Exploring the relationship between Diabetes and its Risk Factors using Generalized Additive Models (GAMs)

Christopher Odoom, Denis Folitse, Sandani Kumanayake, Owen Gallagher

2023-05-16

Background of Study

- Diabetes is a chronic condition that affects how the body processes blood sugar (glucose). It occurs when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces.
- Type 1 diabetes, also known as juvenile diabetes, is an autoimmune disease that usually develops in childhood or adolescence. It occurs when the body's immune system mistakenly attacks and destroys the cells in the pancreas that produce insulin.
- Type 2 diabetes, which accounts for 90-95% of all cases, is usually diagnosed in adults, but it is becoming more common in children and adolescents due to rising rates of obesity. It occurs when the body becomes resistant to the effects of insulin or when the pancreas cannot produce enough insulin to meet the body's needs.

Background of Study

- The risk factors for type 2 diabetes include physical inactivity, a family history of diabetes, high blood pressure and high cholesterol.

Motivation

- Diabetes is a leading cause of blindness, kidney failure, amputations, heart attacks, and stroke.
- According to Mayo Clinic(2021), There is no cure for diabetes, but it can be managed through lifestyle changes, such as eating a healthy diet, getting regular exercise, and monitoring blood sugar levels, as well as medications like insulin and oral hypoglycemic drugs.
- According to the World Health Organization, the number of people with diabetes has risen from 108 million in 1980 to 422 million in 2014.

Motivation

- In addition to the physical health complications of diabetes, it can also have a significant impact on mental health. People with diabetes are at increased risk of depression, anxiety, and other mood disorders.
- Research is ongoing to better understand the causes of diabetes and to develop new treatments and prevention strategies. Generalized additive models (GAMs) are one tool that can be used to explore the relationship between diabetes and its risk factors.

Objectives

- To understand the relationship between the risk factors of diabetes and their impact on getting diabetes using a GAM logistic regression.
- To consider the possibility of three-way interaction using tensor product smoothing, that way we can see the interactive effects of risk factors on diabetes.
- To compare the predictions from the GAM fit with a selected best performing machine learning algorithm for predicting diabetes.

Methodology

Data

This data is collected from kaggle, updated by Aksha Gattatray Khare with the objective of predicting whether a patient has diabetes, based on certain diagnostic measurements. In total, the data contains 8 variables listed below

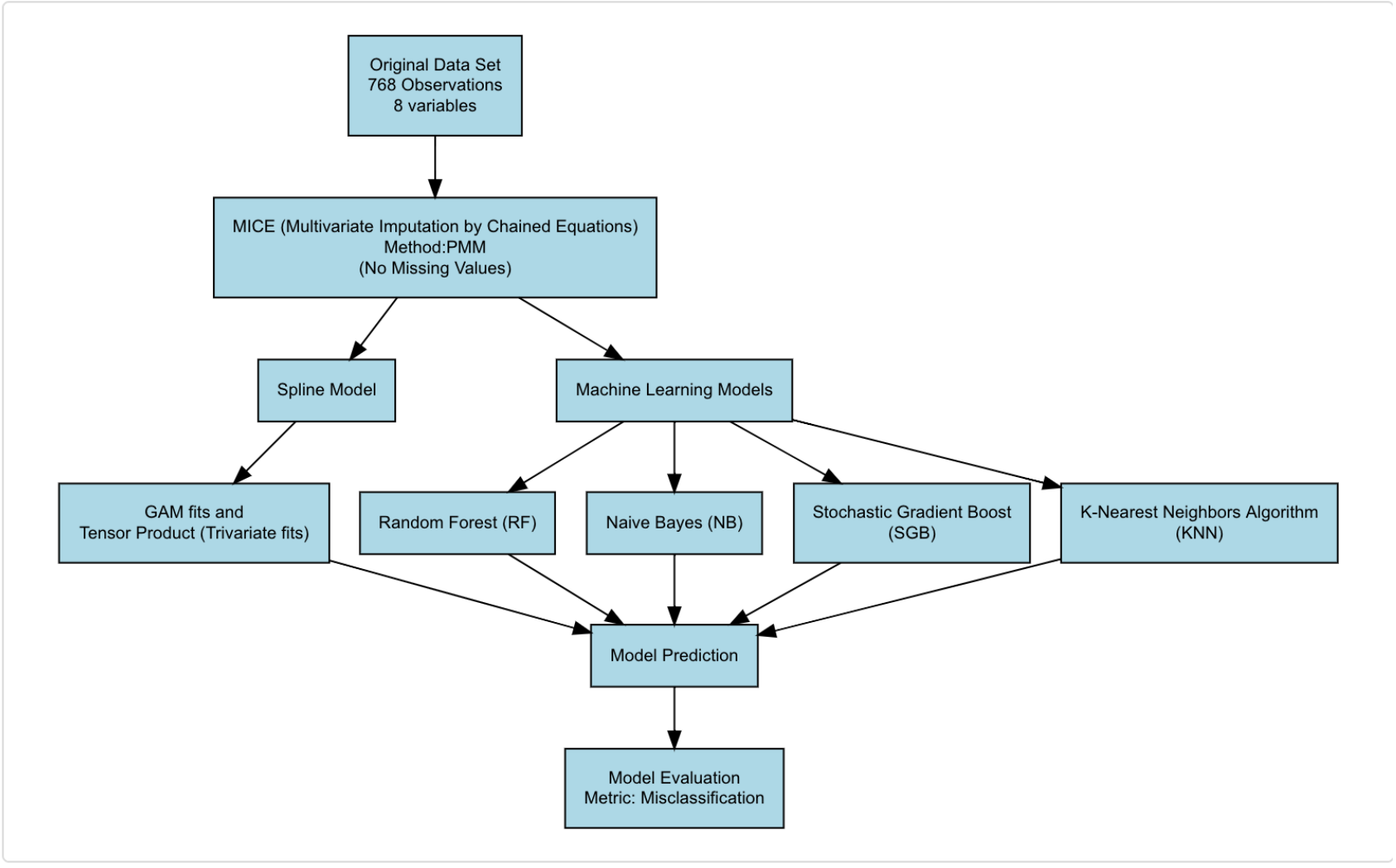
- ***Outcomes:*** To express the final result 1 is Yes and 0 is No
- ***Pregnancies:*** indicates the number of pregnancies
- ***Glucose:*** indicates the plasma glucose concentration
- ***Blood Pressure:*** indicates diastolic blood pressure in mm/Hg
- ***Skin Thickness:*** indicates triceps skinfold thickness in mm

