SYST 542

A Decision Support System for finding Diabetes in Women

Project by:

Srijan Yenumula
Navya Avuthu
Salim Zeibak
Neha Allam

Index

1.0 Project Overview

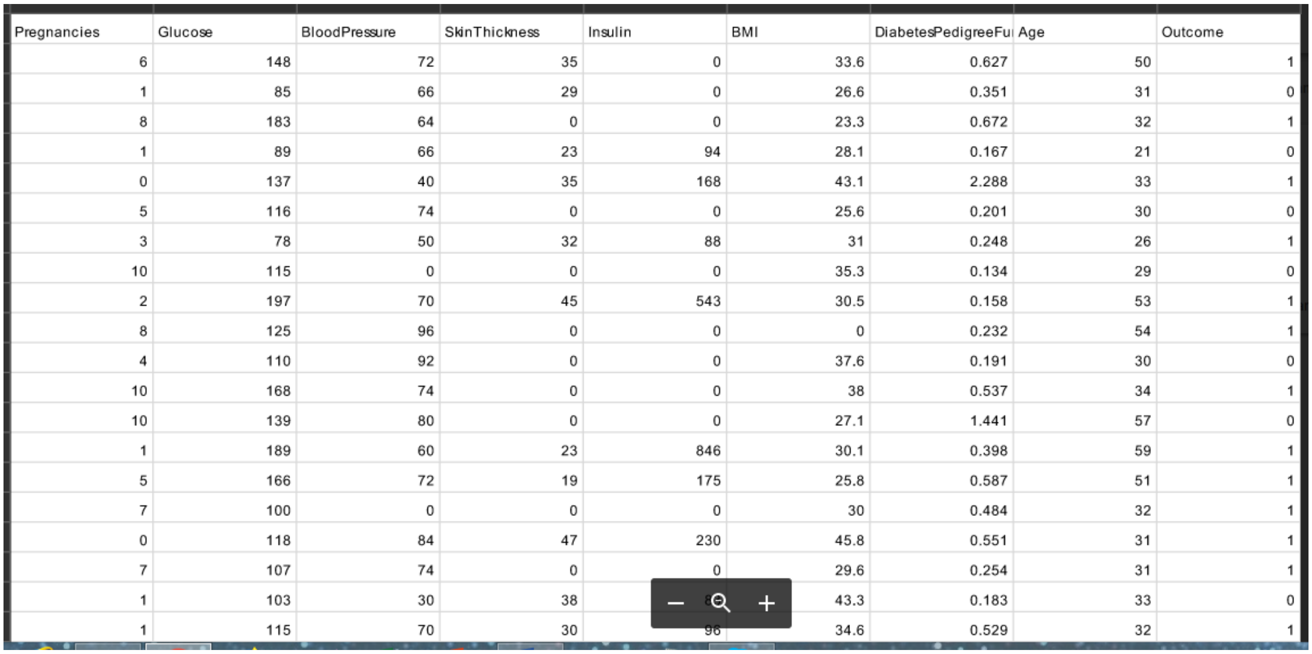
1.1 Problem Definition: Prevalence of Type 2 Diabetes in Pima Indian People

According to authors Baier and Hanson in their piece “Genetic Studies of the Etiology of Type 2 Diabetes in Pima Indians,” it is a well established fact that Native Americans of “Pima Indians” have some of highest Diabetes diagnoses in the world. They have the highest recorded incidents of non-insulin dependant diabetes of any geographically defined population

Throughout the years, they have been carefully studied and under continuous examination for diabetes. They have mostly been living in various parts of Arizona with relatively little contact with outside settlers and Spanish missionaries. This type of isolation has made them an interesting study in their genetics and environmental factors and how that contributes to their likelihood of obtaining a Type 2 diabetes diagnosis.

Because of this history, we would like to explore how a Decision Support System could help in the early detection and screening of many people in this community. Our system will focus on the women of the Pima Indian tribe. They are at high risk for diabetes and this puts them at a disproportionate disadvantage in their quality of life. Our system will look at how we can improve their quality of life and provide them support to prevent them from getting diabetes.

We have collected a dataset from kaggle and a screenshot of it is as below:



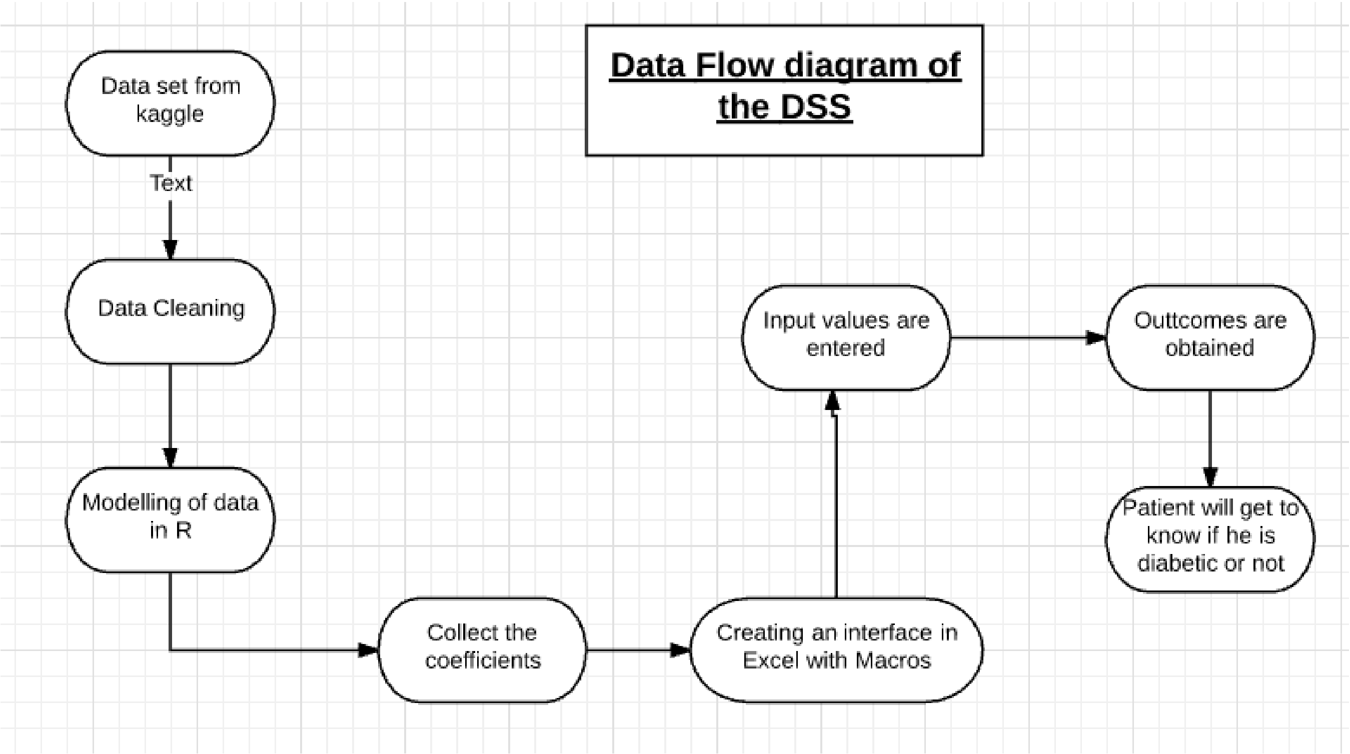
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFu	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	[illegible]	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1

