DSC 323: Data Analysis & Regression
Report Final

Predicting Diabetes in Women

Zaid Akel, Sierra Bagwell, Joshua Calloway, Milin Dharmshibhai Desai,
Vallabh Datta Varma Penmetcha

I.ABSTRACT

Diabetes is a chronic group of metabolic diseases where an individual suffers from an extended length of high levels of blood sugar within their body. This can be due to a number of reasons, based upon the type of diabetes; insulin production within the body can be inadequate or much lower than expected, or the body's metabolic actions can be inhibited due to an improper reaction to insulin. Long term, diabetes begins to wear down on the body's organs and veins, meaning that the earlier diabetes can be found and brought under control in an individual, the better the quality of life they'll be expected to have thereafter.

The objective of our research is to design a predictive algorithm using SAS to find the optimal classifier to give the closest results when compared to the predictive models medical professionals often use.

Goal - build a model to accurately predict whether or not an individual would have diabetes based upon recorded measurements per observation; remove outliers, remove insignificant variables by using selection methods, generate a predictive model that is both accurate and has a high sensitivity.

Findings - Certain variables within the dataset were found to have a high impact upon whether or not an individual will have diabetes. This shows a level of correlation between certain variables in a given observation, allowing us to refine a predictive logistic regression model.

Recommendation - given various attributes, the logistic regression model, and the threshold, we could predict with some confidence if someone would have diabetes. Our generated logistic model would identify the most significant predictors for diabetes.

II.INTRODUCTION

Diabetes is a disease in which blood sugar (glucose) levels in your body are too high. Diabetes can cause serious health problems, including heart attack or stroke, blindness, problems during pregnancy, and kidney failure. About 15 million women in the United States have diabetes, or about 1 in every 9 adult woman [1].

The cause of diabetes can be predicted by various attributes such as diet, fitness, age, pregnancy, and genetics. When statistics are collected for both diabetic and non-diabetic populations, we can apply data science techniques to model and predict the likelihood of any individual of getting diabetes. In other words, we could pose the statement to an individual that "Given your BMI and your age" how likely are you to get diabetes if you get pregnant.

The motivation for preventing diabetes is significant due to the damage that diabetes pose to an individual's health. There are four kinds of diabetes ranging from Type-1, Type-2, Prediabetes, Gestational.[2] Diabetes can cause loss of vision, Kidney damage, neuropathy, liver problems and heart problems. Since much of diabetes is affected by glucose and BMI, an individual can live a healthier lifestyle to reduce the likelihood of diabetes.

[1] https://www.womenshealth.gov/a-z-topics/diabetes
[2] N. Sneha & Tarun Gangil - https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0175-6

III.METHODOLOGY

The data was obtained from the kaggle website.[3]

The project was conducted in the following steps as proposed in the project proposal:

I. Clean Data

A. Check for duplicate observations missing values and unexpected outliers.
   1. NOTES: We found some outliers in BMI, Glucose and blood pressure variables. We will try to remove the most severe outliers in model development stage.

B. Remove unnecessary predictors.

C. Impute any necessary data.

D. Interaction Terms
   1. Test if there is an interaction between terms. Possible ones to test:
      a) Age & Glucose
      b) Glucose & Diabetes Pedigree Full
      c) Glucose & BloodPressure

II. Explore Data

A. We will generate box plots to check the effect of all independent variables on outcome. Generate histogram to check the distribution of independent variables. Since "Outcome" is boolean variable it makes no sense to analyze the Scatter Plot, AND PROC MEANS

B. Examine correlations and check for multicollinearity
   1. Remove variables where necessary.

[3] Kaggle Diabetes at https://www.kaggle.com/saurabh00007/diabetescsv#diabetes.csv

III. Analyze Data

A. Full Model
   1. As “Outcome” is Boolean we will check the frequency of the 0's and 1’s. Perform logistic regression and examine the regression analysis for predictors and y-variable. check R2, GOF, AIC, SC, likelihood Estimates, Beta values, multicollinearity etc.,
   2. Establish outliers and influential points; remove them where necessary.

If linearity established, split the data into 80% training and 20% test sets.
If transformations are needed for predictors, transform and repeat B again.

B. Methods of Selection
   1. Run the selection methods on training data set, fit a Regression Model using; compute new y
      a) Forward
      b) Backward
      c) Stepwise selection methods

C. Modeling
   1. Check for R2, GOF, AIC, SC, Likelihood estimates, multicollinearity(1) etc., Check for(2) outliers, influential points and remove them where necessary.
   2. Compute predictions.
   3. If logistic analysis is not satisfied for the model 1 then create model 2 and repeat the process.

IV. Final Model Preparation

A. Examine performance stats for generated models.

B. Select final model.

V. Final Model

A. Discuss and analyze performance stats using test data set.

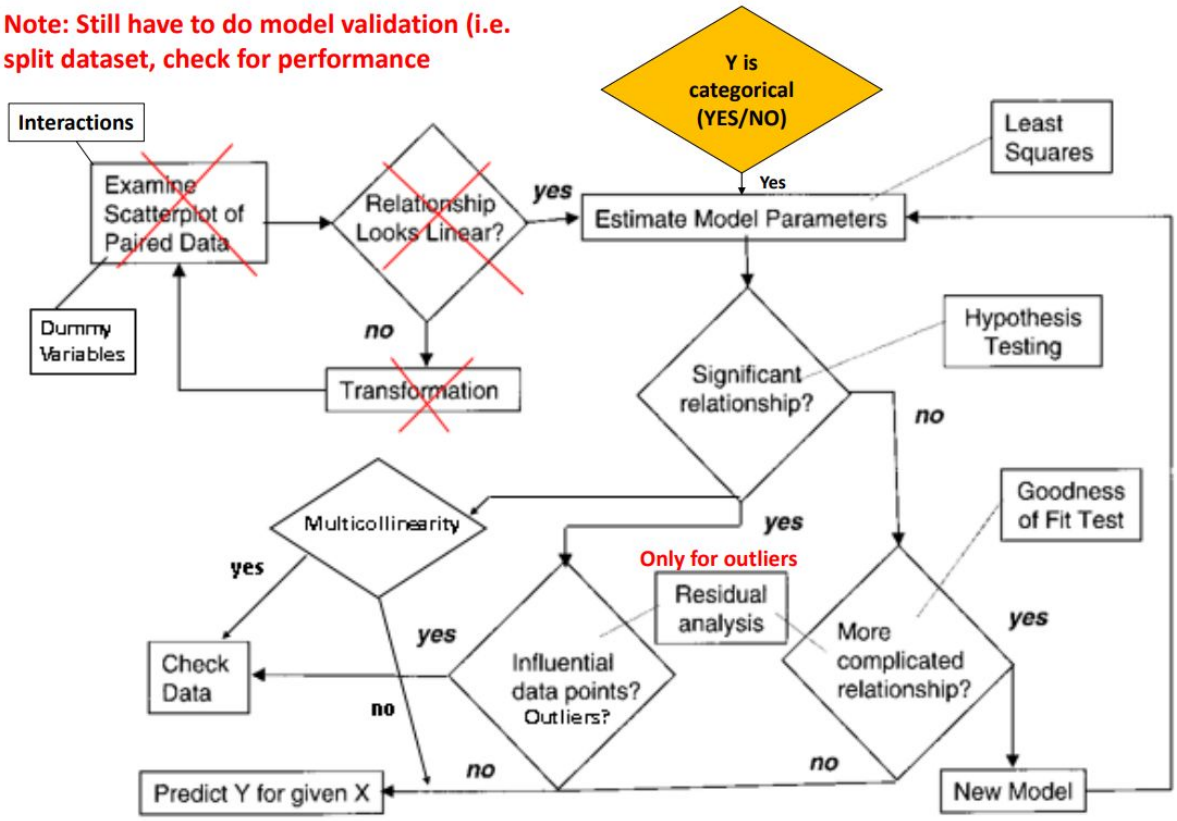
B. Check for instances of collinearity.

C. Check for outliers and influential points.

D. Discuss parameter predictors.

How is a Logistic Regression Analysis done?

Note: Still have to do model validation (i.e. split dataset, check for performance



IV. ANALYSIS, RESULTS & FINDINGS

A. Cleaning Data

The independent variable is the Outcome which is yes if the person got diabetes and no if the person did not get diabetes.

The variables we were working with from the data set were:

- Pregnancies - How many times you have been pregnant; numerical val
- Glucose - Glucose level in the blood; numerical value
- Blood Pressure - Blood Pressure; numerical value
- SkinThickness - Thickness of the skin; numerical value
- Insulin - Insulin level in the blood; numerical value
- BMI - Body Mass Index; numerical value
- Diabetes Pedigree Full; numerical value
- Age; numerical value
- Outcome- 1 = diabetes, 0 = no diabetes ; binary value

We also created three interaction variables:

- Glu_Bp - Glucose and Blood Pressure
- Glu_Dp - Glucose and Diabetes Pedigree Full
- Age_Glu - Age and and Glucose

B. Data Exploration

We got our dataset from Kaggle[4]. The dataset had 768 observations and 9 variables.

[4] Dataset from https://www.kaggle.com/edubrq/diabetes

Frequency Table

Frequency of Outcome

The FREQ Procedure

Outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	500	65.10	500	65.10
1	268	34.90	768	100.00

Running the Frequency Procedure in SAS, we found that an Outcome of not getting diabetes (Outcome = 0) was more likely, composing 65.10% of the values for Outcome. We also found that an Outcome of getting diabetes (Outcome = 1) was less likely, 34.90% of the values for Outcome.

See Appendix A1-3 for exploratory analysis, which includes:

- Boxplots
- Scatterplots
- Histogram

See Appendix A1 for Boxplots:

In our exploratory analysis of the boxplots, we can confirm our findings from the histograms. Looking at the difference from the upper quartile (Q3) from the median and lower quartile (Q1), we see that the difference from the median to the quartiles is not equidistant, indicating that the data is not normally distributed; most of the data is skewed.

In our boxplots, we do not see any outliers. From our preliminary data exploration, we wanted to look especially close at glucose, age and blood pressure which we believed would maintain significant effect on Outcome. Looking at our boxplots, we see that the median Age for a positive Outcome (Outcome = 1; person has diabetes) is higher than the median Age in a negative outcome. This is especially true when looking at the boxplot for Glucose. For a positive Outcome, the median Glucose level is very high, significantly higher than that of a negative Outcome. Because these two variables varied the most with respect to Outcome, we created several interaction terms. Age_Glu serves as the interaction term between Age and Glucose; the boxplot of Age_Glu shows that the median for Age_Glu is also significantly higher in a positive Outcome than negative.

