as probability-based, function-based, rule-based, tree-based, instance-based, etc. The unsurprised learning is the descriptive type learning. This learning is used to describe or summarize the data. The examples of the unsupervised learning algorithms are clustering, association rule mining, etc. The semi-supervised learning is the combination of supervised and unsupervised.

Literature Survey

The Diabetes predication system for pregnant lady is an attempt for helping pregnant ladies to detect their blood sugar level by using Machine learning. In this section other researcher's work which is relevant to the present study are presented.

1. S. R. Surya Msc., MCA. M.Phil. , in his research paper concluded that, In this survey paper mainly focused on predict Diabetes mellitus by using big data analytics. According to this analysis the most of the paper covers the algorithm Hadoop/Map reduce environment to predict the diabetes types prevalent and complications of patient with it. In each of the paper they used different dataset for analytics. For the calculation they used decision tree algorithm. So, the result is quite good. It is possible to future improve the diabetes mellitus to use any machine learning algorithm / for efficiency of the prediction.

2. Sai Poojitha Nimmagadda, Sagar Yeruva, Rakesh Siempu in their research paper stated Improved Diabetes Prediction Model for Predicting Type-II Diabetes concluded, In the present study, the main objective is to find a model that predicts Diabetes Mellitus in people when given inputs and it provides higher accuracy rate than the existing models. In order to compare different models, multiple classification algorithms and clustering algorithms were used and implemented. The models include K-means with logistic regression which has got accuracy 94.6%, Hierarchical clustering with Logistic Regression has got accuracy 94.1%, Hierarchical clustering with SVM has accuracy 94.1, Hierarchical clustering with Decision Tree has accuracy 90% and our proposed model IDPA has got accuracy 96.07%. We can say that IDPA has highest accuracy when compared with other models and other researches models. This model when included in real time applications in healthcare sector can be used to predict diabetes with greater accuracy. The model can be enhanced by using real time dataset or hospital patient's data. It would be beneficial if user gets a mobile application which not only predicts diabetes or no diabetes but also stores the patient information.

3. PREDICTION OF DIABETES MELLITUS USING DATA MINING TECHNIQUES: A REVIEW states, Different approaches for the prediction of Diabetes Mellitus and its types are concentrated in this study. Data mining is a technique used to extract

useful information from existing large volume of data which enable us to gain more knowledge. In this way data mining techniques are applied in health care sector in order to predict various diseases and to find out efficient ways to treat them as well.

4. Artificial intelligence is having more effect is machine realizing, which creates calculations ready to take in examples and choice standards from information. Machine learning calculations have been implanted into information mining pipelines, which can consolidate them with established measurable techniques, to remove learning from information. Inside the EU-financed MOSAIC undertaking, an information mining pipeline has been utilized to determine an arrangement of prescient models of sort 2 diabetes mellitus (T2DM) entanglements in light of electronic wellbeing record information of almost one thousand patients. Such pipeline includes clinical focus profiling, prescient model focusing on, prescient model development and model approval. In the wake of having managed to miss information by methods for irregular woods (RF) and having connected appropriate methodologies to deal with class unevenness, we have utilized Logistic Regression with the stepwise component choice to foresee the beginning of retinopathy, neuropathy, or nephropathy, at various time situations, at 3, 5, and 7 years from the main visit at the Hospital Center for Diabetes (not from the conclusion). Considered factors are sexual orientation, age, time of determination, weight file (BMI), glycated haemoglobin (HbA1c), hypertension, and smoking propensity. Lust models, custom fitted as per the complexities, gave an exact up to 0.838. Diverse factors were chosen for every complexity and time situation, prompting particular models simple to mean the clinical practice.

5. In this paper, analysis of a Pima Indian dataset is done using various classification techniques like Naïve Bayes, Zero R, J48, random forest, MLP, logistic regression. Comparison and prediction whether positive and negative diabetes. Diagnosing diabetes through data mining tool using the WEKA tool, in terms of accuracy and performance MLP is better.

6. Implementation of dietary and lifestyle interventions prior to and early in pregnancy in high risk women has been shown to reduce the risk of gestational diabetes mellitus (GDM) development later in pregnancy. Although numerous risk factors for GDM have been identified, the ability to accurately identify women before or early in pregnancy who could benefit most from these interventions remains limited.

7. Gestational diabetes (GDM) refers to the normal metabolism of glucose before pregnancy and the occurrence of diabetes during pregnancy. This disease is a serious threat to the health of this pregnant woman and infant, so it is important to accurately predict whether the target is a gestational diabetes patient based on various indicators. Based on the measured data of the hospital, this paper uses decision tree, logistic regression and DenseNet to predict the target when the disease is sick or to be sick in the future, and discuss their prediction accuracy separately, which can help doctors make rapid diagnosis and make timely prevention. In the end, it was found that the DenseNet model can better predict whether the target is gestational diabetes or not, and the model flexibility is better.

Problem Identified-

Many people do not know that they are suffering from diabetes and many complications occur if diabetes remains untreated. It is a serious health matter during which the measure of sugar substance cannot be controlled. The early identification is the only remedy to stay away from the complications.

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc.

Gestational diabetes mellitus (GDM) is a severe and neglected threat to maternal and child health. Many women with GDM experience pregnancy-related complications including high blood pressure, large birth weight babies and obstructed labour. Approximately half of women with a history of GDM go on to develop type 2 diabetes within five to ten years after delivery.

The prevalence of high blood glucose (hyperglycaemia) in pregnancy increases rapidly with age and is highest in women over the age of 45.

In 2019

- There were an estimated 223 million women (20-79 years) living with diabetes. This number is projected to increase to 343 million by 2045.
- 20 million or 16% of live births had some form of hyperglycaemia in pregnancy. An estimated 84% were due to gestational diabetes.
- 1 in 6 births was affected by gestational diabetes.
- The vast majority of cases of hyperglycaemia in pregnancy were in low- and middle-income countries, where access to maternal care is often limited.

It is important for women with diabetes in pregnancy or GDM to carefully control and monitor their blood glucose levels to reduce the risk of adverse pregnancy outcomes with the support of their healthcare provider.

Solution Proposed

The diabetes in pregnancy or GDM is such a big threat for both the child and the mother if it is remain untreated. But if the mother gets the idea about the likelihood of her to be diabetic, she can take care of herself.

If by somehow, she got a prior knowledge of her chances to be diabetic and also the ways i.e. either precautions taken by her or contacting doctors for this, she can be treated.

And she can get the idea about her condition, by referring to our website.

We had tried to make a website where the pregnant lady by inputting some of her health-related information will get to know is she diabetic or not, and if she found to be diabetic, then she will be provided some precautions and the way she can contact to a doctor.

Since we are using a machine learning model for prediction of her to be diabetic, and as it is said that,

"All models are wrong, but some are useful."

So if for someone, model has predicted positive result, a proper medical check-up is required for diagnosis and then treatment.

The health related information which is required for filling the form , are those factors who can affect or are directly related to the probability of her being diabetic.

We had first done a comparative study using the dataset and applying different classification machine learning models on it. Accordingly, we selected that model who was giving maximum accuracy and then we used that model in our project.

Methodology Used

5.1 Model diagram

Proposed procedure is given below in the form of model diagram. The figure shows the flow of the research conducted in constructing the model.

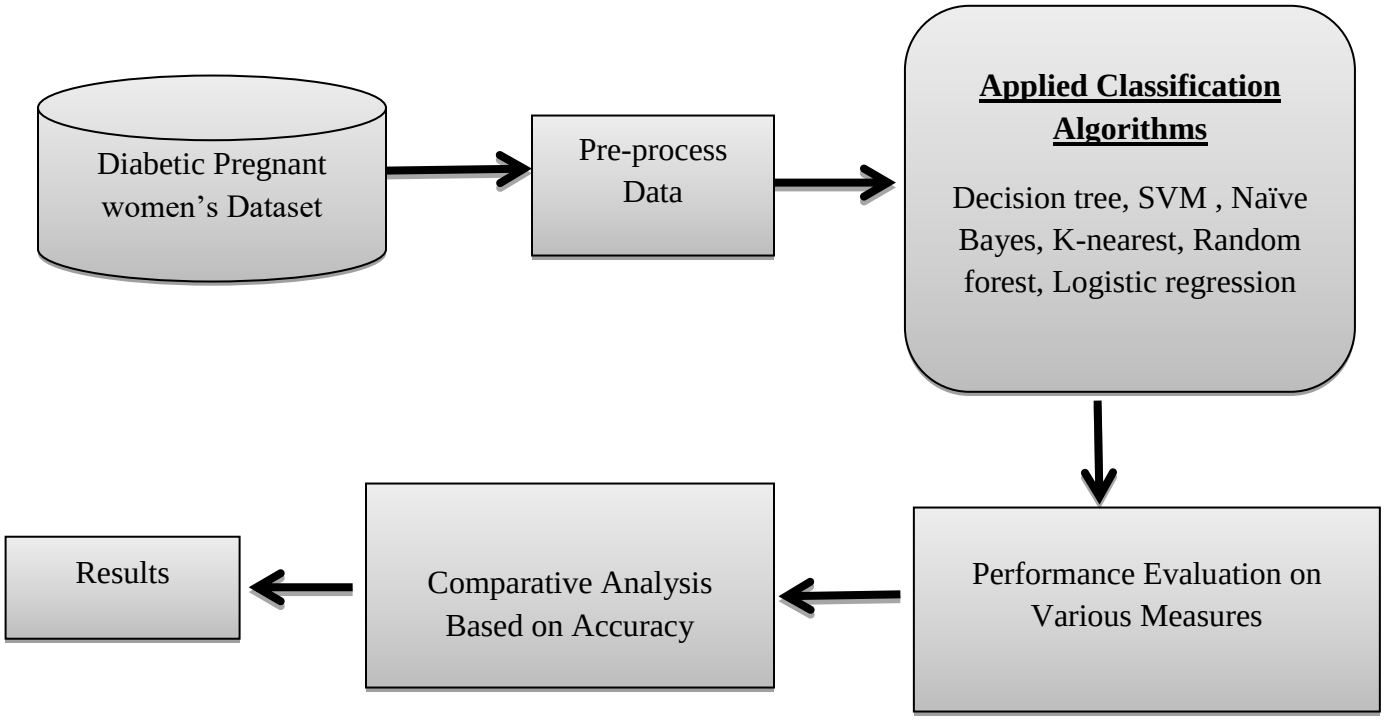


Fig 1: Proposed Model Diagram

5.2 Brief Description of Algorithms Used

5.2.1 Logistic Regression

Logistic regression is a statistical technique used to predict probability of binary response based on one or more independent variables. It uses logistic or sigmoid function for prediction.

The logistic function or the sigmoid function was developed by statisticians so that they can describe the properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment.

It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

Sigmoid Function: p = 1/ (1+e^-y) Or y=ln(p/1-p)

Hence, we can conclude

ln(p/1-p)=b0 + b1*x + b2*x +……

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the input value (x) and ….

After training the logistic regression model on training set of our dataset, we will use the model for predicting the outcome of inputs of test set.

Then we will compare the predicted outcome with the actual outcome mentioned in the test set using confusion matrix.

N=192	Predicted: NO	Predicted: YES
Actual: NO	93	14
Actual: YES	21	26

Table 1: Confusion Matrix of Logistic Regression

5.2.2 K-Nearest Neighbours (K-NN)

The K-Nearest Neighbours or K-NN is a non-parametric method i.e. KNN makes no assumptions about the functional form of the problem being solved can be used. It can be used for both regression and classification.

For building an knn model we first choose any k no. of neighbours , then we determine the k-nearest neighbours of the new data point (using Euclidean distance , Hamming Distance , Manhattan Distance or Minkowski Distance) .Among those k neighbours , it counts the number of data points in each category. Finally, it assigns the new data point to the category where it counted the most neighbours.

K-NN is also called as instance-learning, as raw training instances are used to make predictions. Similarly, it is called as lazy learning because no learning of the model is required and all of the work happens at the time a prediction is requested.

Class probabilities can be calculated as the normalized frequency of samples that belong to each class in the set of K most similar instances for a new data instance. For example, in a binary classification problem (class is 0 or 1):

p (class=0) = count (class=0) / (count(class=0)+count(class=1))

After training the K-NN model on training set of our dataset, we will use the model for predicting the outcome of inputs of test set.

