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

Christopher Odoom, Denis Folitse, Sandani Kumanayake, Owen Gallagher

2023-05-25

Abstract

Diabetes is a worldwide epidemic that is one of the leading causes of kidney failure, heart attack, and strokes. While some Diabetes is, only genetic majority of people living with Diabetes have obtained the disease through poor lifestyle choices, such as inadequate nutrition and poor exercise habits. Therefore, if we could predict whether someone was at risk for Diabetes before they had the disease, we could alter their lifestyle to prevent it. We used data from the National Institute for Diabetes and Digestive and Kidney Diseases that only had women over 21 and of Pima Indian heritage. The dataset had missing values filled in using the MICE package in R using the Predictive Mean Matching method. This study uses a logistic General Additive Model (GAM) to understand the relationship between certain variables and Diabetes. Additionally, we use a three-way interaction tensor product between BMI, Age, and Glucose in Blood to help predict Diabetes. Finally, we test whether the GAM has more predictive power than typical Machine Learning models such as K-Nearest Neighbors, Random Forest, Naive Bayes, and Stochastic Gradient Boosting. The results suggest that the GAM model performed just as well or better than the Machine Learning models in predicting whether someone had Diabetes or not, and GAMs with a tensor product could be a useful tool going forward to predict Diabetes.

Background

Diabetes is an epidemic not only in America but in the entire world. The World Health Organization (WHO) is so concerned with Diabetes that it has not only launched a Global Diabetes Compact in 2021 but has also adopted goals towards controlling and preventing Diabetes (World Health Organization, 2023). The reason for such concern is the rising diabetes rates across the world. In 2021, approximately 537 million adults lived with Diabetes; that number is expected to be 643 million by 2030 and about 783 million by 2045 (International Diabetes Federation, 2021).

The WHO defines Diabetes as "a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces" (World Health Organization, 2023). There are two types of Diabetes, Type I, and Type II. An oversimplification is that genetic factors cause Type I while lifestyle habits cause Type II diabetes. In Type I Diabetes, the white blood cells in the pancreas attack the cells that make insulin. Thus, the person with Type I Diabetes does not have enough insulin to regulate their blood sugar levels. Type I Diabetes typically occurs early in someone's life. No one knows what causes Type I diabetes nor how to prevent it.

Unlike Type I Diabetes which occurs when white blood cells attack, Type II Diabetes occurs when the body cannot use insulin well enough to regulate blood sugar levels (Centers for Disease Control and Prevention, 2023). Type II diabetes typically occurs later in someone's life. It is caused by being overweight or physically inactive; some people can be at a higher risk rate because of genetics. Nonetheless, Type II diabetes is preventable if someone takes the counteracting effects such as exercising and eating healthy. What is most concerning about Type II diabetes is that while it is preventable, over 95% of people with Diabetes have Type II diabetes meaning people either do not know they are at risk or are ignoring the signs leading to health complications.

While Diabetes is typically not life-threatening, it is life-altering and can lead to serious long-term outcomes. For people with Type I Diabetes, insulin injection is necessary to live. While people with Type II diabetes typically take some form of medication, such as metformin, to regulate their blood sugar levels. Diabetes can also lead to more drastic outcomes. According to the WHO, "diabetes is a major cause of blindness, kidney failure, heart attacks, stroke, and lower limb amputation" (World Health Organization, 2023). Moreover, while unlikely, in severe cases, kidney damage can lead to death. Not only does Diabetes affect someone's health, but also their financial health as well.

To live with Diabetes, a patient either needs to take insulin or medication, both of which cost money, turning Diabetes into a massive business for pharmaceutical companies. The CDC reported that the total cost of Diabetes was 327 billion dollars in 2017. (CDC, 2022). At the time of this writing, sixty 500mg tablets of metformin, a medication used by people with Type II diabetes, cost $12.29. The Mayo Clinic recommends a starting dose of 1000mg daily; thus, over a year, the minimum spent would be $147.48 for someone with Type II diabetes. The price will increase if the person requires a higher than 1000mg dosage. While this may not seem like much in the CDC study, "adults with family income below the federal poverty level have the highest prevalence of diabetes" (CDC, 2022). Therefore, part of our motivation for our study is to hopefully catch possible diabetes patients before they get the disease, helping their physical and financial health.

Objectives:

The objectives of our study are to

1. Understand the relationship between the risk factors of Diabetes and their impact on getting Diabetes using a GAM logistic regression.
2. To consider the possibility of three-way interaction using tensor product smoothing, we can see the interactive effects of risk factors on Diabetes.
3. To compare the predictions from GAM fit with a selected best-performing machine learning algorithm for predicting Diabetes. The overall goal is that if we can build a model to catch people at risk for Diabetes early, medical professionals can guide the patient toward a healthier life and avoid the disease.

Methodology

Data Collection

The dataset used in this study was obtained from Kaggle, originally sourced from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset's objective is to predict the presence of diabetes based on diagnostic measurements Specifically, all individuals here are women who are at least 21 years of age and belong to the Pima Indian community. To ensure consistency, specific constraints were applied during the selection of instances from a larger database. The data set comprises of 8 variables and 768 observations. They are described in the appendix.

