Predicting the Onset of Diabetes

Michael DE LA ROSA and Christopher MAO

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 relevant 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.

1. INTRODUCTION

Diabetes Mellitus is a group of diseases characterized by chronic hyperglycemia. Type I is due to a lack of the hormone used to control blood glucose levels, insulin, as a result of damage to the pancreas, the organ that secretes insulin. Type II diabetes is much more common and is caused by insulin resistance and impaired secretion of insulin. The chronic hyperglycemia of diabetes can lead to widespread damage across virtually all parts of the body; it can cause or lead to increased risk of immunosuppression, retinopathy, cardiovascular disease, kidney disease, neuropathy, diabetic ketoacidosis, and life-threatening disease such as diabetic ketoacidosis.

Diabetes has become a significant public health concern due to its increasing prevalence and substantial economic impact. The number of adults living with diabetes worldwide has surpassed 800 million, more than quadrupling since 1990 [4]. In the United States, the total annual cost of diabetes reached \$412.9 billion in 2022, including \$306.6 billion in direct medical costs and \$106.3 billion in indirect costs [3]. This growing incidence and associated financial burden underscore the urgent need for effective public health interventions and policies to address the diabetes epidemic. The combination of a high prevalence of disease and the existence of good treatments to prevent or slow complications means that early detection of diabetes can have significant benefits.

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. Despite its extensive use in machine learning research, there is limited focus on comparing diverse predictive techniques tailored to healthcare settings.

This study aims to evaluate and compare multiple modeling approaches, including logistic regression, tree-based methods, and dimensionality reduction techniques, to identify key predictors and optimize diagnostic accuracy.

2. METHODS

2.1 Pre-processing

To begin the analysis, we first examined the dataset, including the distributions of each predictor as seen in (Figure 1). This clearly demonstrated a lack of any degenerate distributions, but showed a concerning trend of zero values for predictors that physiologically should not be zero such as glucose or BMI. To correct these values, we converted them into missing values (Figure 2) and performed predictive mean matching using the MICE package in R [2] (Figure 3). Afterwards, the data was checked for correlation (Figure 4) and for multicolinearity. Using a correlation threshold of 0.75 and VIF threshold of 10, no offending predictors were found, and thus we could proceed.

After completion of imputation, we proceeded to center and scale the data set. Cook's distance was used to identify outliers, and then any influential points with a cook's distance of $4/(n-k-1)$ were removed.

2.2 Models

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.

For our use of GAMs, we employed three different strategies. The first was using a smooth spline for all terms, one

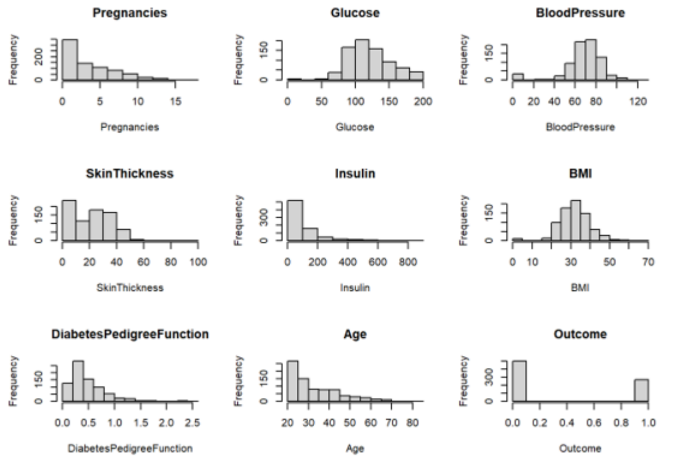


Figure 1: Histogram of each predictor before preprocessing.

