Binary Classification Models on Diabetes Data

Anasah Wawem Christopher

January 12, 2026

Contents

Abstract

This report provides a comparative analysis of diabetes data utilizing two two basis-expansion classifications. The report covers data structure, data wrangling, preprocessing, and analysis. This study aims to build binary regression models using Generalized Additive Model (GAM) with splines smoothers (natural splines) and Radial-Basis Function (RBF) expansion with penalized logistic regression and compare these models to determine which one best fits the diabetes data.

1 Introduction

Diabetes in females is a multifactorial disease influenced by myriad interacting factors, from glucose levels and insulin sensitivity to BMI, age, whether the person is pregnant or not, and even skin thickness. It is not good enough to just have the data, and there is a need for reliable ways to estimate a person's probability of developing the disease. Although classical statistical models, such as logistic regression, are useful, they usually assume linear relationships that may be oversimplifications of real-world patterns and lead to misleading predictions.

To better reflect this nuanced, often nonlinear relationship in health data, researchers have turned to more flexible techniques, such as basis function models. Herein, we compare two such approaches: the Generalized Additive Model (GAM) and Radial Basis Function (RBF) expansion with penalized logistic regression. We will test them on real-world diabetes data to see which one provides clearer insight with stronger predictive performance in the clinical setting.

2 Data Structure

2.1 Data Overview

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases downloaded from kaggle.com. The data was generated from only females patients and at least 21 years old of Pima Indian heritage. There are eieght medical predictor variables and only one target dependent variable (Outcome) recorded on 768 patients. There are about 268 1s (patients with diabetes) and 500 0s (patients without diabetes). Information about dataset attributes is summarized in table1.

Table 1: Dataset Structure and Variable Description

Variable	**Type**	**Unique Values**	**Description**
Pregnancies	Integer	17	To express the Number of pregnancies
Glucose	Integer	136	To express the Glucose level in blood
BloodPressure	Integer	47	To express the Blood pressure measurement
SkinThickness	Integer	51	To express the thickness of the skin
Insulin	Integer	186	To express the Insulin level in blood
BMI	Numeric	248	To express the Body mass index
DiabetesPedigreeFunction	Numeric	517	To express the Diabetes percentage
Age	Integer	52	To express the age
Outcome	Integer	2	To express the final result 1 is Yes and 0 is No

3 Data Preprocessing

3.1 Missing Value and Treatment

There are no missing values in the data set. However, variables such as Glucose, BloodPressure, BMI, and Insulin are possibly to have zero as an entry, which is not medically acceptable. The zeros in these variables have been replaced with "NA" and were imputed after spliting the data using the median.

3.2 Feature Engineering

The following adjustments were made:

- The response variable (output) was converted to factor.
- All the predictors were scaled using z-scores.
- The data was partitioned into a ratio of 80:20 for the train and test datasets, respectively, using stratification.

4 Analysis And Results

4.1 Correlation Analysis

Figure 1 shows how each numeric predictor is linearly associated with other predictors. There is a moderate positive correlation between SkinThickness and BMI, which means as an individual's skin thickness increases, their BMI also increases. Glucose and Insulin levels are moderately correlated positively. The higher the patient's glucose level, the higher the insulin level. Also, Age and Pregnancies have a positive moderate correlation. Older women tend to have more pregnancies than younger women. The rest of the variables show weak correlation. There is no strong correlation, hence no extreme multicollinearity, which is beneficial for GAM and RBF.