Handling Missing Values

Missing values in the dataset were addressed using the Predictive Mean Matching (PMM) method, which involves three steps. First, missing values were predicted using the Multivariate Imputation by Chained Equations (MICE) algorithm. Next, a predicted value close to the estimated value was selected for each missing sample. Finally, the relevant data points were chosen from the original dataset, excluding missing values.

Prediction Methods

The analysis employed five primary prediction methods to classify individuals with or without diabetes. These methods include:

Generalized Additive Models (GAMs):
GAMs are a flexible regression modeling technique that allows for nonlinear relationships between predictors and the response variable. They can capture complex patterns in the data by incorporating smoothing functions.

Tensor Product (Trivariate fits): This method utilizes trivariate functions to model the relationships between three predictor variables simultaneously. It extends the capabilities of traditional regression models by capturing interactions and nonlinear effects more effectively.

Random Forest (RF):
RF is a machine-learning algorithm that constructs multiple decision trees and combines their predictions to achieve a more accurate classification. It is particularly effective for handling high-dimensional data and capturing complex interactions between variables.

Naive Bayes (NB):
NB is a classification algorithm based on Bayes' theorem, assuming independence among predictors. It calculates the probability of each class based on the predictor values and assigns the observation to the class with the highest probability.

Stochastic Gradient Boost (SGB):
SGB is a machine learning algorithm that sequentially trains an ensemble of weak prediction models, typically decision trees, by correcting errors made by previous models. The final prediction is obtained by weighted averaging of the individual models' predictions.

K-Nearest Neighbors Algorithm (KNN):
KNN is a non-parametric classification algorithm that assigns a new data point to the class of its majority k-nearest neighbors in the training set. It determines the class based on the similarity of the new observation to its neighboring points in the feature space.

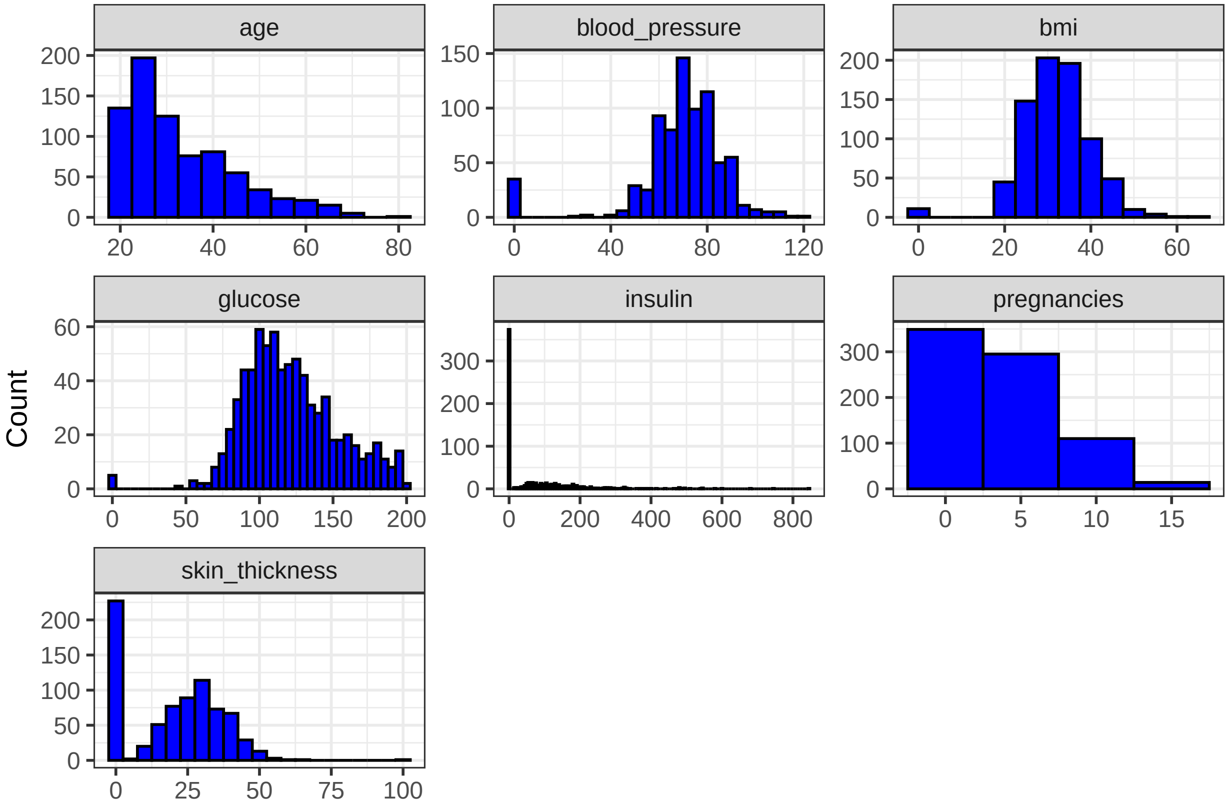
Model Evaluation

The performance of the prediction models was evaluated using misclassification rates, specifically focusing on two types: Diabetic_Misclass (identifying diabetic individuals as having non-diabetes) and Non_Diabetic_Misclass (identifying non-diabetic individuals as diabetic).