Methodology

Data

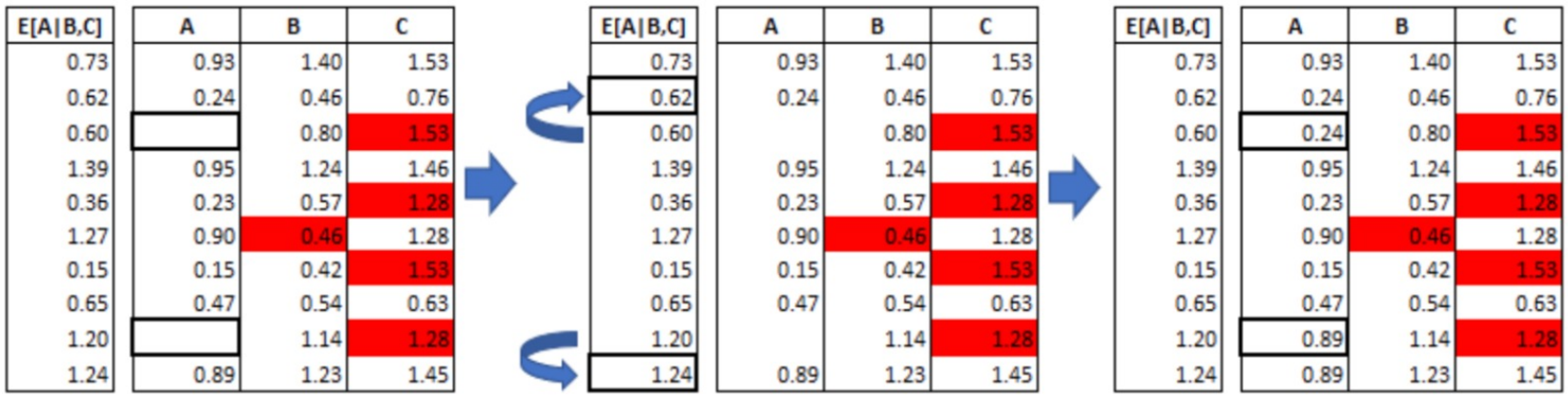
- ***dpfunction:*** indicates the function which scores likelihood of diabetes based on family history
- ***age:*** indicates the age of the person
- ***Insulin:*** indicates insulin in blood (U/mL)
- ***bmi:*** indicates the body mass index in kg/m2

Modeling Workflow



Predictive Mean Matching (PMM)

- Predict the missing values using the MICE algorithm
- chooses a predicted value close to the predicted value of the missing sample.
- Select the relevant data point from the original, non missing data set.



E[A|B,C]	A	B	C
0.73	0.93	1.40	1.53
0.62	0.24	0.46	0.76
0.60		0.80	1.53
1.39	0.95	1.24	1.46
0.36	0.23	0.57	1.28
1.27	0.90	0.46	1.28
0.15	0.15	0.42	1.53
0.65	0.47	0.54	0.63
1.20		1.14	1.28
1.24	0.89	1.23	1.45

E[A|B,C]	A	B	C
0.73	0.93	1.40	1.53
0.62	0.24	0.46	0.76
0.60		0.80	1.53
1.39	0.95	1.24	1.46
0.36	0.23	0.57	1.28
1.27	0.90	0.46	1.28
0.15	0.15	0.42	1.53
0.65	0.47	0.54	0.63
1.20		1.14	1.28
1.24	0.89	1.23	1.45

E[A|B,C]	A	B	C
0.73	0.93	1.40	1.53
0.62	0.24	0.46	0.76
0.60	0.24	0.80	1.53
1.39	0.95	1.24	1.46
0.36	0.23	0.57	1.28
1.27	0.90	0.46	1.28
0.15	0.15	0.42	1.53
0.65	0.47	0.54	0.63
1.20	0.89	1.14	1.28
1.24	0.89	1.23	1.45

Methodology

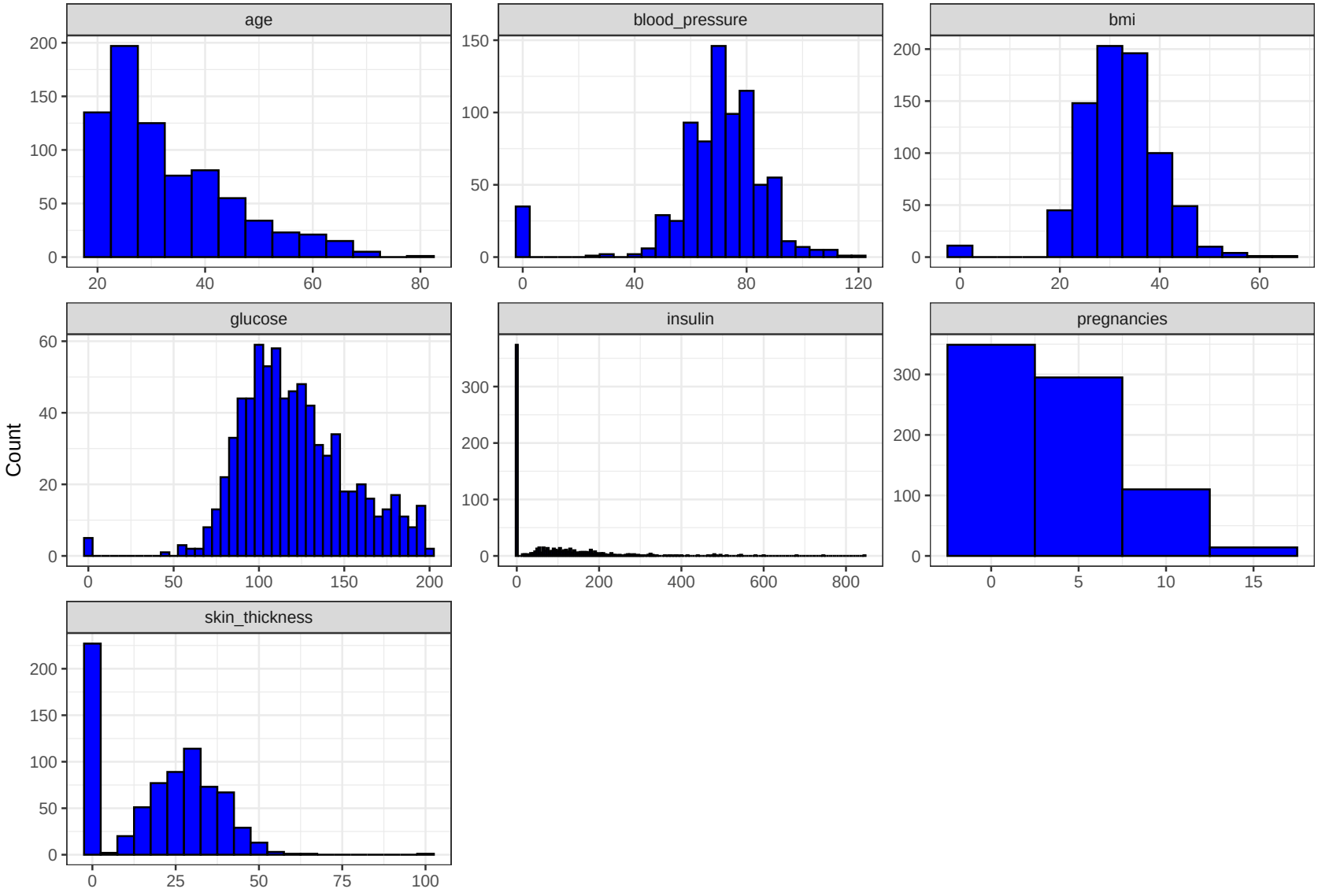
- **GAMs**
- **Tensor Product (Trivariate fits)**
- **Random Forest (RF)**: This is a machine-learning algorithm that builds multiple decision trees and combines them to obtain a more accurate prediction. It is used for classification and regression problems.
- **Naive Bayes (NB)**: This is a classification algorithm based on Bayes' theorem. It assumes that the predictors are independent of each other and calculates the probability of each class based on the predictor values.