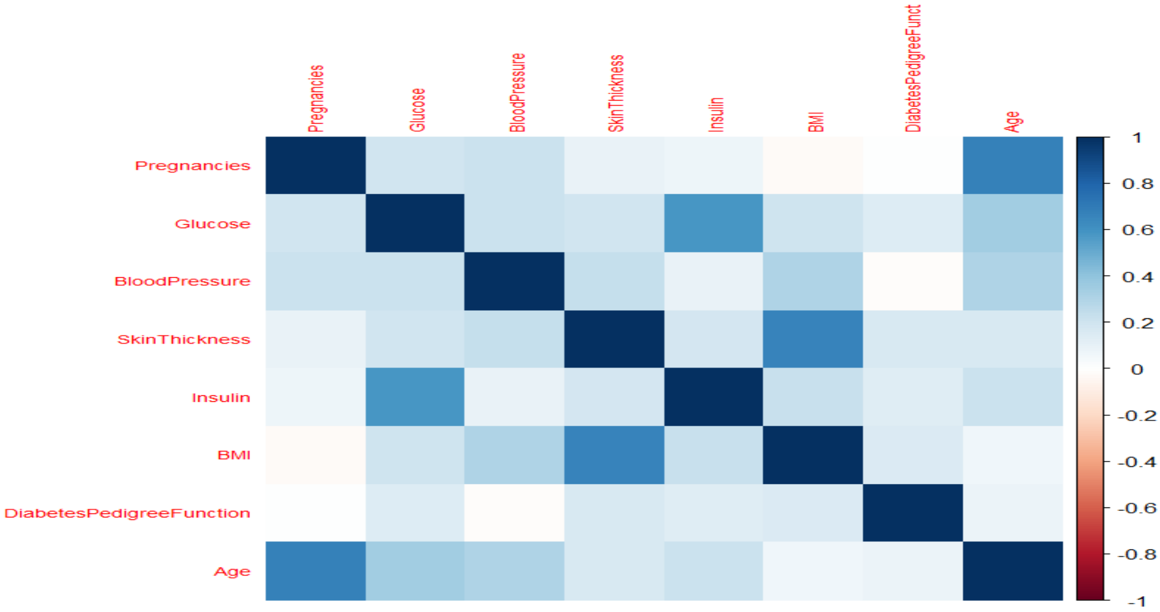

Figure 1: Correlation Plot of Predictors

4.2 Visualization of Predictors

Figure 2 shows the distribution of predictors across different physiological categories, which is representative. Glucose is right-skewed, which means there is a very high level of glucose among some patients. This clearly distinguishes diabetic from non-diabetic patients. BloodPressure is approximately normal and centered around 70 to 80. Pregnancies is highly right skewed. Most women have between 0 and 5 pregnancies, with a few outliers with over 15 pregnancies. SkinThickness is roughly unimodal. Some values are low, which are possibly missing or imputed. The dataset's skewness and non-Gaussian distributions support the use of basis-function models, such as GAM or RBF, which are superior to simple logistic regression at capturing nonlinear effects.