ANALYSIS AND RESULTS

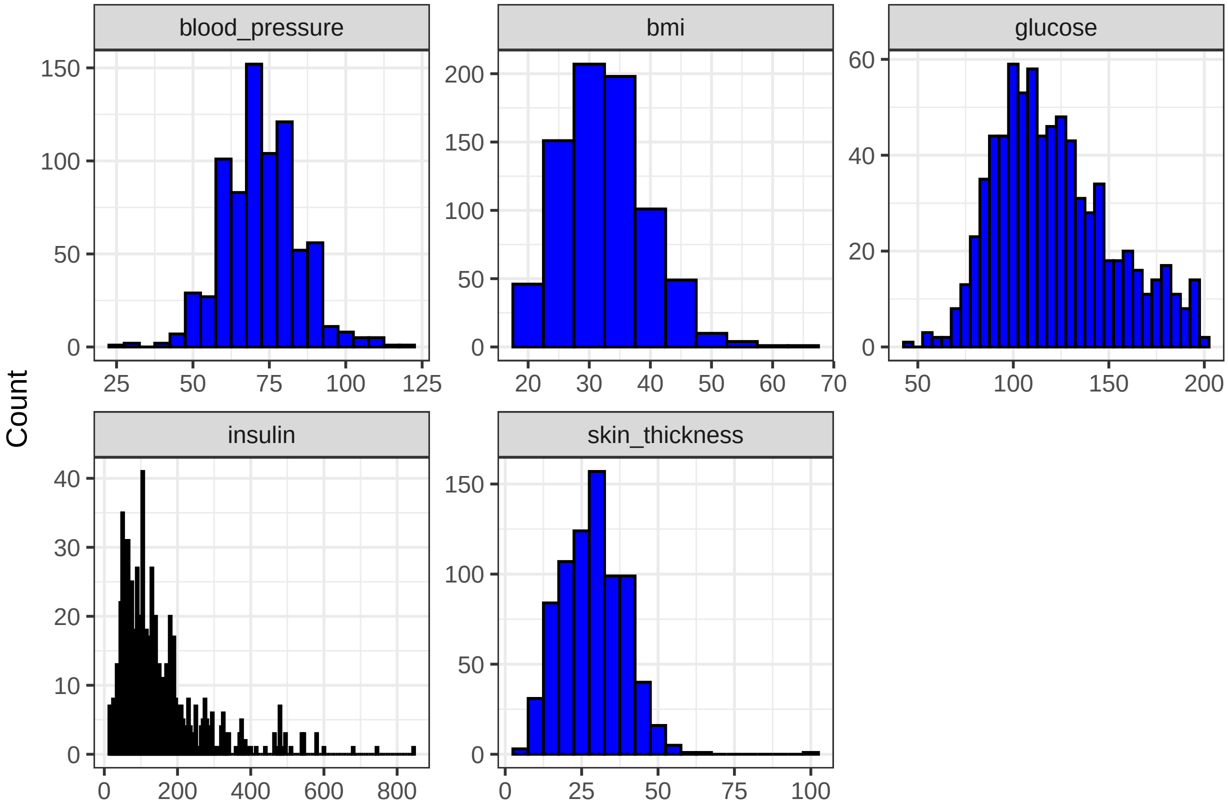
Desrcriptive Analysis

Histogram of Original Variables



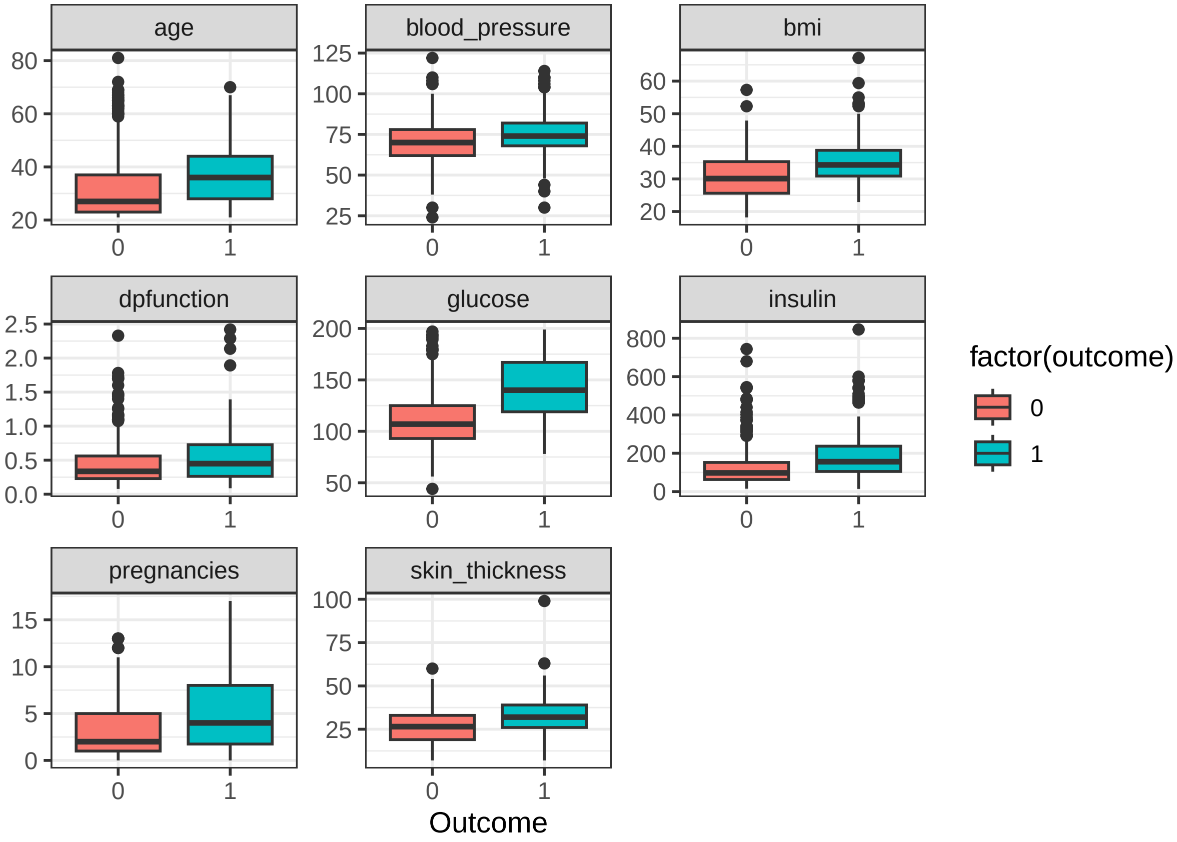
The plots above show the distribution of the variables to be used in our analysis. As Can be seen, variables such as blood pressure, bmi, glucose, insulin and skin thickness contains a lot of 0's. The 0's here are not meaningful and hence are considered as missing values.It is impossible to have skinless human beings and even observe people with zero body mass Index. We therefore imputed for this values using the Mice package (method=pmm)

