TAKE HOME ASSIGNMENT

Categorical Data Analysis

DECEMBER 6, 2023

NAME: P A GUNAWARDANA
STUDENT ID: D/DBA/21/0045

Contents

1. Introduction

Diabetes is a widespread chronic condition impacting millions globally, characterized by high blood glucose levels. Early prediction of Diabetes outcomes is crucial for effective management. Logistic regression, a statistical tool, helps model the relationship between patient information and the likelihood of having Diabetes. This report uses the 'diabetes.csv' dataset with nine variables, including medical predictors like pregnancies, glucose, and BMI. We use R software to build and compare logistic regression models, considering factors like Deviance and AIC. The goal is to identify the best model, interpret results, and offer insights for diabetes management.

Note: Data is not divided into the training set and testing set. The whole data set has been used for the model-building process.

2. Data Exploration

The dataset consists of 768 observations with nine original variables and an additional variable, "Outcome_Label," indicating the presence or absence of Diabetes. These include pregnancy information, glucose levels, blood pressure, skin thickness, Insulin, BMI, Diabetes pedigree function, Age, and the outcome.

2.1 Summary Statistics

Pregnancies: Ranging from 0 to 17, with a mean of 3.85.

Glucose: Varies from 0 to 199, with an average of 120.9.

Blood Pressure: Classified into "Low," "High," and "Medium."

Skin Thickness: Ranges from 0 to 99, with a mean of 20.54.

Insulin: Varies from 0 to 846, averaging at 79.8.

BMI: Spanning 0 to 67.1, with a mean of 31.99.

Diabetes Pedigree Function: Ranges from 0.078 to 2.42, with a mean of 0.4719.

Age: From 21 to 81, with a mean of 33.24.

Outcome: A binary variable indicating the presence (1) or absence (0) of Diabetes.

Outcome Label: Categorized as "No Diabetes" (500 instances) and "Diabetes" (268 instances).

2.2 Missing Values

No missing values were detected across any variable.

2.3 Age Verification

All recorded ages are 21 or above, ensuring the dataset meets the specified criteria.

3. Categorical Variable Analysis

The team conducted a Chi-Square test to explore the relationship between the categorical variable "Blood Pressure" and the diabetes outcome. The test assumes:

H0 (Null Hypothesis): Blood Pressure and Outcome are independent.

H1 (Alternative Hypothesis): Blood Pressure and Outcome are not independent.

The contingency table reveals the distribution of outcomes based on blood pressure categories:

	High	Medium	Low
No Diabetic	88	119	293
Diabetic	77	39	152

Figure 1:Contingency Table

Pearson's Chi-squared test indicates a significant association between Blood Pressure and Outcome (X-squared = 17.423, df = 2, p-value = 0.0001647). The p-value less than the typical significance level (0.05) suggests evidence to reject the null hypothesis.

This implies that Blood Pressure is a meaningful factor in predicting diabetes outcomes, providing valuable insights for model building.

4. Interpreting Logistic Regression Results

Here, R has fitted the model with all the specified predictors. According to this, it is visible that the skin thickness, Insulin, Age and Blood Pressure are not significant. However, we have not specified a model selection method here. Therefore, the model accuracy might be very low. This might not be the best model which fits the data.

1.1 Significant Predictors:

Glucose, BMI, DiabetesPedigreeFunction, and Pregnancies are statistically significant predictors ($p < 0.05$).

Glucose, BMI and Pregnancy have positive coefficients, indicating that an increase in these variables raises the odds of Diabetes.

1.2 Non-Significant Predictors:

Skin thickness, Insulin, Age, and Blood Pressure are not statistically significant in predicting Diabetes.

1.3 Deviance and AIC

The residual Deviance is 729.17 on 758 degrees of freedom.

The Akaike Information Criterion (AIC) is 749.17, a measure of the model's goodness of fit. Lower AIC values indicate better-fitting models.