1.2 Project Overview

Our system will focus on helping women in the Pima Indian community. The organizational focus will be on the women in this community, the healthcare system and various research groups.

We will utilize a dataset to create the logic for the Decision Support System. This logic will help assess which factors cause diabetes in this community and help them predict who in the future is at risk to develop this disorder.

Our plan to is to develop a subsystem that takes this statistical analysis of training data to develop the decision logic, and then using a validation data set to make predictions about who is at risk to develop diabetes. This tool will then become a useful guide to many in the Pima Community to know whether they are at risk. The system can them make recommendations on which factors specifically put them at risk in their case.



Data Flow Diagram of the Decision Support System

1.3 Expected Impact

We expect that this information will greatly help the screening process for preventing people from getting diabetes in the Pima Indian community. The goal would be to provide better information and decision making and ultimately reducing the incidence of diabetes in that community.

We can evaluate our results by comparing our the results of our model to the validation data result to see how effective our DSS was in predicting the outcome of each patient. The outcome obtained is very helpful for the patient as he gets to know if he is diabetic or not. There are chances that he might be a diabetic in future and our DSS helps a person find that in prior.

Benefits:

The benefits of this Decision Support System is that it allows patient to become aware of him being a diabetic or not. It also gives a future result along with the present status. With the result it tells the reason for the patient being diabetic from the attributes of the dataset. By knowing the reason, the patient will get a chance to improve his health in that particular area. For example, if the patient comes to know that the cause of diabetes is by the increase in the body mass index, the DSS will display the reason. Patient will learn this reason and will have a chance to work on it. The main purpose of this DSS is to make the patient's work more easier. Links provided below are the links that are put into the system, which on clicking will lead the patient to that particular website where he can follow the tips given in that website.

Links for remedies:

1) **Insulin related:**
https://www.healthline.com/nutrition/15-ways-to-lower-blood-sugar

2) **BMI related: https://www.everydayhealth.com/diabetes/0507/lower-your-bmi-to-prevent-diabetes.aspx**

3) **Blood Pressure related:**
https://www.webmd.com/hypertension-high-blood-pressure/guide/high-blood-pre

4) **Glucose related:**
https://www.organicfacts.net/home-remedies/gestational-diabetes.html

2.0 Users of DSS

2.1 Who will use this system for Pima Indians?

The main users for this system will be the women of the Pima Indian community as well as the local Doctors, Nutritionists, Hospitals, Pharmacies and Fitness Centers. All of these different users can engage is monitoring and regular testing of the various factors that are measured to predict the outcome of having diabetes.

2.2 Inputs

For the inputs, we will collect personal information of each patient, but the inputs crucial to the model will be: Age of Patients, BMI, Insulin Level, Skin Thickness, Blood Pressure, Glucose Level and Number of Pregnancies. Also another input is the Diabetes Pedigree Function, this can be determined by a medical professional based on the patient's family history.

2.3 DSS Concept

This concept will be built on an Solver-oriented DSS, where we take these inputs and their probability to produce a Binary and predicted outcome of whether the patient has diabetes or not.

We will use our training data and apply Logistic Regression models and decision trees to find the relationships between each factor. This will be the basis of the probabilities for each factor and allow for the development of predictive algorithms for our validation training set.

Solver Oriented DSS:

In a Solver Oriented DSS the decision is made by solving an algorithm that is proposed for a typical case. The algorithm is simplified and a unique solution is proposed. In case of typical decisions if we have an algorithm, the work gets easier as everything is solved in a method. It reduces a lot of time too.

It is an advantage to have algorithms getting involved and if we have experts who can solve that then it's very efficient to have best decisions. The problem occurs to have that algorithm. If there isn't any proper algorithm for a problem then the work becomes highly tedious. Also if we not have the necessary problem solver tools or ways or experts to decode the algorithm then the decision making will get worse and couldn't be made in the best way.

In this Decision Support System, we approach the models using an algorithm for the R-model development. That is how it implements a Solver Oriented DSS.

2.4 User Requirements

Users will require an easy interface to enter in patient data. Users will collect the various data points which must be complete prior to being able to use the system. This means that any user must have these data points collected before they can use the DSS.

Once a user has all these data points, they can enter all this information in an easy to access interface where the result will be a prediction in whether the person will get diabetes or not, along with some information packets customized for that person.

3.0 Project Management Plan

3.1 The System Architecture

Our system requires various data points on each patient. It is designed to receive all the data points from the inputs mentioned in Section 2. Our data set will be from a project on Kaggle. We will segment this data into a training set (60% of data) and a validation set (40%). The training set will drive the Model Logic and Algorithms produced for the Knowledge Center of the DSS. We will only be using this data set from the Kaggle source.

The data is represented in a CSV file with each line representing patient data. Each patient has the required data points for the model. To obtain new data in this system, it will require new patients to enter in every data point.