Then we will compare the predicted outcome with the actual outcome mentioned in the test set using confusion matrix.

N=192	Predicted: NO	Predicted: YES
Actual: NO	112	18
Actual: YES	26	36

Table 2: Confusion Matrix of KNN

5.2.3 Support Vector Machine (SVM)

SVM is one of the standard set of supervised machine learning model employed in classification. Given a two-class training sample the aim of a support vector machine is to find the best highest-margin separating hyper plane between the two classes. For better generalization hyper plane should not lies closer to the data points belong to the other class. Hyperplane should be selected which is far from the data points from each category. The points that lie nearest to the margin of the classifier are the support vectors.

The SVM finds the optimal separating hyperplane by maximizing the distance between the two decision boundaries. It is not biased by outliers, as well as not sensitive to over fitting but it is not appropriate for non linear problems and is also not the best choice for large number of features as it is quite complex.

After training the SVM model on training set of our dataset, we will use the model for predicting the outcome of inputs of test set.

Then we will compare the predicted outcome with the actual outcome mentioned in the test set using confusion matrix.

N=192	Predicted: NO	Predicted: YES
Actual: NO	117	13
Actual: YES	27	35

Table 3: Confusion Matrix of SVM

5.2.4 Kernel SVM

Kernel SVM or SVM with kernel trick is used to deal non-linear problems by the SVM. Since we know that SVM uses a hyperplane to make a decision boundary between the 2 categories or class of the outcome, but if the points are located as such that it becomes impossible to simple separate out those points using a plane then we use the kernel trick there.

It does it via mapping the non linearly separable data points in the higher dimension and making it linearly separable. These transformations are called kernels. Popular kernels are: Polynomial Kernel, Gaussian Kernel, Radial Basis Function (RBF), Laplace RBF Kernel, Sigmoid Kernel, Anove RBF Kernel, etc. We can select the best for our model using parameter tuning.

It is not biased by outliers and it is not sensitive to over fitting.

But still it is not the best choice for large number of features as it is quite complex.

After training the Kernel SVM model on training set of our dataset, we will use the model for predicting the outcome of inputs of test set.

Then we will compare the predicted outcome with the actual outcome mentioned in the test set using confusion matrix.

N=192	Predicted: NO	Predicted: YES
Actual: NO	117	13
Actual: YES	27	35

Table 4: Confusion Matrix of Kernel SVM

5.2.5. Naive Bayes

Naive Bayes is a classification technique with a notion which defines all features are independent and unrelated to each other. It defines that status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability it is considered as a powerful algorithm employed for classification purpose.

It works well for the data with imbalancing problems and missing values. Naive Bayes is a machine learning classifier which employs the Bayes Theorem. Using Bayes theorem posterior probability P (C|X) can be calculated from PI, P(X) and P(X|C). Therefore,

$$P(C|X) = (P(X|C)\, PI)/P(X)$$

Where,

P (C|X) = Target class's posterior probability.

P (X|C) = Predictor class's probability.

P I = Class C's probability being true.

P (X) = Predictor's prior probability.

After training the Naïve Bayes model on training set of our dataset, we will use the model for predicting the outcome of inputs of test set.

Then we will compare the predicted outcome with the actual outcome mentioned in the test set using confusion matrix.

N=192	Predicted: NO	Predicted: YES
Actual: NO	114	16
Actual: YES	29	33

Table 5: Confusion Matrix of Naïve Bayes

5.2.6. Decision Tree Classification

Decision Tree is a supervised machine learning algorithm used to solve classification problems. The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data.

It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, Decision tree chooses each node by evaluating the highest information gain among all the attributes.

To define information gain precisely, we need to define a measure commonly used in information theory called entropy that measures the level of impurity in a group of examples. Mathematically, it is defined as:

Entropy: $\sum i=1-p*\log2\ (pi)$ **[pi=Probability of class i]**

Gain: $Gain(S,A) = Entropy(S) - \sum v \text{ belongs to } Values\ (A)\ (|Sv|/|S|).Entropy\ (Sv)$

where,

S is a set of instances
A is an attribute,
Sv is the subset of S with A = v
Values (A) is the set of all possible values of A.

After training Decision Tree Classification model on training set of our dataset, we will use the model for predicting the outcome of inputs of test set.

Then we will compare the predicted outcome with the actual outcome mentioned in the test set using confusion matrix.

N=192	Predicted: NO	Predicted: YES
Actual: NO	104	26
Actual: YES	17	45

Table 6: Confusion Matrix of Decision Tree Classification

5.2.7. Random Forest Classification:

In Random Forest Classification, we use the concept of ENSEMBLE learning which means taking multiple machine learning algorithms and put them together to form a final one, hence the final one leveraging different machine learning algorithms. Random forest runs decision tree multiple times.

For making random forest model first it need to pick at random K data points from the Training set and then build the decision tree associated to those K data points. Then choose the number Ntree of tress we want to build and then again repeat the above steps.

For a new data point, each one of the Ntree trees predict the category to which the data points belongs, and assign the new data point to the category that wins the majority vote.

After training Random Forest Classification model on training set of our dataset, we will use the model for predicting the outcome of inputs of test set.

Then we will compare the predicted outcome with the actual outcome mentioned in the test set using confusion matrix.

N=192	Predicted: NO	Predicted: YES
Actual: NO	110	20
Actual: YES	24	38

Table 7: Confusion Matrix of Random Forest Classification

5.3 Dataset

In this work Anaconda Navigator is used for performing the experiment. Anaconda is the most popular data science platform and the foundation of modern machine learning. The biggest advantages of using Anaconda is-

- Collect data from files, databases, and data lakes
- Manage environments with Conda (all package dependencies are taken care of at the time of download)
- Share, collaborate on, and reproduce projects
- Deploy projects into production with the single click of a button

The main aim of this study is the prediction of the pregnant lady is affected by diabetes launching IDE Spyder via Anaconda Navigator by using the medical database diabetes.csv.

Dataset	Rows (Number of Observations)	Columns (Number of Attributes)
diabetes	768	9

Table 8: Description of dataset

For the project work, we split the number of observations in training data (which will train the model) and testing data (on which we will test the model).

We split the data with test size containing 25% of the whole observations.

Hence for training set we have 75% of 768 observations i.e. 576 observations

And for test set we have 25% of 768 observations i.e. 192 observations.

The dataset consists of following attributes-

1. Pregnancies – Number of times pregnant
2. Glucose – Plasma glucose concentration over 2 hours in an oral glucose tolerance test
3. Blood Pressure – Diastolic blood pressure (mm Hg)
4. Skin Thickness – Triceps skin fold thickness (mm)
5. Insulin – 2-Hour serum insulin (mu U/ml)
6. BMI – Body mass index (weight in kg/(height in m)2)
7. Diabetes Pedigree Function- Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
8. Age – Age (years)
9. Outcome – Class variable (0 if non-diabetic, 1 if diabetic)

For our project we find significant attributes which affects the most to the outcome and apply the model on it.

5.4 Accuracy Measures

Measures	Definitions	Formula
1. Accuracy(A)	Accuracy determines the accuracy of the algorithm in predicting instances.	A=(TP+TN) / (Total no of samples)
2. Precision(P)	Classifier's correctness/accuracy is measured by Precision.	P = TP / (TP+ FP)
3. Recall I	To measure the classifier¢¢s completeness or sensitivity, Recall is used.	R =TP / (TP+FN)
4. F-Measure	F-Measure is the weighted average of precision and recall.	F=2*(P*R) / (P+R)

Table 9: Accuracy Measures

Classification Algorithms	Precision	Recall	F-Measure	Accuracy(in %)
1. Logistic Regression	0.65	0.55	0.59	0.79
2. K-NN	0.67	0.58	0.62	0.77
3. SVM	0.73	0.56	0.63	0.79
4. Kernel SVM	0.73	0.56	0.63	0.79
5. Naïve Bayes	0.67	0.53	0.59	0.76
6. Decision tree Classification	0.63	0.72	0.67	0.78
7. Random forest Classification	0.65	0.61	0.63	0.77

Table 10: Comparative Performance of Classification Algorithms on Various Measures

Corresponding classifiers performance over Accuracy, Precision, F-measure and Recall values are listed in above Table and classifiers performance on the basis of classified instances are defined in following table.

Where,
TP = True Positive
TN = True Negative
FP = False positive
FN = False Negative

Total no of instances	Classification Algorithms	Correctly Classified Instances	Incorrectly Classified Instances
192	**1. Logistic Regression**	152	40
	2. K-NN	148	44
	3. SVM	152	40
	4. Kernel SVM	152	40
	5. Naïve Bayes	147	45
	6. Decision tree Classification	151	41
	7. Random forest Classification	148	44

Table 11: Classifier's Performance on The Basis of Classified Instances

5.5 Model Deployment

In this project we deployed our model with the micro web framework called Flask, it is a web framework. This means flask provides you with tools, libraries and technologies that allow you to build a web application. This web application can be some web pages, a blog, a wiki or go as big as a web-based calendar application or a commercial website.

Flask is part of the categories of the micro-framework. Micro-framework is normally framework with little to no dependencies to external libraries. This has pros and cons. Pros would be that the framework is light, there are little dependency to update and watch for security bugs, cons is that some time you will have to do more work by yourself or increase yourself the list of dependencies by adding plug-in. In the case of Flask, its dependencies are

- Werkzeug a WSGI utility library
- jinja2 which is its template engine

The basic workflow of a machine learning model is shown in the Fig. 2

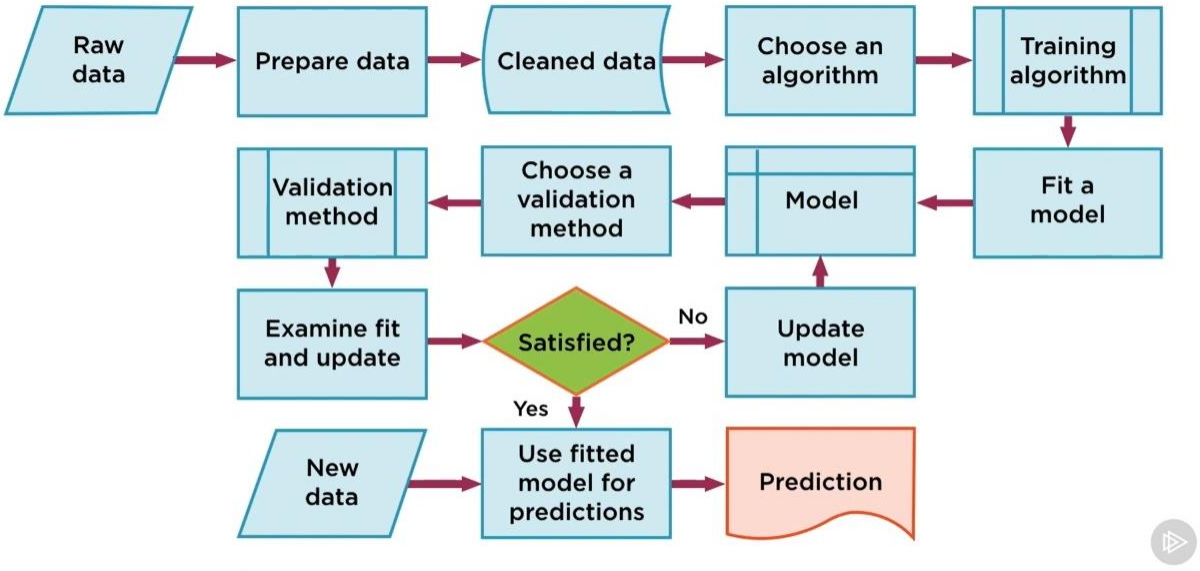


Fig 2: Basic Machine Learning Workflow

5.6 Database for User Information

In this project we have used MySQL server to store the information of registered users who can access our website. By registering themselves and creating a account they can login and can get know about their condition i.e. either they are diabetic or not.
As an input will take the User's Name, Email Id, Mobile Number, State, City and Password, and store all these values in a table and will identify them with an Id number. And if the user wants to login into its account so she can use the email as her user id and the password.