See Appendix A2 for Scatter Plots

Because we used logistic regression to create a model for our binary variable, Outcome, the scatter plot was not beneficial in determining which variables maintain a significant

relationship with Outcome. Looking at the graph, we see this as evident, as the scatterplots with respect to outcome show an indeterminable relationship.

See Appendix A3a for a detailed look at Histograms which includes

- Histograms of every independent variable
- Histograms of interaction variables

In our exploratory analysis of the histograms, we found that none of the variables were exactly normally distributed. The variable BMI is nearly normally distributed, with the mean (31.99258) only slightly less than the median (32); thus each variable maintained a slight skew as indicated by their distribution. Most of the variables are right skewed in their distribution, with their mean greater than their median.

We do not have a histogram for Outcome as Outcome is a binary variable and we would not learn anything by plotting 1 and 0.

See Appendix A4 Full Model Output which includes

- Analysis of Maximum Likelihood
- Goodness of Fit Statistics
- Estimated Correlation Matrix
- Influence Diagnostics

Full Model

Our dependent variable (or y-variable) is “Outcome” which is Boolean. “1” describes the patient as diabetic and “0” describes the patient as not diabetic. Since the dependent variable is Boolean, we should perform logistic regression to predict whether the patient is diabetic or not.

After analysing histograms, box plots and scatter plots for the data we performed logistic regression for full model. There are eight predictors in the data set and we created three interaction variables, so the full model contains eleven variables and they are Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Age_Glu, Glu_Dp, Glu_Bp.

From the likelihood Estimates table [A4.a] we can see that the value of Pr>ChiSq is more than 0.05 for BloodPressure, Skin Thickness, Insulin, Age_Glu and Glu_Bp variables. So, these are insignificant variables and must be addressed. We used selection methods to select the significant variables.

The standardized estimate value is more for “Glucose” and estimate is more for “DiabetesPedigreeFunction”. This indicates that “Glucose” has the strongest influence in predicting the diabetes and “DiabetesPedigreeFunction” has the strongest effect on outcome.

Chand CD et., said that logistic regression model takes formula log[p(x)/1-p(x)] = b1x1 + b2x2 +.... . + bkxk[5]

where X (x1,x2, ... ,xk) represents the vector of k’s risk factors by the logistic regression approach.

The full model equation is

*Log ((Outcome=1)/(Outcome=0)) = -15.0886 + (0.1246*Pregnancies) + (0.0882*Glucose) + (0.0237*BloodPressure) + (0.00152*Skinthickness) – (0.00128*Insulin) + (0.0892*BMI) + (4.2193*DiabetesPedigreeFunction) + (0.0780*Age) – (0.00049*Age_Glu) – (0.0252*Glu_Dp) – (0.00030*Glu_Bp)*

From the model statistics [A4.b] we can say that only 30.87% of variability in outcome can be explained by all the variables. The value of Likelihood ratio is less and values of AIC and SC are more so, we can confirm that this is not a good model.

In the next step we checked for multicollinearity [A4.c] for full model and found that there is a correlation between interaction variables and variable associated with interaction terms hence this can be neglected. So, there is no multicollinearity issue in the model.

In next step we checked for outliers and influential points for full model. From the Pearson residuals [A4.d] there are few extreme outliers i.e., points above and below (_-^+)3 band and they must be deleted from the model in later stage.

The band for influential points is calculated by 2/√n where n is the number of observations. There are 768 observations in full model. Now 2/√n = 0.072. So, all the points above and below (_-^+)0.072 band are considered as influential points. From Dfbetas plots [A4.e] there are many influential points in the data and this is due to incomplete data. For the time being we are going to ignore them and check them again in fitted model and final model.

After full model we split the data into 80% training and 20% test sets. Now there are 615 observations in training set and 153 observations in test set.

[5] Chang CD, Wang CC, Jiang BC. Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. Expert Syst Appl 2011; 38:5507e13

Selection Methods

Stepwise, Forward and Backward selection methods are performed to find the optimal model with significant predictors. Model #2 (Fitted model) is created based on output from three selection methods.

1. Stepwise Selection Method

The Stepwise selection method [A5.a] selected five variables and they are Glucose, BMI, Pregnancies, DiabetesPedigreeFunction and Glu_Bp. It excluded six other variables. From the Likelihood estimates all the variables are significant. The AIC and SC values are significantly improved. R-square and likelihood ratio reduced a bit but we can ignore it.

2. Forward Selection Method

The Forward selection method [A5.a] selected five variables and they are Glucose, BMI, Pregnancies, DiabetesPedigreeFunction and Glu_Bp.

3. Backward Elimination Method

The Backward Elimination method [A5.a] selected five variables and they are Glucose, BMI, Pregnancies, DiabetesPedigreeFunction and Glu_Bp.

All the three selection methods gave the same output. So, model two should be created with five variables (Glucose, BMI, Pregnancies, DiabetesPedigreeFunction and Glu_Bp).

Fitted Model

The fitted model is built form the outputs of the three selection methods. This model has five variables and they are Glucose, BMI, Pregnancies, DiabetesPedigreeFunction and Glu_Bp.

From the likelihood Estimates table [A6.a] we can see that the value of Pr>ChiSq is less than 0.05 for all variables. Hence we can reject the null hypothesis and say that there is at least one variable that has significant influence on outcome.

The standardized estimate value is more for “Glucose” and estimate is more for “DiabetesPedigreeFunction”. This indicates that “Glucose” has the strongest influence in predicting the diabetes and “DiabetesPedigreeFunction” has the strongest effect on outcome.

From the model statistics [A6.b] we can say that 30.82% of variability in outcome can be explained by Glucose, BMI, Pregnancies, DiabetesPedigreeFunction and Glu_Bp variables. The value of Likelihood ratio is decreased from 283.5612 to 226.5596 and values of AIC and SC improved from 733.923 to 586.342 and 789.648 to 612.872 respectively. We can say that this model is better than full model because there is significant improvement in AIC, SC values and error reduced.

Next we checked for multicollinearity for full model and from correlation matrix [A6.c] found

that there is no multicollinearity issue in the model.

Fitted Model Equation

Log ((Outcome=1)/(Outcome=0)) = -9.0938 + (0.1532 Pregnancies) + (0.045*Glucose) + (0.0832* BMI) + (0.9775*DiabetesPedigreeFunction) - (0.00010*Glu_Bp)*

The y-variable is transformed into logarithmic so, to find the effect of each variable on outcome, it should be retransformed.

With increase in number of pregnancies by one, the odds of diabetes will increase by 16.55%. With increase of person's glucose level by one unit, the odds of diabetes will increase by 4.6%. With increase in body mass index of a person by one unit, the odds of diabetes will increase by 8.676%. With increase in diabetes pedigree function by one unit, the odds of diabetes will increase by 165.7%. With increase in interaction between glucose and blood pressure by one unit, the odds of diabetes will decrease by 0.01%.

Outliers and Influential Points

From the Pearson residuals [A6.d] there are three points above +3 band and two points below -3 bands. There are few points on the line and we do not consider them as outliers. The observations 7, 10 350, 623 and 745 are outliers and are removed from the model.

The band for influential points is calculated by $2/\sqrt{n}$ where n is the number of observations. There are 615 observations in fitted model. Now $2/\sqrt{n} = 0.08$. So, all the points above and below (_-^+)0.08 band are considered as influential points. From Dfbetas plots [A6.e] there are many influential points in the data and this is due to incomplete data. We didn't delete them because there are many influential points and if we delete them we will be left with few observations. They don't have much effect on our analysis.

After deleting the outliers we again performed three selection methods to build third model.

Selection Methods

1. Stepwise Selection Method

The Stepwise selection method [A5.b] selected the same variables as in fitted model and they are Glucose, BMI, Pregnancies, DiabetesPedigreeFunction and Glu_Bp. From the Likelihood estimates all the variables are significant. The AIC and SC values are reduced. R-square and likelihood ratio increased.

2. Forward Selection Method

The Forward selection method [A5.b] selected Glucose, BMI, Pregnancies, DiabetesPedigreeFunction and Glu_Bp variables.

3. Backward Elimination Method

The Backward elimination method [A5.b] selected Glucose, BMI, Pregnancies, DiabetesPedigreeFunction and Glu_Bp variables.

All the three selection methods gave the same output. So, third model should have the same variables as fitted model (Glucose, BMI, Pregnancies, DiabetesPedigreeFunction and Glu_Bp). Re run the model after deleting the outliers and computing the analysis again and check for all statistics, multicollinearity and outliers.

Final Model

Our Final model is created after deleting the outliers and there are 610 observations in the model. The variables Glucose, BMI, Pregnancies, DiabetesPedigreeFunction and Glu_Bp are included in this model. Logistic regression is performed on the training set.

From the likelihood Estimates [A7.a] table we can see that the value of Pr>ChiSq is less than 0.05 for all variables. Hence we can reject the null hypothesis and say that there is at least one variable that has significant influence on outcome.

The standardized estimate value is more for “Glucose” and estimate is more for “DiabetesPedigreeFunction”. This indicates that “Glucose” has the strongest influence in predicting the diabetes and “DiabetesPedigreeFunction” has the strongest effect on outcome.

From the model statistics [A7.b] we can say that 34.09% of variability in outcome can be explained by the variables. The value of Likelihood ratio is increased from 226.5596 to 254.3398 and AIC and SC values decreased from 586.342 to 550.595 and 612.872 to 577.076 respectively. From the above statistics we can say that this model is better than the fitted model.

Then we run “corrp” to check for multicollinearity issue. The estimated correlation matrix [A7.c] shows that all the values are within (_-^+)0.09 range, this indicates that there is no multicollinearity issue with model.

In next step we checked for outliers and influential points. From the Pearson residuals [A7.d] there are six points above +3 band and three points below -3 bands which are considered as outliers. Our goal is to predict whether a patient is diabetic or not and the above patients are the ones who suffer from diabetes and excluding those points will make a bad model. Hence we ignored the outliers.

The band for influential points is calculated by $2/\sqrt{n}$ where n is the number of observations. There are 610 observations in final model. Now $2/\sqrt{n} = 0.0809$. So, all the points above and

below (_-^+)0.0809 band are considered as influential points. From Dfbetas plots [A7.e] there are many influential points in the data and we didn't delete them because we will be left with few observations. They don't have much effect on our analysis.