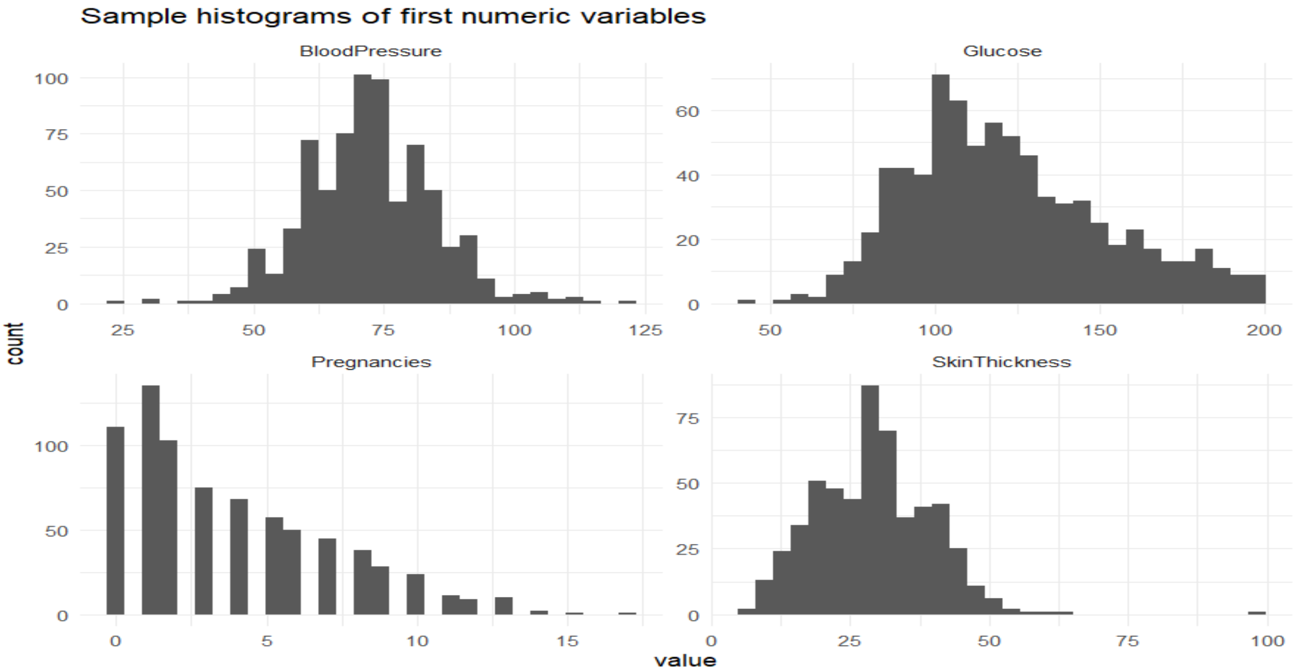

Figure 2: Histogram Plot of Some Predictors

4.3 Generalized Additive Model with Natural Splines

Table 2 presents the summary of the generalized additive model using natural splines. The optimal basis dimension was selected at k=10 via nested cross-validation, where the inner resampling with 5-folds was used to compare the performance of four candidate k values (k = 4, 6, 8, 10) based on the training dataset. Glucose, BMI, DiabetesPedigreeFunction, BloodPressure, and Age all have significant non-linear effects on the risk of developing diabetes (p-value <0.05). Glucose and BMI have the biggest contributions with high values of EDF, which indicates a flexible non-linear pattern. SkinThickness and Insulin did not have any meaningful effects after smoothing penalties were applied. This model explains 33.4% of the deviance, with an adjusted R^2 of 0.368, indicating the model is nonlinearly strongly related to the response variable in a moderate sense.

Table 2: Summary of Generalized Additive Model (GAM)

Term	**EDF**	**Ref.df**	**Chi-Square**	**p-value**
Intercept	–	–	–	1.76×10^{-15}
s(Pregnancies)	0.594	9	1.427	0.1154
s(Glucose)	3.033	9	83.068	$< 2 \times 10^{-16}$ ***
s(BloodPressure)	0.827	9	4.573	0.0181 *
s(SkinThickness)	0.000	9	0.000	0.5215
s(Insulin)	0.327	9	0.483	0.2214
s(BMI)	4.257	9	32.455	1.62×10^{-6} ***
s(DiabetesPedigreeFunction)	1.613	9	7.639	0.0063**
s(Age)	2.523	9	19.012	2.71×10^{-5} ***
Model Information:				
Family: Binomial; Link: Logit				
Adjusted R^2: 0.368; Deviance explained: 33.4%				
REML score: 289.82; Scale estimate: 1; $n = 615$				

Table 3 shows the confusion matrix and the performance metrics of the GAM with natural splines. The GAM model shows excellent performance in classification for the test dataset using Youden's J statistic as the threshold. The overall accuracy is 82.35%, which is way higher than that from the No Information Rate, 65.36%, which was verified by a very small p-value of 0.000002528. That means the model is doing a much better job than always predicting the majority class. The model shows high sensitivity (76.00%), with most of the diabetic cases correctly identified as belonging to class 1 and a very high specificity of 94.34%, indicating this model has a strong ability to detect the non-diabetic cases as class 0.

The kappa value of 0.6435 here suggests moderate agreement between the predicted and actual classes, considering class imbalance. The balanced accuracy is 0.8517, which shows good overall discriminative ability when both classes are weighted equally.

Table 3: Confusion Matrix and Performance Metrics for GAM Model

Metric	Value
True Negatives (TN)	76
False Positives (FP)	24
False Negatives (FN)	3
True Positives (TP)	50
Accuracy	0.8235
95% CI for Accuracy	(0.7537, 0.8804)
No Information Rate (NIR)	0.6536
P-Value [Acc > NIR]	0.000002528
Kappa	0.6435
Mcnemar's Test p-value	0.0001186
Sensitivity	0.7600
Specificity	0.9434
Positive Predictive Value (PPV)	0.9620
Negative Predictive Value (NPV)	0.6757
Prevalence	0.6536
Detection Rate	0.4967
Detection Prevalence	0.5163
Balanced Accuracy	0.8517