Hardware and Software Requirements

There are various requirements to successfully deploy the system. These are mentioned below:

Hardware Requirement

- Processor: Minimum 1 GHz; Recommended 2GHz or more.
- Ethernet connection (LAN) OR a wireless adapter (Wi-Fi)
- Hard Drive: Minimum 32 GB; Recommended 64 GB or more.
- Memory (RAM): Minimum 2 GB; Recommended 4 GB or above.

Software Requirement

- Operating System – Microsoft Windows or others.
- Web BrowserMicrosoft Internet Explorer, Mozilla, Google Chrome or later.

Tools Used

- Platform- Anaconda Navigator(version 3.6)
- IDE- Spyder (version 3.3.4)
- MySQL Server (on Xampp Server)
- Language used- Python 3.7 and its supported libraries.
- Front end – Bootstrap, HTML, CSS, JavaScript

UML Diagrams

Activity Diagram for login page

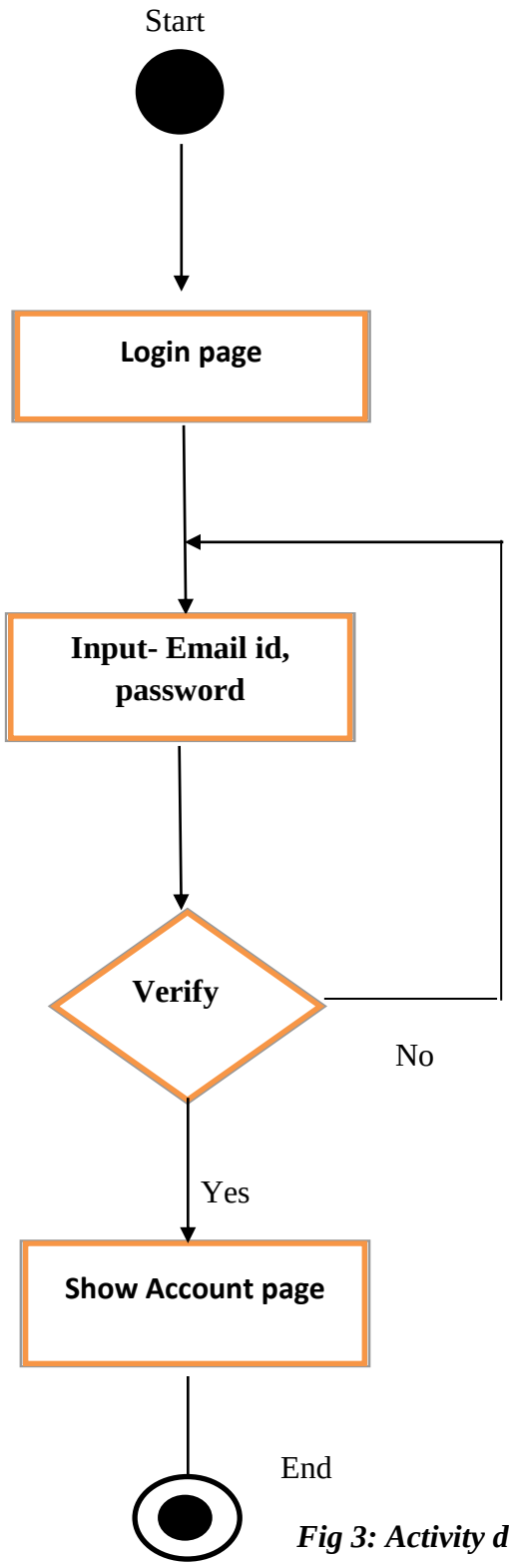


Fig 3: Activity dig. for login form

Activity Diagram for signup page

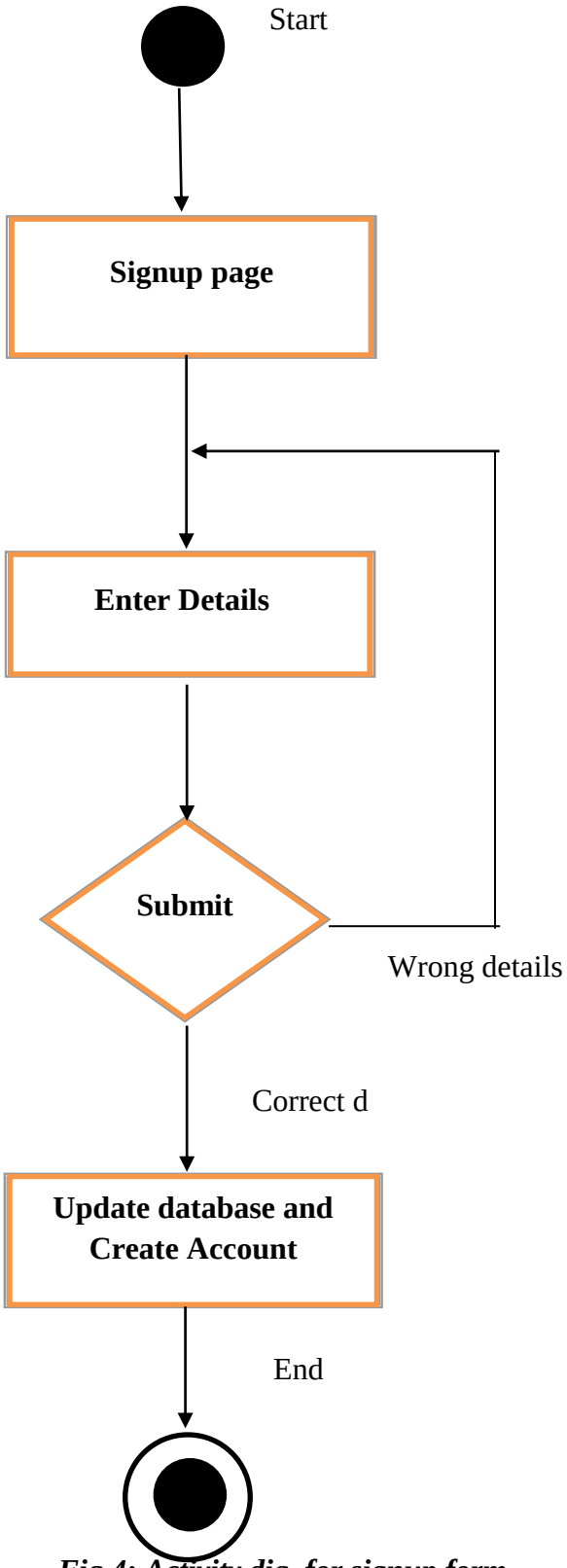


Fig 4: Activity dig. for signup form

Activity Diagram for Prediction form

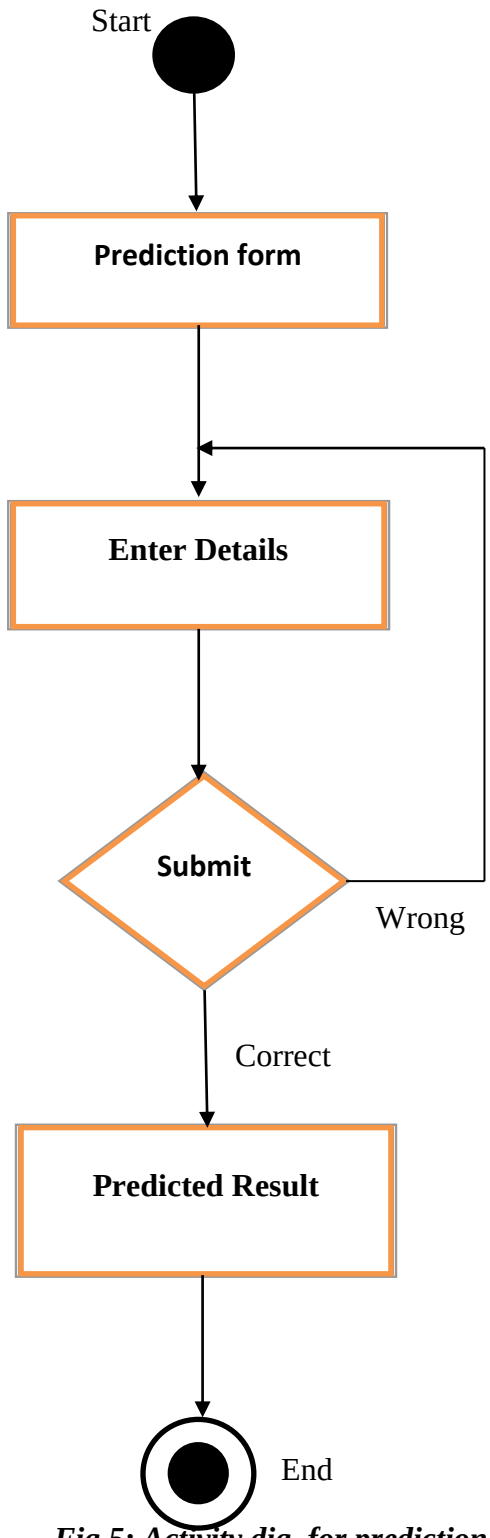


Fig 5: Activity dig. for prediction form

Sequence diagram for login page

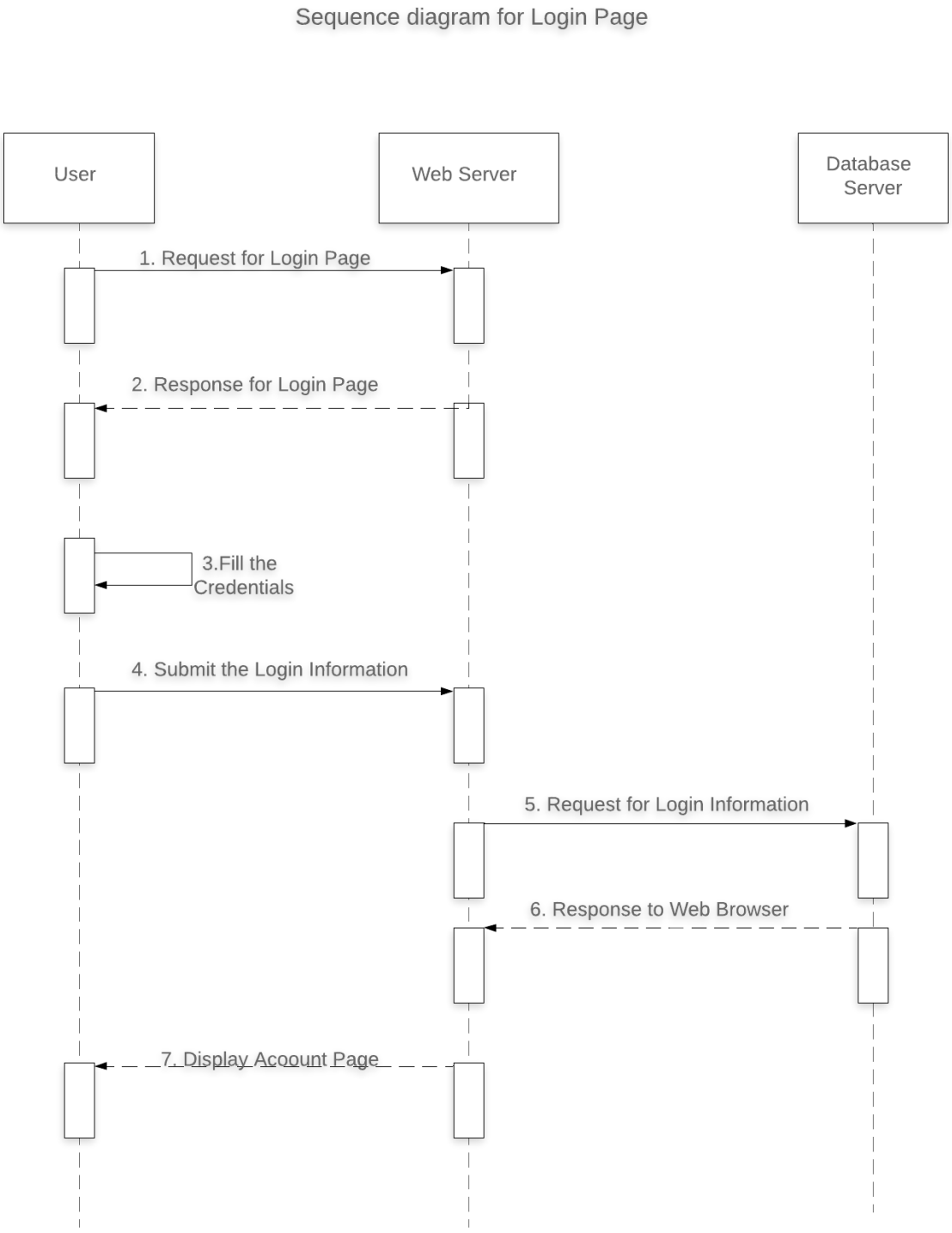


Fig 6: Sequence dig. for login form

Sequence diagram for signup page

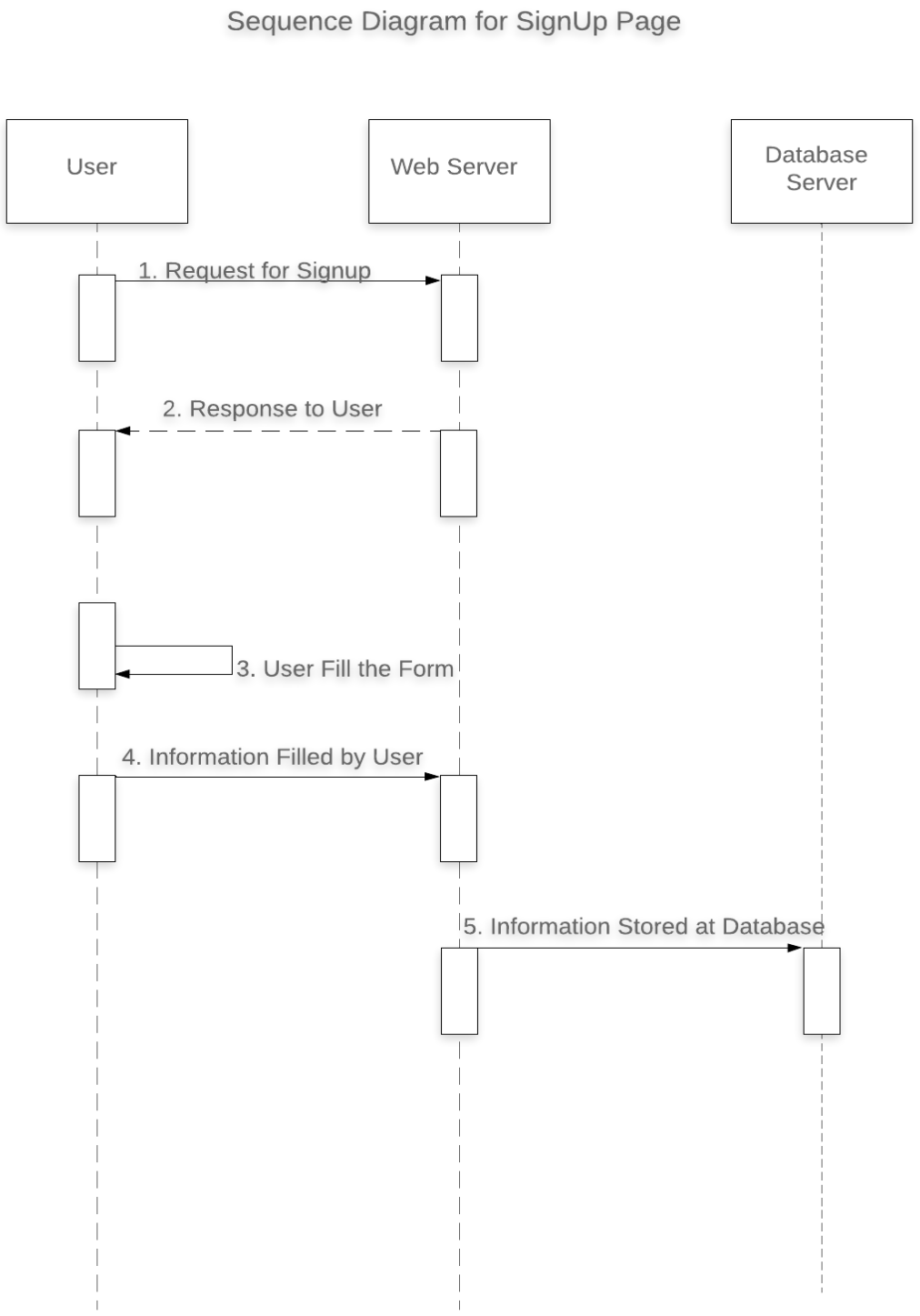


Fig 7: Sequence dig. for signup form

Sequence diagram for Prediction form