3.2 Logistic Regression:

Logistic Regression is a classification algorithm. It is used to predict a binary outcome given a set of independent variables. It predicts the probability of occurrence of an event by fitting data to a logit function. Along with this, in Logistic Regression, we use MLE approach for coefficient estimation.
To evaluate the performance of a logistic regression model, following few metrics are considered.

- AIC (Akaike Information Criteria) –AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value.
- Null Deviance– Null Deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model.
- Residual Deviance - Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model.

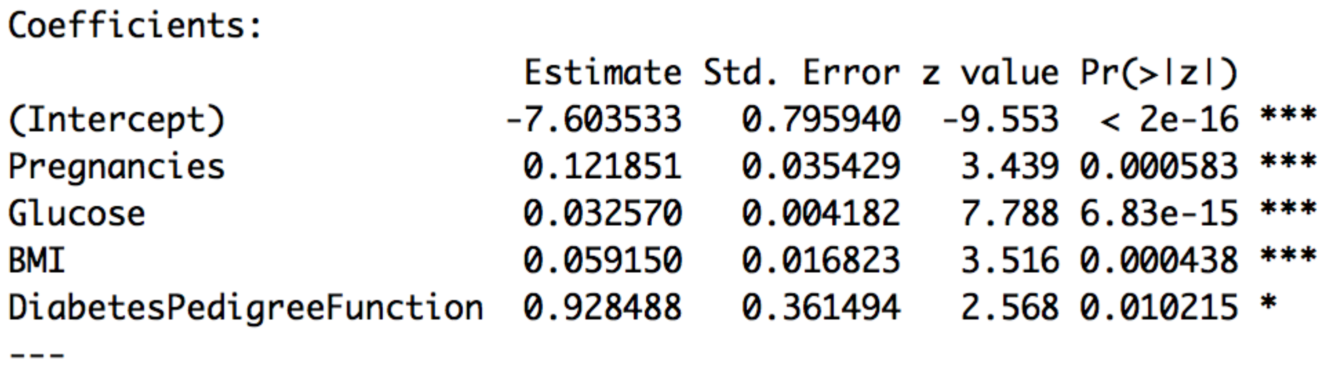
For our model, we had to do some data cleaning. Few values had to be removed from the data set as to obtain efficient results

Following are the values we got for our data set:

AIC Value: 465.69
Null Deviance: 586.51 on 459 degrees of freedom
Residual Deviance: 447.69 in 451 degrees of freedom

a) Feature Selection in backward Logistic Regression:

The available features from the data set are Pregnancies, Glucose, Body Mass Index (BMI), Insulin, Skin Thickness, Blood Pressure, Age, Diabetes Pedigree Function. The model made us choose the best possible combination of feature selection for achieving the highest accuracy.

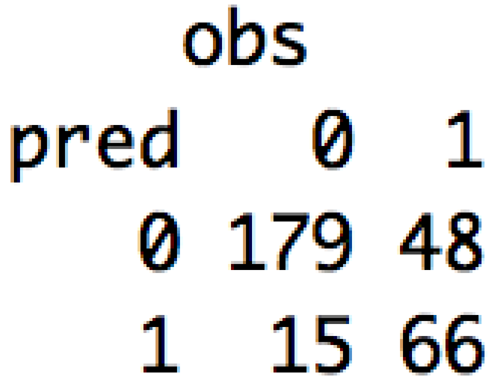


```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -7.603533   0.795940  -9.553  < 2e-16 ***
Pregnancies               0.121851   0.035429   3.439 0.000583 ***
Glucose                   0.032570   0.004182   7.788 6.83e-15 ***
BMI                       0.059150   0.016823   3.516 0.000438 ***
DiabetesPedigreeFunction  0.928488   0.361494   2.568 0.010215 *  
---
```

We know that these are the best values as the p-value for each feature is less than 0.05.

b) Confusion Matrix:

Apart from these values the Logistic Regression model also gives us a confusion matrix that helps us to predict the accuracy. The confusion matrix is as shown in the below diagram.



```
    obs
pred   0  1
   0 179 48
   1  15 66
```

From the above matrix, we can derive the following conclusions. Firstly, this matrix is obtained from the validation set (40%) and the following numbers in the matrix are the number of cases from the observer and predicted values. The row represents the observed values from the data set while the column represents the predicted values from the data set. 0 is said to be a negative case of diabetic or non-diabetic. 1 is said to be a positive case of diabetic.

Through the confusion matrix we majorly require two values that are, True Positive and True Negative. The 1*1, 2*2 elements are True Positive and True Negative respectively. The number of cases in True Negative are 179, that is total number of 179 cases have been true where the prediction and the observed values said the patient is non-diabetic. Similarly total number of 66 cases are recorded to be true positive where both predicted and observed valued said the patient is diabetic. The other two values are false positive and false negative where the predicted and observed values are different from each other. The accuracy of the total model has been up to 79% that is approximately 80%. We used Backward Step Logistic Regression.

c) Accuracy of the Statistical Model from R

We obtain the accuracy from the Confusion Matrix. From the above figure we calculate the accuracy. Accuracy can be computed by the values of true positive and true negative. We have 179 and 66 as true positive and true negative values respectively. By the following formula:

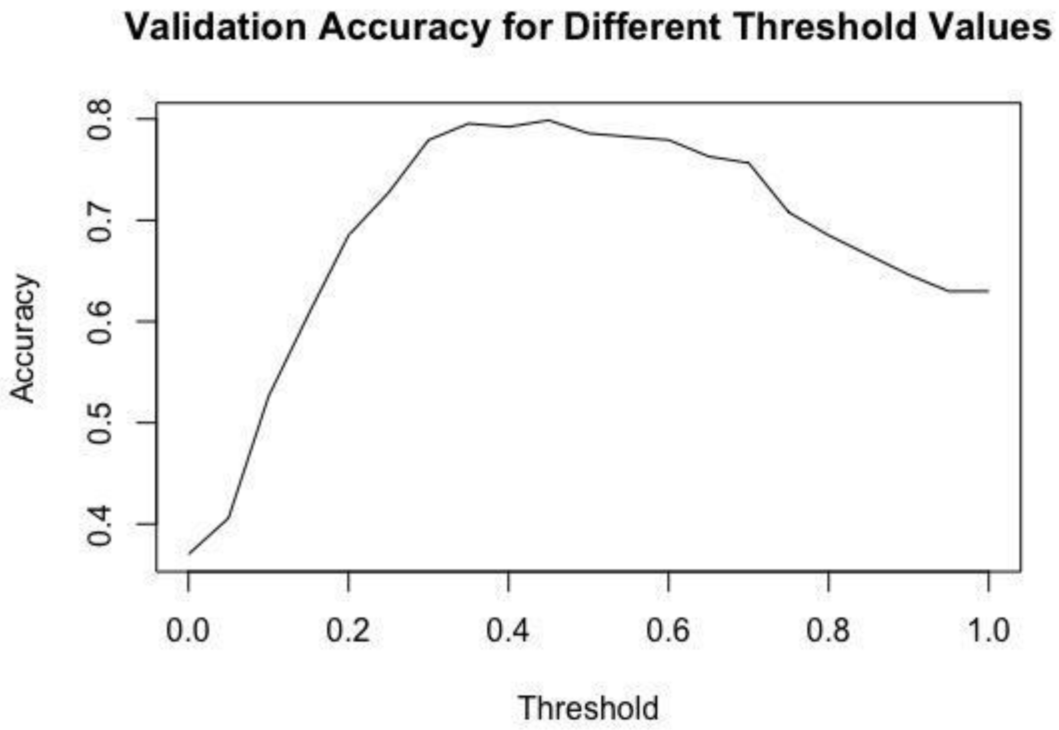
Accuracy = (true positive + true negative) / all the values of the matrix

= (179 + 66) / (179 + 66+ 15+ 48)

= 245 / 305

= 0.795 = 79.5%

Accuracy obtained = 79.5%



From the graph we can say that the highest accuracy is at a threshold of 0.5.

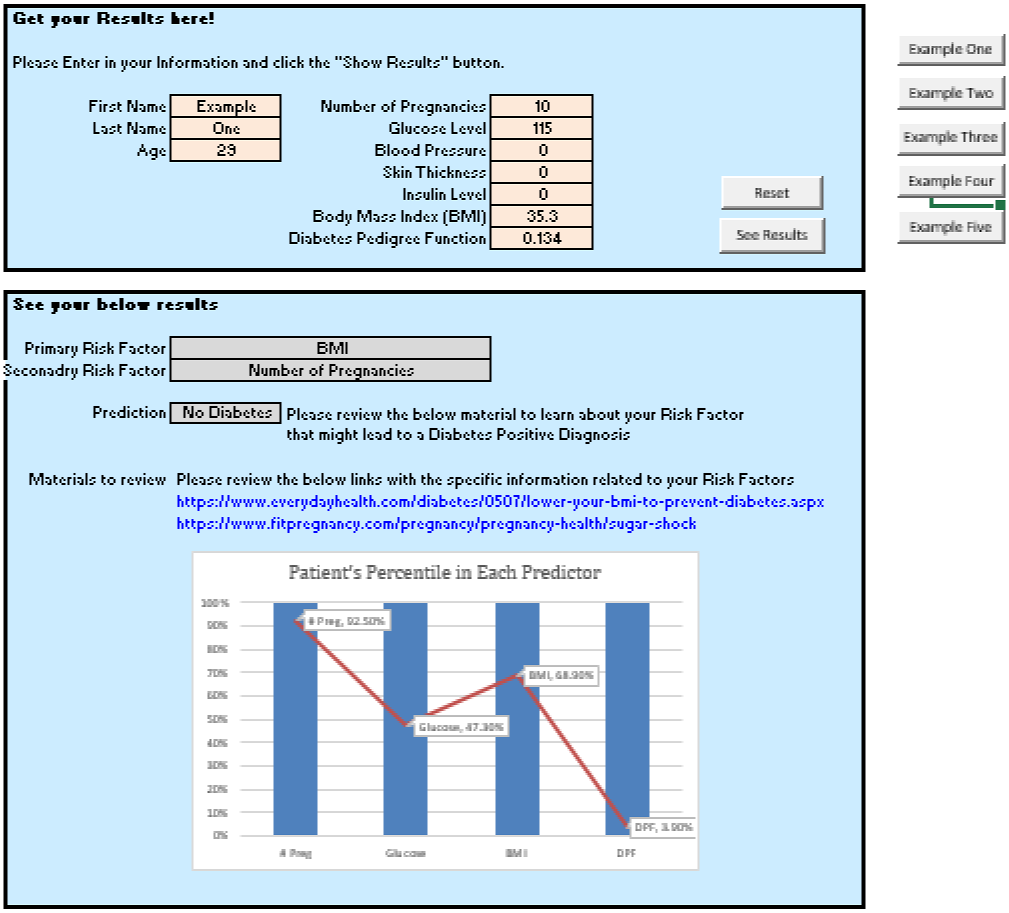
3.2 Walking-through of System

The system will start with a portal where users can enter in each data point. The data will then be processed through the logic model to determine which group the patient belongs in. Once that determination is made, the system will show whether the patient appears to be at risk for diabetes.

If the patient is at risk for diabetes, the system will produce customized packets of information and suggested screening for the patient to continue and avoid the diagnosis. The system will be designed to be very simple where almost anyone that has the data points can use it to see their results.

The patient will then see the Primary and Secondary risk factors that could lead them to a Positive Diagnosis of Diabetes. Along with that information is some guidance to some helpful website links that are targeted based on the Primary and Secondary risk factors.

a) Working of the system and screenshots:



Get your Results here!

Please Enter in your Information and click the "Show Results" button.

