# "To What Extent Do Factors Such as the Number of Pregnancies, Glucose Levels, BMI, and the Diabetes Pedigree Function Predict the Risk of Developing Diabetes?"

## Background:

Diabetes, also known as Diabetes Mellitus is a global health issue characterized by high blood glucose levels over a prolonged period, leading to serious complications if left untreated. The ability to predict the onset of diabetes based on certain predictors can significantly impact managing this disease. Machine learning models, especially logistic regression and ensemble techniques, have shown promise in predicting diabetes' likelihood by analyzing various predictors like the number of pregnancies, glucose levels, BMI, and the Diabetes Pedigree Function.

The use of logistic regression is a common theme across the articles. For example, Tabaei and Herman developed a predictive equation incorporating variables like age, sex, BMI, postprandial time, and plasma glucose levels to screen for undiagnosed diabetes, demonstrating logistic regression's utility in diabetes prediction. Shahram Latifi's work further explores logistic regression alongside ensemble methods, emphasizing the importance of selecting the right predictors and refining the model for accuracy.

The number of Pregnancies is recognized as a significant predictor because it impacts glucose tolerance and insulin resistance. Glucose Level is a critical predictor as it directly measures the body's ability to handle glucose. BMI is indicative of obesity, a known risk factor for diabetes. Diabetes Pedigree Function represents genetic predisposition, offering insights into the hereditary risks of diabetes.

The model will build upon the logistic regression approach, proven to be effective in the articles reviewed. Given the predictors' differing nature and impact on diabetes risk, logistic regression offers the flexibility to understand how each predictor contributes to the likelihood of diabetes. Moreover, considering the enhanced performance with ensemble methods as highlighted by Shahram Latifi, integrating a logistic regression model within an ensemble framework will be considered to leverage the strengths of multiple models for better prediction accuracy.

## Similarities and Differences in Predictors:

Glucose levels and BMI are universally recognized and used across the models for their direct link to diabetes risk. The inclusion of the Diabetes Pedigree Function is not explicitly mentioned by Bahman P. Tabaei and William H. Herman, focusing more on direct measurements like glucose and BMI, but is an integral part of this research, emphasizing genetic factors. Unlike the sole focus on logistic regression, this research is open to integrating ensemble methods to refine the prediction accuracy further, inspired by the advancements discussed by Shahram Latifi.

## Method:

The dataset is sourced from Kaggle, featuring variables like pregnancies, BMI, insulin levels, and age. This dataset includes variables like Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age. The outcome variable is Outcome, indicating the presence or absence of diabetes.

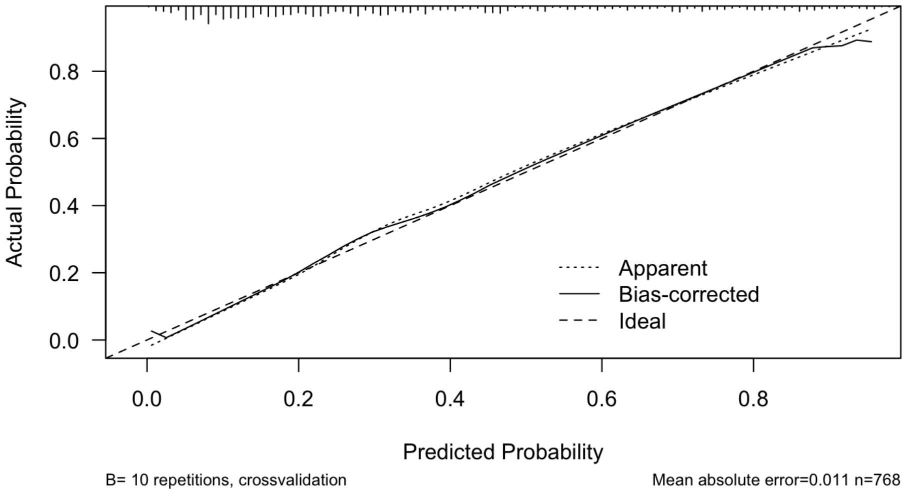
The Outcome variable was transformed into a factor with labels "No Diabetes" and "Diabetes" for clarity in analysis and plotting. Initial data exploration involved generating histograms and boxplots for each predictor variable by the outcome, offering visual insight into their distribution across diabetes conditions.