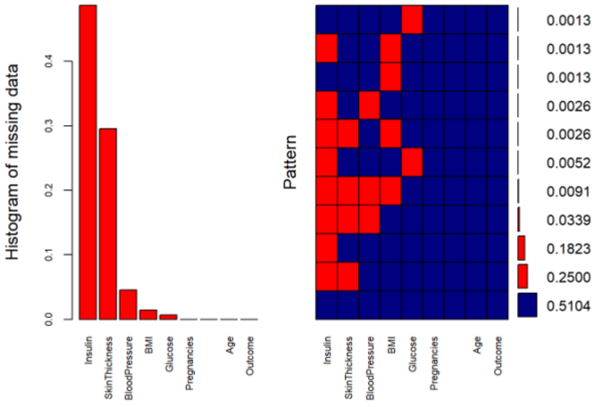


Figure 2: MICE output showing a histogram of missing data by predictor on the left and a graph on the left showing combinations of missing values, with the missing values in red for each.

with a smoothing spline for those that we found evidence of a nonlinear relationship, and one with smoothing spline for those with evidence of a nonlinear relationship as well as interaction terms.

3. RESULTS

Model performance was assessed using metrics such as accuracy, ROC-AUC, and F1-score. Each of these metrics was calculated using the average of all 10 folds in cross-validation. We paid special attention to AUC due to its importance to physicians. These results can be found in Table 1, and a graph of all ROC curves can be found in Figure 6.

The full logistic regression model, incorporating all predictors, achieved an accuracy of 83.0% and an AUC of 0.901, with significant predictors including glucose, BMI, pregnancies, and diabetes pedigree function (Figure 7).

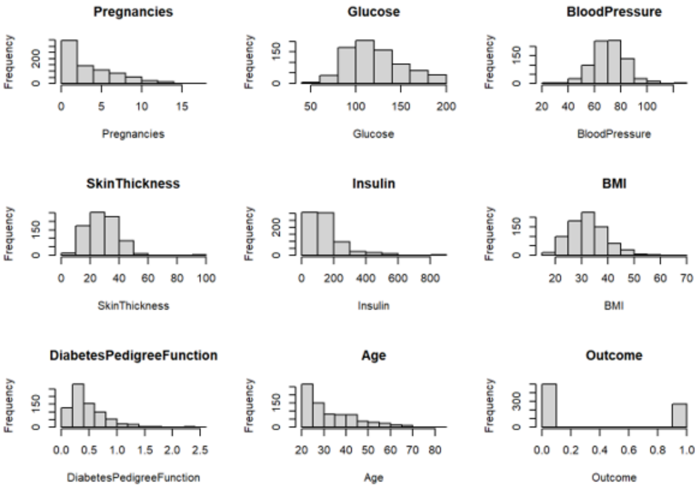


Figure 3: Histogram of each predictor after performing imputation.

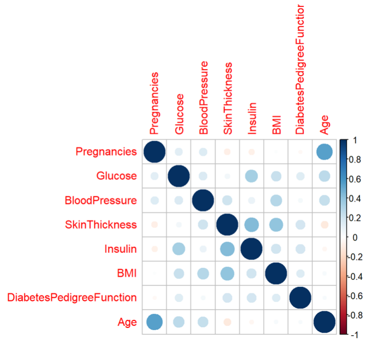


Figure 4: Correlation plot of all predictors in the dataset. Of note, the highest correlation is between age and pregnancy at 0.6, less than the threshold of 0.75

Stepwise logistic regression reduced the model to these key predictors while maintaining an AUC of 0.900 and an accuracy of 83.4%, providing a more interpretable yet robust solution (Figure 8).

Logistic regression with cubic splines resulted in one of the highest AUC values. Using 1 knot there was an accuracy of 81.8% and an AUC of 0.907, and with 2 knots there was an accuracy of 80.1% and an AUC of 0.897.