4.4 Radial Basis Function Expansion with Penalized Logistic

Table 4 provides a summary of the RBF expansion with penalized logistics using Gaussian radial basis functions centered on k-means cluster centroids. RBF centers and bandwidth used are (k = 10, 20, 30) and kernel bandwidth is (γ = 0.01, 0.05, 0.1, 0.5), respectively, and a grid search is used to select the optimal number. The RBF-transformed features are input to L1-penalized regression, and the model performance was evaluated using 5-fold cross-validation AUC. The RBF kernel model yields good performance with k = 10 and $\gamma = 0.05$, and its CV AUC of 0.8381 shows good discrimination ability.

Table 4: Summary of Best Radial Basis Function Kernel Model

Model Setting	Value
Number of RBF Centers (k)	10
Gamma (γ)	0.05
Cross-validated AUC	0.8381

Table 5 shows the confusion matrix and the performance metrics of the RBF wih penalized regression model. From the performance using Youden's J statistic as the threshold, the model shows strong but imbalanced classification performance with 81.05% accuracy and a kappa of 0.62 (moderate agreement beyond chance). It identifies most of class 0 cases correctly (specificity 92.45%) and also identified a quite number of class 1 cases correctly (sensitivity 75.00%). The significant McNemar's test, p = 0.0002, indicates a systematic bias between error types that is class 1 misclassified as class 0 considerably more often than class 0 misclassified as class 1. This is possibly due to class imbalance.

Table 5: Confusion Matrix and Performance Metrics for Best RBF Model

Confusion Matrix	
True Negatives (TN)	75
False Positives (FP)	25
False Negatives (FN)	4
True Positives (TP)	49
Performance Metrics	
Accuracy	0.8105
95% CI	(0.7393, 0.8692)
No Information Rate	0.6536
P-Value [Acc > NIR]	0.0000146
Kappa	0.6171
Mcnemar's Test P-Value	0.0002041
Sensitivity	0.7500
Specificity	0.9245
Positive Predictive Value (Precision)	0.9494
Negative Predictive Value	0.6622
Prevalence	0.6536
Detection Rate	0.4902
Detection Prevalence	0.5163
Balanced Accuracy	0.8373

4.5 Model Comparison

Figure 3 shows the plot of the Receiver Operating Characteristic Curve (ROC) of RAM and RBF models on the test set. Both curves are close to the top-left corner, indicating high true positive rates and low false positive rates, with the RBF + glmnet curve slightly below GAM in some regions. Both models exhibit satisfactory performance, suggesting that nonlinear modeling is beneficial for the dataset. However, compared to RBF + glmnet, GAM is marginally stronger, capturing more complex relationships. The small differences in sensitivity-specificity trade-offs between the modeling methodologies are highlighted by the curves' minor divergence. Compared to random guessing, both seem significantly better (the diagonal reference line).