A logistic regression model (glm) was initially fitted with Outcome as the response variable and all other variables as predictors using a binomial family, indicative of a binary outcome. Variable selection was performed using the step function with both forward and backward directions, based on the Bayesian Information Criterion, to identify the most significant predictors.

The refined logistic regression model was then fitted with the selected predictors. The model's summary provided coefficients, standard errors, z-values, and p-values for each predictor, indicating their significance.

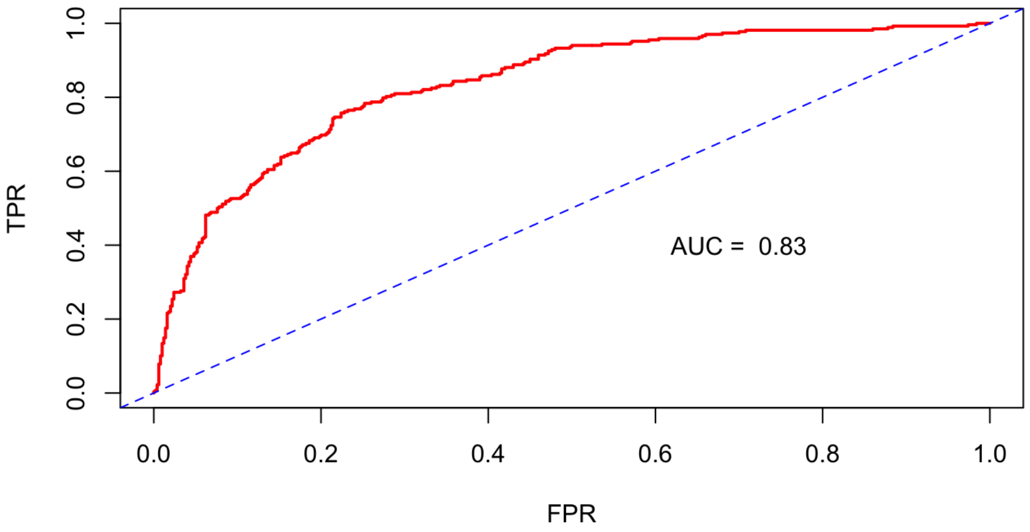
Further analysis included the use of the lrm function from the rms package for logistic regression modeling, focusing on the selected predictors. This step is likely aimed at refining the model fit and assessing its predictive performance. Model calibration was assessed through cross-validation, and ROC curves were plotted to evaluate the model's discriminatory ability, represented by the area under the curve (AUC). Dfbetas and deviance residuals were analyzed to identify influential points and assess the model's fit across different levels of the predictors.

Results:
Histograms and boxplots visualized the distribution of each predictor by the outcome, highlighting key trends and outliers that informed subsequent modeling decisions (graphs distributed in the appendix). A logistic regression model was initially fitted with all variables, but variable selection using the Bayesian Information Criterion pinpointed Pregnancies, Glucose, BMI, and DiabetesPedigreeFunction as significant predictors. This focused approach reduced model complexity without sacrificing predictive power.



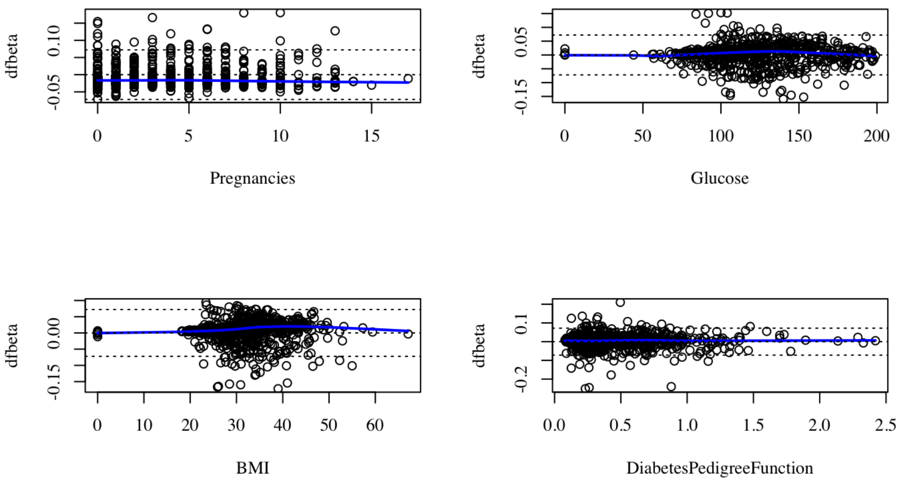
Graph 1: Model calibration with cross-validation and bootstrap