Methodology

- **Stochastic Gradient Boost (SGB)**: This is a machine learning algorithm that builds an ensemble of weak prediction models (e.g. decision trees) and sequentially trains them to correct the errors of the previous model. The final prediction is based on the weighted average of all the models.
- **K-Nearest Neighbors Algorithm (KNN)**: This is a non-parametric classification algorithm that assigns a new data point to the class of the majority of its k-nearest neighbors in the training set.

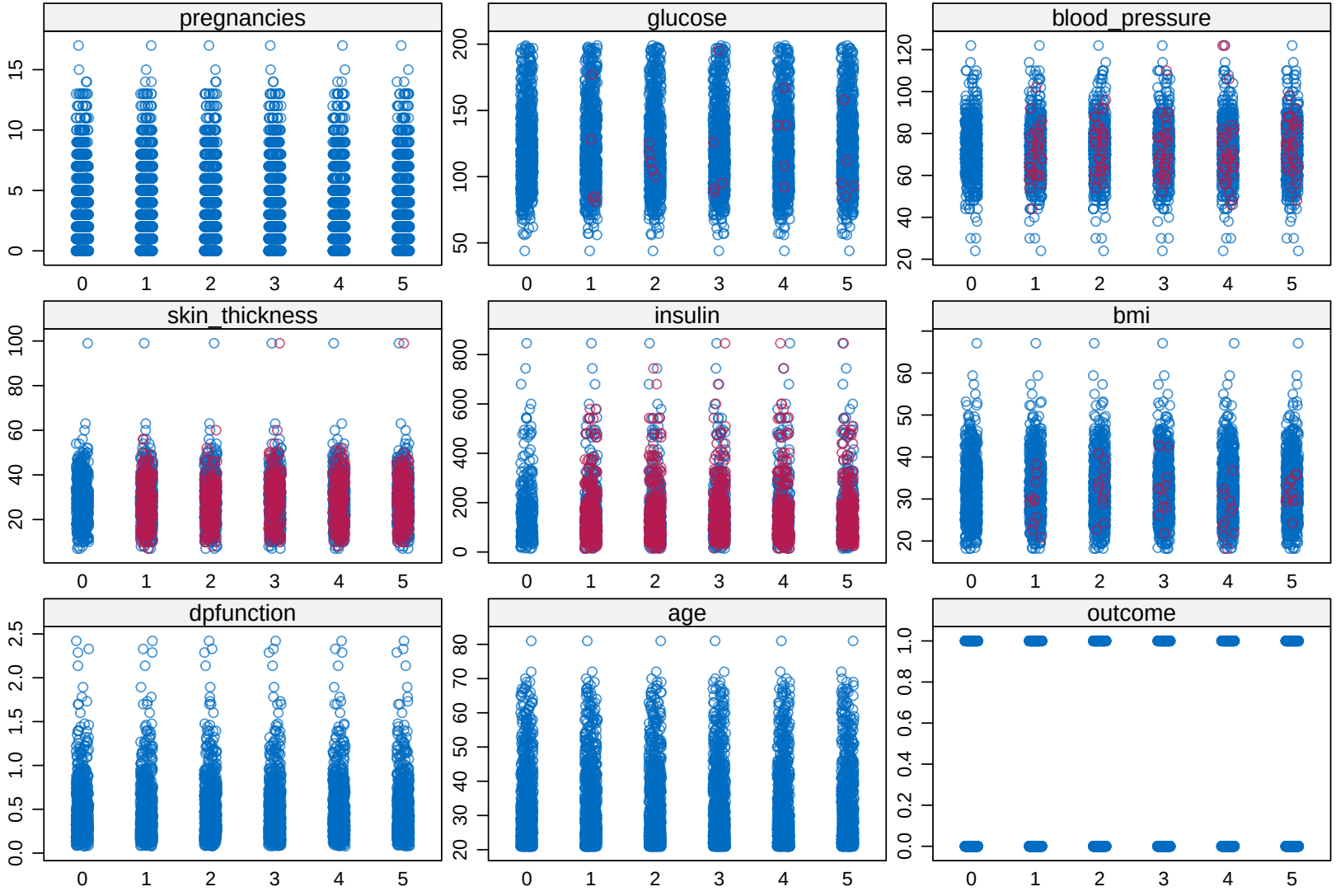
Descriptive Analysis

Histogram of Original Variables



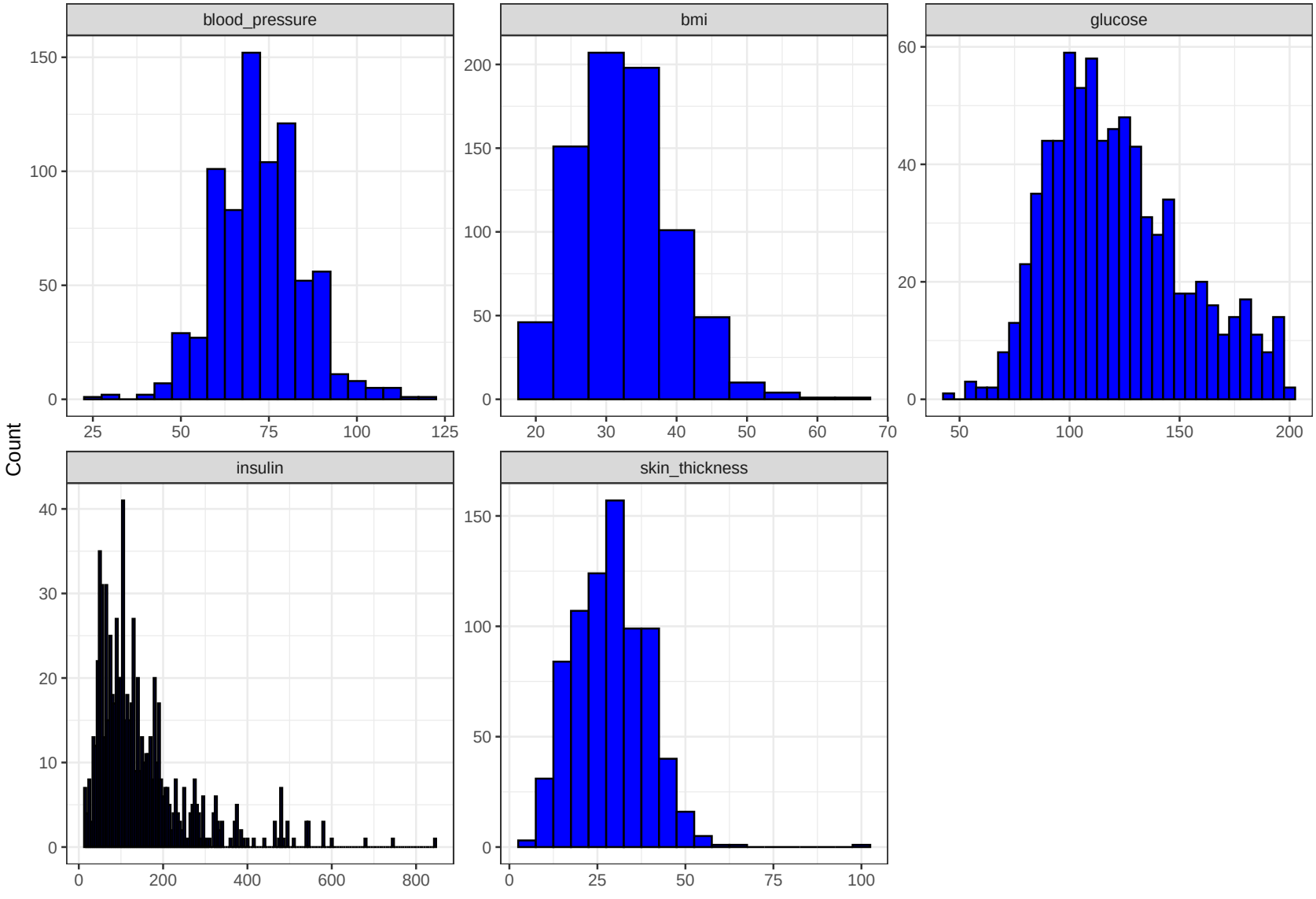