V. RESULTS

The variables that are finally included in model are Glucose, BMI, Pregnancies, DiabetesPedigreeFunction and Glu_Bp. Our regression model is able to explain 34.09% of variability in outcome using these variables.

Predictions are done on testing set. Testing test contains 153 observations. From classification table [A8.b] the value of Sensitivity+Specificity is more for 0.3 hence we performed classification metrics using the threshold value as 0.3.

From predicted probabilities [A8.a] predicted probability when number of pregnancies are 1, glucose is 89, BMI 28.1, Diabetes pedigree function was 0.167 and Glu_Bp was 5874 is 0.03248. Since, predicted probability is below 0.3 for person with above stats, we can conclude that she is not diabetic.

95% confidence interval is (0.01913, 0.05464). So, 95% of time, the predicted probability will fall within 0.01913 and 0.05464. This indicates that, the odds of a person having diabetes when number of pregnancies are 1, glucose is 89, BMI 28.1, Diabetes pedigree function was 0.167 and Glu_Bp was 5874 will increase between 1.93% and 5.616%.

Then we performed frequency for prediction and from the table [A8.c] number of true positives are 41, number of true negatives are 68, number of false positives are 36 and number of false negatives are 8.

Lavrac, et.al [6]Sensitivity measures the fraction of positive cases that are classified as positive. Specificity measures the fraction of negative cases that are classified as negative. Sensitivity for our regression model is 83%. The probability that a person has diabetes given that she is diabetic is 83%. Kaoru, et. al [7] It is also given by ROC curve [A8.d]. Specificity for the model is 65.38%. The probability that a person doesn't have diabetes given that he is not diabetic is 65.38%. Accuracy and precision of the model are 71.24% and 53% respectively. F-metric is 64.7%

[6] Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med 1999;16:3e23.

[7] The Logistic Regression and ROC Analysis of Group-based Screening for Predicting Diabetes Incidence in Four Years.

- http://www.lib.kobe-u.ac.jp/repository/81000093.pdf

Our goal is to predict the probability of a person having diabetes based on several attributes, we should optimize the model for more sensitivity. Our model has high sensitivity than specificity and is good but, we got only 83% sensitivity. For health care the model should have high accuracy and sensitivity (above 96%). So, this cannot be used practically. We should do additional work to improve the model and discussed them in future work section.

VI. FINDINGS

From our logistic regression model we found that Glucose has the strongest influence in predicting the diabetes and DiabetesPedigreeFunction has the strongest effect on outcome.

Log (Outcome=1/Outcome=0) = -10.2161 + (0.1662 Pregnancies) + (0.0499*Glucose) + (0.0940* BMI) + (1.2635*DiabetesPedigreeFunction) - (0.00011*Glu_Bp)*

With increase in number of pregnancies by one, the odds of diabetes will increase by 18.08%. With increase of person's glucose level by one unit, the odds of diabetes will increase by 5.116%. With increase in body mass index of a person by one unit, the odds of diabetes will increase by 9.85%. With increase in diabetes pedigree function by one unit, the odds of diabetes will increase by 253.77%. With increase in interaction between glucose and blood pressure by one unit, the odds of diabetes will decrease by 0.011%.

It is observed that with increase in glucose levels and diabetes pedigree function a person is most likely to be affected by diabetes.

FUTURE WORK

For future work, we need more data to work with and maybe different datasets that have more variables. We could then try different interaction variables to see if we get a better model. With more data, we would need to extend more data cleaning, handling missing data, removing outliers and influential points. Finally, we could do more background research on diabetes to get a better business understanding. There are four severity levels of diabetes ranging from Gestational to Prediabetes, to Type-2, and Type-1.[8] It would be interesting to explore datasets and how they predict the different severities of diabetes. In addition to Logistical Regression

[8]N. Sneha & Tarun Gangil - https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0175-6

Models, we could apply the other data science models of decision trees, K nearest neighbors, or random forests as described in by N. Sneha & Tarun Gangil.[9]

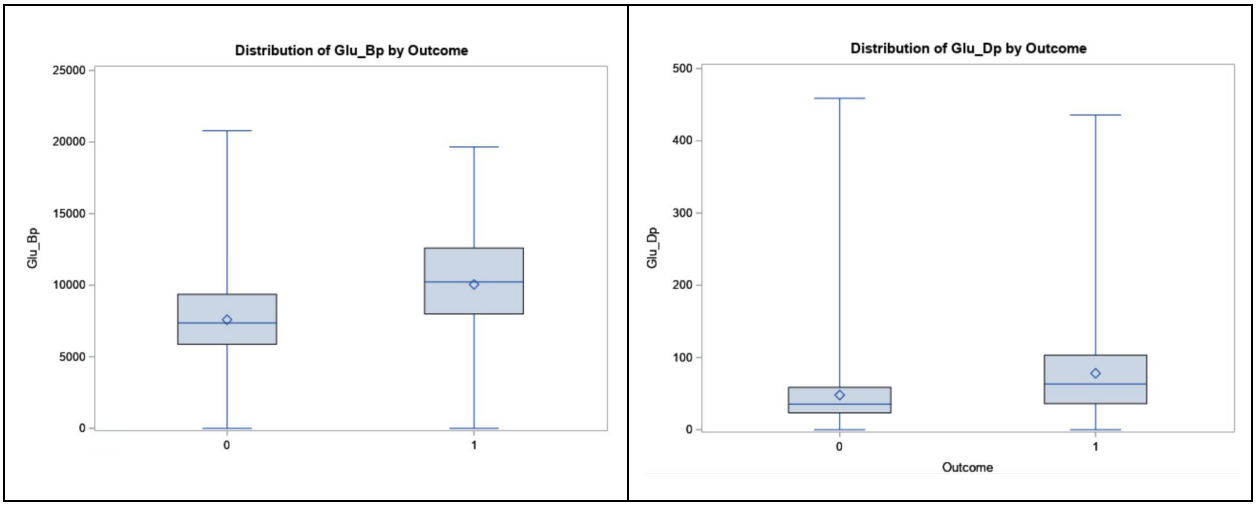
REFERENCES

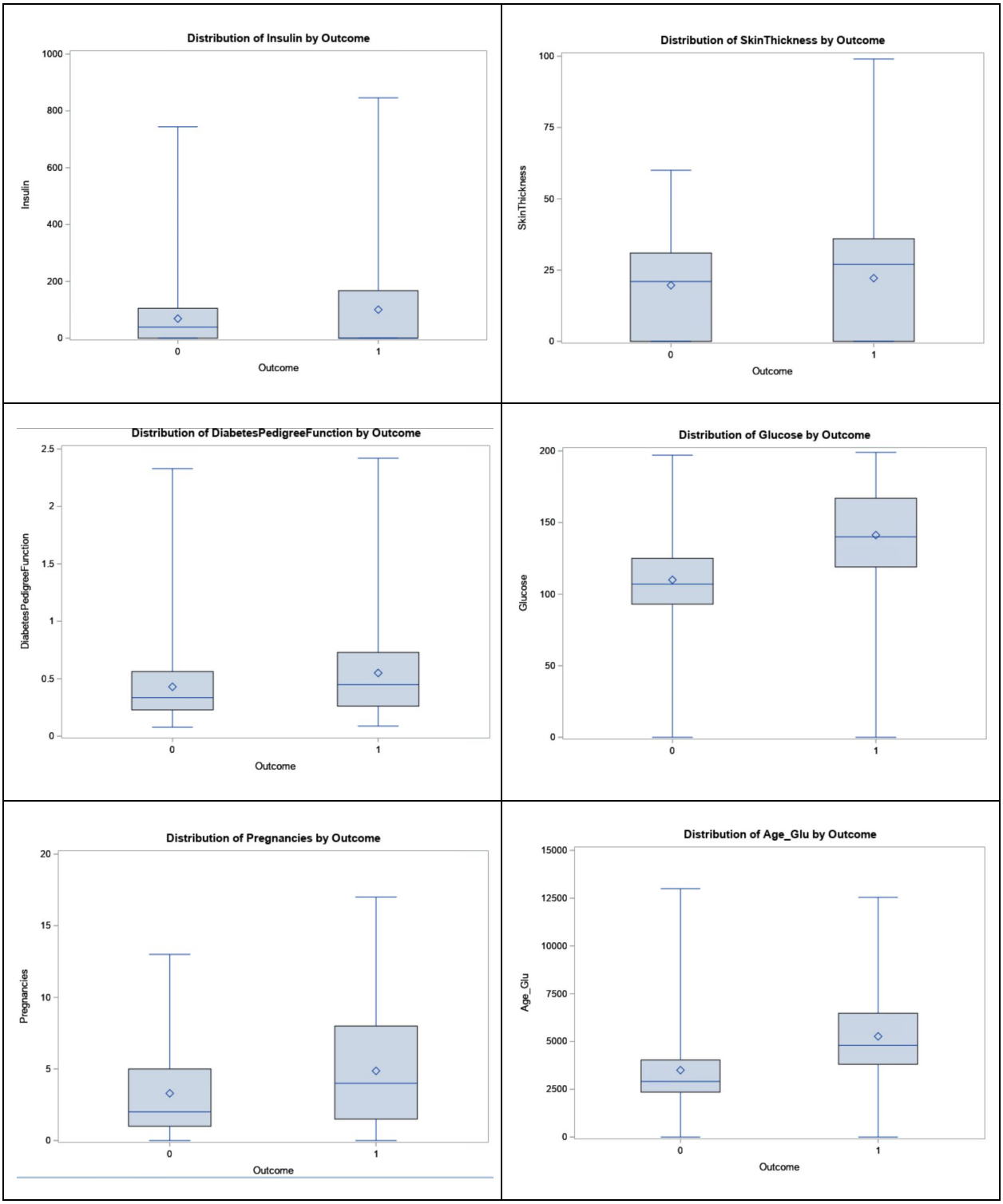
- Chang CD, Wang CC, Jiang BC. Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. Expert Syst Appl 2011; 38:5507e13.
- The Logistic Regression and ROC Analysis of Group-based Screening for Predicting Diabetes Incidence in Four Years. http://www.lib.kobe-u.ac.jp/repository/81000093.pdf
- N. Sneha & Tarun Gangil - https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0175-6
- Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med 1999;16:3e23.
- https://www.womenshealth.gov/a-z-topics/diabetes.