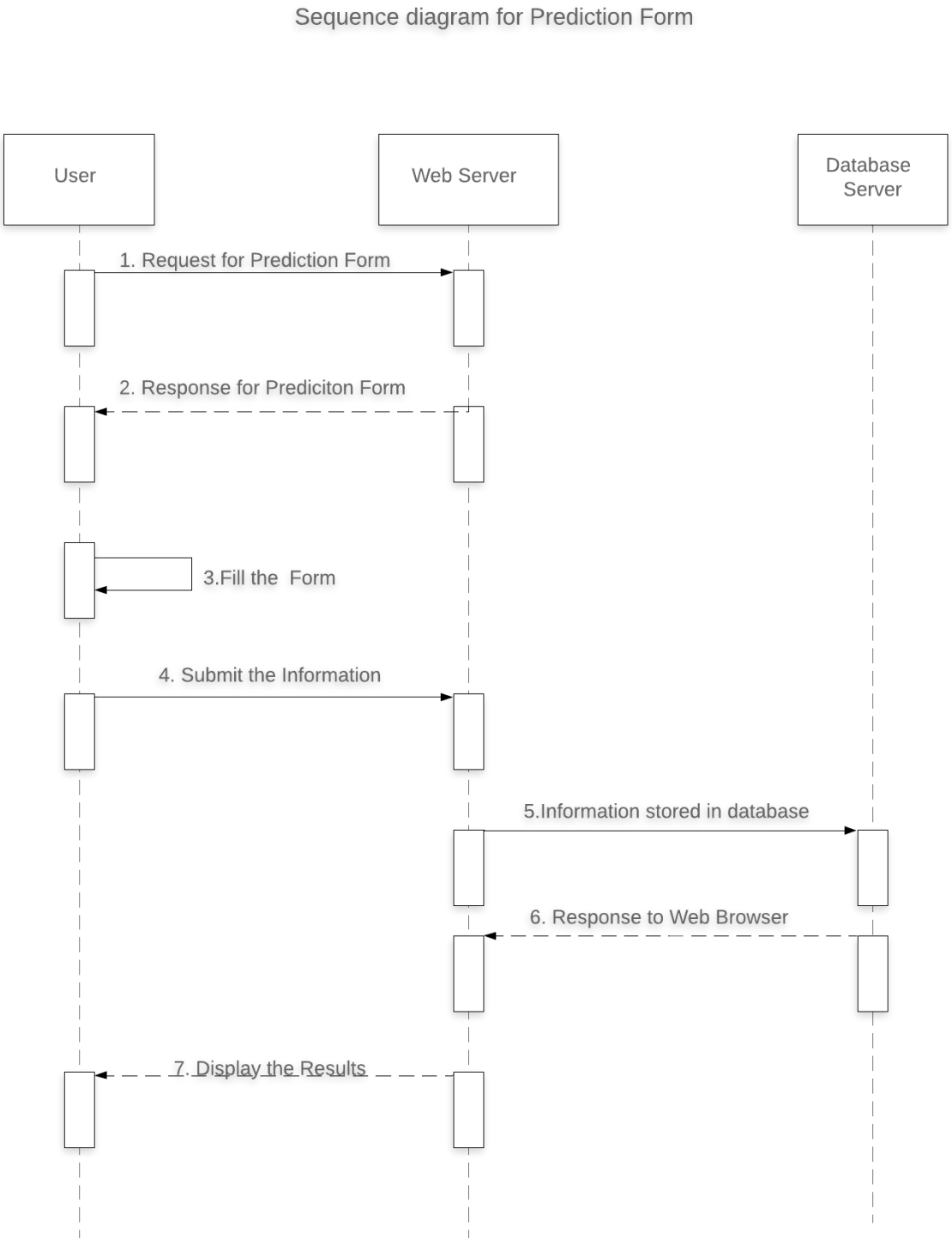


Fig 8: Sequence dig. for prediction form

Use-Case Diagram

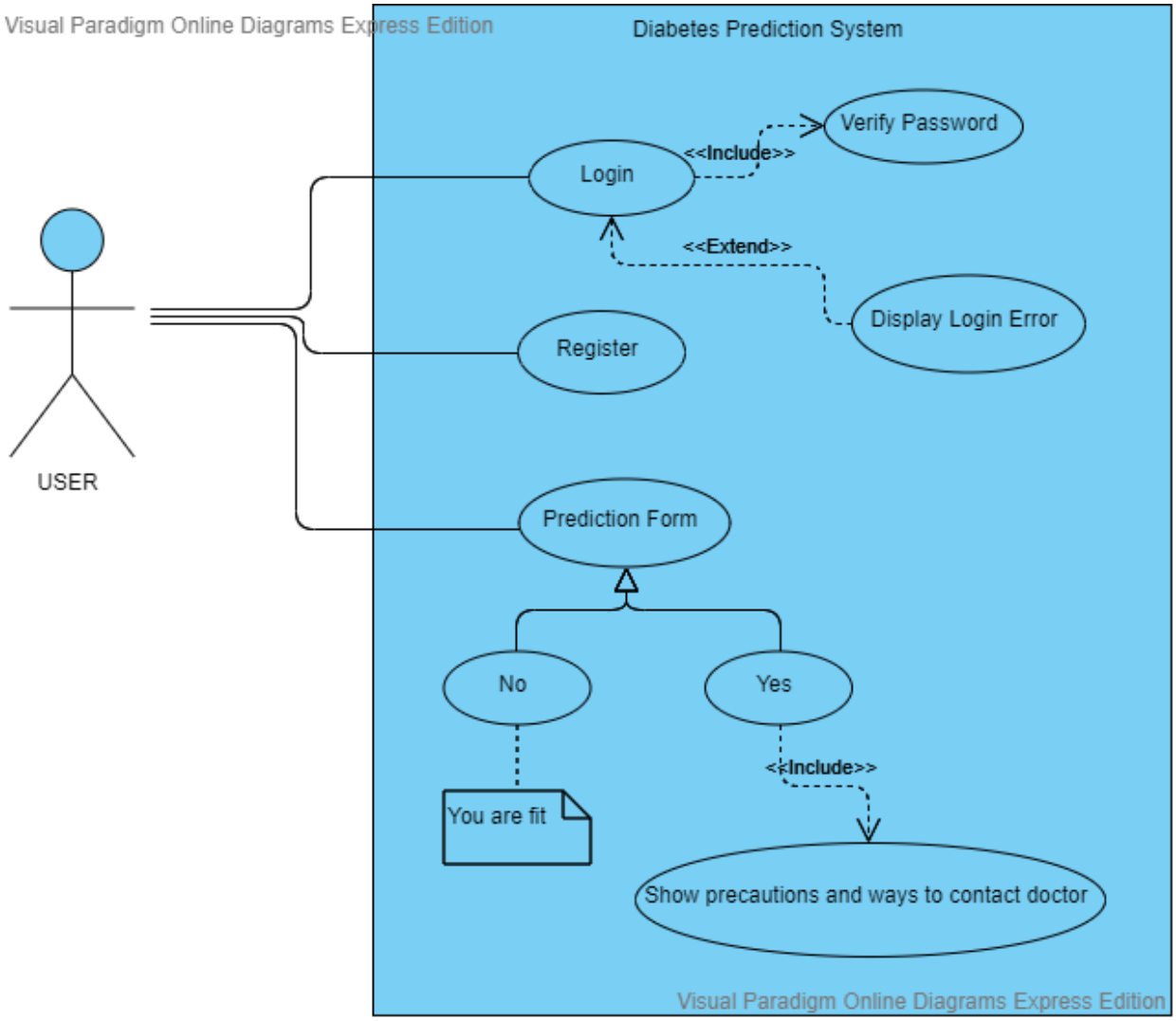


Fig 9: Use-case diagram

E-R diagram

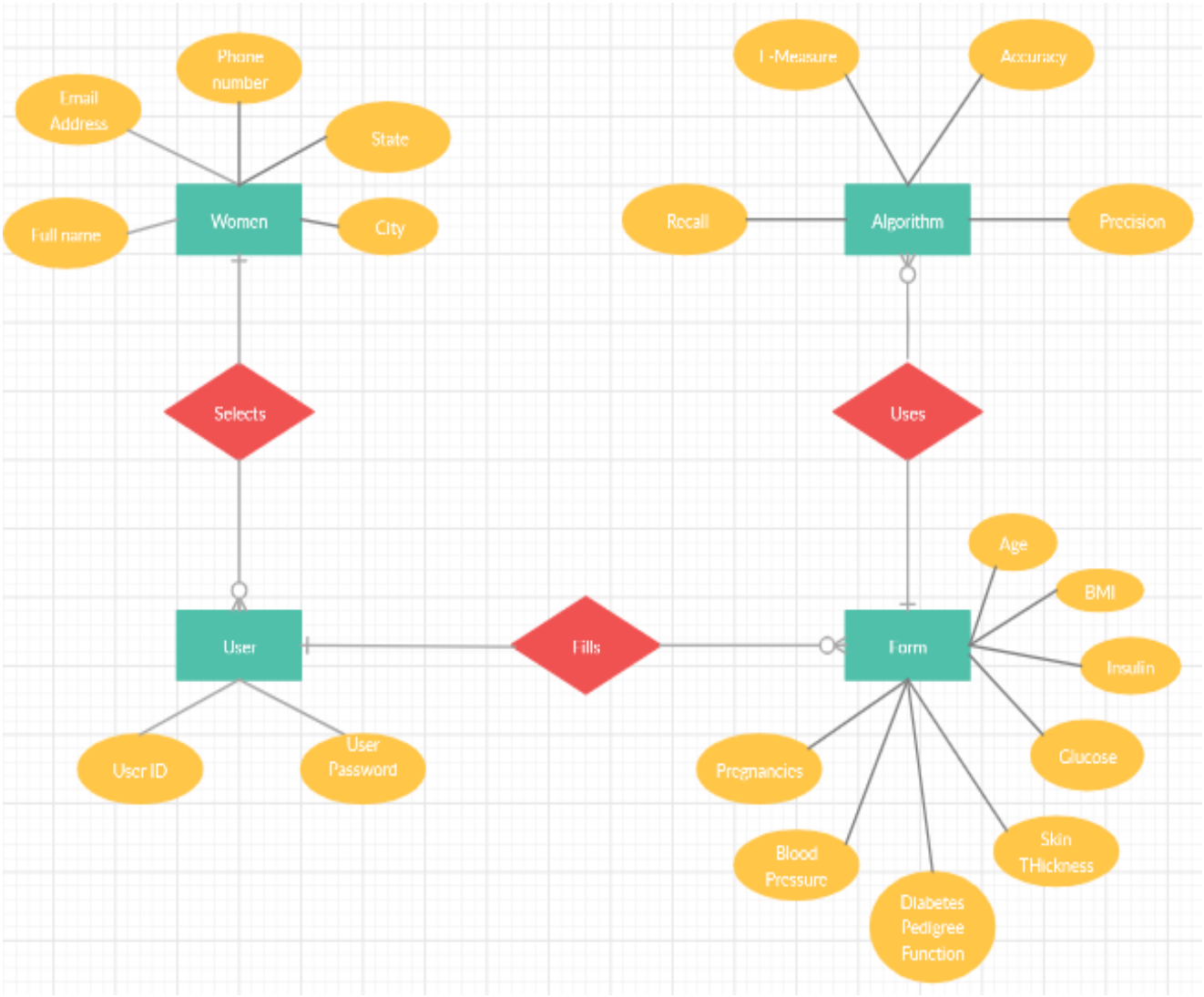


Fig 10: ER Diagram

Results

As the two tables above represents the different performance values of all classification algorithms calculated on various measures.

Analysing the tables we can conclude that Support Vector Machine algorithm is giving maximum accuracy. Although SVM and kernel SVM along with Logistic regression was giving similar values but since it is clear from other measures that logistic regression is somewhere lacking than SVM. Hence, we choose SVM for our project.

On analysing the attributes we have found that only 7 attributes are significant which we will use in our model and they are as follows :

1. Pregnancies – Number of times pregnant
2. Glucose – Plasma glucose concentration over 2 hours in an oral glucose tolerance test
3. Blood Pressure – Diastolic blood pressure (mm Hg)
4. Insulin – 2-Hour serum insulin (mu U/ml)
5. BMI – Body mass index (weight in kg/(height in m)2)
6. Diabetes Pedigree Function- Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
7. Age - Age (years)

Performances of all classifier's based on various measures are plotted via a graph in the following figures:

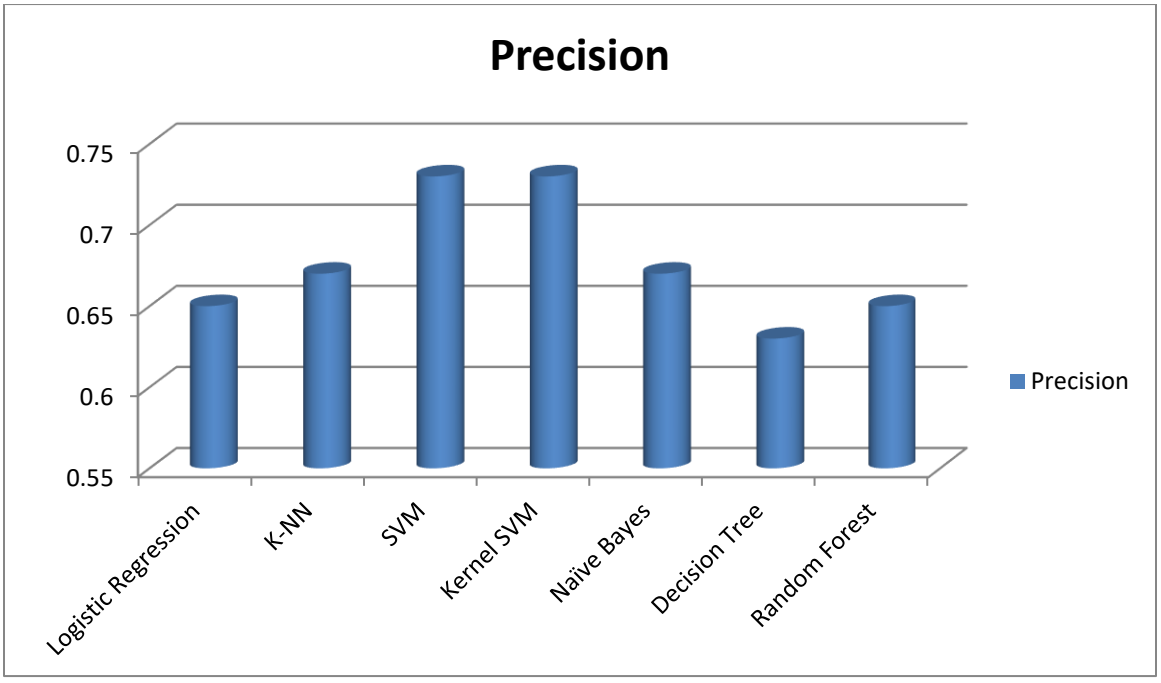


Fig.11: Classifier's performance based on Precision

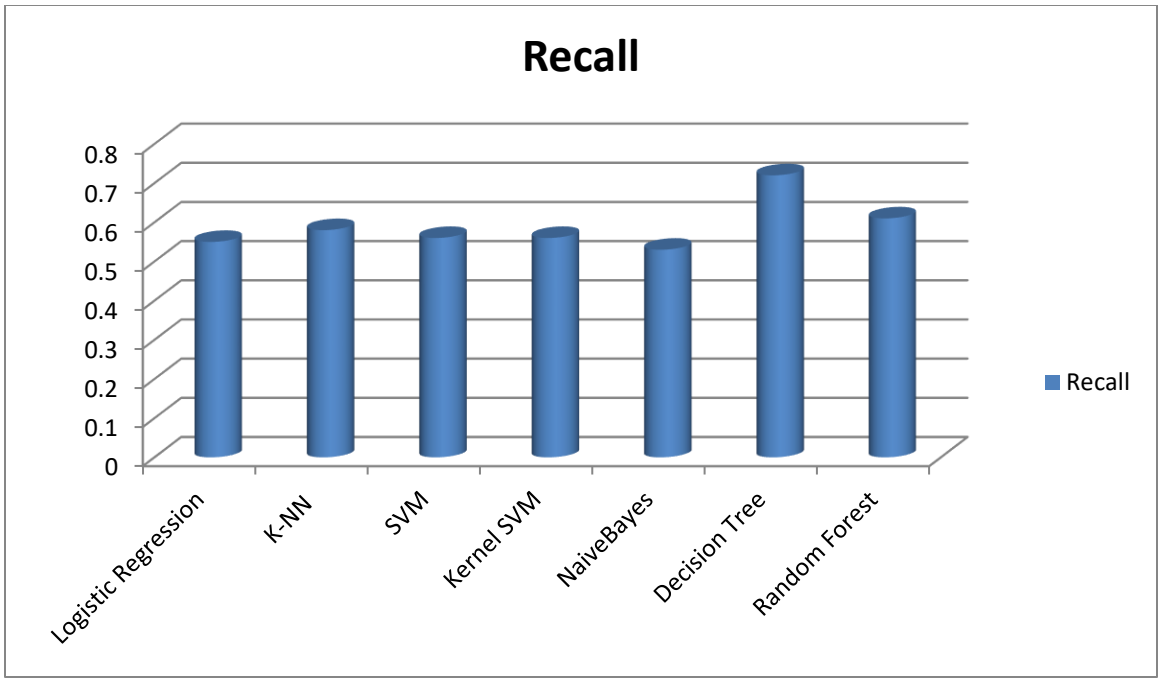


Fig.12: Classifier's performance based on Recall

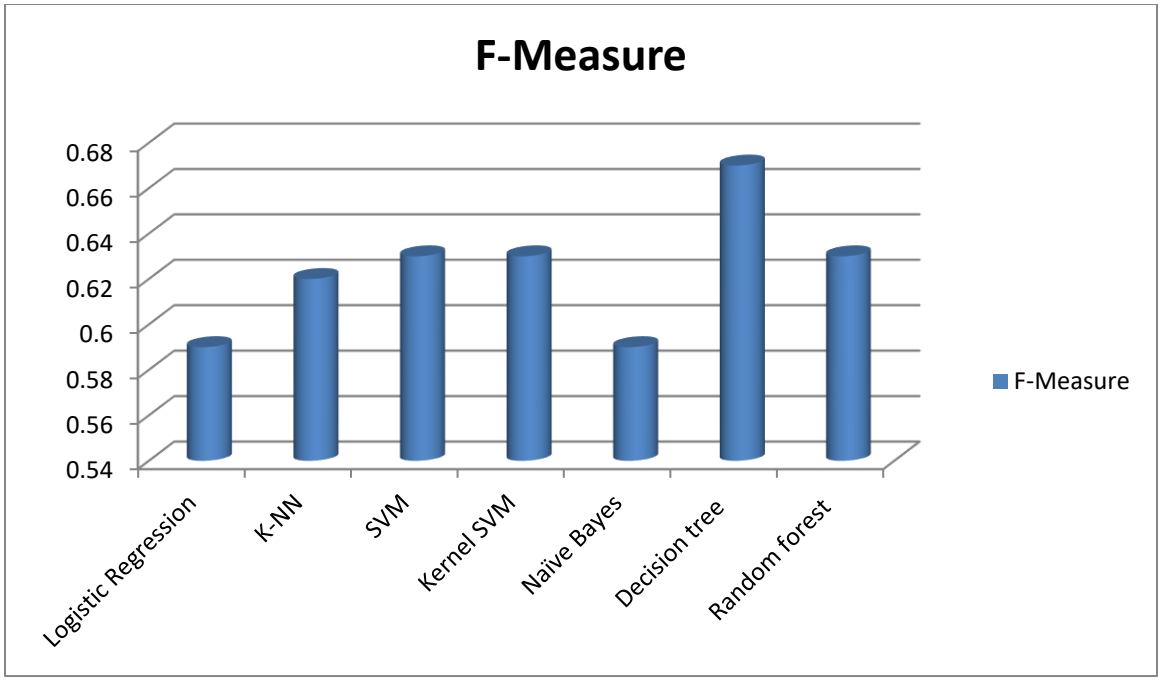