Descriptive Analysis

Stripplot



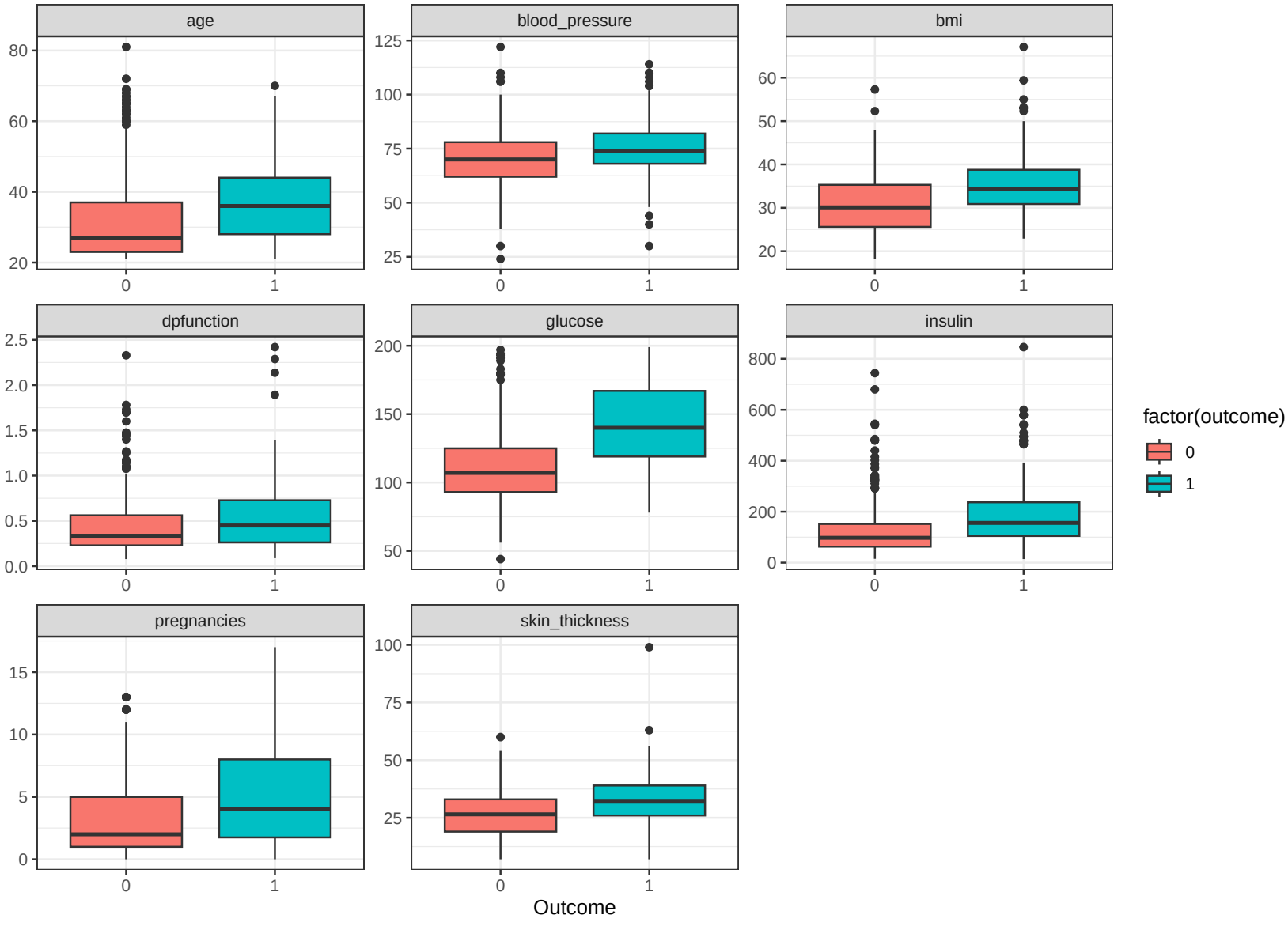
Descriptive Analysis | continue

Histogram of Variables



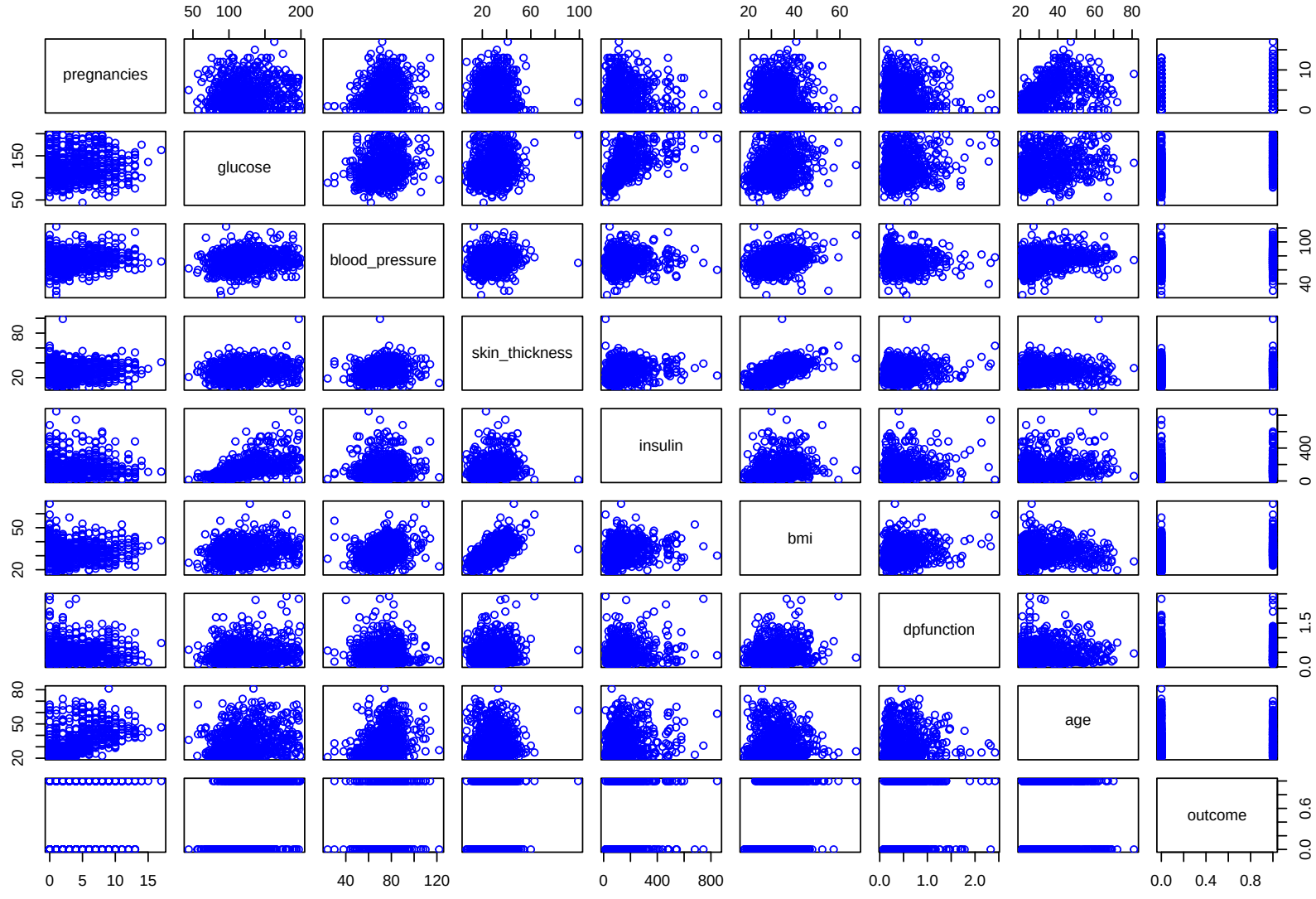
Descriptive Analysis

Boxplot of Variables by Outcome



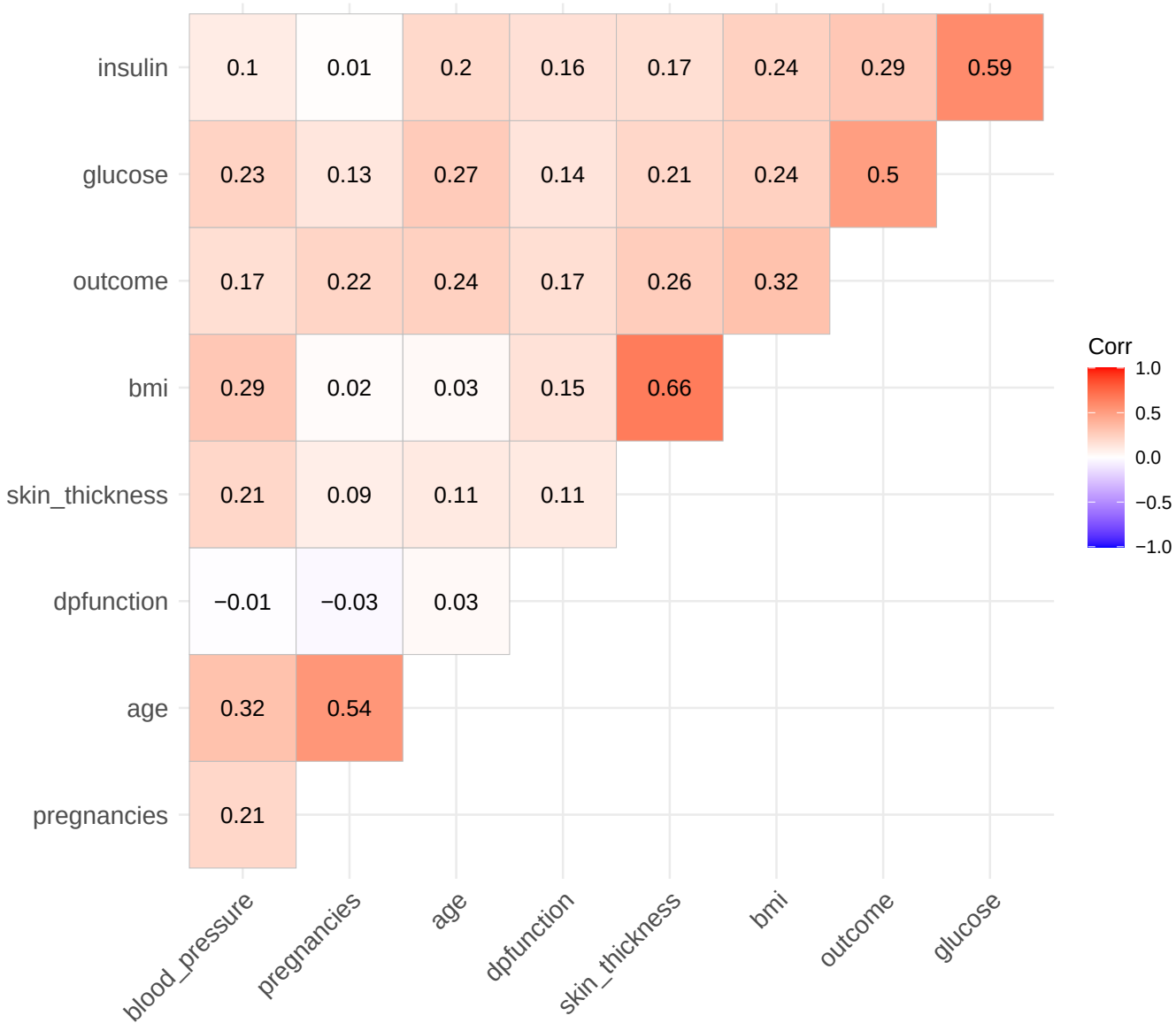
Descriptive Analysis

Pair Plot



Descriptive Analysis

Correlation Matrix

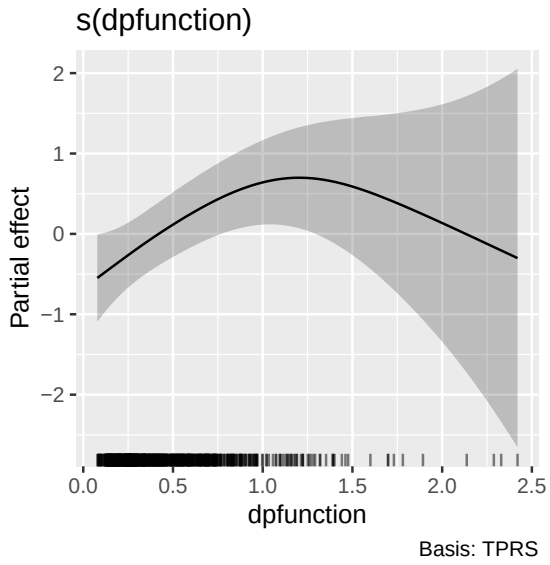


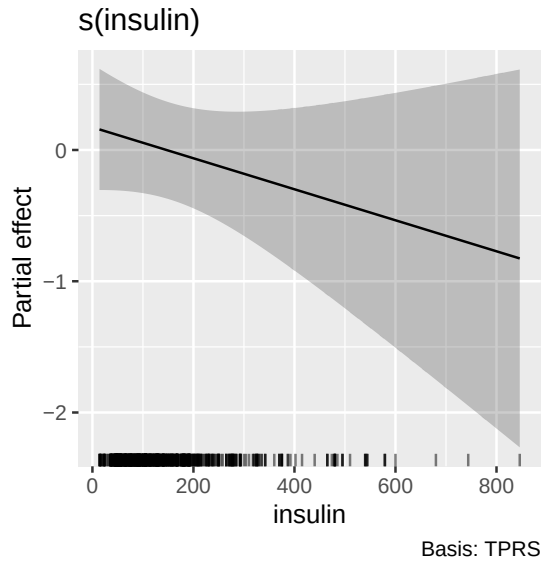
Model Fit

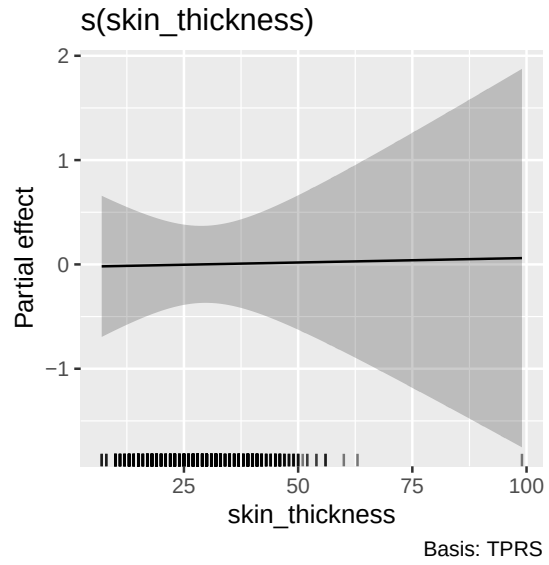
```
Family: binomial
Link function: logit

Formula:
outcome ~ s(dpfunction) + s(insulin) + s(skin_thickness) + s(blood_pressure) +
    s(pregnancies) + te(age, glucose, bmi)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2079     0.1884   -6.41 1.46e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                       edf Ref.df  Chi.sq p-value
s(dpfunction)        2.194  2.764  12.648 0.00381 **
s(insulin)           1.000  1.000   1.333 0.24834
s(skin_thickness)    1.000  1.000   0.004 0.94753
s(blood_pressure)    1.000  1.000   1.829 0.17631
s(pregnancies)       5.945  6.893   9.387 0.22874
te(age,glucose,bmi) 22.894 25.715 121.232 < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.381   Deviance explained = 36.4%
UBRE = -0.086284  Scale est. = 1         n = 768
```