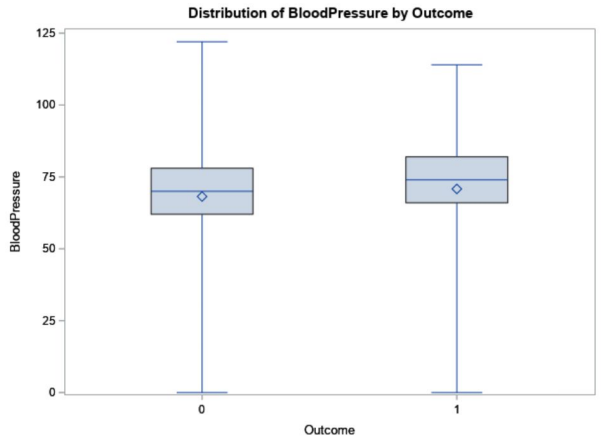
[9]N. Sneha & Tarun Gangil - https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0175-6

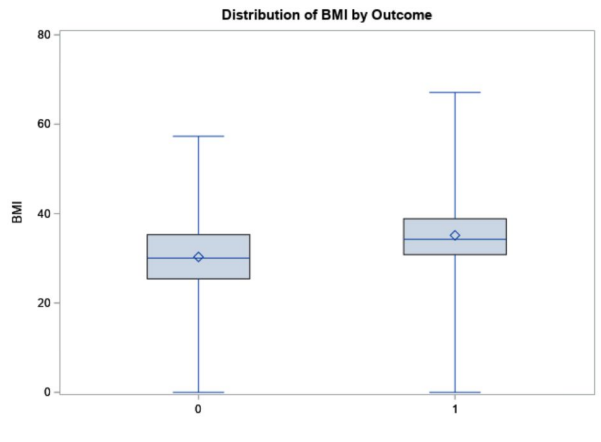
VII. APPENDIX

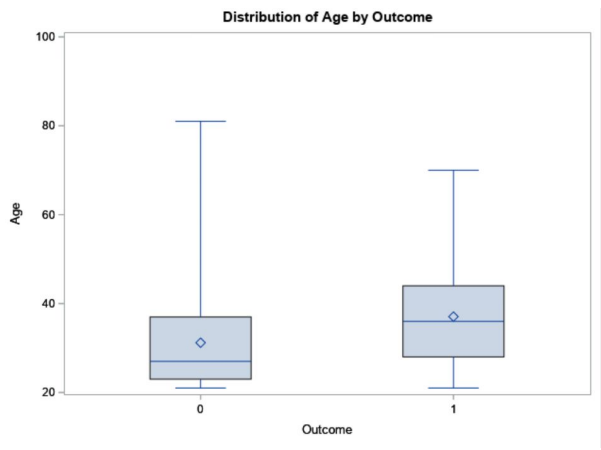
A1. Box Plots



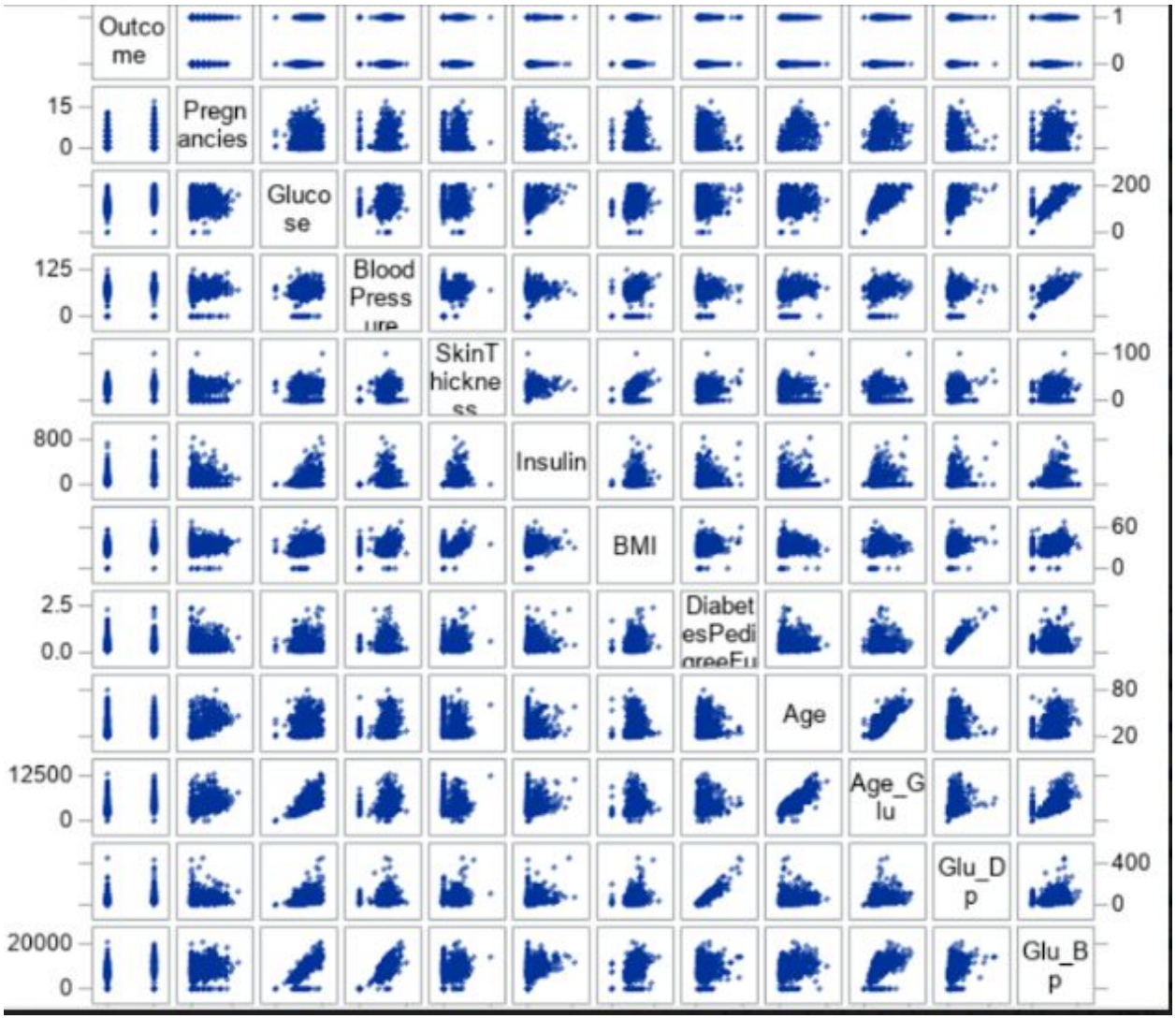

Distribution of Insulin by Outcome
Insulin
1000
800
600
400
200
0
0
1
Outcome
Distribution of SkinThickness by Outcome
SkinThickness
100
75
50
25
0
0
1
Outcome
Distribution of DiabetesPedigreeFunction by Outcome
DiabetesPedigreeFunction
2.5
2
1.5
1
0.5
0
0
1
Outcome
Distribution of Glucose by Outcome
Glucose
200
150
100
50
0
0
1
Distribution of Pregnancies by Outcome
Pregnancies
20
15
10
5
0
0
1
Outcome
Distribution of Age_Glu by Outcome
Age_Glu
15000
12500
10000
7500
5000
2500
0
0
1
Outcome


Distribution of BloodPressure by Outcome
BloodPressure
125
100
75
50
25
0
0
1
Outcome


Distribution of BMI by Outcome
BMI
80
60
40
20
0
0
1

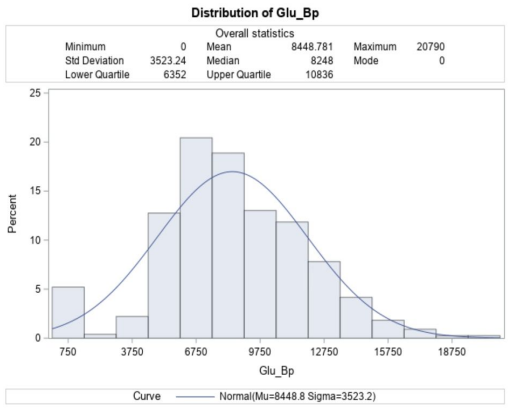

Distribution of Age by Outcome
Age
100
80
60
40
20
0
1
Outcome

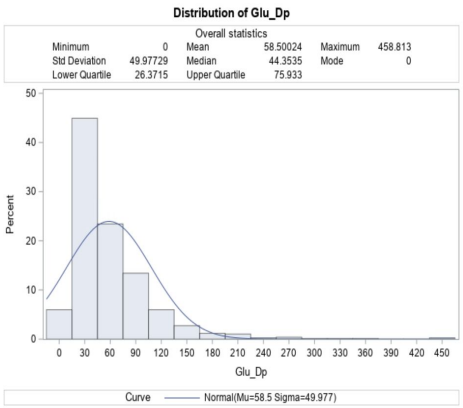
A2. Scatter Plots

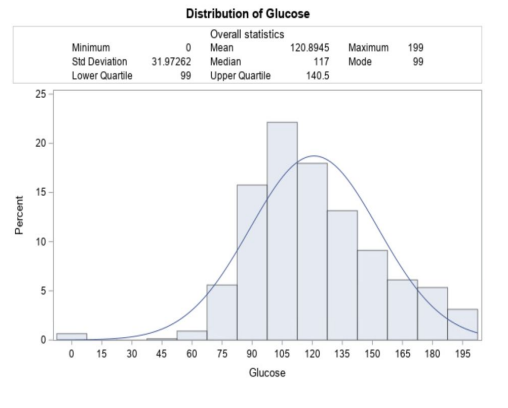


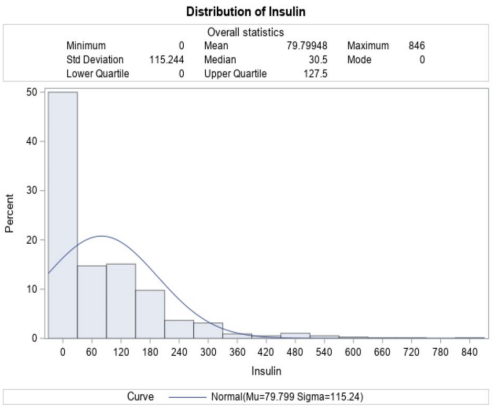
A3. Histograms

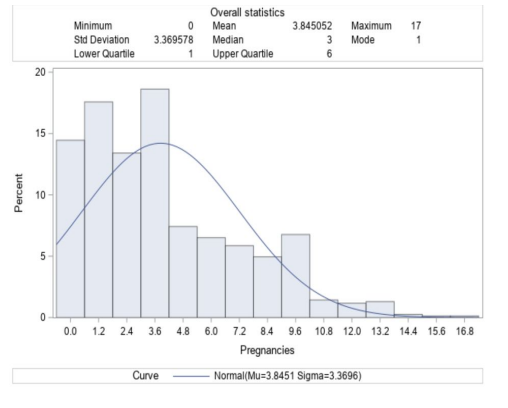
a. Distributions of Dependent Variables and Interaction Variables