First Name	Example	Number of Pregnancies	10
Last Name	One	Glucose Level	115
Age	29	Blood Pressure	0
		Skin Thickness	0
		Insulin Level	0
		Body Mass Index (BMI)	35.3
		Diabetes Pedigree Function	0.134

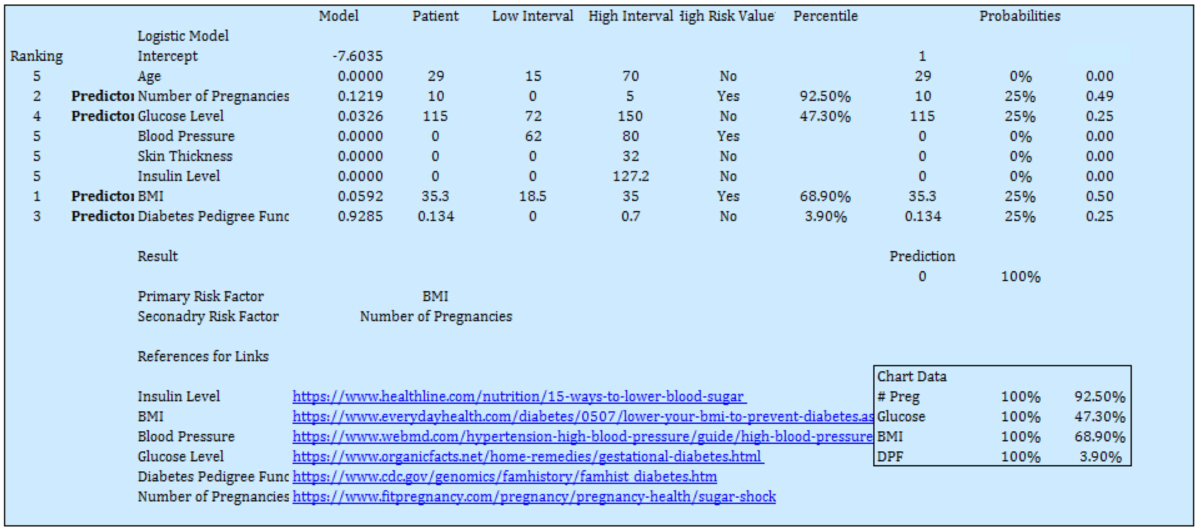
See your below results

Primary Risk Factor: BMI
Secondry Risk Factor: Number of Pregnancies

Prediction: No Diabetes — Please review the below material to learn about your Risk Factor that might lead to a Diabetes Positive Diagnosis

Materials to review: Please review the below links with the specific information related to your Risk Factors
https://www.everydayhealth.com/diabetes/0507/lower-your-bmi-to-prevent-diabetes.aspx
https://www.fitpregnancy.com/pregnancy/pregnancy-health/sugar-shock

Excel Model for DSS



Ranking			Model	Patient	Low Interval	High Interval	ligh Risk Value	Percentile	Probabilities		
		Logistic Model									
Ranking		Intercept	-7.6035						1		
5		Age	0.0000	29	15	70	No		29	0%	0.00
2	**Predictoı**	Number of Pregnancies	0.1219	10	0	5	Yes	92.50%	10	25%	0.49
4	**Predictoı**	Glucose Level	0.0326	115	72	150	No	47.30%	115	25%	0.25
5		Blood Pressure	0.0000	0	62	80	Yes		0	0%	0.00
5		Skin Thickness	0.0000	0	0	32	No		0	0%	0.00
5		Insulin Level	0.0000	0	0	127.2	No		0	0%	0.00
1	**Predictoı**	BMI	0.0592	35.3	18.5	35	Yes	68.90%	35.3	25%	0.50
3	**Predictoı**	Diabetes Pedigree Func	0.9285	0.134	0	0.7	No	3.90%	0.134	25%	0.25

Result — Prediction: 0, 100%

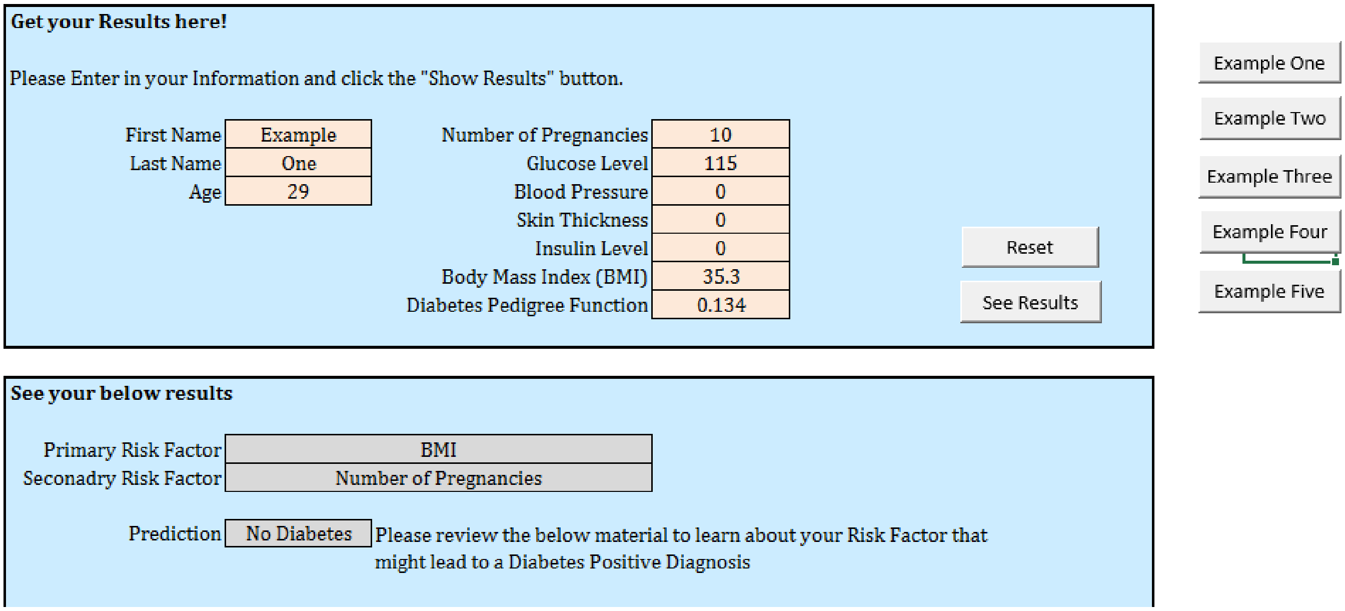
Primary Risk Factor: BMI
Seconadry Risk Factor: Number of Pregnancies

References for Links

Insulin Level	https://www.healthline.com/nutrition/15-ways-to-lower-blood-sugar
BMI	https://www.everydayhealth.com/diabetes/0507/lower-your-bmi-to-prevent-diabetes.as
Blood Pressure	https://www.webmd.com/hypertension-high-blood-pressure/guide/high-blood-pressure
Glucose Level	https://www.organicfacts.net/home-remedies/gestational-diabetes.html
Diabetes Pedigree Func	https://www.cdc.gov/genomics/famhistory/famhist_diabetes.htm
Number of Pregnancies	https://www.fitpregnancy.com/pregnancy/pregnancy-health/sugar-shock

Chart Data		
# Preg	100%	92.50%
Glucose	100%	47.30%
BMI	100%	68.90%
DPF	100%	3.90%

Logic model



Get your Results here!