Fig.13: Classifier's performance based on F-Measure

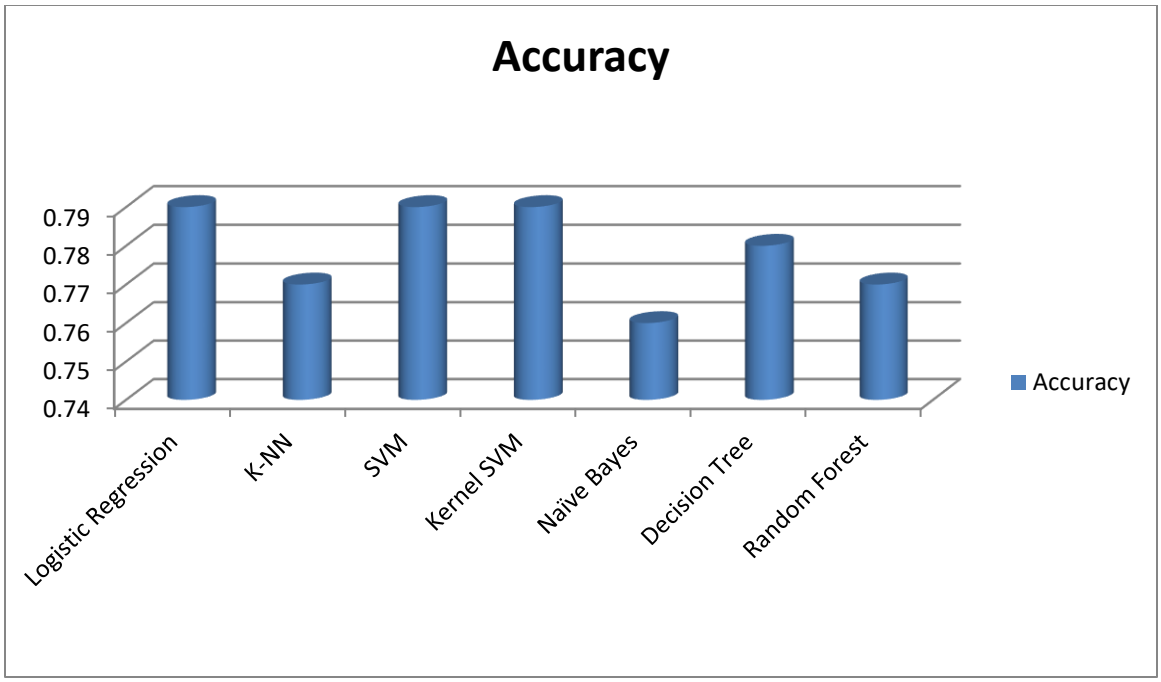


Fig.14: Classifier's performance based on Accuracy

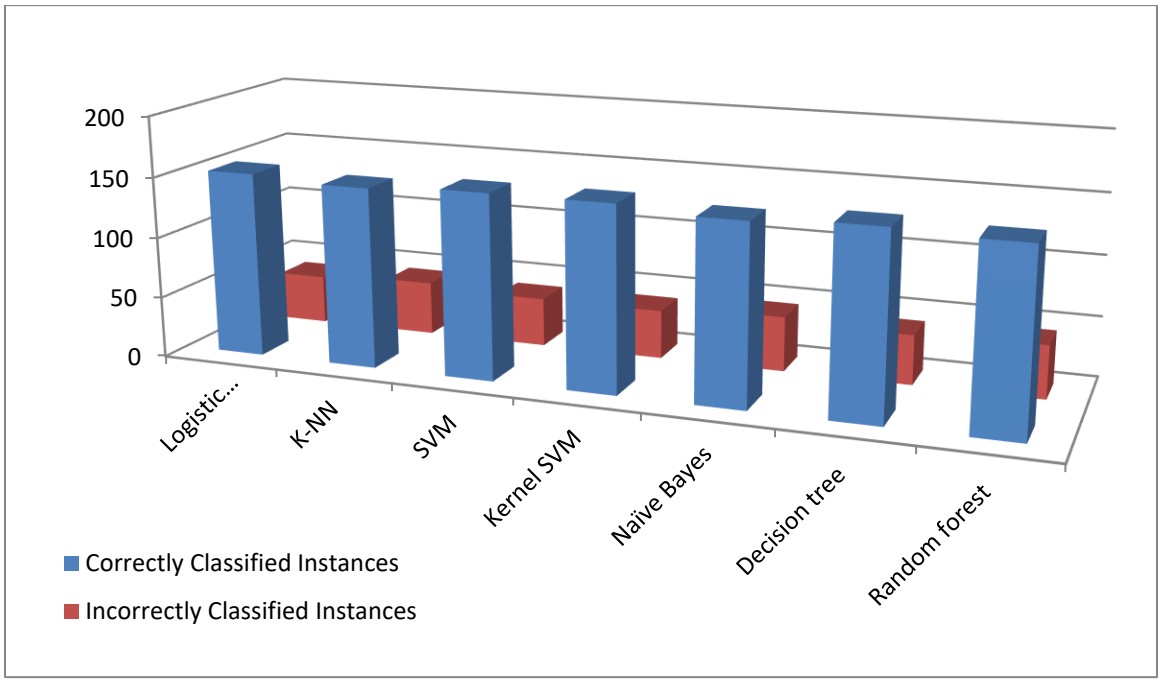
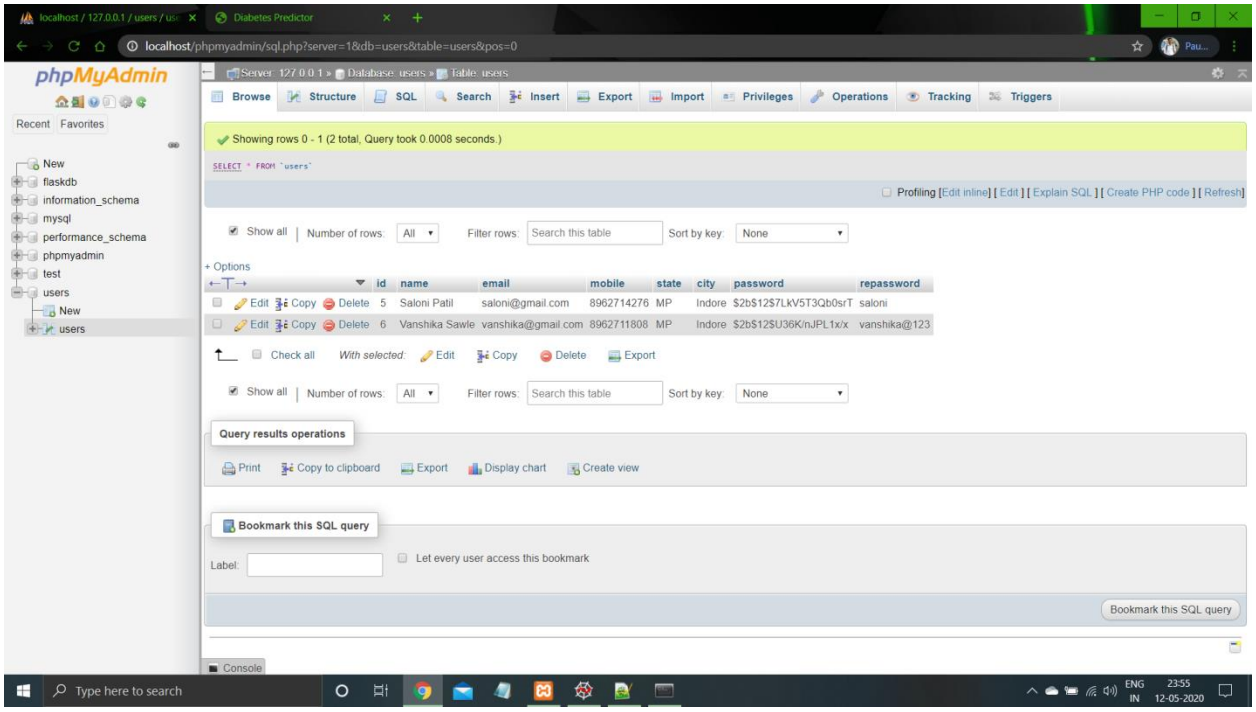
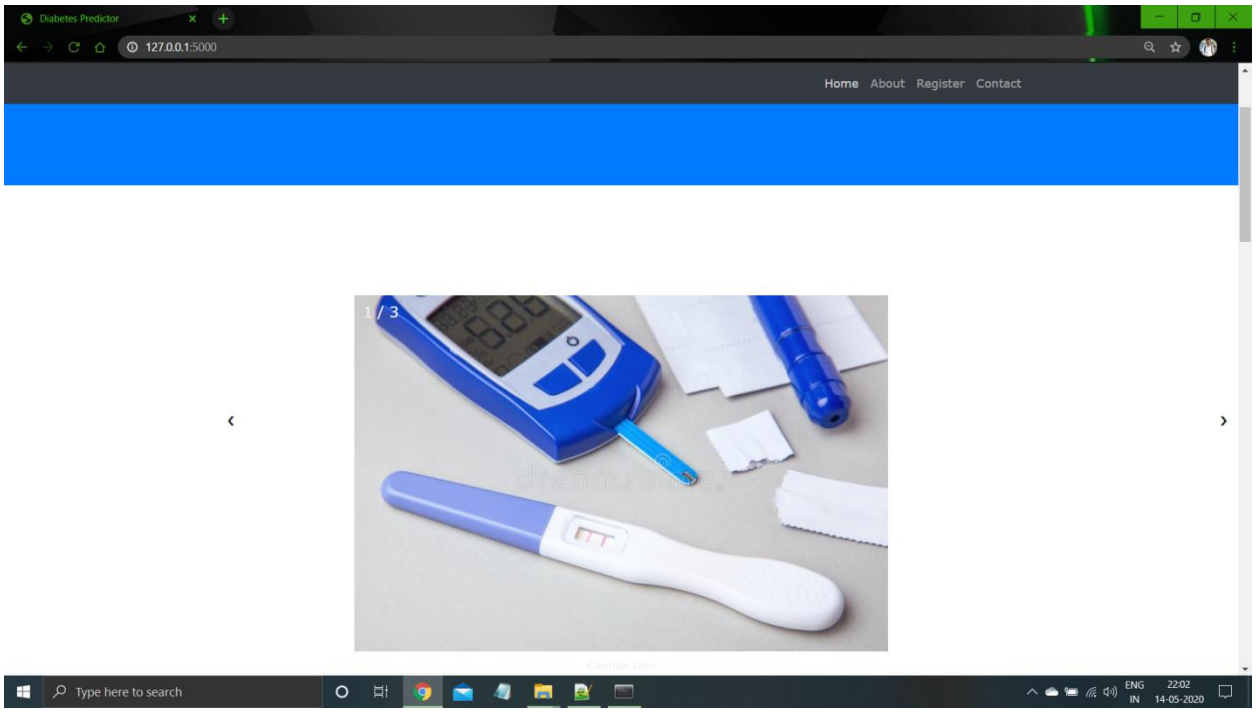