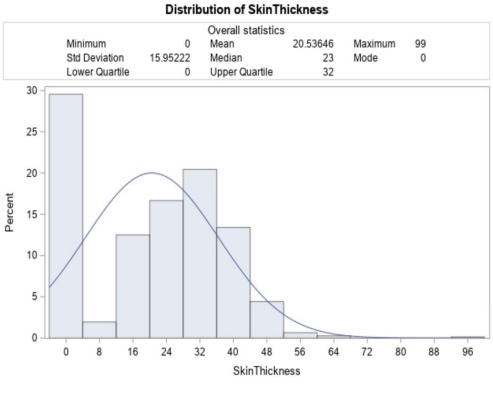

Distribution of Glu_Bp
Overall statistics
Minimum 0 Mean 8448.781 Maximum 20790
Std Deviation 3523.24 Median 8248 Mode 0
Lower Quartile 6352 Upper Quartile 10836
Percent
Glu_Bp
Curve Normal(Mu=8448.8 Sigma=3523.2)

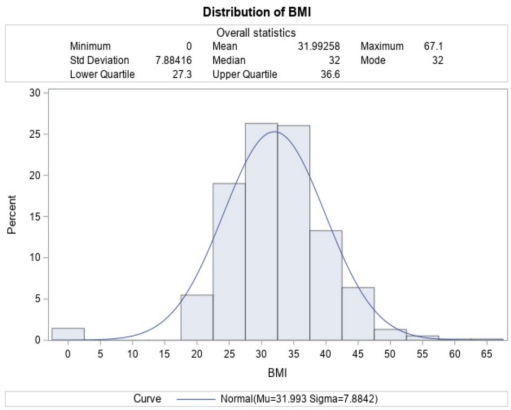

Distribution of Glu_Dp
Overall statistics
Minimum 0 Mean 58.50024 Maximum 458.813
Std Deviation 49.97729 Median 44.3535 Mode 0
Lower Quartile 26.3715 Upper Quartile 75.933
Percent
Glu_Dp
Curve Normal(Mu=58.5 Sigma=49.977)

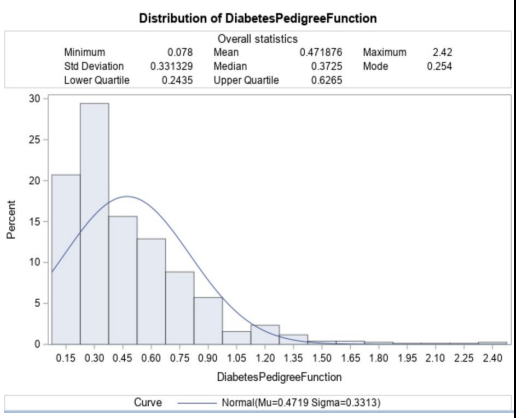

Distribution of Glucose
Overall statistics
Minimum 0 Mean 120.8945 Maximum 199
Std Deviation 31.97262 Median 117 Mode 99
Lower Quartile 99 Upper Quartile 140.5
Percent
Glucose

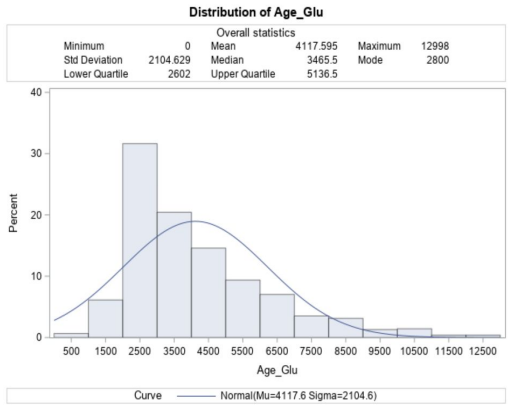

Distribution of Insulin
Overall statistics
Minimum 0 Mean 79.79948 Maximum 846
Std Deviation 115.244 Median 30.5 Mode 0
Lower Quartile 0 Upper Quartile 127.5
Percent
Insulin
Curve Normal(Mu=79.799 Sigma=115.24)

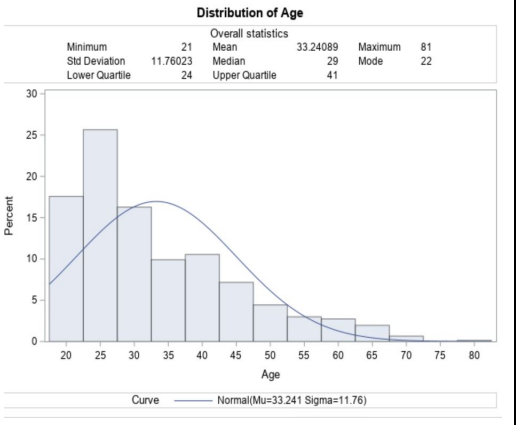

Overall statistics
Minimum 0 Mean 3.845052 Maximum 17
Std Deviation 3.369578 Median 3 Mode 1
Lower Quartile 1 Upper Quartile 6
Percent
Pregnancies
Curve Normal(Mu=3.8451 Sigma=3.3696)

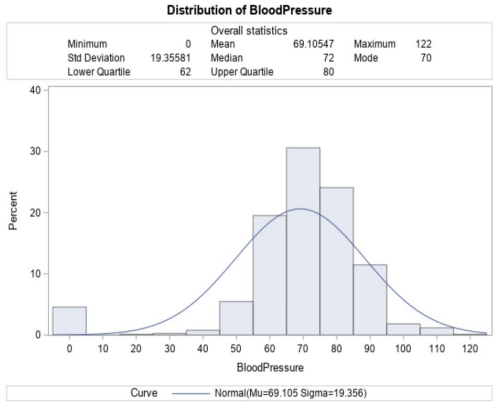

Distribution of SkinThickness
Overall statistics
Minimum 0 Mean 20.53646 Maximum 99
Std Deviation 15.95222 Median 23 Mode 0
Lower Quartile 0 Upper Quartile 32
Percent
SkinThickness











A4. Full Model Output

a. Analysis of Maximum Likelihood

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-15.0886	2.4189	38.9109	<.0001	
Pregnancies	1	0.1246	0.0322	14.9489	0.0001	0.2315
Glucose	1	0.0882	0.0187	22.2102	<.0001	1.5554
BloodPressure	1	0.0237	0.0288	0.6795	0.4098	0.2533
SkinThickness	1	0.00152	0.00687	0.0490	0.8248	0.0134
Insulin	1	-0.00128	0.000883	2.1101	0.1463	-0.0815
BMI	1	0.0892	0.0154	33.3711	<.0001	0.3876
DiabetesPedigreeFunc	1	4.2193	1.1416	13.6601	0.0002	0.7708
Age	1	0.0780	0.0390	4.0036	0.0454	0.5056
Age_Glu	1	-0.00049	0.000284	2.9943	0.0836	-0.5702
Glu_Dp	1	-0.0252	0.00829	9.2548	0.0023	-0.6948
Glu_Bp	1	-0.00030	0.000225	1.7767	0.1826	-0.5820

b. Goodness of Fit Statistics

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	995.484	733.923
SC	1000.128	789.648
-2 Log L	993.484	709.923

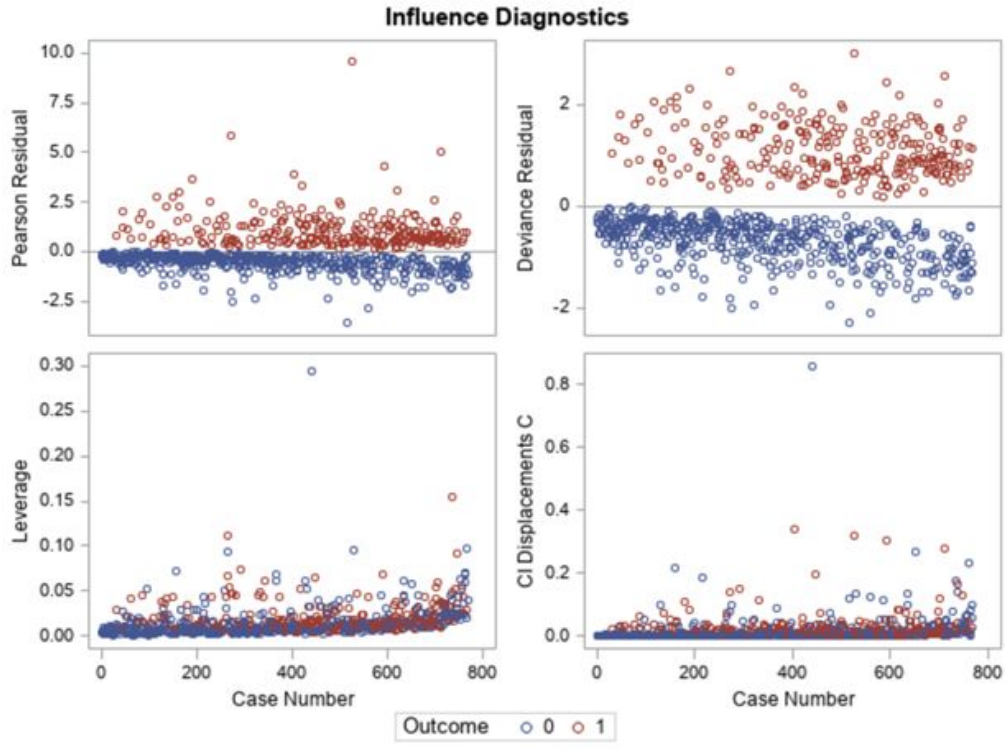
R-Square	0.3087	Max-rescaled R-Square	0.425