5. Scatter Plots for Non-Categorical Variables

5.1 Glucose vs Outcome:

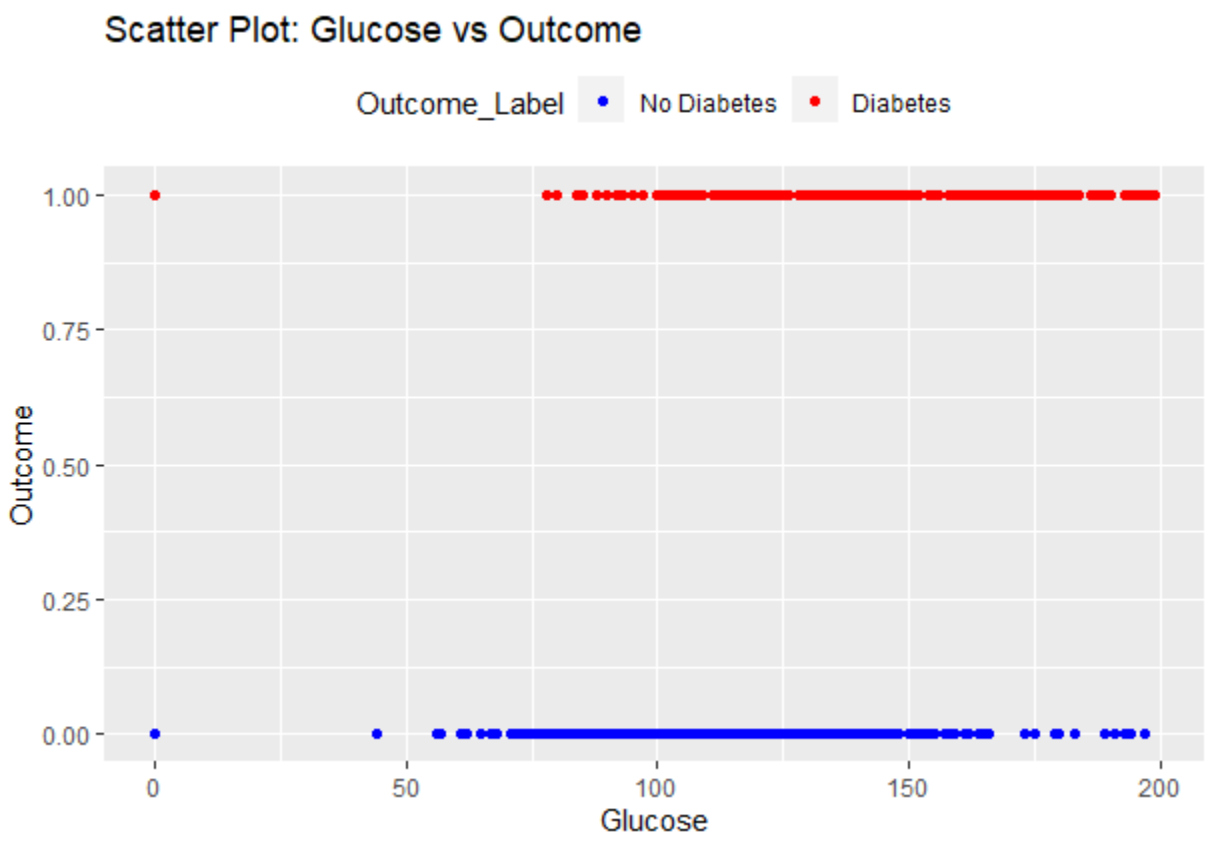

Figure 2: Scatter Plot- Glucose vs Outcome

The scatter plot demonstrates a positive association between glucose levels and diabetes outcome, indicating that individuals with higher glucose levels are more likely to have Diabetes.

5.2 Skin Thickness vs Outcome

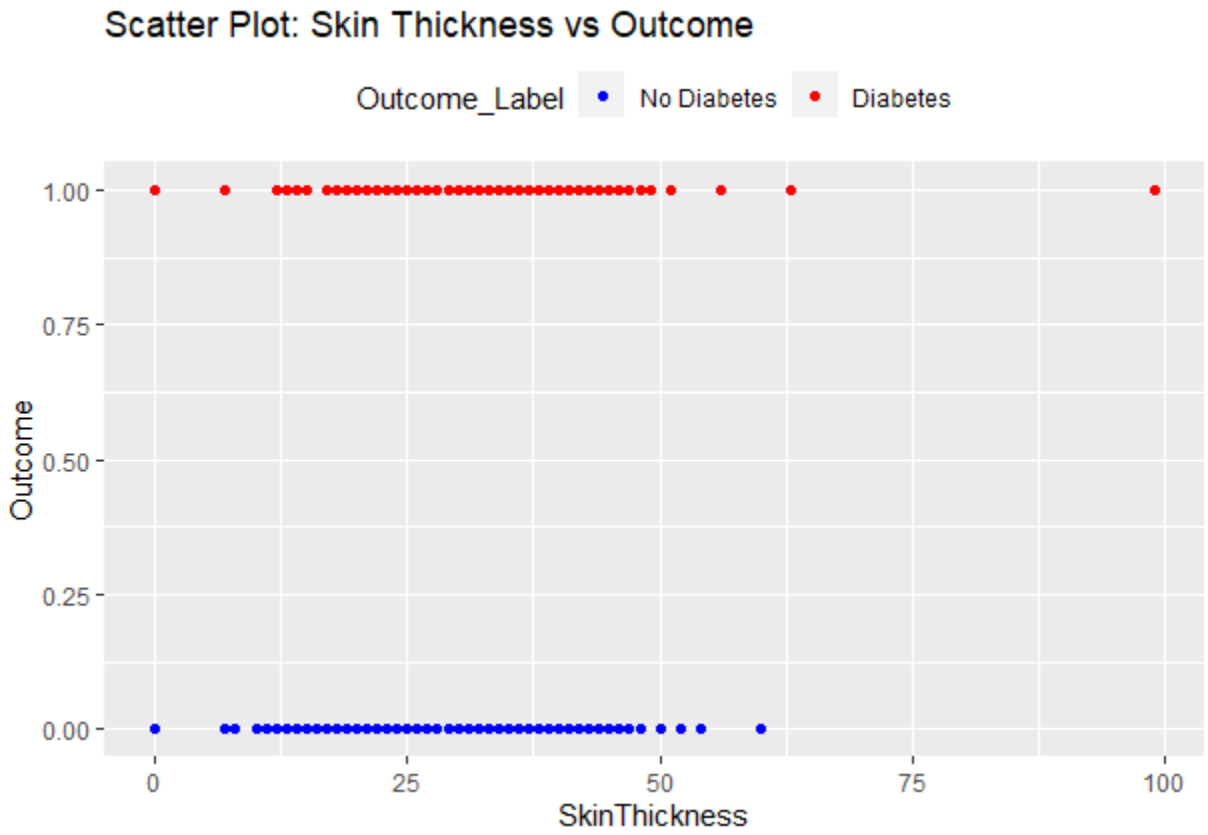

Figure 3: Scatter Plot- Skin Thickness vs Outcome

According to the scatter plot, there is no clear evidence to comment on the relationship between the Skin Thickness and Outcome.

