1. Introduction:

Diabetes is one of the concerning diseases of all today. As part of this research, Pima Indians Dataset is being analyzed to get into the details of different attributes taking part for testing positive with the Diabetes.

Research Goal:
Include an investigation through the level of prediction of diabetes presence using this dataset in accordance with the given attributes. During this process, build a relationship between how other factors affect having diabetes in Pima Indians.

Dataset Intro:
768 entries of Pima Indians Diabetes Dataset have been given. A binomial class variable called diabetes is given with 1 as the diagnosis of diabetes. 8 other numerical variables giving the number of times pregnant, concentration of glucose at two hours in an oral glucose test, blood pressure in mm Hg, skin fold thickness of triceps in mm, 2-hour serum insulin in mm U/ml, obesity level, pedigree, age in years for aiding the analyzing process. There is no missing data for this.

2. Methods:

Firstly, the dataset given went through the check of missing values. A linear model of each variable with each other variables have been applied to check how each attribute is being related to each other.
Secondly, Pima Indians Dataset includes a binomial outcome if the instance is diabetic by depending on few other dependent attributes. For getting an analysis using logistic regression, probabilities are transformed to a real line modelling the log odds of the probability. A positive value gives that its more likely to happen while a negative is contra. So, GLM is fitted for this dataset to maximize the likelihood of the attributes. Deviance residuals are considered. Thus, as Logistic regression has the characteristics mentioned before this method that is needed for this dataset, it is used as part of this research for getting a basic relationship dependance between variables.
Lastly, Receiving Operating Characteristics curve is the next method that is used to make prediction outcomes. During this process, a threshold must be chosen for which above that will be an outcome of 1 and below is 0. Sensitivity and Specificity will be included in this curve. In between, a table is constructed which will give the outcome of diabetes using the predicted probability with the threshold.

3. Results:

Firstly, Considering the GLM model, considering an alpha value of 0.05, intercept, Pregnancy, Glucose and BMI of instances are considered significant predictors in determining the diabetes in Pima Indians while triceps, insulin and age are not. In

addition to it, A negative relationship is seen between the outcome and pressure, outcome and insulin that says it is less likely for diabetes in Pima Indians.
Further, it is evident that on deduction of (767-759 = 8) independent instances, the residual deviance has been decreased by 270.03. There is not much deviance in the residuals.
Secondly, taking the log odds of probabilities into consideration, being pregnant or old, having high glucose concentration, BMI, triceps skin thickness, is more likely to prone to diabetes in Pima Indians while blood pressure and insulin is less likely.
Last but not the least, an AUC of 0.83 using the ROC curve tells that this is not the perfect model but can be considered. From the figure 1, a threshold of 0.2 has been taken as part of research of the Pima Indian Diabetes Dataset. The reason for this will be that considering this diabetes dataset minimization of the prediction of a negative test report being a positive in this dataset as diagnosis of diabetes is necessary from the early stages. While making the table (figure 2) using the predicted probabilities, this can be checked. The aim is to always have a maximized true positive rate against false positive rate based on the data set. For this research, it is clearly shown from the figure 1 that the true positive rate is 0.89 while the false positive rate is 0.44 at an alpha value of 0.2.

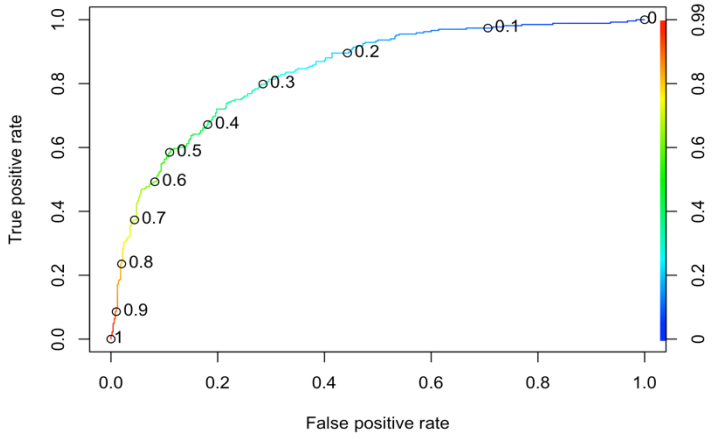

Figure 1

	neg	pos
No	279	28
Yes	221	240

Figure 2

Further investigation into AIC can be informative in order to see how the model is the best fit for the dataset. Well, here the AIC is a bit higher.
A little more interpretation on deviance residuals can be done here.
Confusion matrix can be interpreted and applied more for this dataset.

4. Conclusion:

To conclude, for, Pima Indians Diabetes Dataset, how many years old people are, pregnancy in women and weight of a person determine the presence of diabetes. Further, the Receiving Operating Characteristics Curve does give some sense in output but the output from the ROC curve is neither fully satisfied nor maximized with true positive rate.