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	283.5612	11	<.0001
Score	235.1453	11	<.0001
Wald	168.7843	11	<.0001

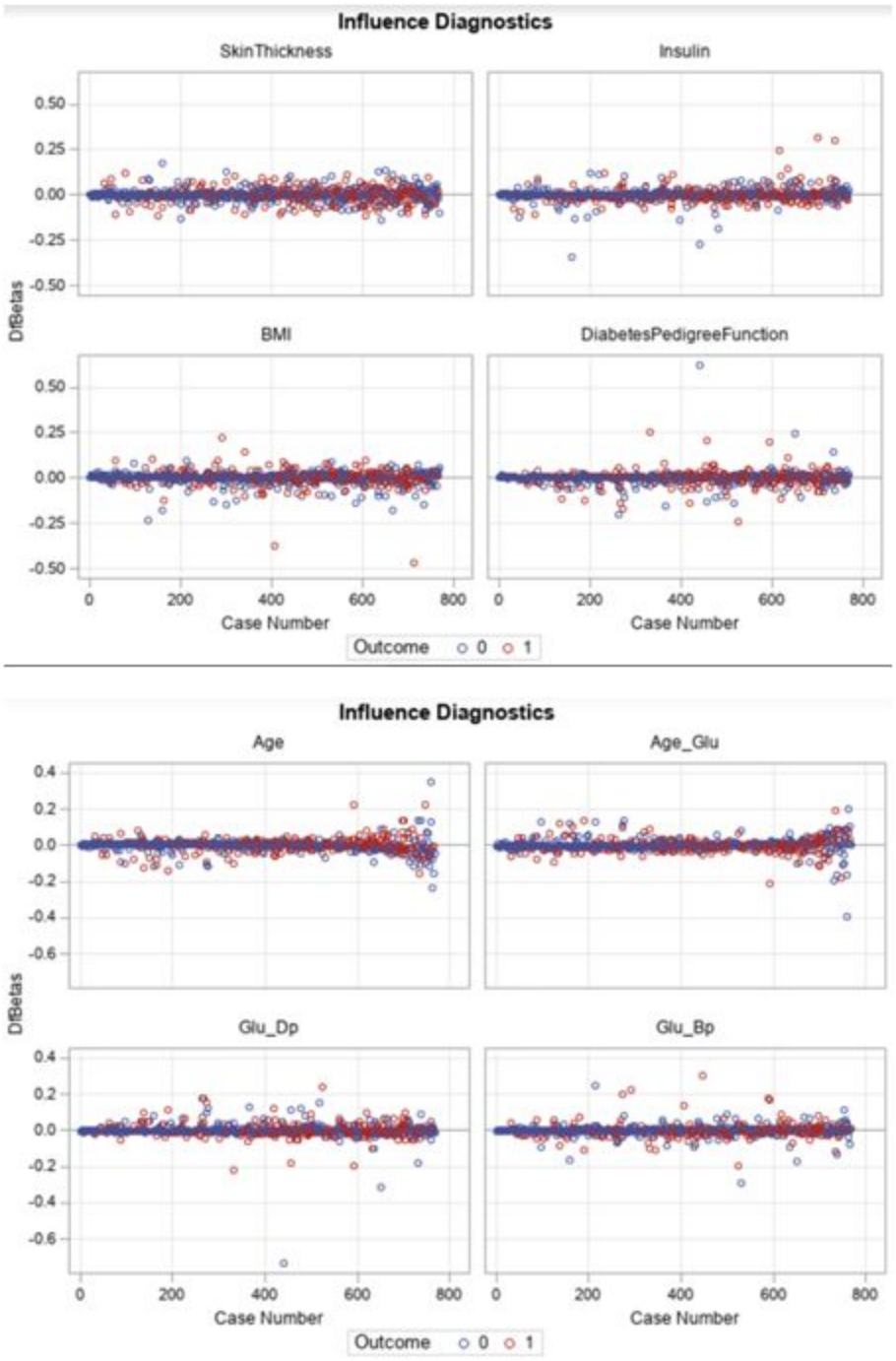
c. Estimated Correlation Matrix

Estimated Correlation Matrix												
Parameter	Intercept	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Age_Glu	Glu_Dp	Glu_Bp
Intercept	1.0000	-0.0342	-0.9643	-0.7138	-0.0444	0.1277	-0.1665	-0.3378	-0.3711	0.3748	0.3299	0.7204
Pregnancies	-0.0342	1.0000	0.0330	0.0520	0.0007	0.0213	0.0841	0.0975	-0.2280	0.1201	-0.0812	-0.0675
Glucose	-0.9643	0.0330	1.0000	0.7266	0.0784	-0.1570	-0.0147	0.3070	0.3545	-0.3830	-0.3108	-0.7504
BloodPressure	-0.7138	0.0520	0.7266	1.0000	-0.0040	-0.0689	-0.0299	-0.0057	-0.2481	0.2403	0.0177	-0.9813
SkinThickness	-0.0444	0.0007	0.0784	-0.0040	1.0000	-0.4242	-0.2908	-0.0340	0.0722	-0.0515	0.0139	-0.0269
Insulin	0.1277	0.0213	-0.1570	-0.0689	-0.4242	1.0000	0.0127	0.0176	-0.0225	0.0388	-0.0431	0.0839
BMI	-0.1665	0.0841	-0.0147	-0.0299	-0.2908	0.0127	1.0000	0.0861	-0.0480	0.0698	-0.0858	-0.0068
DiabetesPedigreeFunction	-0.3378	0.0975	0.3070	-0.0057	-0.0340	0.0176	0.0861	1.0000	0.0895	-0.1085	-0.9706	0.0005
Age	-0.3711	-0.2280	0.3545	-0.2481	0.0722	-0.0225	-0.0480	0.0895	1.0000	-0.9719	-0.1046	0.2384
Age_Glu	0.3748	0.1201	-0.3830	0.2403	-0.0515	0.0388	0.0698	-0.1085	-0.9719	1.0000	0.1231	-0.2386
Glu_Dp	0.3299	-0.0812	-0.3108	0.0177	0.0139	-0.0431	-0.0858	-0.9706	-0.1046	0.1231	1.0000	-0.0127
Glu_Bp	0.7204	-0.0675	-0.7504	-0.9813	-0.0269	0.0839	-0.0068	0.0005	0.2384	-0.2386	-0.0127	1.0000

d. Outliers



e. Influential Points



A5. Selection Methods

a. Selection methods used on Full Model

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	802.902	586.342
SC	807.324	612.872
-2 Log L	800.902	574.342

R-Square	0.3082	Max-rescaled R-Square	0.4232

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	226.5596	5	<.0001
Score	193.5665	5	<.0001
Wald	138.4479	5	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
9.1147	6	0.1672

Summary of Stepwise Selection							
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Glucose		1	1	162.8863		<.0001
2	BMI		1	2	29.3631		<.0001
3	Pregnancies		1	3	21.1527		<.0001
4	DiabetesPedigreeFunc		1	4	13.8437		0.0002
5	Glu_Bp		1	5	5.2757		0.0216

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-10.2161	0.8616	140.5998	<.0001
Pregnancies	1	0.1662	0.0337	24.2734	<.0001
Glucose	1	0.0499	0.00583	73.2567	<.0001
BMI	1	0.0940	0.0166	32.0840	<.0001
DiabetesPedigreeFunc	1	1.2635	0.3441	13.4812	0.0002
Glu_Bp	1	-0.00011	0.000048	5.1542	0.0232

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	802.902	586.342
SC	807.324	612.872
-2 Log L	800.902	574.342

R-Square	0.3082	Max-rescaled R-Square	0.4232

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	226.5596	5	<.0001
Score	193.5665	5	<.0001
Wald	138.4479	5	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
9.1147	6	0.1672

Summary of Forward Selection					
Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	Glucose	1	1	147.7589	<.0001
2	BMI	1	2	25.8285	<.0001
3	Pregnancies	1	3	19.8359	<.0001
4	DiabetesPedigreeFunc	1	4	9.1635	0.0025
5	Glu_Bp	1	5	5.0597	0.0245

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-9.0938	0.7818	135.3126	<.0001
Pregnancies	1	0.1532	0.0323	22.5024	<.0001
Glucose	1	0.0450	0.00548	67.3403	<.0001
BMI	1	0.0832	0.0156	28.3037	<.0001
DiabetesPedigreeFunc	1	0.9775	0.3268	8.9475	0.0028
Glu_Bp	1	-0.00010	0.000046	4.9418	0.0262

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	802.902	586.342
SC	807.324	612.872
-2 Log L	800.902	574.342

R-Square	0.3082	Max-rescaled R-Square	0.4232

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	226.5596	5	<.0001
Score	193.5665	5	<.0001
Wald	138.4479	5	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
9.1147	6	0.1672

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	SkinThickness	1	10	0.3274	0.5672
2	BloodPressure	1	9	0.7838	0.3760
3	Insulin	1	8	0.8226	0.3644
4	Glu_Dp	1	7	1.8447	0.1744
5	Age_Glu	1	6	2.7453	0.0975
6	Age	1	5	2.5214	0.1123

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-9.0938	0.7818	135.3126	<.0001
Pregnancies	1	0.1532	0.0323	22.5024	<.0001
Glucose	1	0.0450	0.00548	67.3403	<.0001
BMI	1	0.0832	0.0156	28.3037	<.0001
DiabetesPedigreeFunc	1	0.9775	0.3268	8.9475	0.0028
Glu_Bp	1	-0.00010	0.000046	4.9418	0.0262

b. Selection Methods used on Fitted Model after removing Outliers

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	794.935	550.595
SC	799.349	577.076
-2 Log L	792.935	538.595

R-Square	0.3409	Max-rescaled R-Square	0.4687

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	254.3398	5	<.0001
Score	212.1500	5	<.0001
Wald	144.3911	5	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
6.5255	6	0.3670

Summary of Stepwise Selection							
	Effect			Number	Score	Wald	
Step	Entered	Removed	DF	In	Chi-Square	Chi-Square	Pr > ChiSq
1	Glucose		1	1	162.8863		<.0001
2	BMI		1	2	29.3631		<.0001
3	Pregnancies		1	3	21.1527		<.0001
4	DiabetesPedigreeFunc		1	4	13.8437		0.0002
5	Glu_Bp		1	5	5.2757		0.0216