5.3 Insulin vs Outcome

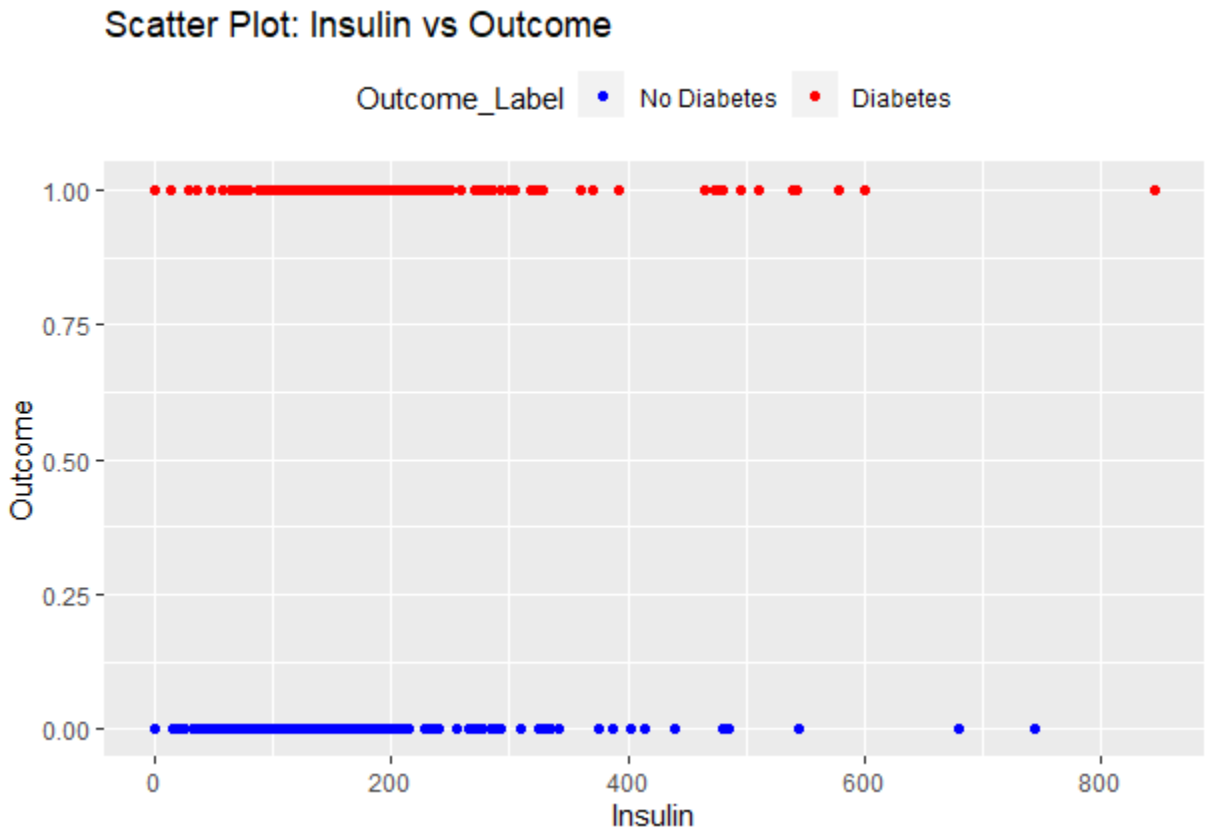

Figure 4: Scatter Plot- Insulin vs Outcome

The Scatter Plot shows no visible relationship between Insulin and the Outcome. Therefore, with the plot's result, we can comment that there is no connection between the Insulin level and the likelihood of having Diabetes.