Histogram of Variables



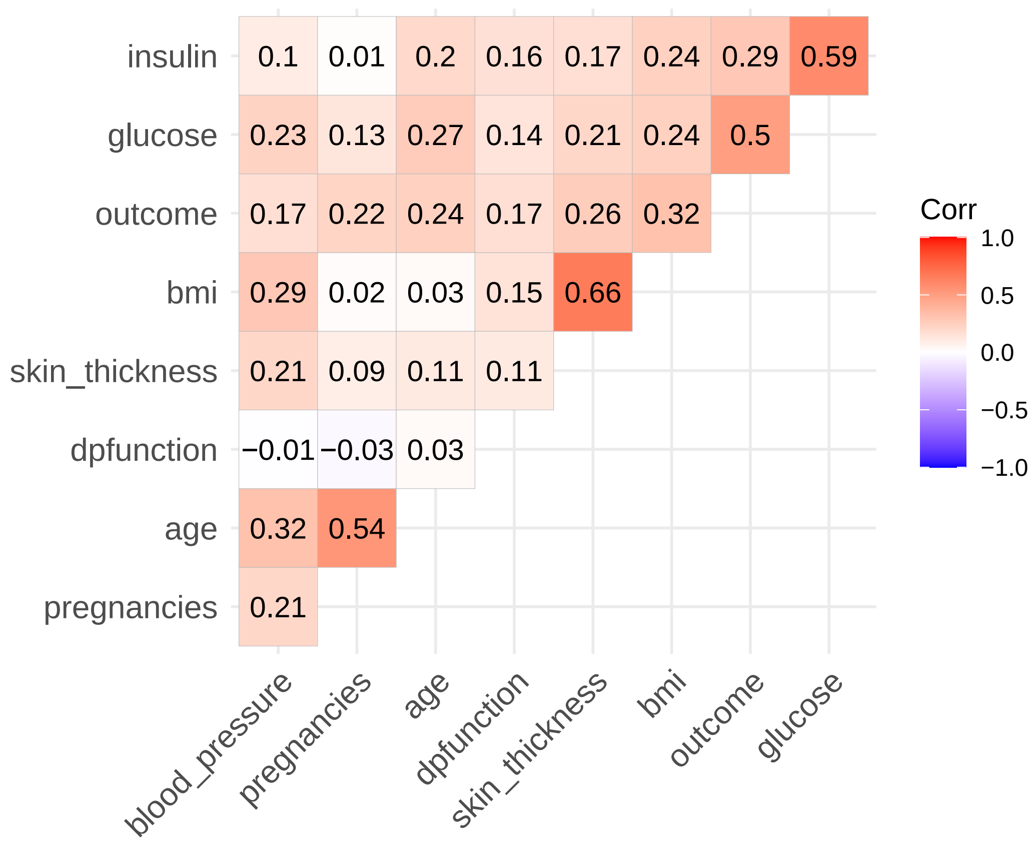
The Histograms above are for the various variables after imputing for the missing values. A careful look at the plot shows that the distribution remains the same as the original data (containing missing values). More importantly, the non-meaningful 0's are no longer a problem. The histogram was created using one of the imputed dataset, these plots are consistent with the four remaining datasets.