The GAM models were the highest performing models in this project. For all smooth terms we found an accuracy of 81.1% with an AUC of 0.916. We found a evidence of a nonlinear relationship for Glucose, BMI, Age, and DiabetesPedigreeFunction by plotting the smooth terms (Figure 9). The majority of the GAM models generated in cross validation also agreed on only these four variables being statistically significant using the Anova for Nonparametric Effects, providing further evidence towards a nonlinear relationship. This mixed model resulted in the best performing

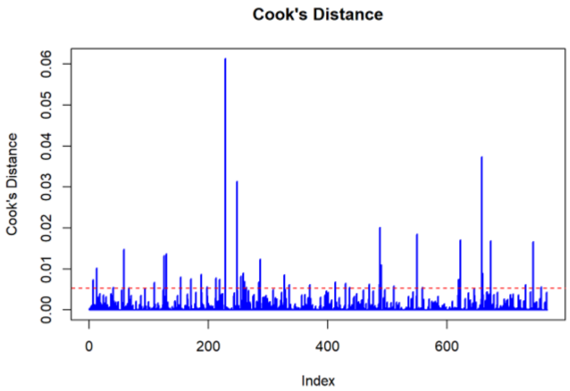


Figure 5: Plot of cook's distance for all points in the dataset, with the threshold of $4/(n-k-1)$ depicted with the red line.

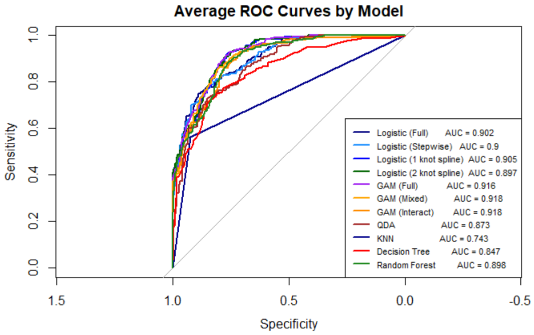


Figure 6: Averaged ROC curves over all 10 folds for each model.

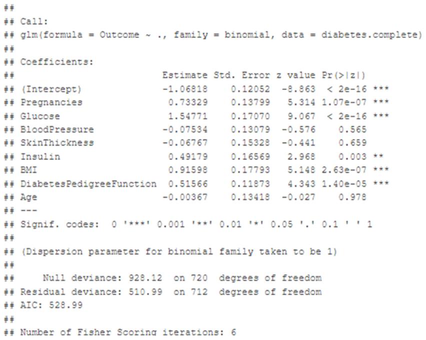

```
##
## Call:
## glm(formula = Outcome ~ ., family = binomial, data = diabetes.complete)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.06818    0.12052  -8.863  < 2e-16 ***
## Pregnancies               0.73329    0.13799   5.314 1.07e-07 ***
## Glucose                   1.54771    0.17070   9.067  < 2e-16 ***
## BloodPressure            -0.07534    0.13079  -0.576    0.565
## SkinThickness            -0.06767    0.15328  -0.441    0.659
## Insulin                   0.49179    0.16569   2.968    0.003 **
## BMI                       0.91598    0.17793   5.148 2.63e-07 ***
## DiabetesPedigreeFunction  0.51566    0.11873   4.343 1.40e-05 ***
## Age                      -0.00367    0.13418  -0.027    0.978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 928.12  on 720  degrees of freedom
## Residual deviance: 510.99  on 712  degrees of freedom
## AIC: 528.99
##
## Number of Fisher Scoring iterations: 6
```

Figure 7: Output from the summary function in R when called on the full logistic regression model.

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

Table 1. Comparison of accuracy, AUC, and F1 scores for each model.

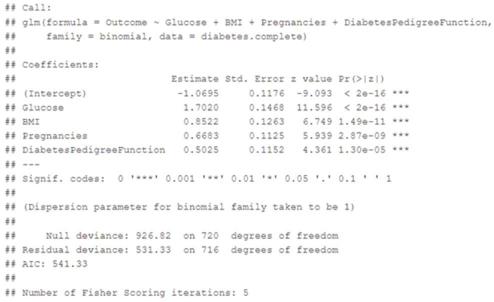

```
## Call:
## glm(formula = Outcome ~ Glucose + BMI + Pregnancies + DiabetesPedigreeFunction,
##     family = binomial, data = diabetes.complete)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -1.0695     0.1176  -9.093  < 2e-16 ***
## Glucose                    1.7020     0.1468  11.596  < 2e-16 ***
## BMI                        0.8522     0.1263   6.749 1.49e-11 ***
## Pregnancies                0.6683     0.1125   5.939 2.87e-09 ***
## DiabetesPedigreeFunction   0.5025     0.1152   4.361 1.30e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 926.82  on 720  degrees of freedom
## Residual deviance: 531.33  on 716  degrees of freedom
## AIC: 541.33
##
## Number of Fisher Scoring iterations: 5
```