5.4 BMI vs Outcome

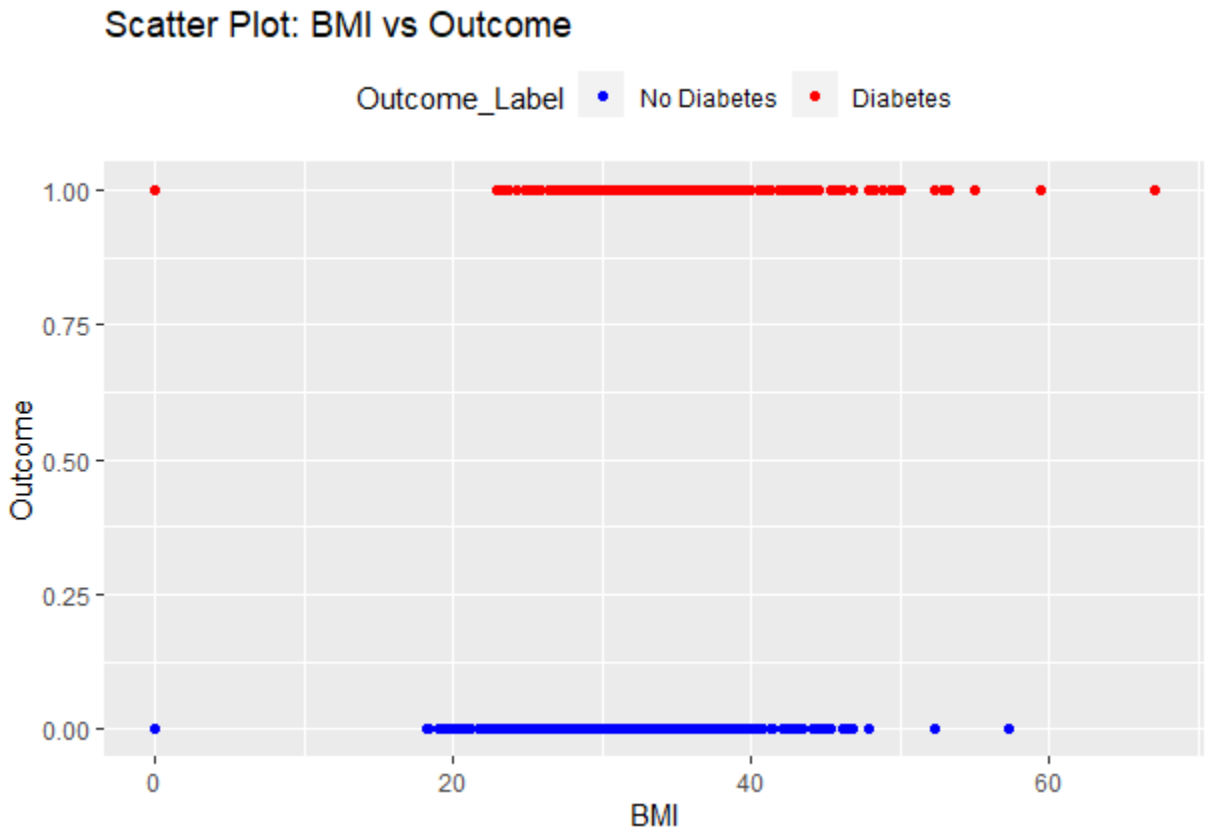

Figure 5:Scatter Plot- BMI vs Outcome

The Scatter Plot shows no clear relationship between the BMI and Outcome. However, there is a slight increase in diabetes patients with an increased BMI.

5.5 DiabetesPedigreeFunction vs Outcome

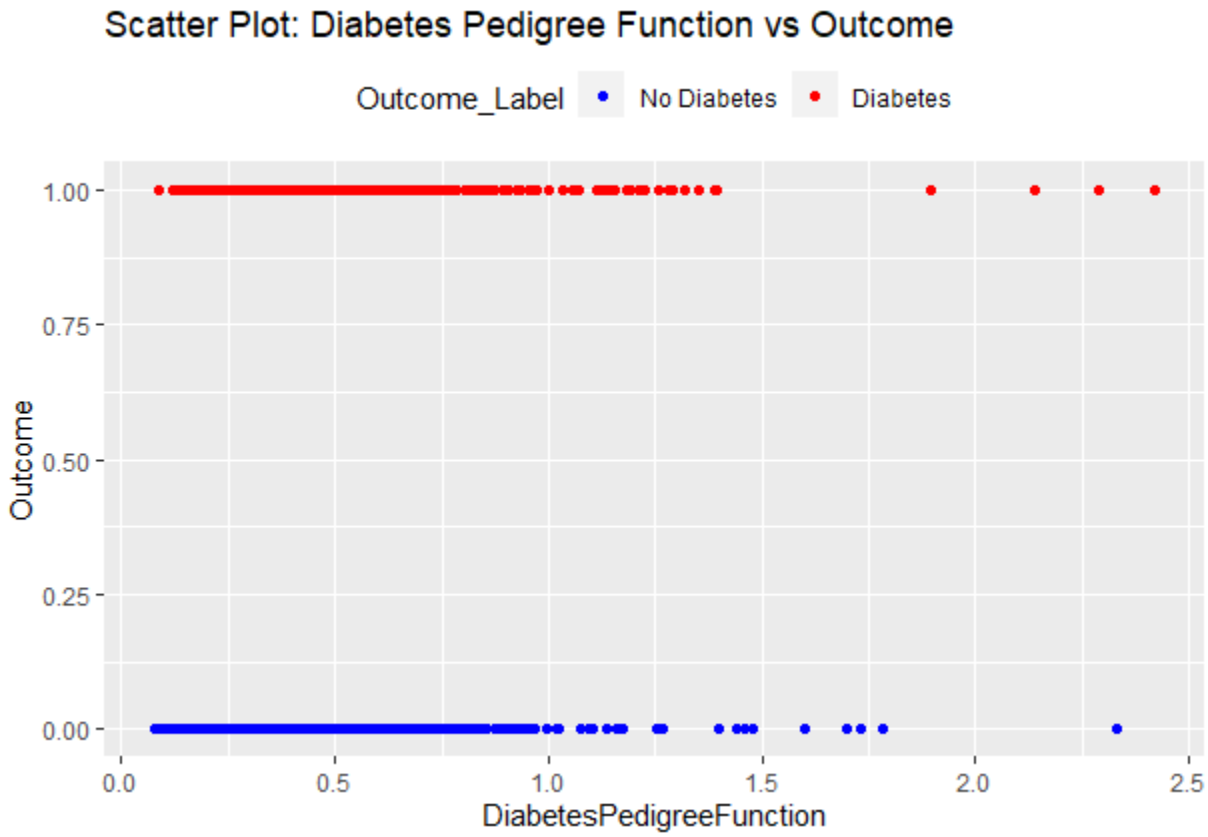

Figure 6: Scatter Plot- DiabetesPedigreeFunction vs Outcome

The scatter plot shows no connection between Diabetes Pedigree Function and Outcome. However, after 1.0, the number of diabetes patients increased compared to the range between 0 and 1.0.

