A Study To Predict Diabetes In Women

Vikas Reddy Bodireddy, Preethi Bommneni, Naveen Kumar Kaparaju.

1.Introduction:

Diabetes is a chronic health condition that interferes with how food is broken down into glucose, which the body uses as energy. Serious health issues like renal failure, eyesight loss, heart attacks, and early death can result from elevated glucose levels.

A study was conducted to examine the influence of various factors, such as pregnancies, age, diabetes pedigree function, skin thickness, blood pressure, BMI, insulin, and glucose levels, on diabetes development. The findings aimed to enhance understanding of these characteristics' impact on the likelihood of developing diabetes.

2.Data Description:

This is a Pima Indian diabetes dataset from kaggle and was originally collected by National Institute of Diabetes in 1980 and is updated every 2 years. This is one of the most widely used dataset to analyze diabetes using machine learning algorithms. Native American Indians known as the Pima reside along the Salt and Gila Rivers in Southern Arizona. Using this dataset, we examine the factors to see if they could increase the risk of developing diabetes.

The dataset has 768 Observations and 9 Variables with 8 variables Pregnancy, Glucose, Blood Pressure (BP), Skin Thickness, Insulin, Body Mass Index(BMI), Diabetes Pedigree Function, and Age as predictor variables and its description is given in *Table i*.

Variables	Type	Description
Pregnancies	Numeric	Frequency of Pregnancy
Glucose	Numeric	Concentration of plasma Glucose(mg/dL)
Blood Pressure	Numeric	Person's Diastolic Blood Pressure (mm Hg)
Skin Thickness	Numeric	Tricep skinfold thickness (mm)
Insulin	Numeric	Two hour serum insulin (µU/ml)
BMI	Numeric	Body Mass Index or Body to mass Ratio (Kg/m^2)
DiabeticPedigreeFunction	Numeric	The likelihood of diabetes based on family history.
AGE	Numeric converted to Categorical	Age of person as per standards (Young Age 21-40 years, Middle Age 41-60 years, Elder and Wise Age 61- 85 years)
Outcome (Diabetes)	Categorical	Diabetes status of person (1- Positive & 0 Negative).

Table i

With one response variable Outcome contains a value of 1(Positive) for patients diagnosed with Type 2 diabetes and 0(Negative) for the person who is not diagnosed with diabetes. In our dataset, there are 268 patients who are diagnosed with positive diabetes and 500 patients with negative diabetes. Which is clearly mentioned in *Table ii*.

Outcome	Observations
Total	768
Positive(1)	268
Negative(0)	500

Table ii

It is important to note that the dataset contains numerous 0 values for insulin levels, which are likely missing or incomplete data points. Although this presents a practical implausibility, the small size of the dataset restricts the removal of these values as it would significantly reduce the available data.

The summary statistics of the variables is provided in *Table iii.*

Variables	Summary			
	Min	Max	Mean	Median
Pregnancy	0.00	17.00	3.84	3.00
Glucose	0.00	199.0	120.9	117.0
Blood Pressure	0.00	122.0	69.11	72.00
Skin Thickness	0.00	99.00	20.54	23.00
Insulin	0.00	846.0	79.8	30.5
BMI	0.00	67.10	31.99	32.00
Diabetes Pedigree Function	0.078	2.42	0.4719	0.3725
Age: n (%)	Young Age:574 (74.7%)	Middle Age : 167 (21.7%)	Elder & Wise : 27 (3.6%)	
Outcome: n (%)	Positive(1) : 268 (34.8%)	Negative (0) : 500 (65.2%)		

Table iii

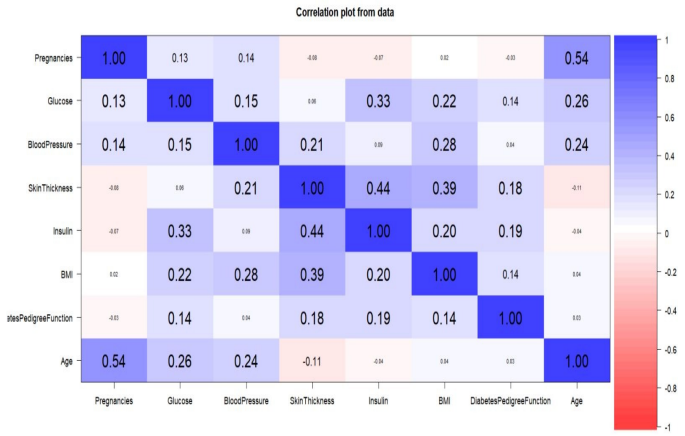
3. Methods:

3.1 Data Cleaning and Modification:

In the Pima Indian Diabetes dataset, the age variable, initially ranging from 0 to 85 years, has been categorized into three age groups: Young Age (20-40 years), Middle Age (41-60 years), and Elder & Wise Age (61-85 years) based on standard conventions. This categorization provides a more meaningful representation of age for analysis.Additionally, the Outcome variable, which originally had categorical values of 0 and 1 representing absence or presence of diabetes, has been transformed into more descriptive categories: "Negative" for individuals without diabetes and "Positive" for individuals diagnosed with diabetes. This change enhances the interpretability and clarity of the variable's meaning in the dataset.

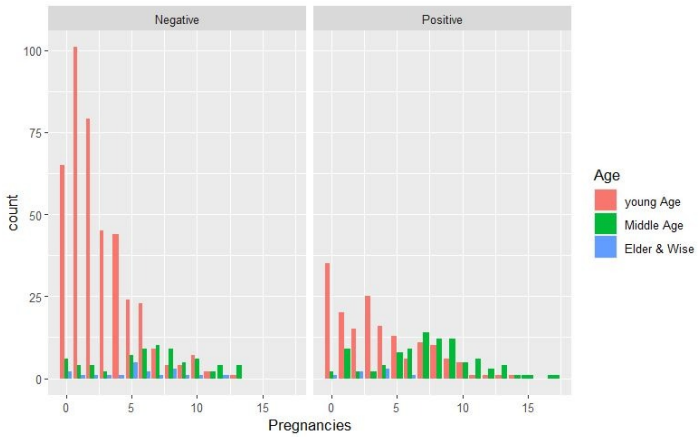
3.2 Graphical Analysis:

The correlation graph indicates a strong positive correlation between age and pregnancies, with a correlation coefficient greater than 0.5. This suggests that a change in age will likely result in a corresponding change in the number of pregnancies. See *Graph i.*



Graph i

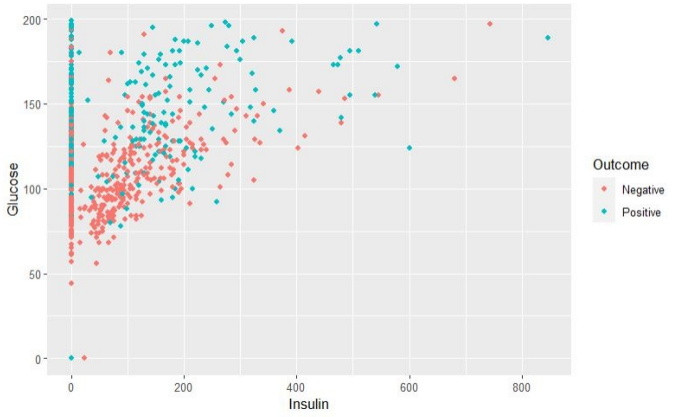
Based on the pregnancies, age, and outcome graph, it appears that middle-aged individuals (between 41 and 60 years old) have a higher likelihood of being diagnosed with diabetes, while younger individuals (between 20 and 40 years old) have a lower likelihood of developing the disease. See *Graph ii.*



Graph ii

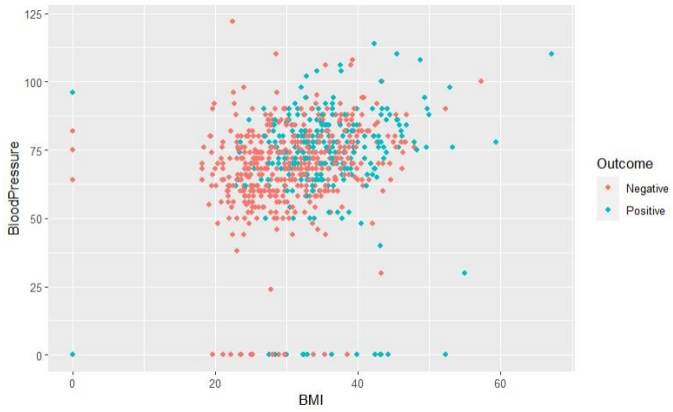
The graph depicting the relationship between glucose and insulin levels suggests that there is a higher

likelihood of developing diabetes when insulin levels decrease and glucose levels increase. However, there are some cases where increased levels of both glucose and insulin still result in a positive diagnosis of diabetes, so this relationship is not entirely conclusive. See *Graph iii.*