Figure 8: Output from the summary function in R when called on the stepwise logistic regression model.

model with an accuracy of 81.6% and AUC of 0.919. For the interaction model we used these smooth terms as well as interactions between Glucose and Age, and BMI and DiabetesPedigreeFunction. This resulted in a close second with an accuracy of 83.0% and an AUC of 0.918.

Decision Trees performed moderately well with an accuracy 79.0% with an AUC of 0.833. Cost complexity pruning was performed in order to minimize both error and complexity of the tree to maximize performance and prevent overfitting.

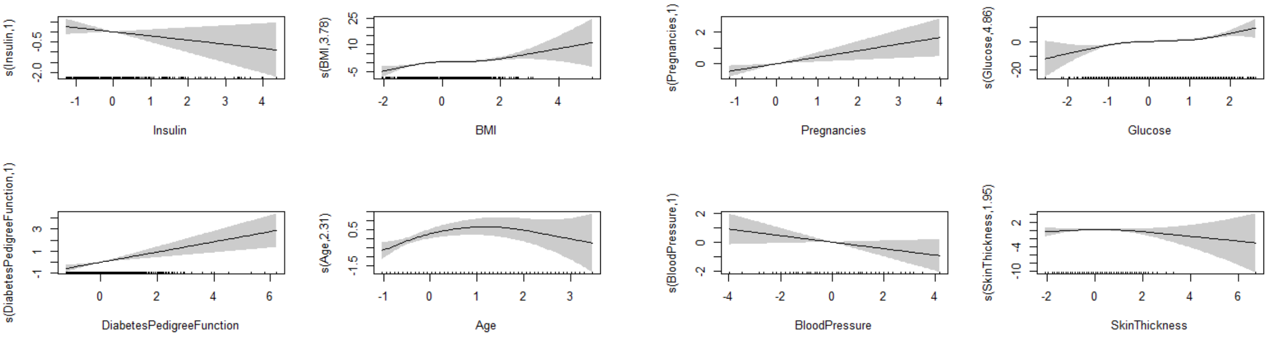
Random Forest outperformed Decision Trees with an AUC of 0.898 and an accuracy of 80.1%. It identified glucose, BMI, insulin, and age as the most influential predictors, further validating their importance in diabetes prediction (Figure 10).

KNN achieved an accuracy of 80.0% and an AUC of 0.743. Performance was unstable with different numbers of K, but the optimal performance was using K=16 (Figure 12).

QDA provided competitive performance with an AUC of 0.874 and an accuracy of 80.5%. The model effectively captured non-linear relationships, as illustrated by its curved decision boundaries (Figure 11).

4. DISCUSSION

Key predictors like glucose and BMI were consistently significant across all models, highlighting their strong asso-



(a) (b)

Figure 9: Plots of the smooth terms in the GAM model with all smooth terms.

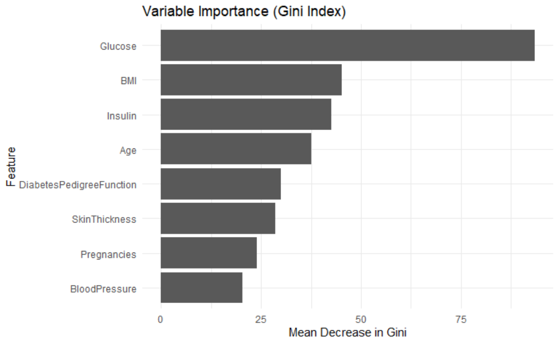


Figure 10: Plot of the Gini index. This shows the mean decrease of Gini impurity by variable, which is a measure of feature importance.