5.6 Pregnancies vs Outcome

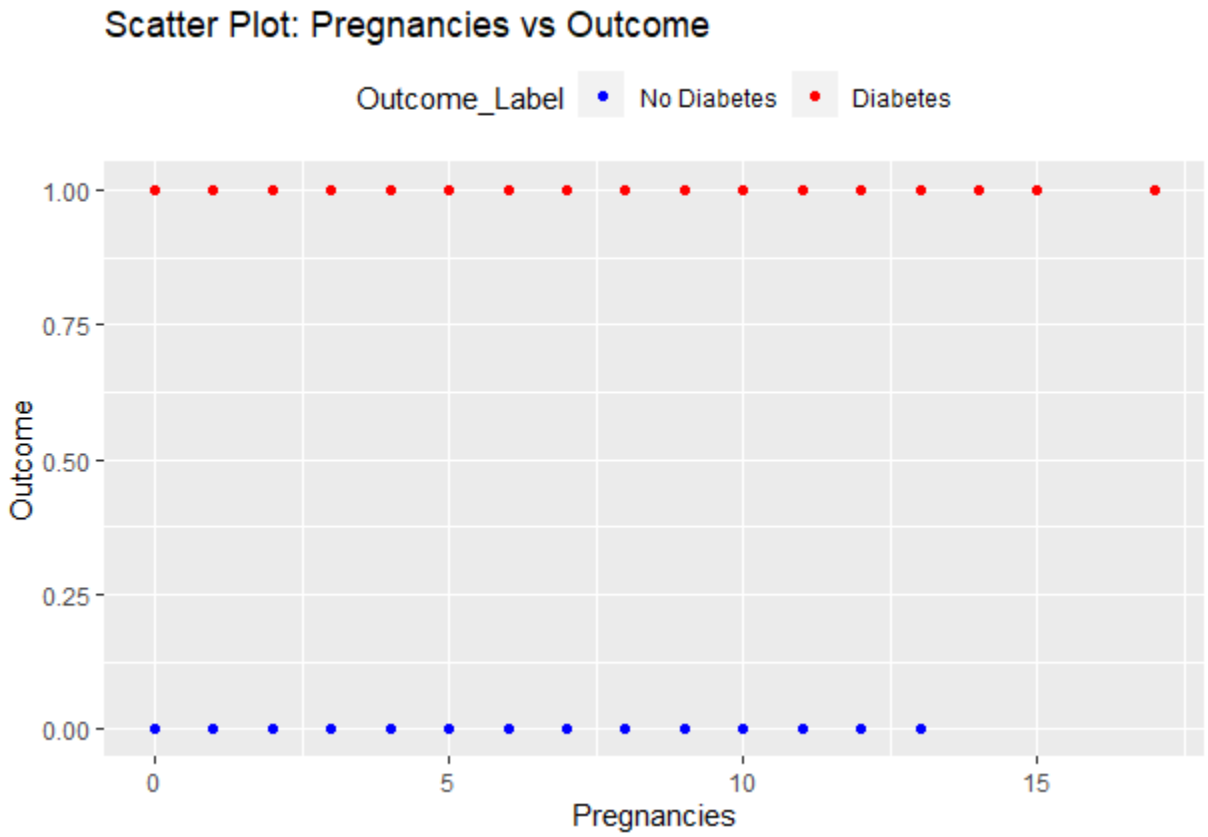

Figure 7: Scatter Plot- Pregnancies vs Outcome

According to the scatter plot, it is visible that there is a slight increase in the likelihood of having Diabetes with the increased number of pregnancies. However, we cannot exactly comment on the relationship between the two variables.