Boxplot of Variables by Outcome



The first plot here indicates that, the ages of diabetic people are averagely higher than that of non-diabetic patients with some few exceptions. As can be seen, diabetic people has an averagely high amount of body mass index, glucose, insulin and skin thickness than non diabetic patients. They also tend to carry a lot more pregnancies.

Correlation Matrix



We explored the correlation among the variables using the correlation matrix above. In terms of Direction, dpfunction has a negative relationship with blood pressure and pregnancies. Regardless, the magnitude indicates that there is no linear relationship between them.The largest correlation coefficient recorded is .66 (body mass index vs skin thickness). This shows that there is a moderate linear relationship between these two variables. Overall, the matrix shows that multicollinearity is not something to worry about.

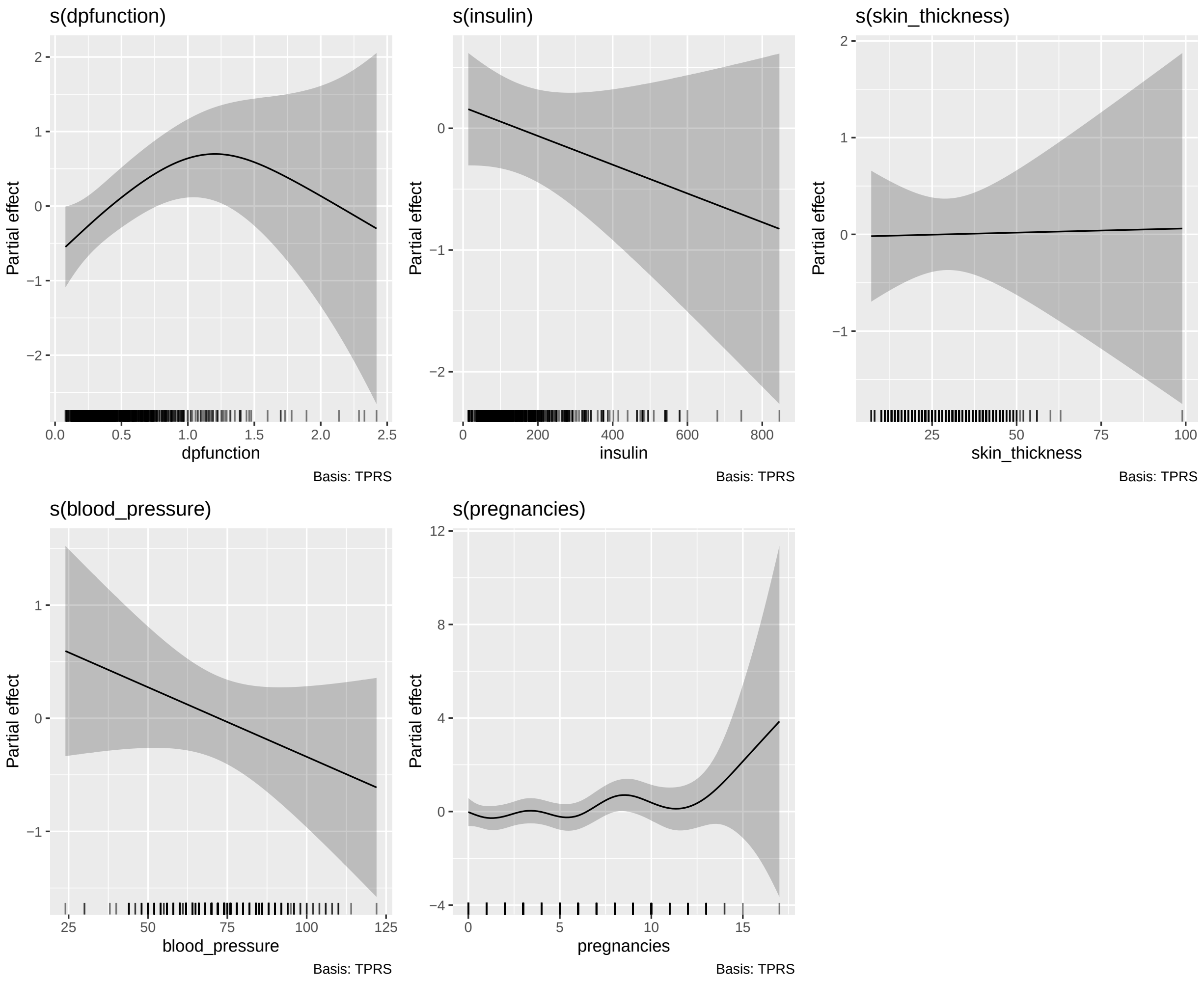
Model Fit | GAMs splines

Table 1: GAMs Spline Trivariate fit

	edf	Ref.df	Chi.sq	p-value
s(dpfunction)	2.194	2.764	12.65	0.003808
s(insulin)	1	1	1.333	0.2483
s(skin_thickness)	1	1	0.004375	0.9475
s(blood_pressure)	1	1	1.829	0.1763
s(pregnancies)	5.945	6.893	9.387	0.2287
te(age,glucose,bmi)	22.89	25.71	121.2	0