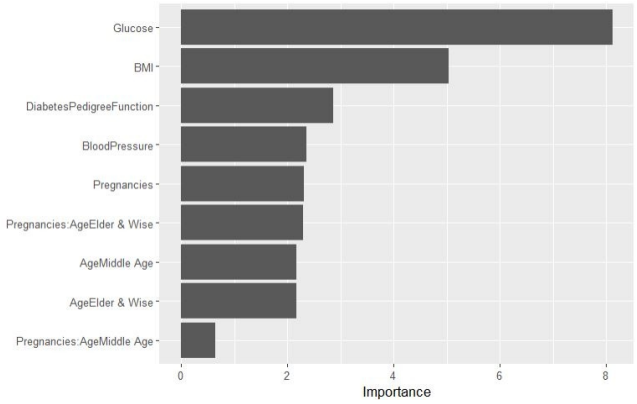
Graph iii

According to the scatter plot of Blood Pressure and BMI, most data points are between 20 and 45 for BMI and 50 to 125 for Blood Pressure. According to the plot, diabetes positive rises with BMI and Blood Pressure. However, a large proportion of patients in the same BMI and Blood Pressure range do not have diabetes, therefore the plot is not conclusive. Therefore, while the plot provides some insight into the association between BMI, Blood Pressure, and diabetes, it is not definite. See *Graph iv.*



Graph iv

The bar graph displays the significance of various variables in the final model. Glucose is the most significant variable, indicating that it has a strong predictive power for diabetes. Other variables such as BMI, DiabetesPedigreeFunction, Blood Pressure, Pregnancies, and interactions between pregnancies and age also show significance in predicting diabetes. See *Graph v.*



Graph v

3.3 Primary analysis:

Multiple logistic regression (MLR) is a statistical method used to model the relationship between a binary outcome variable and multiple predictor variables. The outcome variable in MLR is binary, taking the values of 0 and 1 or categorical, while the predictor variables can be continuous, categorical or binary. The goal of MLR is to identify the relationship between the outcome variable and the predictor variables while controlling for the effects of confounding variables.

The output of an MLR model is a logit equation that estimates the probability of the outcome variable given the predictor variables. MLR is widely used in various fields, including healthcare, social sciences, and business, for predicting and explaining binary outcomes such as disease diagnosis, customer behavior, and election results. It is a powerful statistical tool that enables researchers to identify significant predictors of a binary outcome and make

informed decisions based on the model's predictions.The output is a logit equation used to predict the reliability of the model.

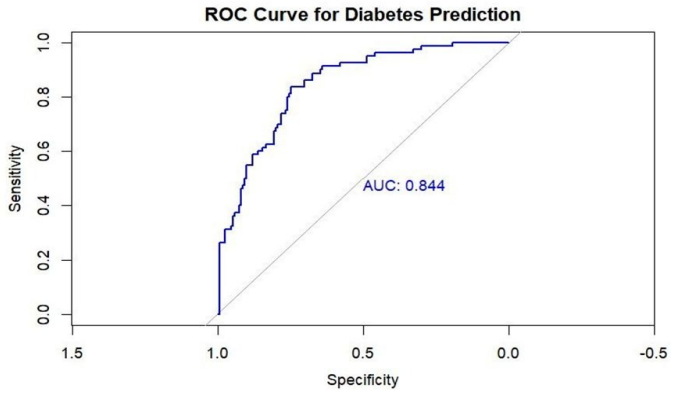
The initial analysis of the model that included all eight predictor variables, namely Pregnancies, Glucose, BMI, Age, Insulin, SkinThickness, Diabetes Pedigree Function, and Blood Pressure, concluded that Skin Thickness and Insulin Levels were highly insignificant in determining the chances of getting Diabetes. However, Age showed a nearly significant relationship with a value of 0.11. To select the best possible model for the dataset, a variable selection criterion called Akaike Information Criteria (AIC) was employed. AIC is a statistical method that compares different models and selects the one with the best balance between model complexity and goodness of fit. By using AIC, we aimed to determine the most significant predictors of Diabetes while minimizing the risk of overfitting.