Please Enter in your Information and click the "Show Results" button.

First Name	Example	Number of Pregnancies	10
Last Name	One	Glucose Level	115
Age	29	Blood Pressure	0
		Skin Thickness	0
		Insulin Level	0
		Body Mass Index (BMI)	35.3
		Diabetes Pedigree Function	0.134

See your below results

Primary Risk Factor	BMI
Seconadry Risk Factor	Number of Pregnancies

Prediction	No Diabetes	Please review the below material to learn about your Risk Factor that might lead to a Diabetes Positive Diagnosis

Primary and secondary risk factors

4.0 Implementation Plan and Subsystems

4.1 Implementation Plan

We have picked a data set from kaggle that has various predictor variables responsible for causing diabetes such as Pregnancies, Glucose, Blood Pressure, Insulin, BMI, Skin Thickness, Age that leads to a response variable called outcome where the result is shown.

The dataset that we have is uncleaned data which has many missing values and since there isn't any use of such rows we clean the data set manually. Once the cleaned data set is obtained we now divide this data into two sets, training set (60%) where we use the values to build a model and the rest as validation set (40%) where the built model is tested for the data. A model is built using Decision Trees in the Excel and Logistic Regression in R.

Depending on the accuracy from both the methods the best one is chosen and then validation set in tested. Finally we will be having the model where the input value is given according to various predictor variables and we get a response variable telling us whether the user will be a diabetic or not in the future.

Since the user finds it difficult to enter the values into the model directly, we have an interface where the user can seamlessly enter the data. In this interface, the user will have a pleasing look and feel interface where they directly enter the values which will direct to the backend, where the model that we built will take them, calculate a response and generate an outcome. This outcome is again retrieved and showed to the user. The outcome will not only show if the user will be a diabetic in the future but also tells the reasons responsible for his diabetes. This helps the user take preventive measures way forward so that he can stay fit and healthy.

The input from the model goes through a logic system that processes the data and outputs all the necessary statistics to inform the patient on how to best proceed. The information given is based on recommendations for which Primary and Secondary factors the patient should be concerned about, as well as letting the patients know how their Predictive factors compare to the rest of those surveyed.

4.2 Model Description and Subsystem

The data entered in from each patient will be stored in a table and collected as raw input. This will be known as the **Data Model**. It will be encouraged that each patient has complete data with no missing fields. Based on results from the Logistic Regression exercise from the Training Data set, we will form different groups where the patients tend to cluster based on their reading. The coefficients from the Logistic Regression will be used against each Patient's numbers to produce a Diagnosis Prediction. Then the data will be observed to see if each Factor's value falls within a High or Low Risk range. After that determination, the scoring system based on each Factor's probability will rank each Factor, resulting in a Primary Factor and Secondary Factor contributing to the Diagnosis.

Based on the various factors involved, each patient might gravitate towards one or two particular factors that will increase their likelihood of getting a diabetes diagnosis. These will be ranked and shown in the output, along with the relevant packet information the patient can follow for more directions on how to proceed.

The data will be taken with from the Data Model to the **Logic Model**. This model will group a patient based on the data points and factors into One or Two Factors. Based on a scoring system. The Cluster Model will have logic that places.

We see the Predictive Factors as follows:

- Pregnancies Factor: Those patients who have had a higher number of pregnancies will be educated about how this factor contributes to possibly developing diabetes in the future.

- Glucose Factor: These patients will be clustered into a group where the patient's level of Glucose in blood is a reflection of diet and the body's ability to proceed Glucose in the Bloodstream.

- BMI Factor: BMI is a weight measurement for the patient adjusted for other factors such as height. Scoring higher on this factor tends to show the patient is overweight and possibly obese, leading to an increased risk of diabetes.

- Diabetes Pedigree Function Factor: This number accounts for the genetic history of the Patient. Scoring high in this field denotes the patient has a high incidence in family history of being Type II diabetic.

Once all these patients are more strongly associated with one or two factor, a customized packet with specific information and guidance will be issued for that particular patient. This will be known as the **Output Model**. That will guide them on the specific steps that are required to move ahead with preventing diabetes. Below are two examples that illustrate this process.