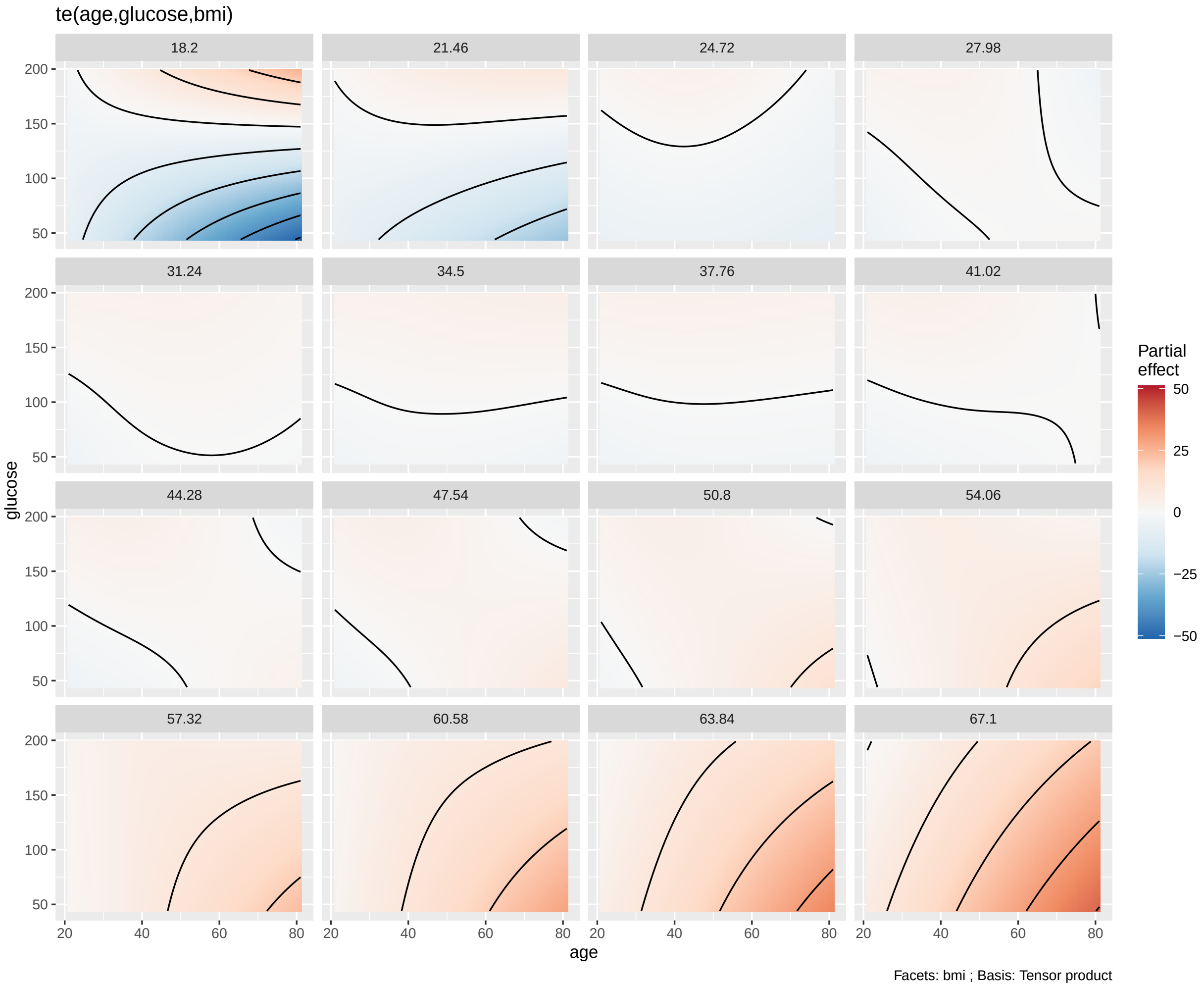
The table above reports the GAMs Spline trivariate fit (tensor product). In this fit, all variables are fitted as a spline term with age, glucose and bmi entering the model as trivariate tensor product. The s(dpfunction) and te(age,glucose,bmi) are both highly significant in the model.The model above has a deviance explained of about 37%.