From the plot, we can see some discrepancy between the predicted probabilities and the actual outcomes, as indicated by the bias-corrected line. However, the discrepancy is minor as shown by the mean absolute error of 0.011, suggesting the model's predictions are fairly well calibrated. The plot also indicates that the model may be slightly overpredicting the likelihood of diabetes for lower probabilities and underpredicting for higher probabilities.



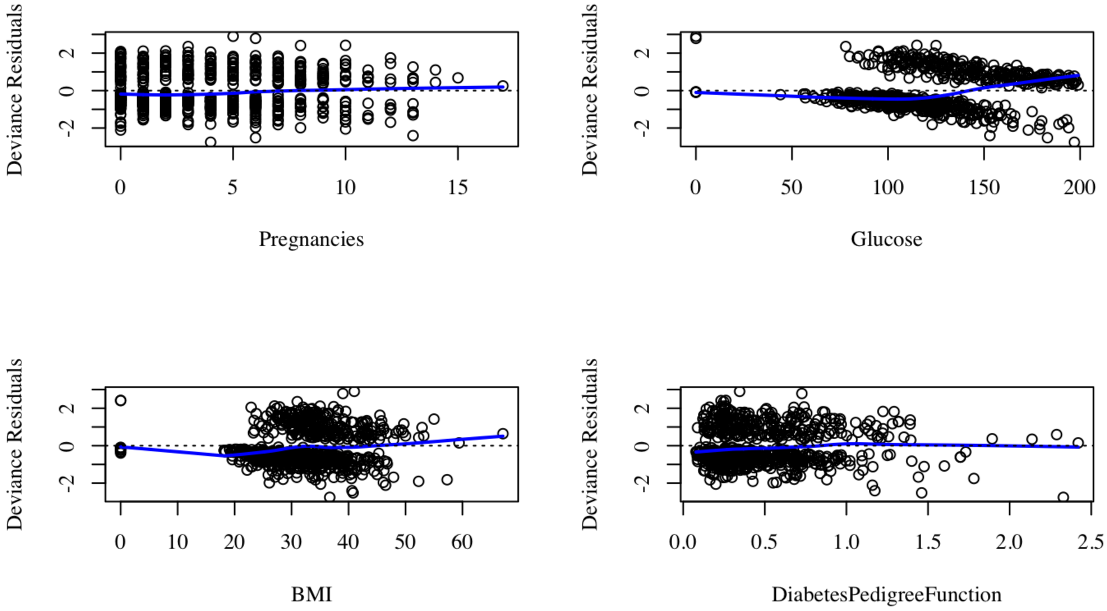
Graph 2: Discrimination with ROC curve

The area under the curve (AUC) is = 0.8345. This means that the model can discriminate between people diagnosed with diabetes and those without diabetes 83.45% of the time. That is, there is an 83.45% chance that the model will be able to distinguish between people with diabetes and those without diabetes.



Graph 3: dfbetas

The dfbetas for Pregnancies are mostly clustered around zero, indicating that individual observations are not exerting undue influence on the model's coefficient for this predictor. The spread of dfbetas for Glucose is larger, but the majority of observations still have a relatively low influence on the coefficient estimate, as they are close to zero. Similar to Pregnancies, the dfbetas for BMI are concentrated around zero. This suggests that removing any single data point does not substantially change the coefficient estimate for BMI. There's a bit more variation in the influence of individual observations on the DiabetesPedigreeFunction coefficient, but most values are still close to zero.