Fig.15: Classified Instances

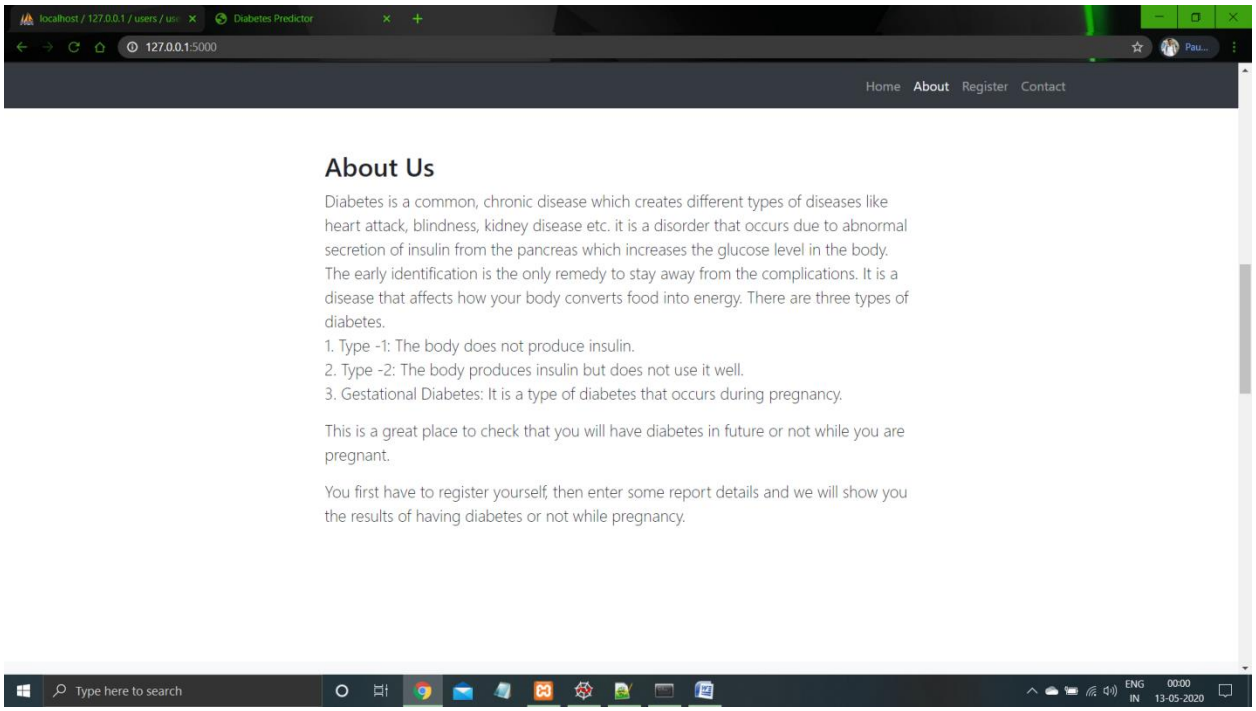
Screenshots



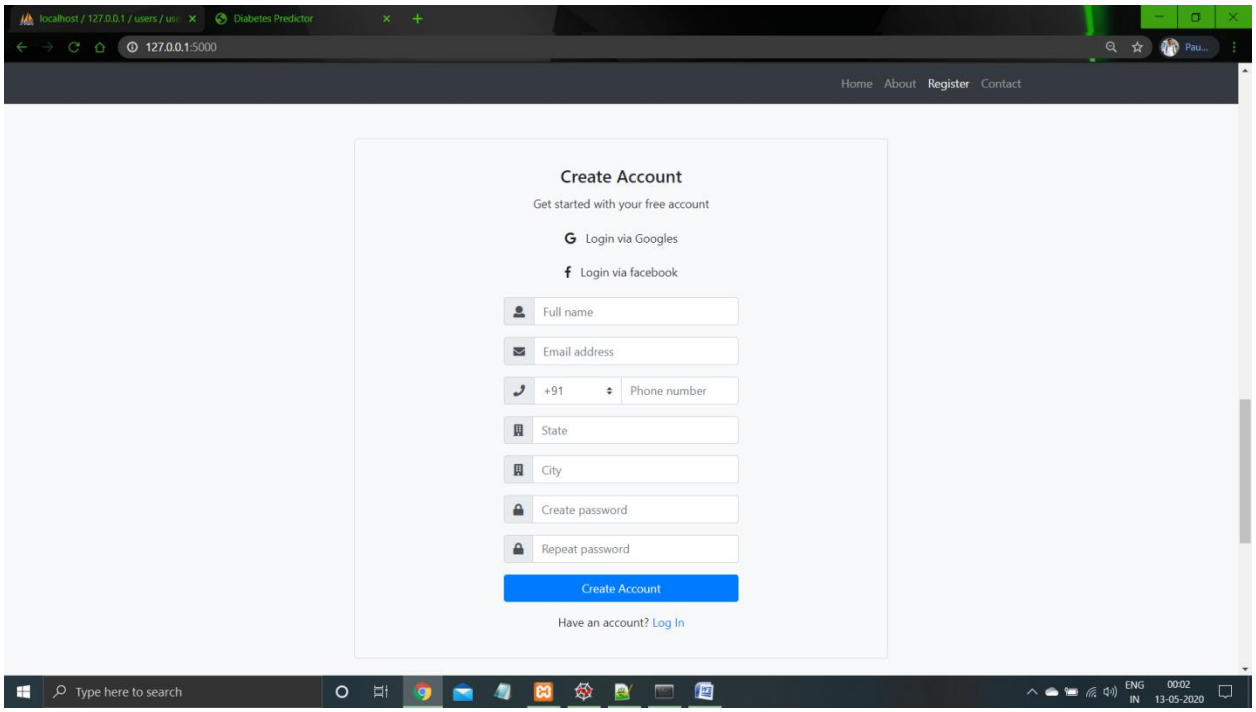
Database



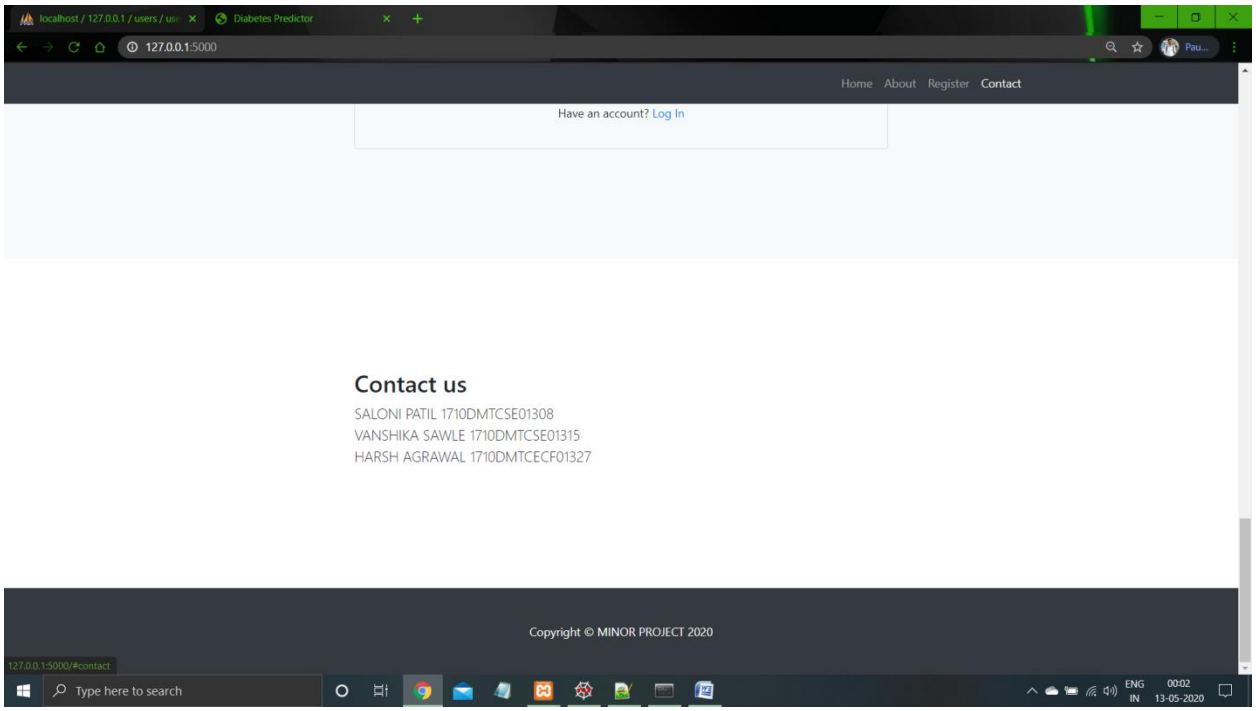
Index page- Home section



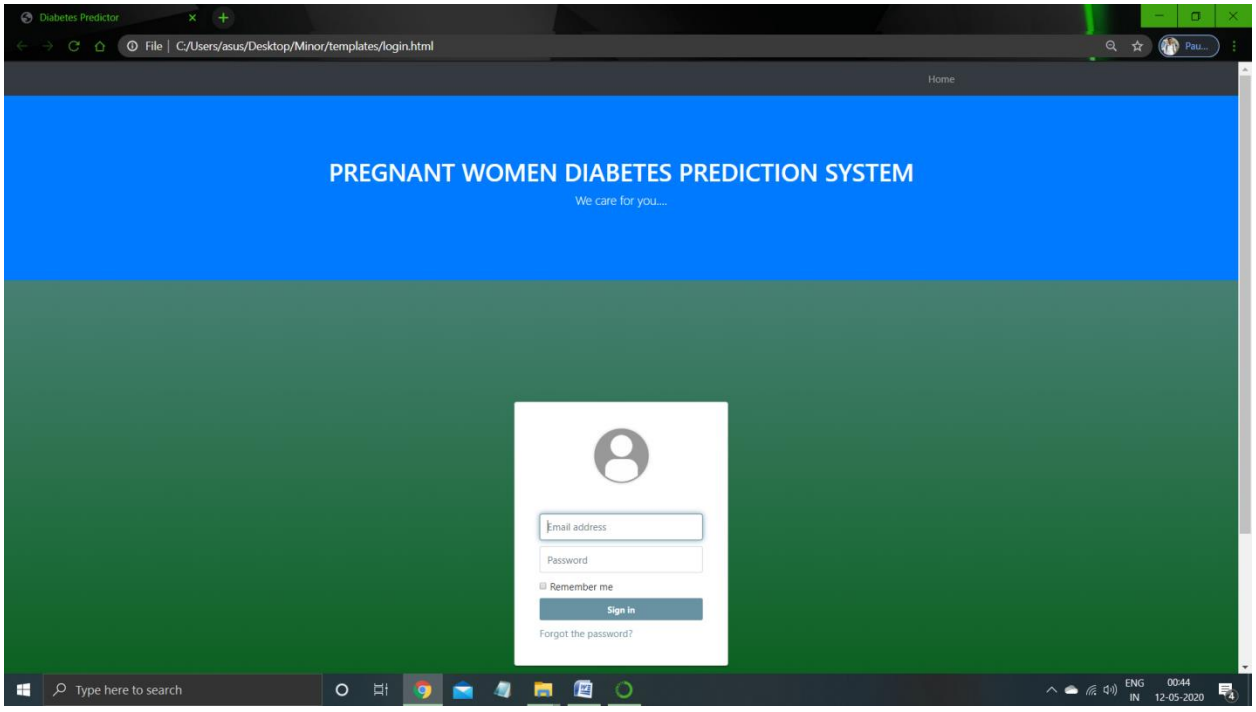
Index page – About section



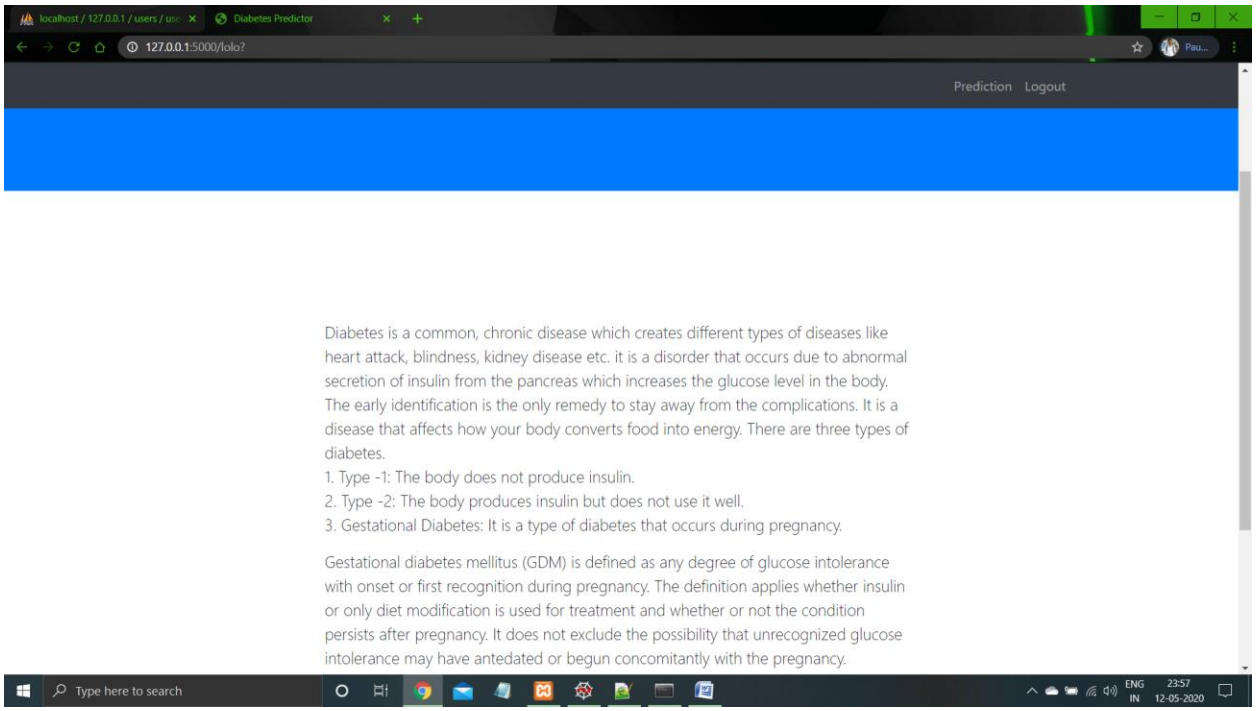
Index page- Register section



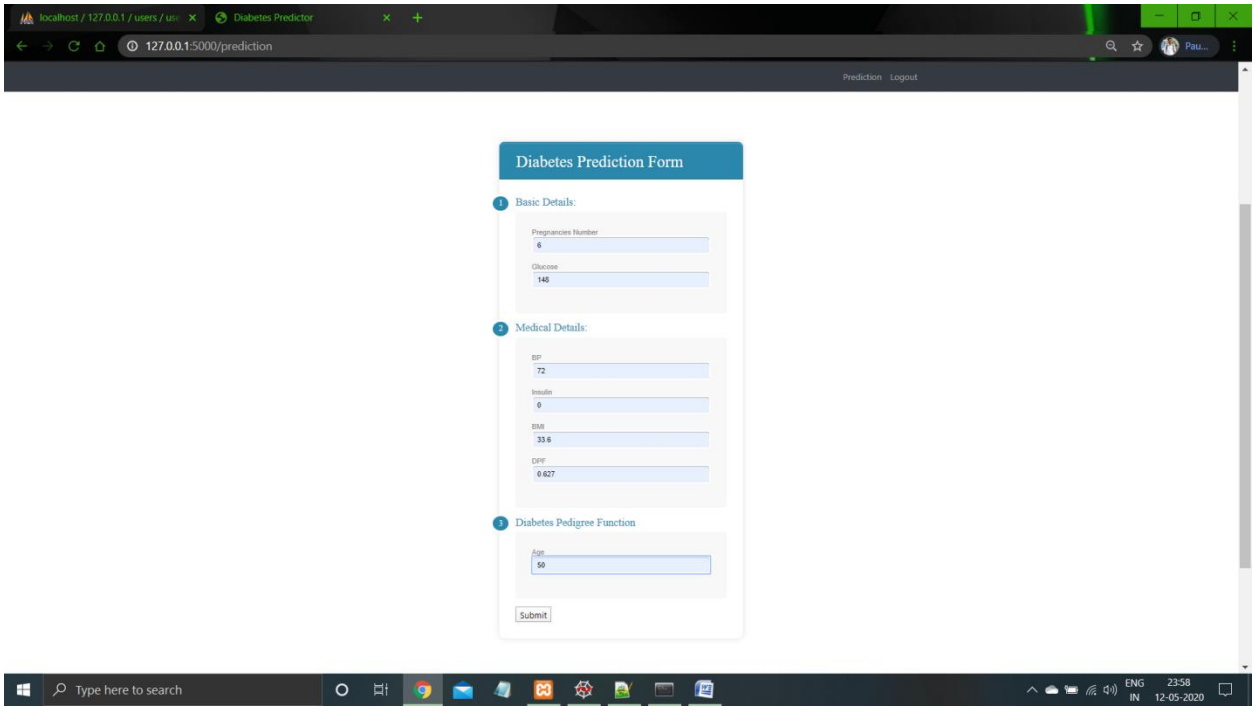
Index page- Contact section



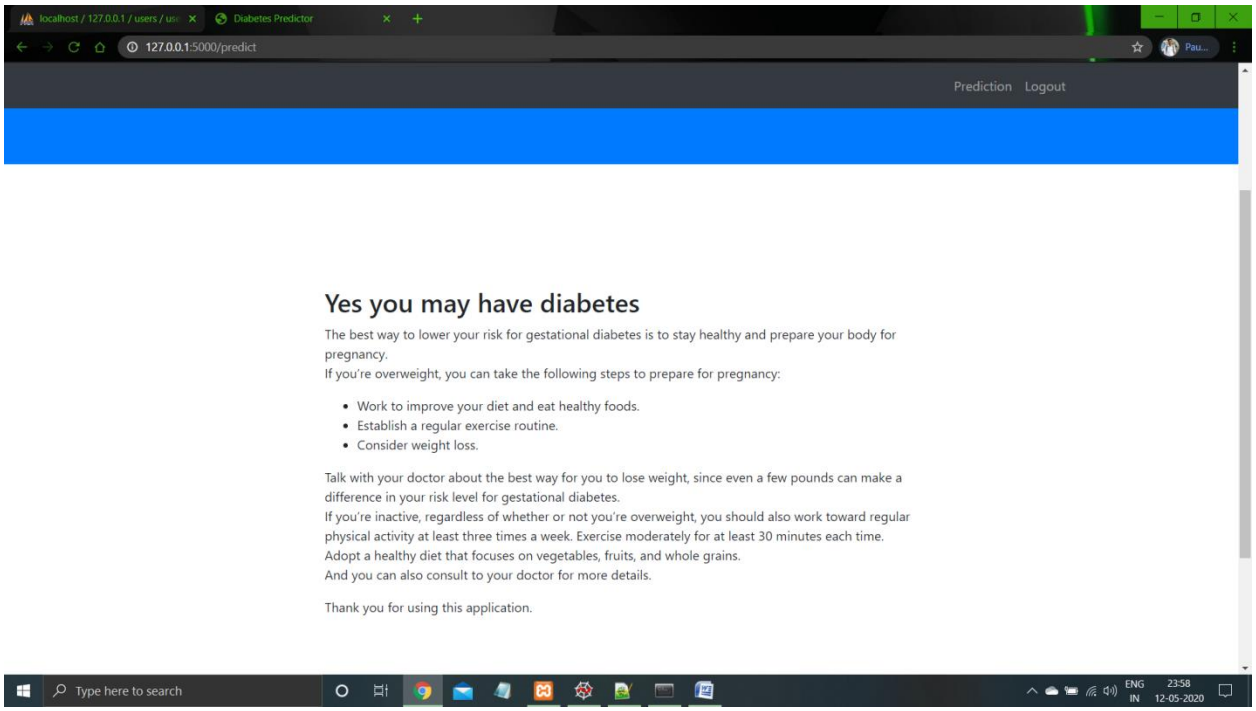
Login Page



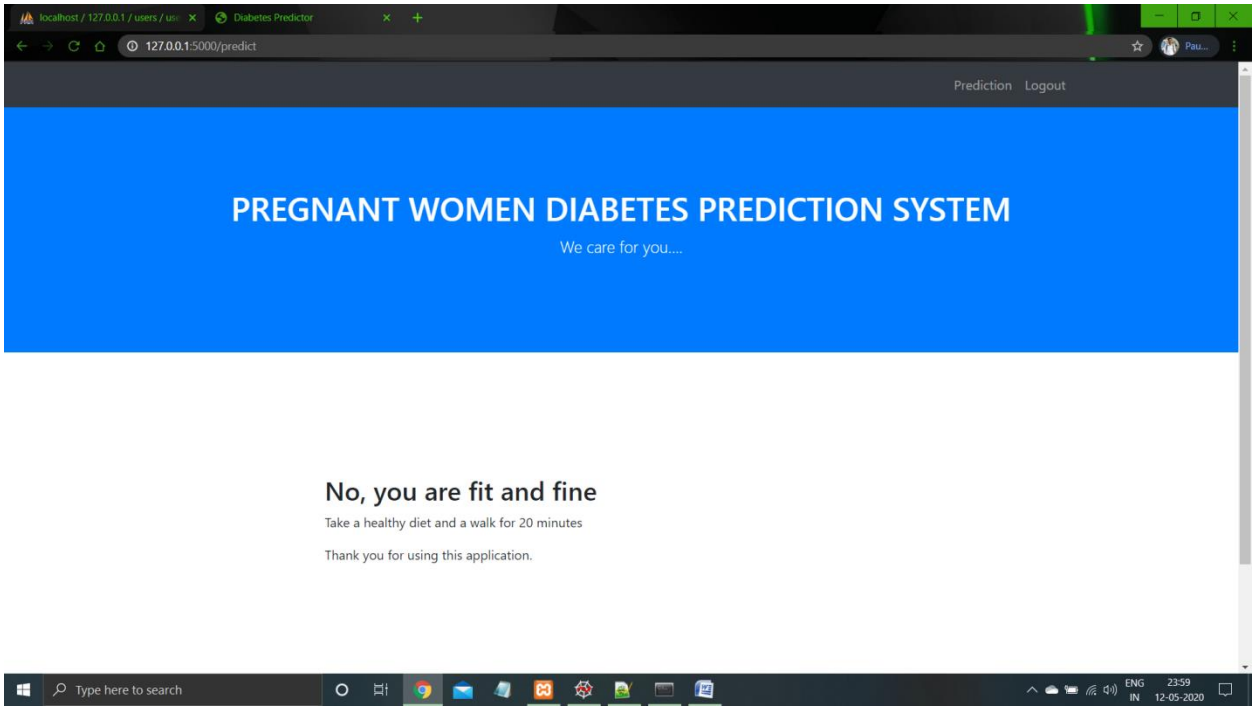
After logging or registering



Prediction Form



Result Page – If the result is yes



Result page- If the result is No

Conclusion

Detection of Gestational Diabetes at its early stage is one of the most important real-world medical problems as it can affect the women a lot. In this project, systematic efforts are made in designing a system which results in the prediction of gestational diabetes for a pregnant lady.

During this work, various machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on a diabetes dataset with 768 observations and 8 attributes from which 6 came as significant attributes.

Experimental results determine the adequacy of the designed system with an achieved accuracy of 79% using the SVM (Support Vector Machine) classification algorithm.

In future the work can be extended and improved for the automation of diabetes analysis for a pregnant lady, including some other machine learning algorithms.

References