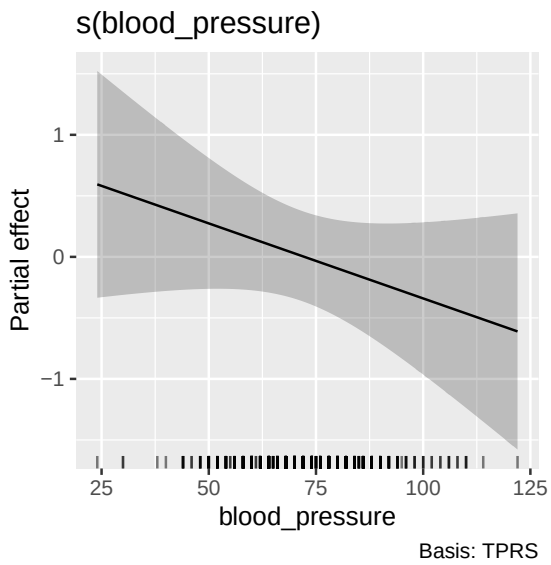
Comparative Study

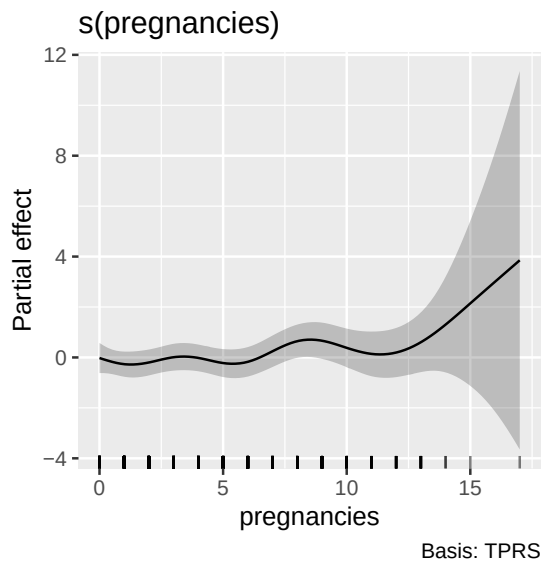
Partial-Effect Plots





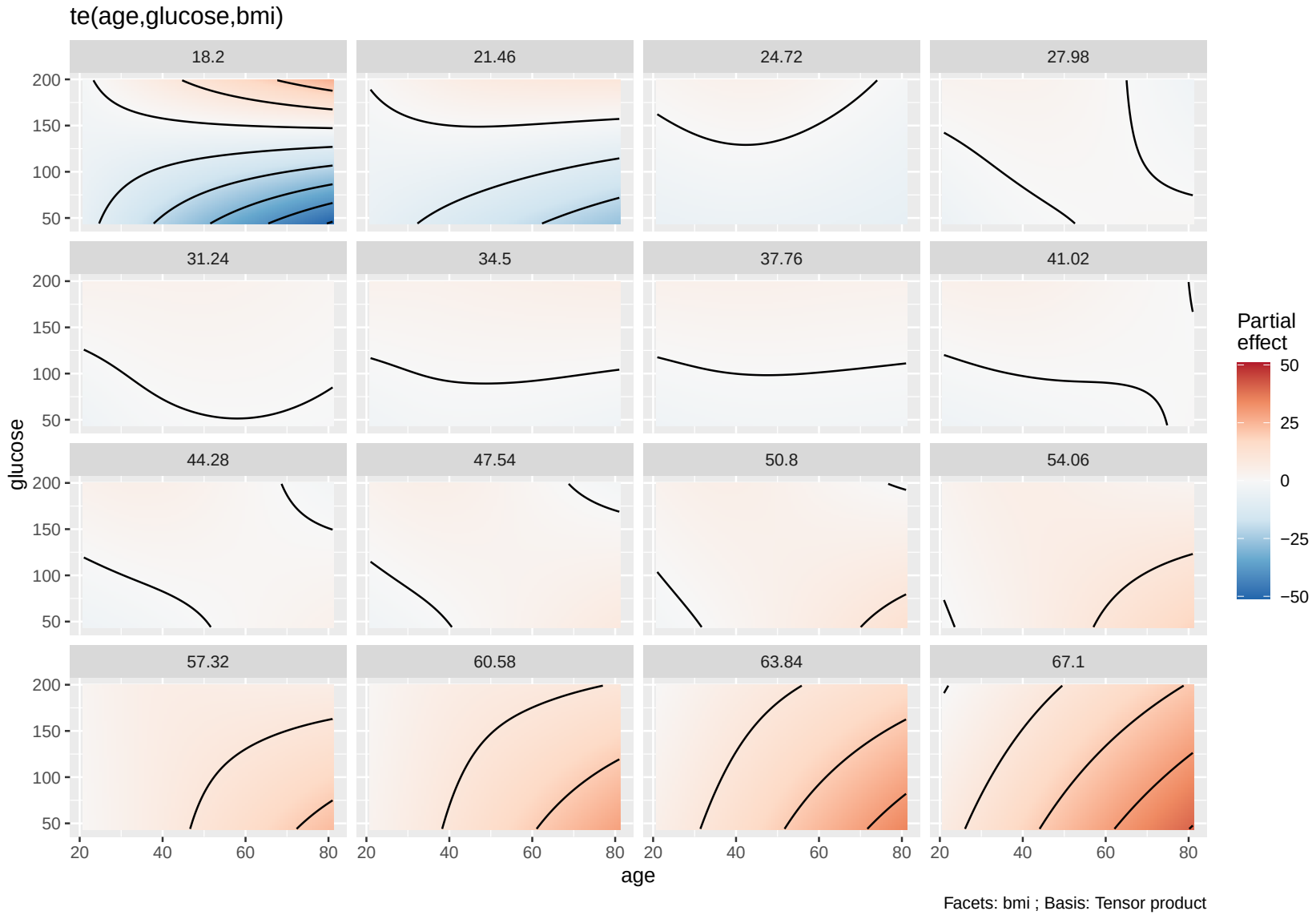






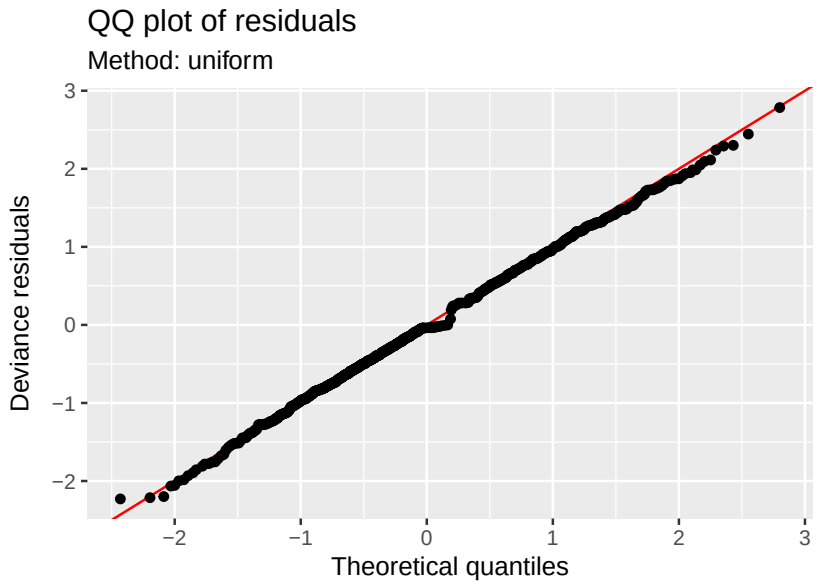
Comparative Study

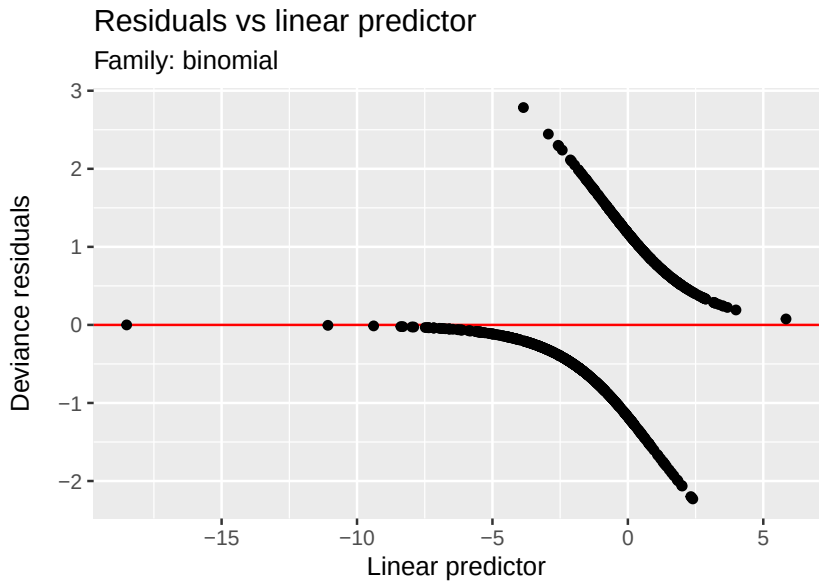
Tensor Product Plot

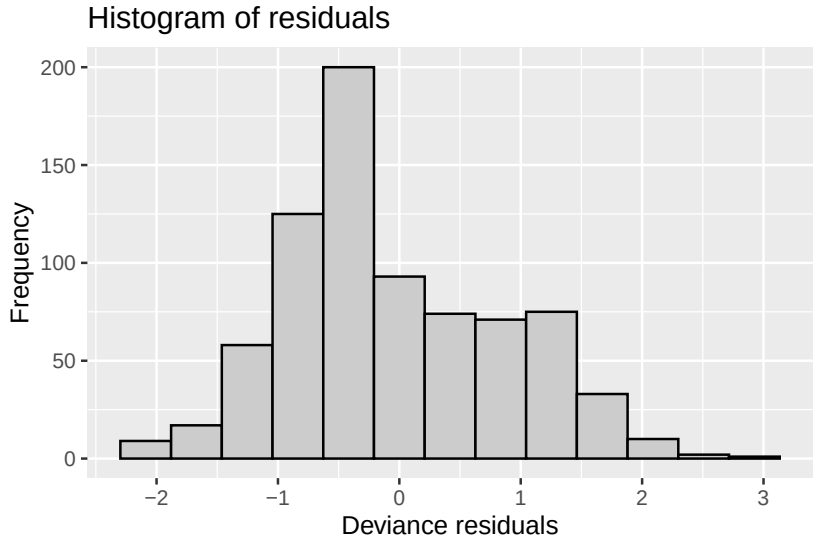


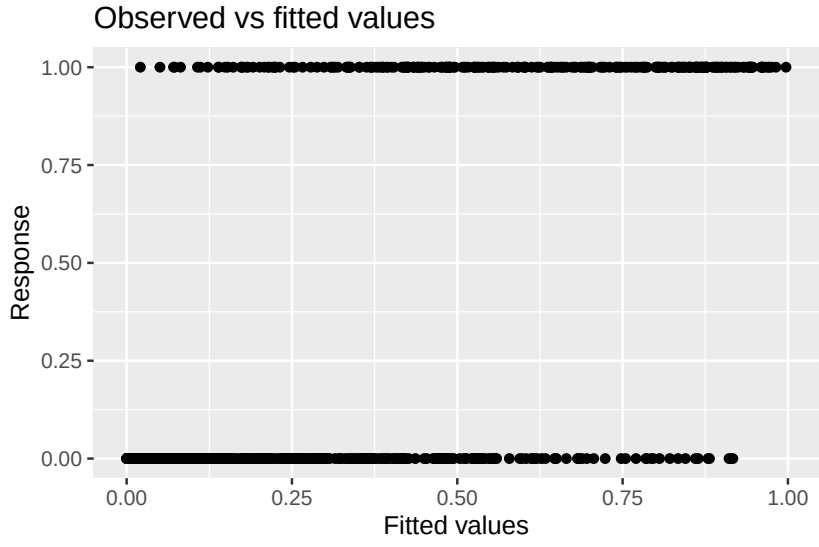
Comparative Study

Appraisal

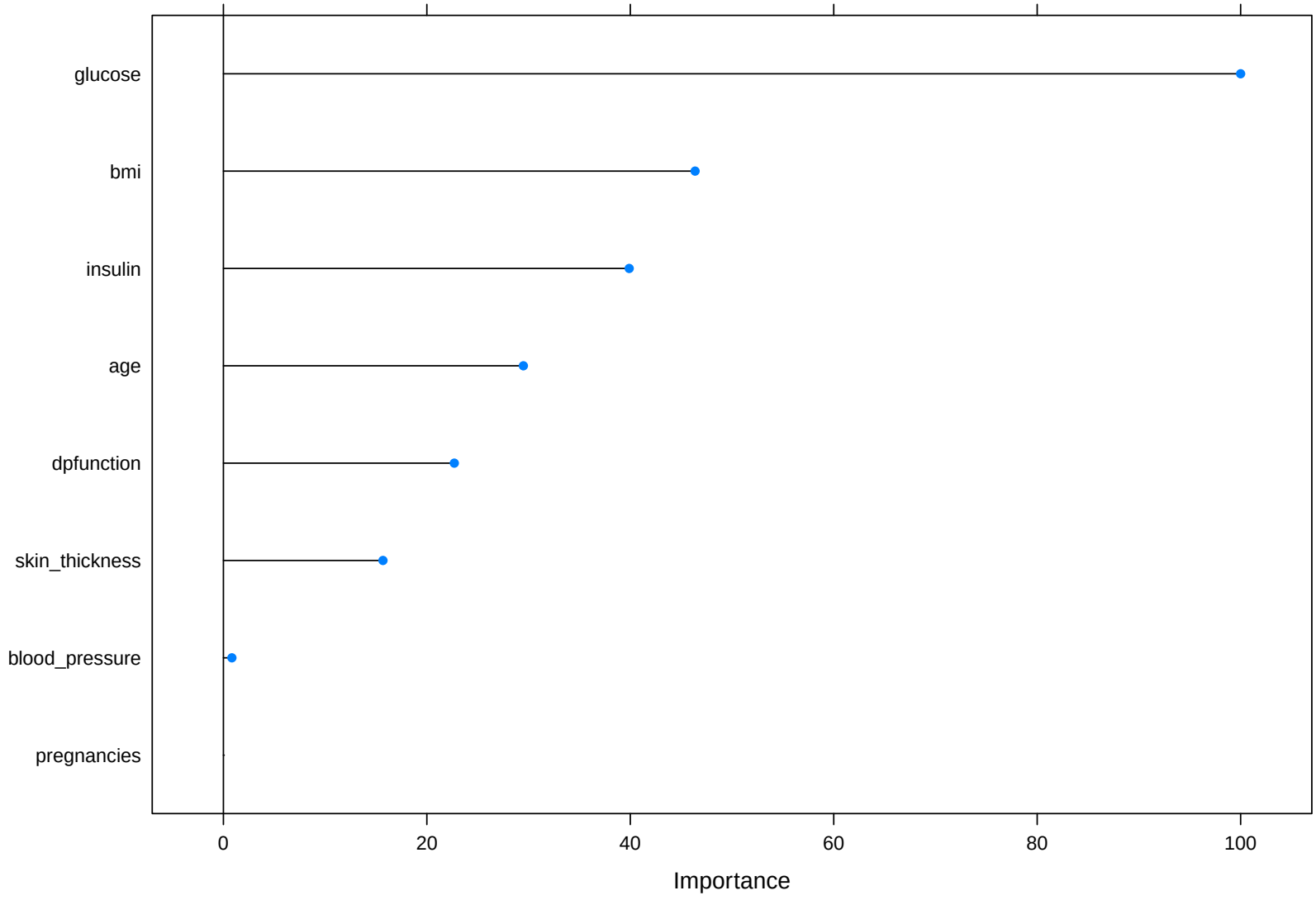








Random Forest Output



Comparative Study

Model	Diabetic_Misclass	Non_Diabetic_Misclass
GAMs	0.31250	0.0666667
NB	0.31250	0.1777778
RF	0.37500	0.1111111
SGB	0.40625	0.0888889
KNN	0.50000	0.1111111

References

- American Diabetes Association. (n.d.). What is Diabetes? https://www.diabetes.org/diabetes/what-is-diabetes
- World Health Organization. (2016). Global report on diabetes. https://www.who.int/publications/i/item/9789241565257
- American Diabetes Association. (n.d.). Diabetes Complications. https://www.diabetes.org/diabetes/complications
- Mayo Clinic. (2021). Type 1 diabetes. https://www.mayoclinic.org/diseases-conditions/type-1-diabetes/symptoms-causes/syc-20353011
- Centers for Disease Control and Prevention. (2021). Type 2 Diabetes. https://www.cdc.gov/diabetes/basics/type2.html

References

- Mayo Clinic. (2021). Diabetes. https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451
- Centers for Disease Control and Prevention. (2021). Risk Factors for Type 2 Diabetes. https://www.cdc.gov/diabetes/basics/risk-factors.html
- American Diabetes Association. (n.d.). Mental Health. https://www.diabetes.org/diabetes/mental-health
- Wood, S. N. (2017). Generalized Additive Models: An Introduction with R (2nd ed.). CRC Press.