Graph 4: Deviance Residuals

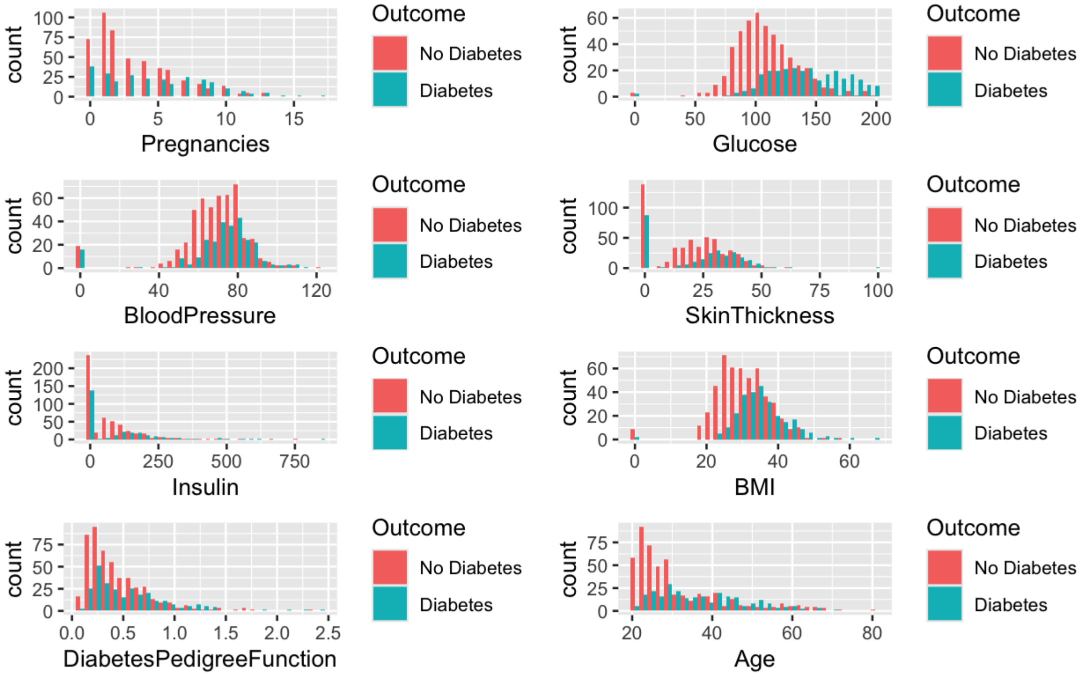
The residuals for Pregnancies, BMI, and DiabetesPedigreeFunction are relatively symmetrically distributed around zero without any obvious patterns or extreme values. The residuals for Glucose seem to have a slight curve, suggesting that the relationship between glucose and the outcome might not be perfectly linear or that other factors might influence this relationship that the model does not capture.

The research question has been addressed through a rigorous statistical analysis employing logistic regression. This model was chosen based on its frequent use and proven effectiveness in similar studies, as well as its flexibility in evaluating the impact of multiple predictors on diabetes risk. Key findings show that pregnancies, glucose levels, BMI, and the Diabetes Pedigree Function are significant predictors of diabetes, aligning with established medical understanding that these factors are instrumental in the progression of this disease.