Partial-Effect Plots



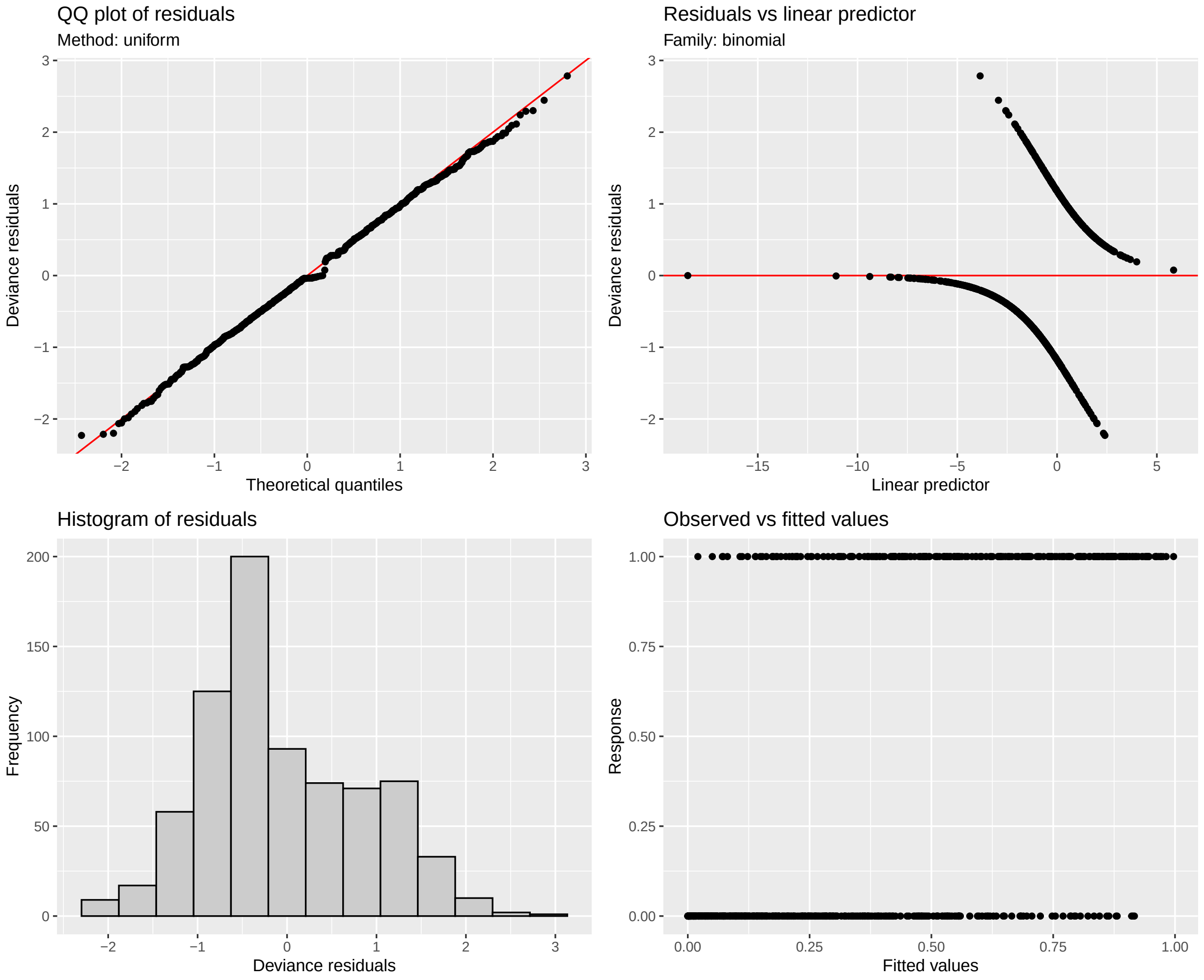
In exploring the partial effects of these variables on diabetes, we can see that blood pressure and insulin have negative linear relationship with Outcome with skin thickness having close to no relationship. This indicates that the more insulin the subjects have in their body, the less likely it is to have diabetes. At the larger values of the insulin, the variability band is quite large. This is similar for blood pressure. People with a very low blood pressure are more likely to have diabetes and vice versa. dpfunction has a curvilinear effect on diabetes. This effect increases as dpfunction increases, hit a peak at dpfunction of 1.25 and then decreases. This shows people with a dpfunction between 1 and 1.5 are more likely to have diabetes. Pregnancies from 1 to 12 tends to have close to 0 effect on diabetes. But this effect becomes positive for pregnancies above 12. People who carry pregnancies above 12 are more likely to have diabetes (This comes with a large variability band)

Tensor Product Plot



This plot indicates that at a bmi of 18.2, people above 50, with glucose level above 160 are likely to have to have diabetes. This is rather opposite for this same group of people with glucose level below 100 while all others with a bmi of 18.2 has close to no relationship with diabetes. These effects seem the same for people with a bmi of 21.46 (only less likly for each of the groups). People with bmi of 24.72 across all ages and glucose level tends to be less prone to diabetes. As bmi increase from 27.96 to 47.54, there seems to be close to no relationship for all glucose level and age (this effect increases bits by bits as bmi increases).People above 40 with bmi's between 50.8 to 67.1, regardless of their glucose level are more likely to have diabetes. This effects increases as bmi is increases,glucose decreases and age increases. People who are more likely to have diabetes are people with bmi of 67.1 aged above 60 and have glucose below 100.