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-10.2161	0.8616	140.5998	<.0001
Pregnancies	1	0.1662	0.0337	24.2734	<.0001
Glucose	1	0.0499	0.00583	73.2567	<.0001
BMI	1	0.0940	0.0166	32.0840	<.0001
DiabetesPedigreeFunc	1	1.2635	0.3441	13.4812	0.0002
Glu_Bp	1	-0.00011	0.000048	5.1542	0.0232

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	794.935	550.595
SC	799.349	577.076
-2 Log L	792.935	538.595

R-Square	0.3409	Max-rescaled R-Square	0.4687

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	254.3398	5	<.0001
Score	212.1500	5	<.0001
Wald	144.3911	5	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
6.5255	6	0.3670

Summary of Forward Selection					
Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	Glucose	1	1	162.8863	<.0001
2	BMI	1	2	29.3631	<.0001
3	Pregnancies	1	3	21.1527	<.0001
4	DiabetesPedigreeFunc	1	4	13.8437	0.0002
5	Glu_Bp	1	5	5.2757	0.0216

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-10.2161	0.8616	140.5998	<.0001
Pregnancies	1	0.1662	0.0337	24.2734	<.0001
Glucose	1	0.0499	0.00583	73.2567	<.0001
BMI	1	0.0940	0.0166	32.0840	<.0001
DiabetesPedigreeFunc	1	1.2635	0.3441	13.4812	0.0002
Glu_Bp	1	-0.00011	0.000048	5.1542	0.0232

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	794.935	550.595
SC	799.349	577.076
-2 Log L	792.935	538.595

R-Square	0.3409	Max-rescaled R-Square	0.4687

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	254.3398	5	<.0001
Score	212.1500	5	<.0001
Wald	144.3911	5	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
6.5255	6	0.3670

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	SkinThickness	1	10	0.0750	0.7842
2	BloodPressure	1	9	0.3259	0.5681
3	Glu_Dp	1	8	0.6776	0.4104
4	Insulin	1	7	1.5702	0.2102
5	Age_Glu	1	6	2.1868	0.1392
6	Age	1	5	1.6053	0.2052

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-10.2161	0.8616	140.5998	<.0001
Pregnancies	1	0.1662	0.0337	24.2734	<.0001
Glucose	1	0.0499	0.00583	73.2567	<.0001
BMI	1	0.0940	0.0166	32.0840	<.0001
DiabetesPedigreeFunc	1	1.2635	0.3441	13.4812	0.0002
Glu_Bp	1	-0.00011	0.000048	5.1542	0.0232

A6. Fitted Model Output

a. Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-9.0938	0.7818	135.3126	<.0001	
Pregnancies	1	0.1532	0.0323	22.5024	<.0001	0.2789
Glucose	1	0.0450	0.00548	67.3403	<.0001	0.7809
BMI	1	0.0832	0.0156	28.3037	<.0001	0.3721
DiabetesPedigreeFunc	1	0.9775	0.3268	8.9475	0.0028	0.1791
Glu_Bp	1	-0.00010	0.000046	4.9418	0.0262	-0.1984

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Pregnancies	1.166	1.094	1.242
Glucose	1.046	1.035	1.057
BMI	1.087	1.054	1.121
DiabetesPedigreeFunc	2.658	1.401	5.043
Glu_Bp	1.000	1.000	1.000

Goodness of Fit statistics

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	802.902	586.342
SC	807.324	612.872
-2 Log L	800.902	574.342

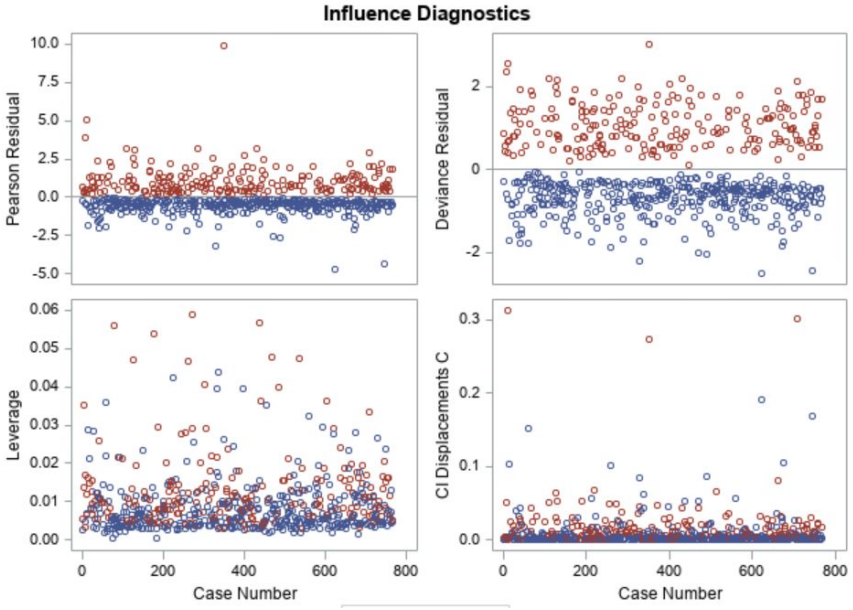
R-Square	0.3082	Max-rescaled R-Square	0.4232

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	226.5596	5	<.0001
Score	193.5665	5	<.0001
Wald	138.4479	5	<.0001

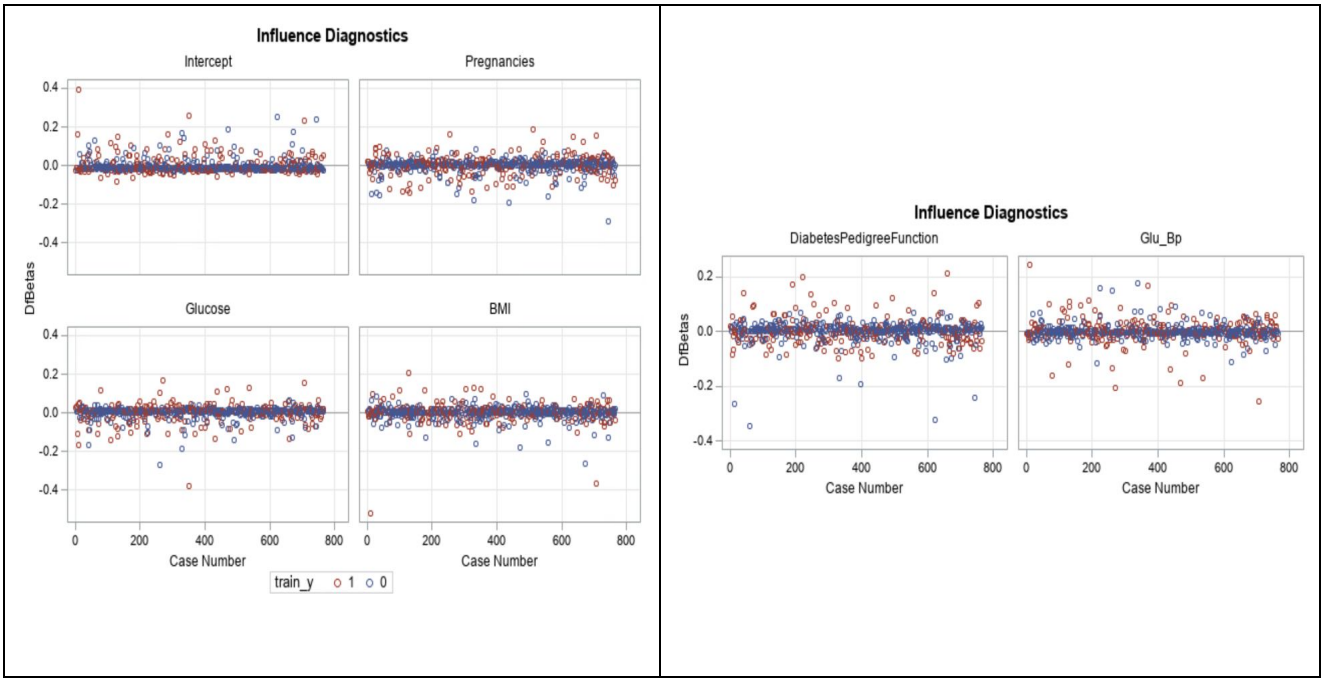
b. Estimated Correlation Matrix

Estimated Correlation Matrix						
Parameter	Intercept	Pregnancies	Glucose	BMI	DiabetesPedigreeFunction	Glu_Bp
Intercept	1.0000	-0.3105	-0.6497	-0.6952	-0.2467	0.2944
Pregnancies	-0.3105	1.0000	0.1287	0.1648	0.0738	-0.1884
Glucose	-0.6497	0.1287	1.0000	0.1551	0.0592	-0.7029
BMI	-0.6952	0.1648	0.1551	1.0000	-0.0117	-0.2487
DiabetesPedigreeFunction	-0.2467	0.0738	0.0592	-0.0117	1.0000	-0.0175
Glu_Bp	0.2944	-0.1884	-0.7029	-0.2487	-0.0175	1.0000

c. Outliers



e. Influential points



A7. Final Model Output

a. Analysis of Maximum Likelihood Estimates

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	254.3398	5	<.0001
Score	212.1500	5	<.0001
Wald	144.3911	5	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-10.2161	0.8616	140.5998	<.0001	
Pregnancies	1	0.1662	0.0337	24.2734	<.0001	0.3014
Glucose	1	0.0499	0.00583	73.2567	<.0001	0.8544
BMI	1	0.0940	0.0166	32.0840	<.0001	0.4157
DiabetesPedigreeFunc	1	1.2635	0.3441	13.4812	0.0002	0.2297
Glu_Bp	1	-0.00011	0.000048	5.1542	0.0232	-0.2091

b. Goodness of fit and Statistics

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	794.935	550.595
SC	799.349	577.076
-2 Log L	792.935	538.595