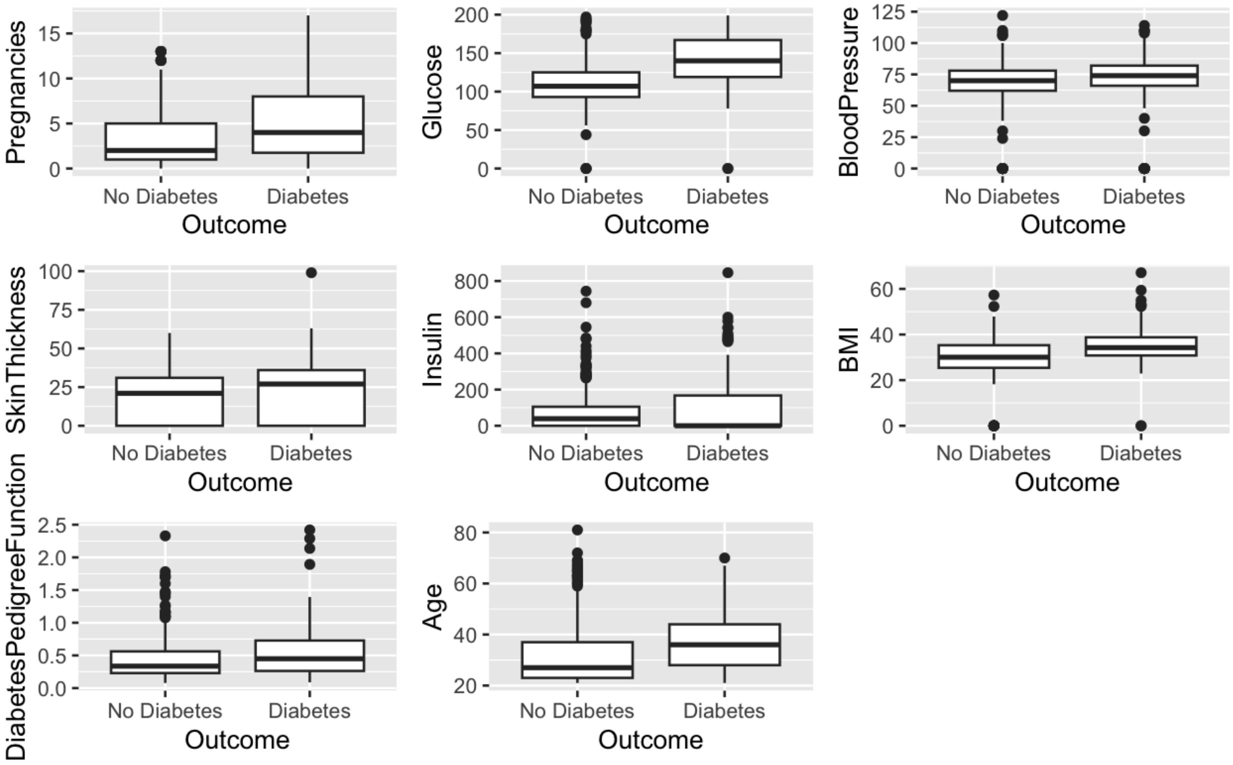
Our model's strength lies in its simplicity and predictive power, a balance achieved by employing variable selection informed by the Bayesian Information Criterion, reducing complexity without undermining model performance. Cross-validation and ROC curve analysis have corroborated the model's robustness, with an AUC of 0.8345 denoting strong predictive ability. The calibration plots, dfbetas, and deviance residuals analyses also support the model's validity, indicating that our predictions are well-calibrated and that no single observation unduly influences the model's estimates.

Despite these strengths, it's important to acknowledge the model's limitations. The slight discrepancy in the model's calibration for lower and higher probability predictions suggests room for improvement. This could potentially be addressed by exploring more complex models or incorporating interaction terms. The curve observed in the residuals for glucose also points to the possibility of non-linear relationships or other influencing factors not captured by the model, which could be explored in future research. Furthermore, the study relies on the assumption that the dataset from Kaggle is representative of the broader population, and findings may not be generalizable to all demographics.

Appendix



Appendix 1: Histogram generated for EDA



Appendix 2: Boxplot generated for EDA

| Coefficients: | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -8.415851 | 0.656908 | -12.811 | < 2e-16 |
| Pregnancies | 0.141926 | 0.027105 | 5.236 | 1.64e-07 |
| Glucose | 0.033826 | 0.003345 | 10.112 | < 2e-16 |
| BMI | 0.078097 | 0.013771 | 5.671 | 1.42e-08 |
| DiabetesPedigreeFunction | 0.901294 | 0.291696 | 3.090 | 0.002 |

Appendix 3: glm summary

References

Article 1:

Tabaei, B. P., & Herman, W. H. (2002, November 1). A multivariate logistic regression equation to screen for diabetes : Development and validation. American Diabetes Association. https://diabetesjournals.org/care/article/25/11/1999/24732/A-Multivariate-Logistic-Regression-Equation-to

Article 2:

Talukder, A., & Hossain, Md. Z. (2020, June 24). Prevalence of diabetes mellitus and its associated factors in Bangladesh: Application of two-level logistic regression model. Nature News. https://www.nature.com/articles/s41598-020-66084-9

Article 3:

Elsevier. (2021, October 25). Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update. https://www.sciencedirect.com/science/article/pii/S2666990021000318

Data:

Devisangeetha. (2017, July 3). Which factor causes diabetes. Kaggle. https://www.kaggle.com/code/devisangeetha/which-factor-causes-diabetes/input