5.7 Age vs Outcome

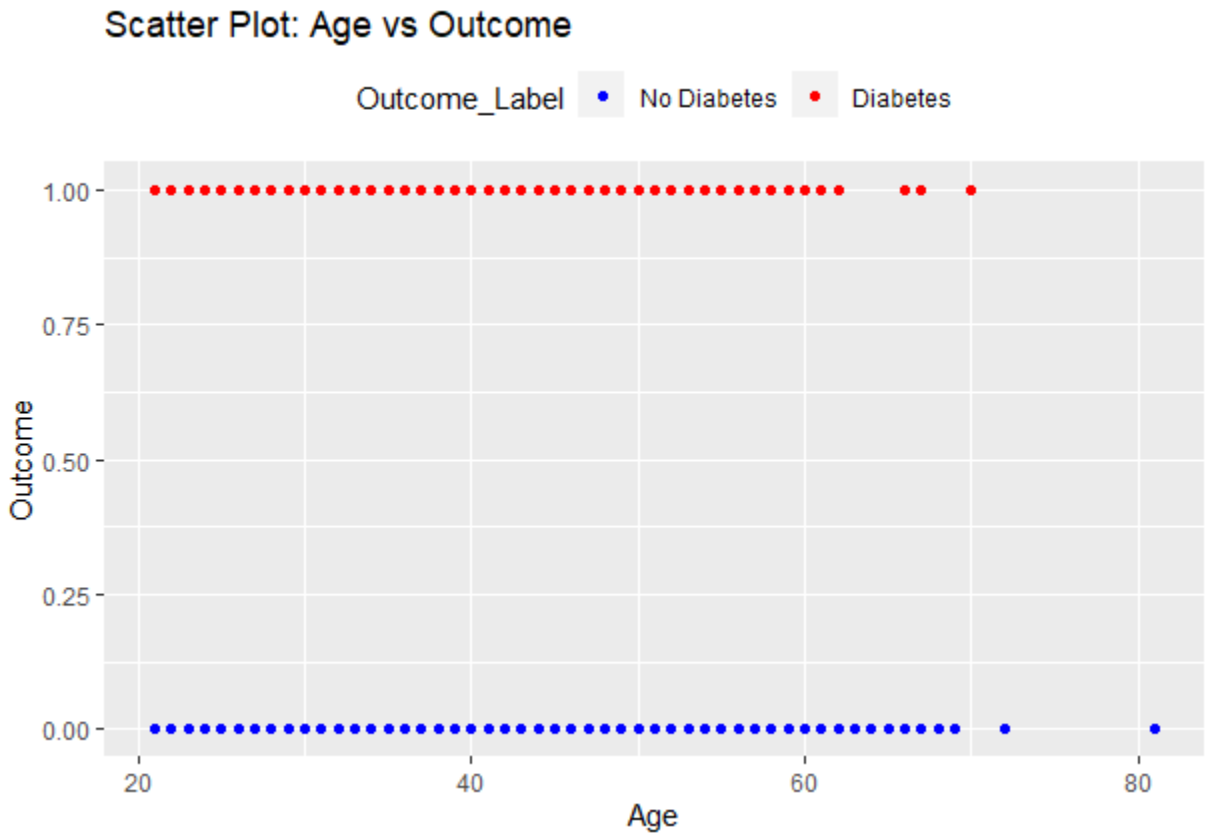

Figure 8: Scatter Plot- Age vs Outcome

The scatter plot between Age vs Outcome shows no significant relationship between Age and Outcome. Therefore, we cannot comment on the connection between Age and getting Diabetes.

6. Logistic Regression Modeling

There are several possible ways to fit the logistic regression model. These methods eliminate or add variables to fit the best logistic regression model. This modelling method uses three main methods to find the best model.

1. Forward Selection
2. Backward Elimination
3. Stepwise Selection

6.1 Method Summaries

Table 1: Summary of Final Models

Method	AIC	Residual Deviance (df=762)
Forward Selection	743.51	731.51
Backward Elimination	743.51	731.51
Stepwise Selection	743.51	731.51

All three methods suggest a model with the same predictor variables. Therefore, all three final models' AIC values and Residual Deviance are the same. Therefore, we can proceed with a model with the predictor variables **Glucose, BMI, Pregnancies, DiabetesPedigreeFunction, and Insulin.**

Therefore, for further proceeding, the **stepwise_model** is selected.

6.2 Final Model Summary

Table 2: Summary of Final Model

Coefficients	**Estimate**	**Std Error**	**Z-value**	**Pr(>|z|)**	**2.5%**	**97.5%**
(Intercept)	-8.6345052	0.6768354	-12.757	2.843e-37	-1.001e+01	-7.355
Glucose	0.0356370	0.0035584	10.015	1.311e-23	2.885e-02	0.043
Insulin	-0.0013438	0.0008034	-1.673	0.09442	-2.930e-03	0.000
BMI	0.0811622	0.0139878	5.802	6.539e-09	5.443e-02	0.109
DiabetesPedigreeFunction	0.9589709	0.2950271	3.250	1.152e-03	3.868e-01	1.544
Pregnancies	0.1370867	0.0273313	5.016	5.283e-07	8.400e-02	0.191

It is identifiable that the p-value of Insulin is 0.09442, which is greater than the 0.05 significance level. Therefore, we do not have enough evidence to say Insulin is significant in the model. When analyzing the middle models that have been built before they reach the final model in each method, it is visible that Insulin acts on the borderline, and it slightly helps to decrease the AIC value. By

considering both the AIC value and the established domain knowledge about the effect of Insulin on Diabetes, it has decided to include Insulin in the model.

6.3 Equation of the final model

Logit(pi) = -8.6345052 + 0.0356370*Glucose - 0.0013438*Insulin + 0.0811622*BMI + 0.9589709*DiabetesPedigreeFunction + 0.1370867*Pregnancies

7. Model Diagnostics

7.1 Hosmer and Lemeshow goodness of fit (GOF) test

H0: The model fits the data well.

H1: The model does not fit the data well.