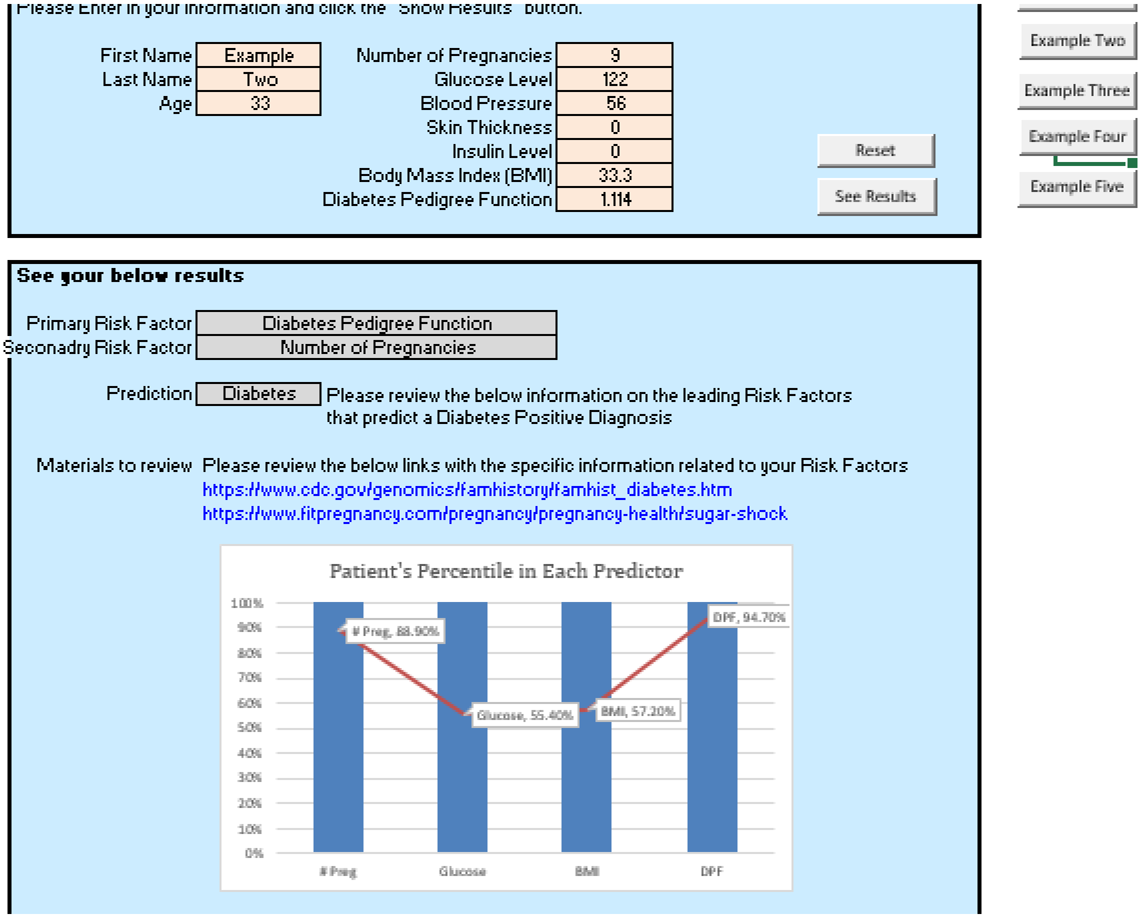
Output 1 Example

For the first image labeled as “Output 1”, we see a Patient named “Example Two”, She has had 9 Pregnancies in her life and has a Diabetes Pedigree Function of “1.114”. These two readings already indicate she is in a high range than most patients. We can see in the Chart that for Pregnancies, she is in the 88.9% Percentile of Patients and with the Diabetes Pedigree Function, she is in the 94.7% Percentile. The model is prediction she is already diabetic. The output gives the relevant information on these two big factors and informs the patients the two biggest factors leading to that prediction.

Output 2 Example

For the second image labeled as “Output 2”, we see a Patient named “Example Three”, her main two risk factors are listed as BMI and Glucose. These two readings already indicate she is in a high range than most patients for these two readings. We can see in the Chart that for BMI, she is in the 88.2% Percentile of Patients and with Glucose, she is in the 83.1% Percentile. The model is prediction she is already diabetic. The output gives the relevant information on these two big factors and informs the patients the two biggest factors leading to that prediction.

Output for Excel model:



Please Enter in your information and click the "Show Results" button.

First Name	Example	Number of Pregnancies	9
Last Name	Two	Glucose Level	122
Age	33	Blood Pressure	56
		Skin Thickness	0
		Insulin Level	0
		Body Mass Index (BMI)	33.3
		Diabetes Pedigree Function	1.114

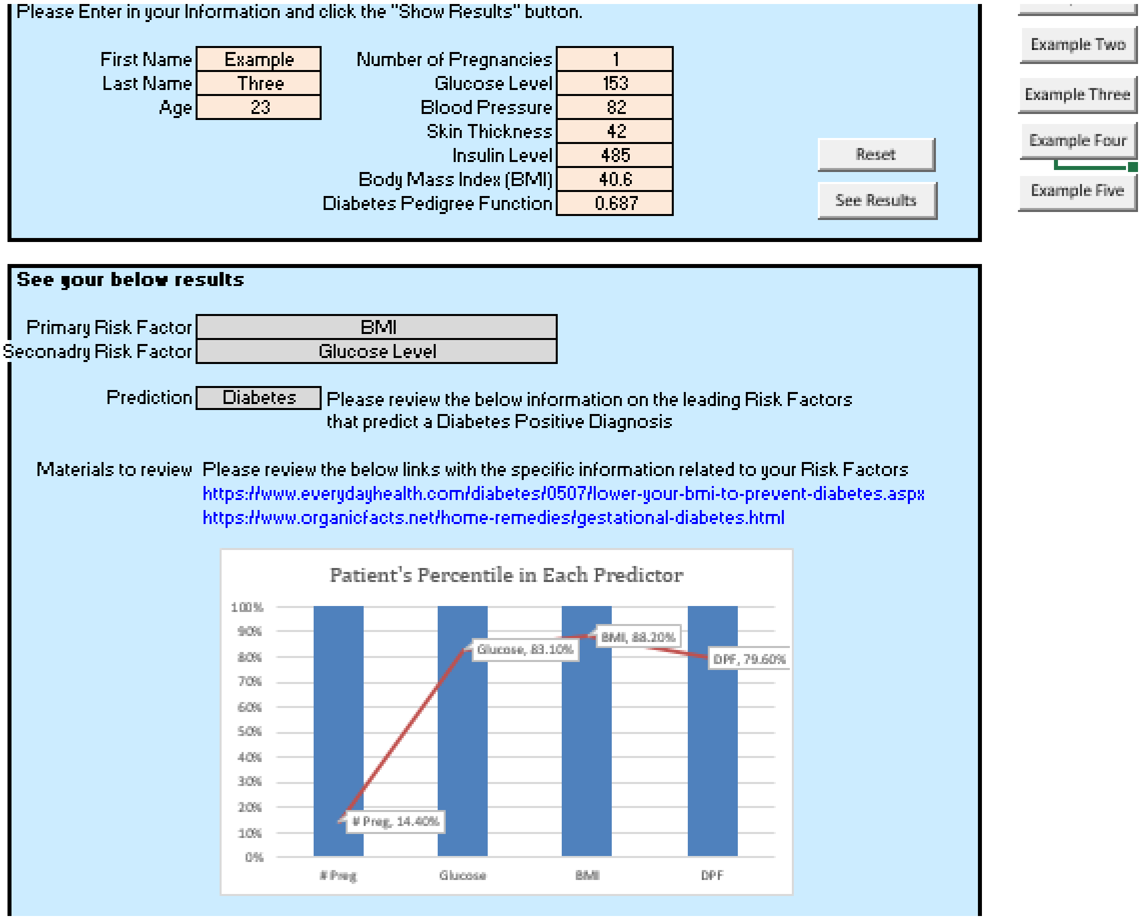
See your below results

Primary Risk Factor	Diabetes Pedigree Function
Seconadry Risk Factor	Number of Pregnancies