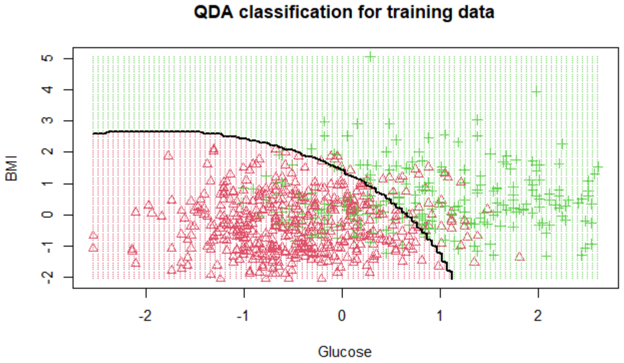


Figure 11: Distribution boundary for the QDA model for glucose vs. BMI.

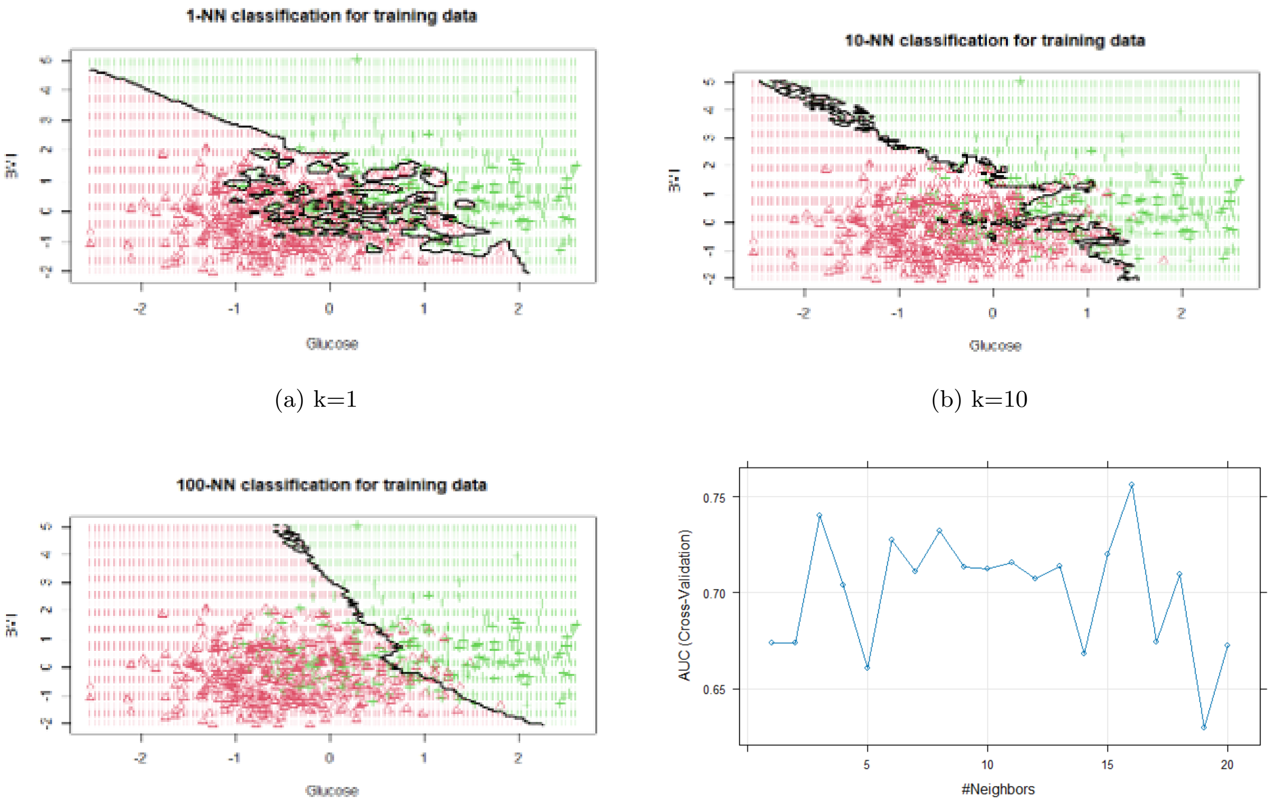
ciation with diabetes outcomes. Random Forest further emphasized the importance of insulin and age, while logistic regression provided a clear ranking of feature significance.