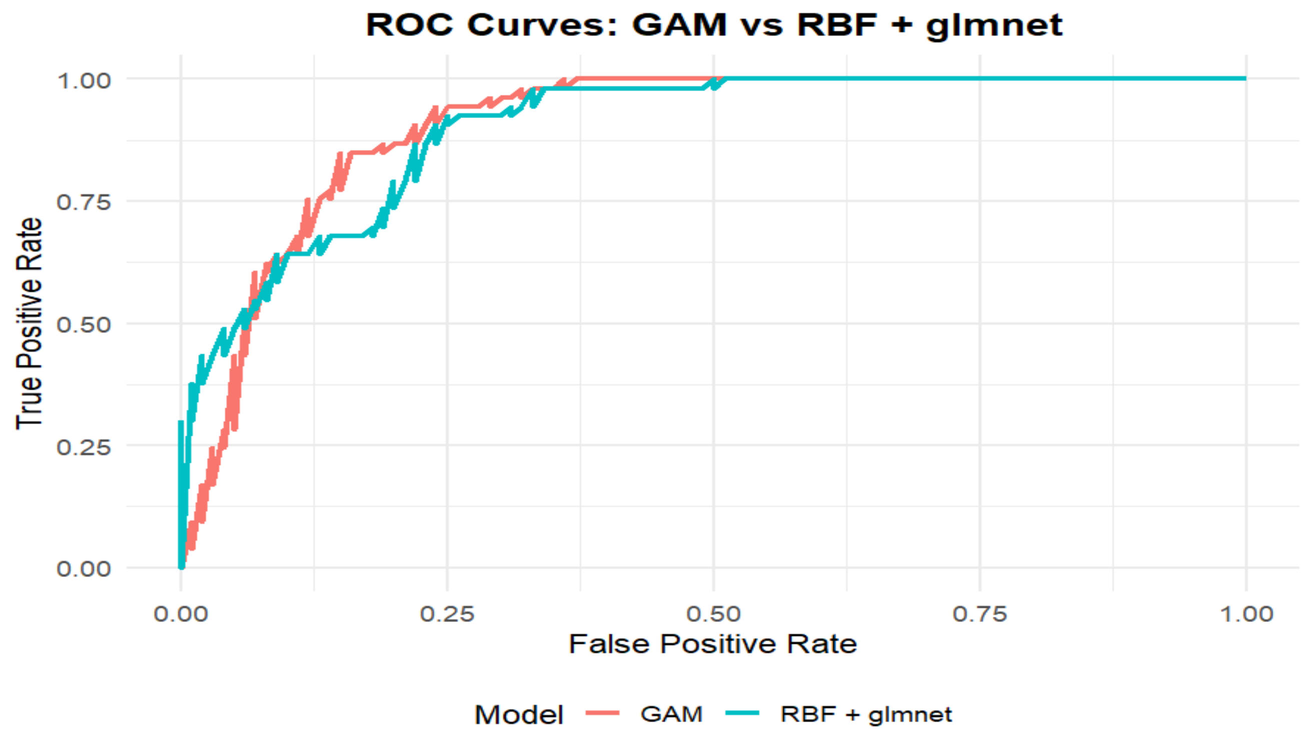

Figure 3: Receiver Operating Characteristic Curve of GAM and RBF Models

Table 6 presents the test AUC and accuracy of the two models. Based on the test-set AUC values, the generalized additive model GAM achieved an AUC of 0.9058, while the RBF + glmnet model achieved a slightly lower AUC value of 0.8992. Again, GAM has the highest accuracy of about 82%, and RBF has about 81%. Although both models reveal strong predictive power, the overall best discrimination between diabetic versus non-diabetic patients for unseen test data is provided by the GAM. The difference is modest, but it does suggest that the important nonlinear relationships were more effective when modeled with the spline-based smooth functions in the GAM compared to the radial basis function expansion combined with the penalized logistic regression.

Table 6: Test Set Model Comparison

Model	**Test AUC**	**Accuracy**
GAM (Spline-Based)	0.9058	0.8235
RBF + glmnet	0.8992	0.8105

5 Conclusion

The aim of this analysis was to compare nonlinear classification models in predicting the diabetes outcome for adult females from eight clinical predictors. After preprocessing, imputation with zeros, scaling of features, and partitioning of the data, two basis-function approaches were deployed. Generalized Additive Model with spline smoothers and an RBF expansion regularized by penalized logistic regression. Both techniques provided excellent predictive power and thus confirm that nonlinear effects in the data play a key role in modeling the risk of diabetes.

The flexible relationships captured via smooth spline functions in the GAM resulted in a test AUC of 0.9058, indicative of excellent discrimination performance. Significant nonlinearities were recognized, particularly in Glucose, BMI, and Age, depicting their strong effects on diabetes outcomes. RBF + glmnet also resulted in competitive performance, with a test AUC of 0.8992 and slightly higher accuracy, reflecting the ability to model complex surfaces in transformed feature space.

Both methods were better at finding non-diabetic cases than diabetic cases, which is a common pattern in medical datasets where the positive signals are weaker. However, the overall balanced accuracy for GAM was 0.8235, and RBF was 0.8105.

In summary, while both models were powerful, the GAM was the stronger and more interpretable approach in this data. Its smooth terms allow more intuitive understandings about how predictors influence diabetes risk and, for that reason, it is not only useful for prediction but also for clinical understanding. The findings support using flexible nonlinear modeling techniques when considering biomedical variables interacting in complex ways.

6 Data Source

National Institute of Diabetes and Digestive and Kidney Diseases downloaded from kaggle.com.