1. LITERATURE SURVEY ON DIABETES MELLITUS USING PREDICTIVE ANALYTICS OF BIG DATA
   http://ijaerd.com/papers/finished_papers/LITERATURE%20SURVEY%20ON%20DIABETES%20MELLITUS%20USING%20PREDICTIVE%20ANALYTICS%20OF%20BIG%20DATA-IJAERDV05I0246907.pdf
2. Improved Diabetes Prediction Model for Predicting Type-II Diabetes
   https://www.ijitee.org/wp-content/uploads/papers/v8i12/L35941081219.pdf
3. PREDICTION OF DIABETES MELLITUS USING DATA MINING TECHNIQUES: A REVIEW
   https://www.researchgate.net/publication/273023827_PREDICTION_OF_DIABETES_MELLITUS_USING_DATA_MINING_TECHNIQUES_A_REVIEW
4. Machine Learning and Data Mining Methods in Diabetes Research
   https://www.sciencedirect.com/science/article/pii/S2001037016300733?via%3Dihub
5. Analyzing Diabetes Datasets using Data Mining
   https://www.lifescienceglobal.com/independent-journals/journal-of-basic-and-applied-sciences/volume-13/84-abstract/jbas/2923-abstract-analyzing-diabetes-datasets-using-data-mining
6. Development and validation of a clinical model for preconception and early pregnancy risk prediction of gestational diabetes mellitus in women
   https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0215173#abstract0
7. Prediction of pregnancy diabetes based on machine learning
   https://ieeexplore.ieee.org/document/8903352/authors#authors
8. International Diabetes Federation
   https://www.idf.org/our-activities/care-prevention/gdm
9. Dataset:
   https://github.com/datasets/dermatology