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  observed_predicted$Observed, observed_predicted$Predicted
X-squared = 9.6476, df = 8, p-value = 0.2906
```

Figure 9: Hosmer and Lemeshow test

Since the p-value (0.2906) is greater than the significance level of 0.05, we fail to reject the null hypothesis.

No significant evidence suggests a lack of fit, indicating that the model adequately fits the data according to the Hosmer-Lemeshow test.

7.2 The area under the ROC curve

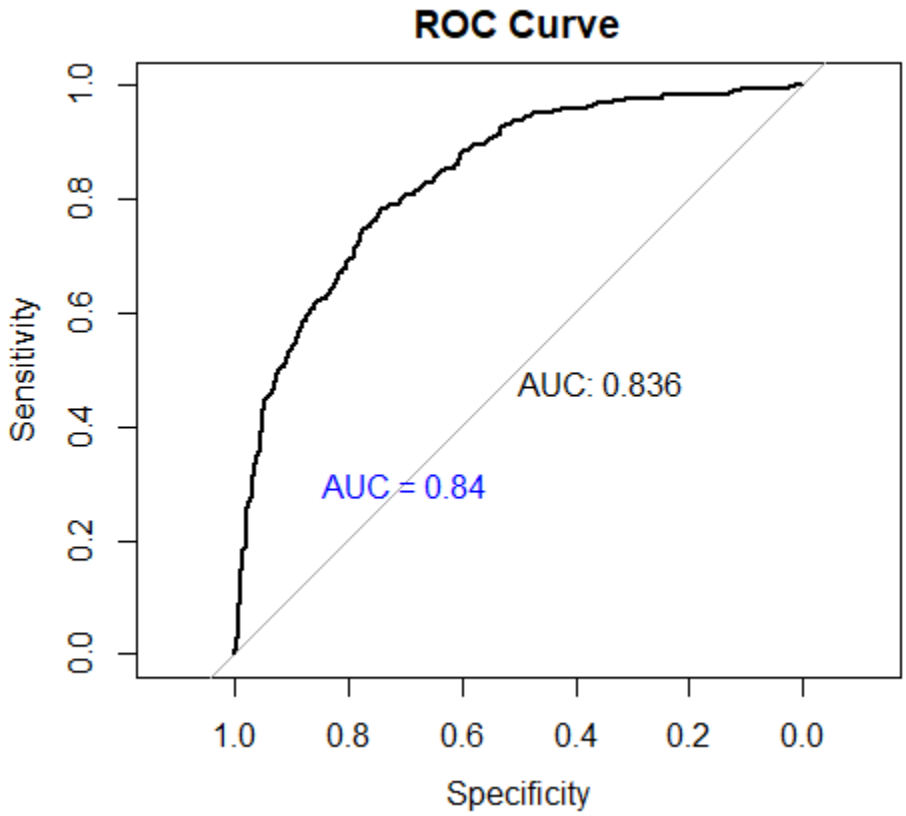

Figure 10: Area under the ROC curve

The area under the ROC curve 0.836 is considered **excellent** since it lies between 0.8 and 0.9. An AUC value of 0.836 indicates that the model has a strong 83.6% probability of correctly distinguishing between individuals with Diabetes and those without, based on their predicted probabilities.

7.3 Deviance Residual Test

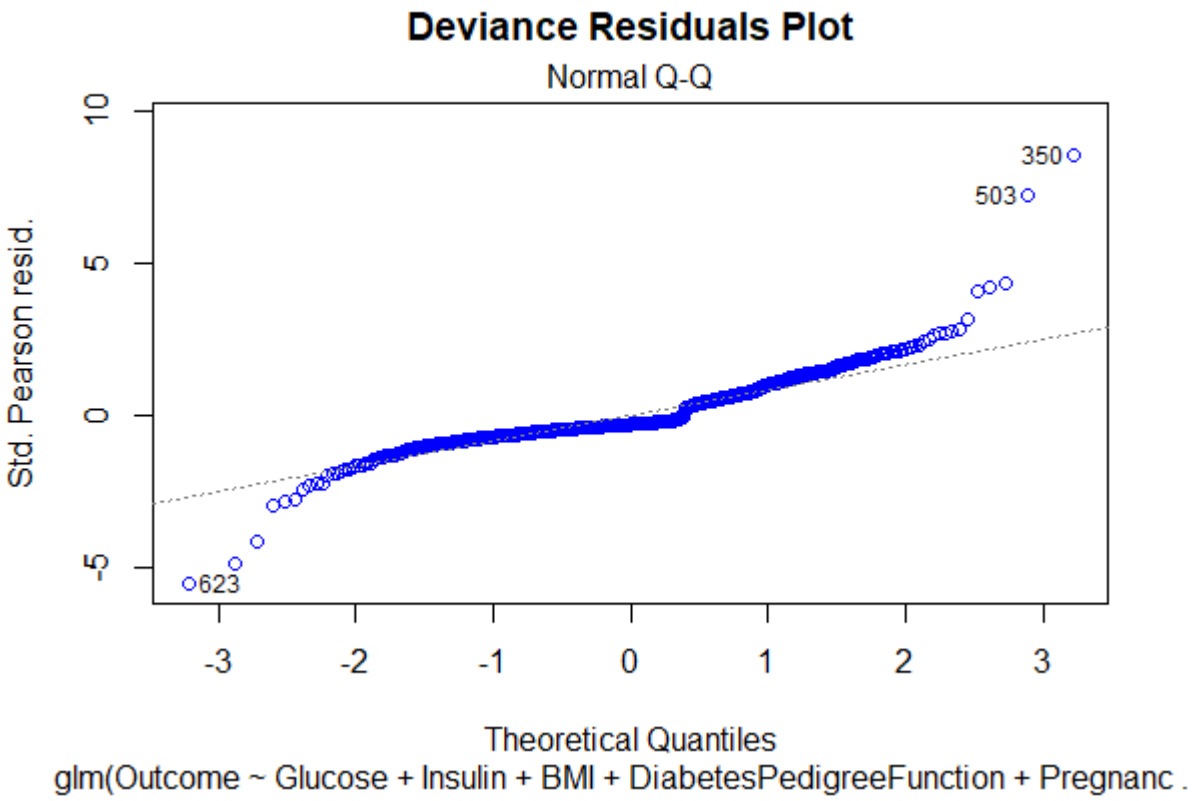

Figure 11: Deviance Residual Plot

According to the deviance residual plot, even though the values form a straight line, they are not exactly parallel to the x-axis. This could occur due to overfitting of the model. However, this model demonstrates a reasonable fit to the data.

7.4 Deviance Test

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			767	993.48		
Glucose	1	184.764	766	808.72	< 2.2e-16	***
Insulin	1	0.950	765	807.77	0.329664	
BMI	1	39.983	764	767.79	2.562e-10	***
DiabetesPedigreeFunction	1	10.237	763	757.55	0.001377	**
Pregnancies	1	26.044	762	731.51	3.338e-07	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Figure 12: Deviance Test Results

Hypothesis

Ho=Model adequately fit the data

H1=Model does not adequately fit the data

The initial model with no predictors has a deviance of 993.48.

When Glucose, BMI, DiabetesPedigreeFunction, and Pregnancies predictors are added sequentially, they significantly reduce the Deviance, leading to improved model fit. Their p-values are highly significant, which is less than 0.05.

Insulin has a p-value of 0.329664, suggesting that it is not statistically significant in predicting the outcome.

8. Discussion

The analysis of the diabetes dataset with 768 observations and nine variables revealed essential findings. No missing values and age verification above 21 were identified.

The Chi-Square test revealed a meaningful connection between "Blood Pressure" categories and diabetes outcomes.

Scatter Plots for non-categorical variables showed clear patterns for "Glucose" and "Pregnancies," while others were less conclusive.

In the initial logistic regression with all the predictors, "SkinThickness," "Insulin," "Age," and "Blood Pressure" were insignificant. Further modelling using backwards, forward and stepwise methods highlighted the importance of "Glucose," "BMI," "DiabetesPedigreeFunction," and "Pregnancies," with "Insulin" just above 0.05. However, Insulin helped reduce the AIC value of the model. Therefore, Insulin was included in the best-fitted model.

A final model was selected from model selection methods by selecting "Glucose," "BMI," "Pregnancies," "DiabetesPedigreeFunction," and "Insulin" as predictor variables of the final model.

Model diagnostics, like the Hosmer and Lemeshow test and ROC curve, indicated a reasonable fit. However, the deviance residual plot showed some deviation from the best results. Nevertheless, the fitted model is acceptable.

9. Conclusion

In conclusion, this exploration of the diabetes dataset provided important insights. It has found key factors like "Glucose," "BMI," "DiabetesPedigreeFunction," and "Pregnancies" to be significant in predicting diabetes outcomes. The model fit reasonably well, but a potential issue with "Insulin" was noticed, which requires further investigation. The findings contribute to understanding diabetes risk factors and can guide future research and interventions.

10. Appendix

Codes

- Model Selection Procedure