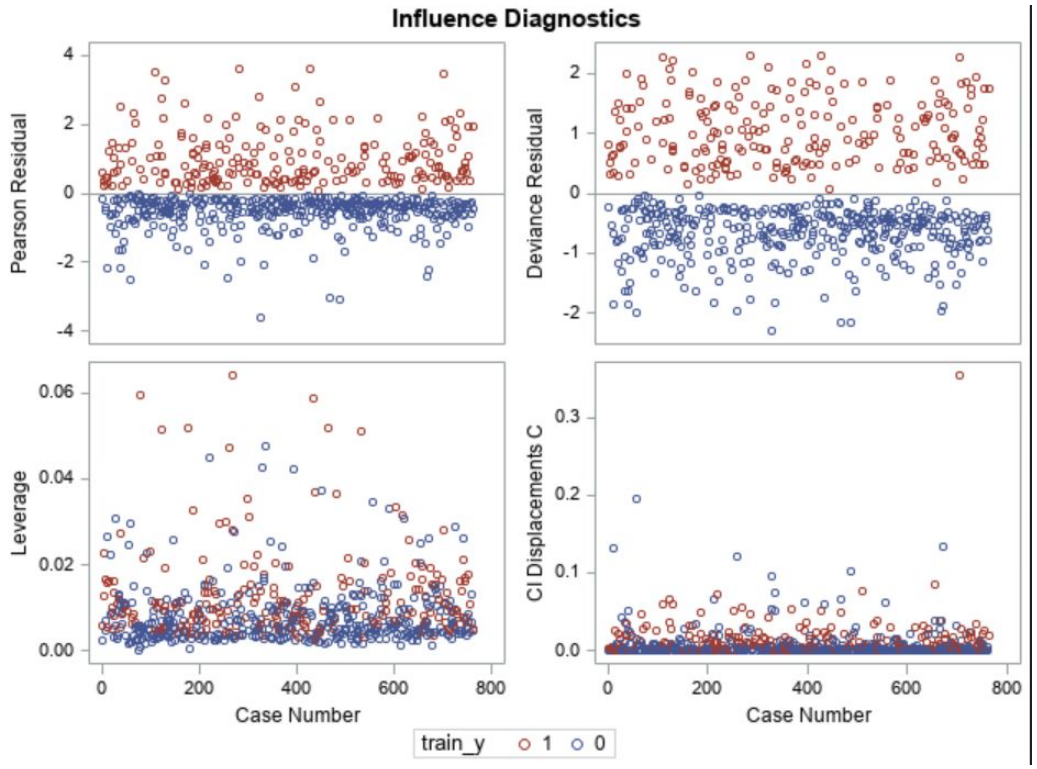
R-Square	0.3409	Max-rescaled R-Square	0.4687

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	254.3398	5	<.0001
Score	212.1500	5	<.0001
Wald	144.3911	5	<.0001

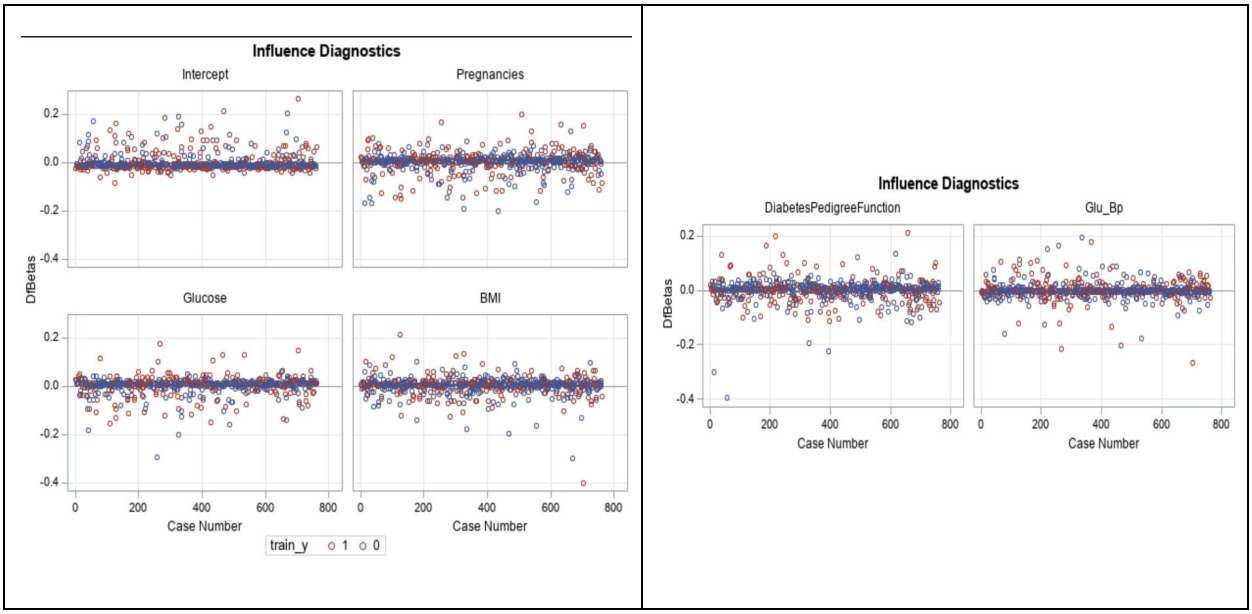
c. Correlation Matrix

Estimated Correlation Matrix						
Parameter	Intercept	Pregnancies	Glucose	BMI	DiabetesPedigreeFunction	Glu_Bp
Intercept	1.0000	-0.3295	-0.6655	-0.7056	-0.2957	0.2920
Pregnancies	-0.3295	1.0000	0.1482	0.1807	0.0975	-0.1900
Glucose	-0.6655	0.1482	1.0000	0.1821	0.1035	-0.6981
BMI	-0.7056	0.1807	0.1821	1.0000	0.0226	-0.2491
DiabetesPedigreeFunction	-0.2957	0.0975	0.1035	0.0226	1.0000	-0.0271
Glu_Bp	0.2920	-0.1900	-0.6981	-0.2491	-0.0271	1.0000

d. Outliers



e. Influential Points



A8. Predictions

a. Prediction and Confidence Intervals

Obs	Selected	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	Age_Glu	Glu_Dp	Glu_Bp
1	0	1	89	66	23	94	28.1	0.167	21	0	1869	14.863	5874
2	0	10	115	0	0	0	35.3	0.134	29	0	3335	15.410	0
3	0	5	166	72	19	175	25.8	0.587	51	1	8466	97.442	11952
4	0	7	100	0	0	0	30	0.484	32	1	3200	48.400	0
5	0	7	107	74	0	0	29.6	0.254	31	1	3317	27.178	7918
6	0	1	115	70	30	96	34.6	0.529	32	1	3680	60.835	8050
7	0	5	109	75	26	0	36	0.546	60	0	6540	59.514	8175
8	0	0	180	66	39	0	42	1.893	25	1	4500	340.740	11880
9	0	2	71	70	27	0	28	0.586	22	0	1562	41.606	4970
10	0	1	101	50	15	36	24.2	0.526	26	0	2626	53.126	5050
11	0	5	88	66	21	23	24.4	0.342	30	0	2640	30.096	5808
12	0	7	150	66	42	342	34.7	0.718	42	0	6300	107.700	9900
13	0	7	187	68	39	304	37.7	0.254	41	1	7667	47.498	12716
14	0	0	105	64	41	142	41.5	0.173	22	0	2310	18.165	6720
15	0	4	146	85	27	100	28.9	0.189	27	0	3942	27.594	12410
16	0	13	126	90	0	0	43.4	0.583	42	1	5292	73.458	11340
17	0	7	83	78	26	71	29.3	0.767	36	0	2988	63.661	6474

Outcome	Age_Glu	Glu_Dp	Glu_Bp	train_y	_FROM_	_INTO_	IP_0	IP_1	_LEVEL_	phat	lcl	ucl	pred_y	threshold
0	1869	14.863	5874	.	.	0	0.96752	0.03248	1	0.03248	0.01913	0.05464	0	0.3
0	3335	15.410	0	.	.	1	0.33727	0.66273	1	0.66273	0.42580	0.83889	1	0.3
1	8466	97.442	11952	.	.	1	0.31707	0.68293	1	0.68293	0.57502	0.77419	1	0.3
1	3200	48.400	0	.	.	0	0.65206	0.34794	1	0.34794	0.20340	0.52722	1	0.3
1	3317	27.178	7918	.	.	0	0.81303	0.18697	1	0.18697	0.13737	0.24929	0	0.3
1	3680	60.835	8050	.	.	0	0.77983	0.22017	1	0.22017	0.17176	0.27765	0	0.3
0	6540	59.514	8175	.	.	0	0.68133	0.31867	1	0.31867	0.26173	0.38160	1	0.3
1	4500	340.740	11880	.	.	1	0.02154	0.97846	1	0.97846	0.93357	0.99323	1	0.3
0	1562	41.606	4970	.	.	0	0.97091	0.02909	1	0.02909	0.01675	0.05005	0	0.3
0	2626	53.126	5050	.	.	0	0.93202	0.06798	1	0.06798	0.04314	0.10554	0	0.3
0	2640	30.096	5808	.	.	0	0.94778	0.05222	1	0.05222	0.03268	0.08246	0	0.3
0	6300	107.700	9900	.	.	1	0.17848	0.82152	1	0.82152	0.74939	0.87632	1	0.3
1	7667	47.498	12716	.	.	1	0.05937	0.94063	1	0.94063	0.89396	0.96751	1	0.3
0	2310	18.165	6720	.	.	0	0.83011	0.16989	1	0.16989	0.10966	0.25377	0	0.3
0	3942	27.594	12410	.	.	0	0.65901	0.34099	1	0.34099	0.25650	0.43695	1	0.3
1	5292	73.458	11340	.	.	1	0.13990	0.86010	1	0.86010	0.74673	0.92764	1	0.3
0	2988	63.661	6474	.	.	0	0.86884	0.13116	1	0.13116	0.08406	0.19892	0	0.3

b. Classification Table

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	Pos Pred	Neg Pred
0.200	190	230	164	26	68.9	88.0	58.4	53.7	89.8
0.250	180	261	133	36	72.3	83.3	66.2	57.5	87.9
0.300	175	293	101	41	76.7	81.0	74.4	63.4	87.7
0.350	161	315	79	55	78.0	74.5	79.9	67.1	85.1
0.400	151	325	69	65	78.0	69.9	82.5	68.6	83.3
0.450	142	333	61	74	77.9	65.7	84.5	70.0	81.8
0.500	130	346	48	86	78.0	60.2	87.8	73.0	80.1
0.550	124	352	42	92	78.0	57.4	89.3	74.7	79.3
0.600	116	365	29	100	78.9	53.7	92.6	80.0	78.5

c. True Positives and True Negatives of Outcome

Performing Prediction

The FREQ Procedure

Frequency	Table of Outcome by pred_y		
	pred_y		
Outcome	0	1	Total
0	68	36	104
1	8	41	49
Total	76	77	153

d. Receiver Operating Characteristic (ROC) curve