Predicting the Onset of Diabetes

Christopher Mao and Michael De La Rosa
The University of Texas at San Antonio, San Antonio TX, 78249

christopher.mao@my.utsa.edu
michael.delarosa@my.utsa.edu

Abstract

This study analyzes the Pima Indians Diabetes Database to develop predictive models for diagnosing diabetes based on diagnostic measurements such as glucose, BMI, insulin levels, and number of pregnancies. The dataset includes medical data from female patients of Pima Indian heritage aged 21 or older. We explored multiple predictive techniques, including logistic regression, stepwise logistic regression, logistic regression with cubic splines, generative additive models, tree-based methods, k-nearest neighbors, quadratic discriminant analysis, and dimensionality reduction approaches. The models were evaluated using accuracy, ROC-AUC, and other metrics, with cross-validation ensuring robustness. Key predictors like glucose and BMI were consistently significant across models. Generative additive models and logistic regression with cubic splines emerged as top performers in overall predictive accuracy. The study underscores the importance of data-driven approaches in healthcare and highlights actionable insights for improving diabetes diagnostics.

Background

Diabetes is a growing public health concern, requiring accurate diagnostic tools to identify at-risk individuals. The Pima Indians Diabetes Database, developed by the National Institute of Diabetes and Digestive and Kidney Diseases, provides an opportunity to explore predictive models for diabetes diagnosis using clinical and demographic data.

Objective

This study aims to evaluate and compare multiple modeling approaches to identify key predictors and optimize diagnostic accuracy.

Methods

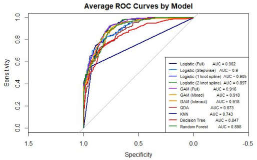
We first examined the dataset for problematic predictors, including multicollinearity, degenerate distributions, and missing values. While no issues were found with multicollinearity or degenerate distributions, several entries had implausible default values (e.g., a blood pressure of 0). These were re-coded as missing and imputed using MICE with predictive mean matching (PMM), which replaces missing values with observed values from individuals with similar predicted values.

After imputation, we centered and scaled the dataset, then used Cook's distance to detect outliers. Applying a cutoff of 4/(n–k–1), we removed any influential points.

We evaluated multiple models: logistic regression with all predictors, stepwise logistic regression (AIC-based), logistic regression with 1- and 2-knot splines, generalized additive models (GAMs), decision trees, random forests, QDA, and KNN (k = 1 to 20). Models were optimized for AUC and accuracy using 10-fold cross-validation.

Results

We obtained the following graphs using the previously described methods. Using stepwise logistic regression with AIC criteria, the variables kept in the model were glucose, BMI, pregnancies, and diabetes pedigree function. For GAMs we found evidence of a nonlinear relationship with Glucose, BMI, Age, and DiabetesPedigreeFunction. Interactions were between Glucose and Age, and BMI and DiabetesPedigreeFunction.



Model	Accuracy	AUC	F1
Logistic (Full)	0.830	0.901	0.722
Logistic (Stepwise)	0.834	0.900	0.730
Logistic (1 knot spline)	0.818	0.907	0.705
Logistic (2 knot spline)	0.801	0.897	0.682
GAM (All Smooth)	0.811	0.916	0.698
GAM (Mixed)	0.816	0.919	0.705
GAM (Interaction)	0.830	0.918	0.728
QDA	0.805	0.874	0.702
k-NN	0.800	0.743	0.655
Decision Tree	0.790	0.833	0.674
Random Forest	0.801	0.898	0.678

Figure 1. Plot of Average ROC of all models across 10 fold cross validation.
Table 1. A table of metrics from the models used in the analysis to compare the accuracy, AUC, and F1 scores of each. KNN uses the best k of k=16.

Results - con't

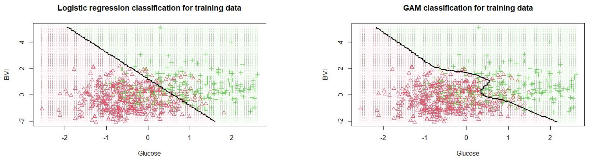


Figure 2. Logistic Regression decision boundary curve (left) and GAM decision boundary curve (right), using glucose and BMI as x and y axis, respectively.

Conclusions

- GAMs with mixed features had the best performance, with GAMS with interactions being the second best performing model. The next type of model was logistic regression with cubic splines.
- Glucose, BMI, pregnancies, and diabetes pedigree function are important predictors of diabetes.

References

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data

Acknowledgements

Assigned as part of the Spring 2025 STA 6933 Course instructed by Professor Min Wang.