```
#forward selection
#we consider all variables in the dataset as potential predictors
initial_model1 <- glm(Outcome ~ 1, data = diabdata, family = "binomial")
forward_model <- step(initial_model, direction = "forward",scope = formula(~ Glucose + Blood.Pressure + SkinThickness + Insulin + BMI + DiabetesPedigreeFunction + Age + Pregnancies), data = diabdata)
summary(forward_model)


#backward selection
initial_model2 <- glm(Outcome~Glucose + Blood.Pressure + SkinThickness + Insulin + BMI + DiabetesPedigreeFunction + Age + Pregnancies, data = diabdata, family = "binomial")
backward_model <- step(initial_model2, direction = "backward")
summary(backward_model)

#step wise selection method
stepwise_model <- step(glm(Outcome ~ Glucose + Blood.Pressure + SkinThickness + Insulin + BMI + DiabetesPedigreeFunction + Age + Pregnancies, data = diabdata, family = "binomial"), direction = "both")
summary(stepwise_model)
```

- Model Diagnostics – Hosmer and Lemeshow test

```
# Get predicted probabilities from the model
predicted_probs <- predict(stepwise_model, type = "response")

# Create a data frame with observed and predicted values
observed_predicted <- data.frame(Observed = diabdata$Outcome, Predicted = predicted_probs)
# Perform the Hosmer-Lemeshow test
# HO- Model adequately fit the data
# H1- Model does not adequately fit the data
hosmer_lemeshow_test <- hoslem.test(observed_predicted$Observed, observed_predicted$Predicted)
# Print the test result
print(hosmer_lemeshow_test)
```

- Model Diagnostics – ROC curve

```
# Create a ROC curve
roc_curve <- roc(diabdata$Outcome, predicted_probs)
# Plot the ROC curve
plot(roc_curve, main = "ROC Curve", print.auc = TRUE)
# Add AUC value to the plot
text(0.7, 0.3, paste("AUC =", round(auc(roc_curve), 2)), col = "blue")
```

- Model Diagnostics – Deviance test and deviance residual plot

```
# Create deviance residual plot
plot(stepwise_model, which = 2, col = "blue", main = "Deviance Residuals Plot")

# Perform deviance test
#Hypothesis
#Ho=Model adequately fit the data
#H1=Model does not adequately fit the data

deviance_test <- anova(stepwise_model, test = "Chisq")
print(deviance_test)
```

- Full R Code Link:- https://drive.google.com/file/d/1WDM-z24ecy6eVn-cqJZ4raXSPAE_5zNQ/view?usp=sharing

11. References

- Diabetes Research, Education, Advocacy | ADA