Prediction Diabetes Please review the below information on the leading Risk Factors that predict a Diabetes Positive Diagnosis

Materials to review Please review the below links with the specific information related to your Risk Factors
https://www.cdc.gov/genomics/famhistory/famhist_diabetes.htm
https://www.fitpregnancy.com/pregnancy/pregnancy-health/sugar-shock

Output 1



Please Enter in your Information and click the "Show Results" button.

First Name	Example	Number of Pregnancies	1
Last Name	Three	Glucose Level	153
Age	23	Blood Pressure	82
		Skin Thickness	42
		Insulin Level	485
		Body Mass Index (BMI)	40.6
		Diabetes Pedigree Function	0.687

See your below results

Primary Risk Factor: BMI
Seconadry Risk Factor: Glucose Level

Prediction: Diabetes — Please review the below information on the leading Risk Factors that predict a Diabetes Positive Diagnosis

Materials to review: Please review the below links with the specific information related to your Risk Factors
https://www.everydayhealth.com/diabetes/0507/lower-your-bmi-to-prevent-diabetes.aspx
https://www.organicfacts.net/home-remedies/gestational-diabetes.html

Output 2

5.0 Evaluation Plan

The prior confusion matrix predicted the model is almost 80% accurate in predicting diabetes. Based on this analysis we use the most accurate logistic model found to predict results.

Meanwhile in the process the data will be visualised and presented for the clear understanding of the user. From the obtained final data we will be in a position to calculate and predict the factors that are majorly responsible for causing diabetes in different cases. We will have the data that tells us what are the preventive measures that the patient needs to take care to avoid becoming a victim of diabetes.

The main point for evaluating this model is its' effectiveness on reducing the Positive Diabetes Diagnosis rates among the Pima Indian community. By Educating and Informing the whole population on their unique and specific risks, it could lead to a cultural shift the reduces the numbers of incidents.

While these results can be difficult to quantify in terms of its’ helpfulness in the community, one should look at the improved life many in the community may have by avoiding this difficult disease all together.

6.0 Conclusion

Through this project with the already present data set for factors causing diabetes we build a function that can take the input value and predict whether the patient is likely to be a diabetic or not. This helps many hospitals to calculate the results within minutes, instead of waiting for the lab reports for hours and days.

It is helpful to know the reasons responsible for diabetes for that particular person so that he/she can take measures well in advance. The Decision Support Systems built will help many hospitals calculate results especially in emergency situations, that saves a lot of time during medication. This system provides a simple interface almost anyone can use and gives actionable intelligence that leads each patient to take personal action and care of their fates.

This system will be a strong predictive tool for helping many women in the Pima Indian community. Giving people specific and concise information to direct them to change daily habits, be aware of potential challenges will greatly reduce the incidence of diabetes in these communities.

Appendices:

R-Code:

```
mydata<-read.csv("diabetes.csv",header=TRUE)

# Summarize the dataset
summary(mydata)

# Find the number of rows of the dataset
row<-nrow(mydata)

set.seed(12345)
trainindex <- sample(row, 0.6*row, replace=FALSE)
training <- mydata[trainindex,]
validation <- mydata[-trainindex,]

# Logistic regression: USe glm instead of lm
# Set 'family=binomial' for a logistic regression model
mylogit<-glm(Outcome ~.,data=training, family=binomial)
# Summary of the regression
summary(mylogit)

# Model coefficients
coef(mylogit)

# stepwise logistic regression with backward selection
mylogit.step = step(mylogit, direction='backward')
summary(mylogit.step)

#############################################################################
########################################

# Prediciting the validation dataset
mylogit.probs<-predict(mylogit.step,validation,type="response")
mylogit.probs
# Check some of the predicted values in validation set
# Note these predicted values are probabilities
mylogit.probs[1:5]

# Create confusion matrix by specifying a cutoff value
mylogit.pred = rep("Non-diabetic", 0.4*row)
mylogit.pred[mylogit.probs > 0.5] = "diabetic"
table(mylogit.pred, validation$Outcome)




#########################################################################

# Confusion Matrix and some accuracy measures with R packages
install.packages("SDMTools")
library(SDMTools)
# note you can specify different values of threshold value from 0 to 1
matrix = confusion.matrix(validation$Outcome,mylogit.probs,threshold=0.5)
matrix

AccuMeasures = accuracy(validation$Outcome,mylogit.probs,threshold=0.5)
# Extracting specific values from accuracy table
AccuMeasures
```

```
# Next we show how to choose the best threshold value with highest accuary
# creating a range of values to test for accuracy
thresh=seq(0,1,by=0.05)

# Initializing a 1*20 matrix of zeros to save values of accuracy
acc = matrix(0,1,20)

# computing accuracy for different threshold values from 0 to 1 step by 0.05
for (i in 1:21){
  matrix = confusion.matrix(validation$Outcome,mylogit.probs,threshold=thresh[i])
  acc[i]=(matrix[1,1]+matrix[2,2])/nrow(validation)
}
# print and plot the accuracy vs cutoff threshold values
print(c(accuracy= acc, cutoff = thresh))
plot(thresh,acc,type="l",xlab="Threshold",ylab="Accuracy", main="Validation Accuracy for
Different Threshold Values")
```

*** THE END ***