GAMs and Logistic Regression with 1 knot cubic splines emerged as the top-performing models, sacrificing some interpretability for better performance. These models were able to capture some of the nonlinear relationships more effectively than the other models, which proved to provide significant gains on improvement. We can see with the decision boundaries of logistic regression, logistic regression with splines, and GAMs, how this increase in flexibility is able to more effectively account for nonlinear relationships (Figure 13). However, it is important to note that there is still a relatively small difference in AUC between normal logistic regression and GAMs; in cases where interpretability is desired above performance, it may be advisable to use normal logistic regression.

While Random Forest identified insulin as an important predictor, logistic regression did not include it as statistically significant. This discrepancy likely arises from the difference in how these models handle relationships within the data—Random Forest captures non-linear interactions, whereas logistic regression relies on linear associations and statistical significance. If glucose and BMI already explain most of the variance, the additional predictive power of insulin becomes negligible for the logistic regression model. This can also be seen in the GAMs, which did not find any evidence of an important non linear relationship between Outcome and Insulin.

Moreover, insulin is not a primary diagnostic criterion for diabetes. Diabetes is typically diagnosed based on: Fasting blood glucose levels (≥126 mg/dL), HbA1c levels (≥6.5%), and Oral glucose tolerance test (OGTT) [1].

Insulin levels are rarely included because they vary widely within and among individuals, increasing after meals and decreasing during fasting periods. This variability makes it challenging to obtain a single measurement that accurately reflects an individual's typical insulin status.