Appraisal


QQ plot of residuals

Method: uniform

Residuals vs linear predictor

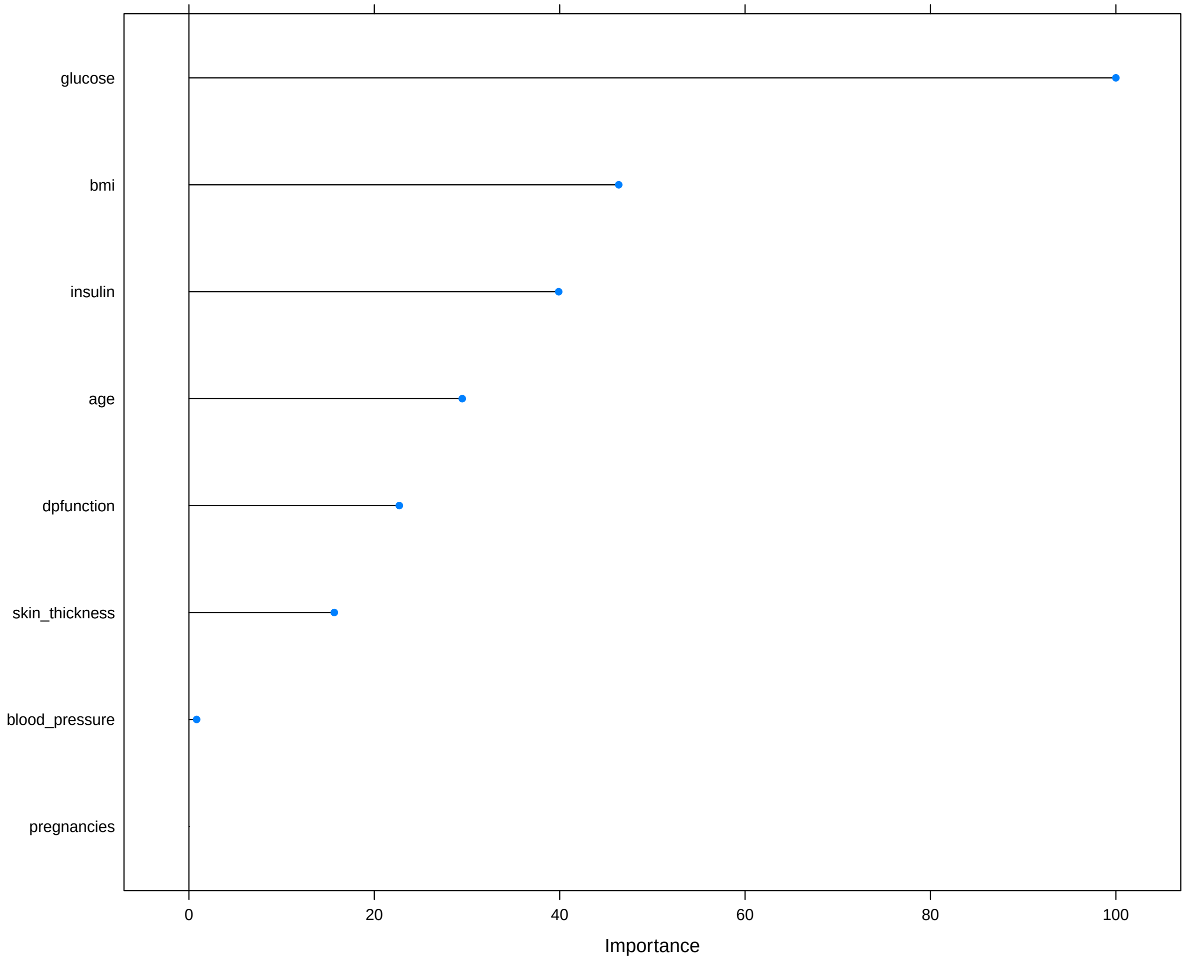
Family: binomial

Histogram of residuals

Observed vs fitted values

The QQ plots residual and histogram of residuals indicates that our residuals does not deviate too much from normality hence we can proceed with our model.

Random Forest Output



The plot above present the order of importance of the variables as used by the Random forest algorithm. The plot shows that glucose level has the largest effect on outcome. It also picked body mass index, insulin and age respectively. The plot indicates that pregnancies is not considered as important in classifying whether a person is diabetic or not.

Result

Table 2: Model Evaluation

Model	Diabetic_Misclass	Non_Diabetic_Misclass
GAMs	0.3125	0.06667
NB	0.3125	0.1778
RF	0.375	0.1111
SGB	0.4062	0.08889
KNN	0.5	0.1111

We divided the data into training and testing(80%-20% respectively). We then trained this data using the the GAMs, NB, RF, SGB and KNN.We applied this model to the test data. The table above shows the performance of each of these models on the data. GAMs and NB have the best performance rate for Diabetic Missclassification. This indicates that, these two models classified only 31% of diabetic people as non-diabetic. The KNN performed the worst with a missclassification rate of 50%. Interestingly, GAMs performed much better in terms of false positive classification. It missclassified only 6% of non diabetics as diabetic. The NB performed the worst (11%).

Conclusions

We drew the following conclusions from the analysis

1. The trivariate spline term tends out to be very significant and useful in predicting the incident of diabetes.
2. Not all the variables turned up to be a useful predictor for diabetes, such as number of pregnancies, insulin concentration, and skin thickness. This result is not consist with ideal situations, however, this finding is possible.
3. Using Gam fit as a predictive model is appropriate since the study shows how well it competed with traditional classification algorithms.

Further Studies

- Study the influence of different methods of Multiple Imputations by chained equations (mice) on the GAMs result.
- Pool results from GAM fits on multiple mice-imputed dataset

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
- 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.

Appendices

Data Description

- ***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
- ***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