3.4 Secondary Analysis:

Initially, we used multiple logistic regression to identify the key predictors of diabetes. However, we found that two of the predictor variables, skin thickness(mm) and insulin levels((*μ*u/mL), were highly insignificant and cannot explain if a person is prone to diabetes or not. Therefore, we performed a variable selection technique called Akaike Information Criteria (AIC) to identify the best model for predicting diabetes.

The AIC analysis helped us to determine the most significant predictors, including pregnancies, age, glucose, BMI, diabetes pedigree function, and blood pressure. However, we observed a correlation between the predictor variables of age and pregnancies. To address this issue, we performed an interaction analysis to identify the effect of the interaction between age and pregnancies. We found that the interaction between pregnancies and elderly persons was significant, while the middle-aged group did not have a significant interaction effect.

This allowed us to create a final and reliable model that can accurately predict the chances of getting diabetes, with an accuracy of 78%. The model has a pretty good ROC(Receiver Operating Characteristic) curve and an AUC value of 84.4 % as seen in the graph. Our findings demonstrate the importance of carefully selecting and analyzing predictor variables to improve the accuracy of predictive models**.** See *Graph vi*



Graph vi

3.5 Other Models:

As there is always a possibility that one model outperforms another model, we have tried to check if the Random Forests Model and XG boost model can provide a better output. When the full model containing all the eight predictor variables is run on the both the algorithms Random forests model gave an accuracy of 78% and XG boost gave an accuracy of 73 %. But after considering Interaction effect between Pregnancies and Age The accuracy of the random Forests model has increased to 80% and That of XG boost increased to 74%. Which provides evidence that the random Forests model is also a best learning algorithm to get a precise output for this dataset. While XG boost is not reliable to work on this dataset.

4. Results:

With a good increase in the accuracy of the model with the selected parameters, pregnancies, age, glucose, BMI, diabetes pedigree function, and blood pressure and the interaction between Pregnancies and Age we get the final model as logit equation.

The logit Equation of the model is

$$log(\frac{\hat{p}(x_1,x_2,x_3,x_4,x_5,x_6,x_7)}{1-\hat{p}(x_1,x_2,x_3,x_4,x_5,x_6,x_7)}) = -8.02 + 0.03 * Glucose - 0.01 * BloodPressure + 0.87 * DiabetesPedigreeFunction + 0.15 * Pregnancies + 0.08 * BMI + 1.05 * AgeMidleAge + 1.7 * AgeOldAge - 0.07 * Pregnancies : AgeMidleAge - 0.548 * Pregnancies : AgeOld Age$$

The model showed an overall accuracy of 78%, sensitivity of 60%, specificity of 88%, and an area under the curve of 0.84.

5. Conclusion:

The study aimed to predict diabetes risk in women using a multivariable logistic regression model. However, caution should be taken while solely using it as a diagnostic tool. The limited dataset size of 768 observations and the presence of high insulin levels in some observations might affect the statistical power and generalizability of the results. Additionally, the dataset considered only a limited number of factors, while several other factors like diet, sedentary lifestyle, stress, and smoking may also impact insulin levels and the risk of diabetes.

Moreover, the dataset only includes women from the Pima tribe living in Southern Arizona, so generalizing the findings to other populations must be done with care. The study used three models, logistic regression, random forests, and XGBoost, to predict the outcomes, and random forests proved to be the best model with an accuracy of 80%. However, further work is required to enhance the accuracy, and using multiple models can help predict outcomes in similar datasets.

6. Appendix:

GitHUB ((R-Code) :
https://github.com/vikasvr1997/Stat_632/blob/main/Project_reconstruct.Rmd

7. Reference: Int J Environ Res Public Health. 2021 Jul; 18(14): 7346.Published online 2021 Jul 9. doi: 10.3390/ijerph18147346