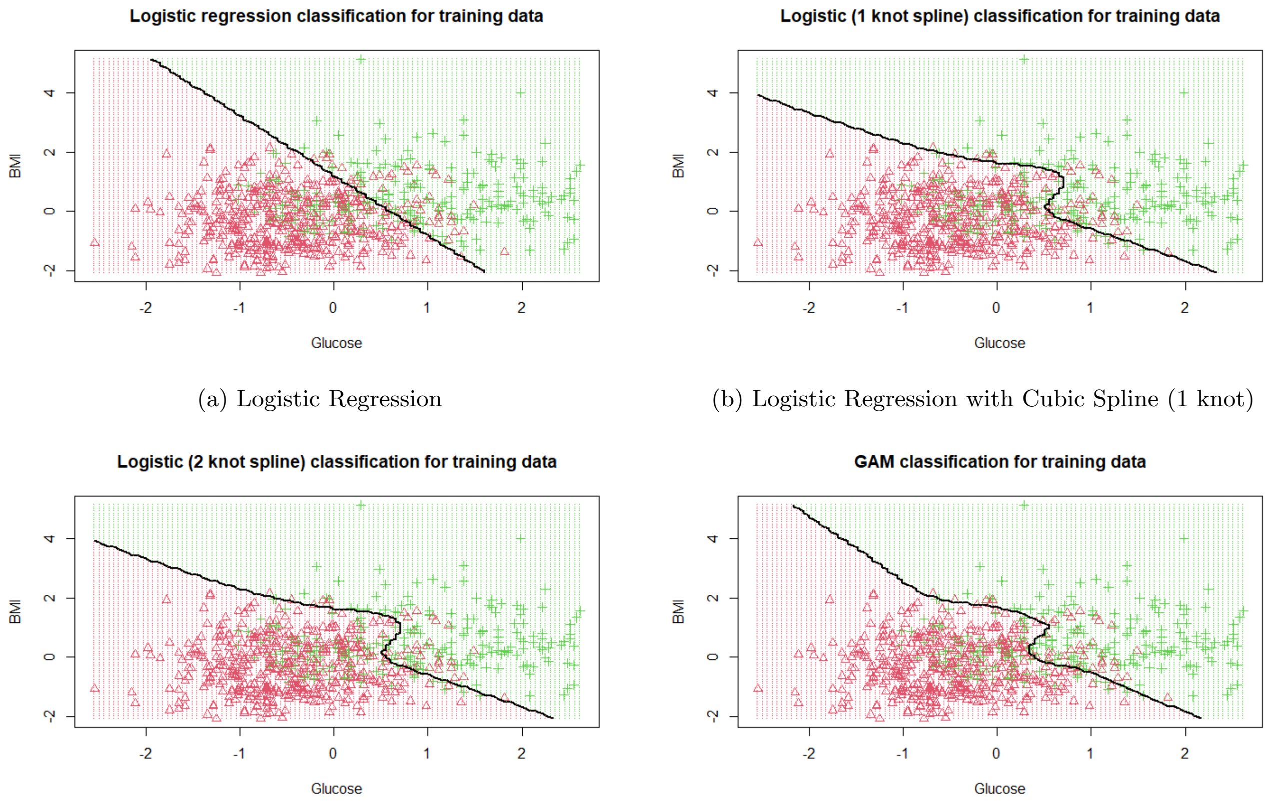
(a) k=1

(b) k=10

(c) k=100

(d) Accuracy Graph

Figure 12: Plots of decision boundaries for different K values (k=1, k=10, k=100) alongside a graph showing the accuracy for each KNN model by their k value.



(a) Logistic Regression

(b) Logistic Regression with Cubic Spline (1 knot)

(c) Logistic Regression with Cubic Spline (2 knot)

(d) Generalized Additive Model

Figure 13: Plots of decision boundaries of Glucose and BMI for logistic regression, logistic regression with cubic splines for 1 and 2 knots, and GAMs.

5. CONCLUSION

By applying robust evaluation metrics and prioritizing model interpretability, this study provides actionable insights for clinicians and researchers seeking effective diabetes screening strategies. Our results emphasize the importance of glucose and BMI as key predictors of diabetes across all models.

Future work should explore additional preprocessing techniques, such as **winsorization** to address extreme outliers and **principal component analysis (PCA)** to reduce multicollinearity among predictors. Advanced modeling approaches, including **support vector machines (SVM)** and **neural networks**, could be investigated to potentially improve predictive accuracy and better handle complex, non-linear relationships in the data. Furthermore, incorporating external datasets and addressing potential biases specific to the Pima Indian population could enhance the generalizability of the findings.

This study highlights the value of data-driven approaches in healthcare and underscores the need for continuous improvement in diabetes screening, ensuring more precise and actionable outcomes for clinical applications.

REFERENCES

[1] American Diabetes Association Professional Practice Committee (2025). Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2025. *Diabetes Care* **48**(Supplement_1) 27–49. https://doi.org/10.2337/dc25-S002.

[2] Anderson, D. and Kurtz, T. Continuous time Markov chain models for chemical reaction networks. http://www.math.wisc.edu/~kurtz/papers/AndKurJuly10.pdf. Accessed 27 July 2010.

[3] Association, A. D. (2022). *New American Diabetes Association Report Finds Annual Costs of Diabetes Exceed $400 Billion.* https://diabetes.org/newsroom/press-releases/new-american-diabetes-association-report-finds-annual-costs-diabetes-be.

[4] Organization, W. H. (2024). *Urgent Action Needed as Global Diabetes Cases Increase Four-Fold Over Past Decades.* https://www.who.int/news/item/13-11-2024-urgent-action-needed-as-global-diabetes-cases-increase-four-fold