Predicting HbA1c Levels and Type II Diabetes Diagnosis with Demographic, Biometric, and Medical History Data

Thaina Gomez & Maxwell Miller-Golub STAT 627 – American University

Introduction/Context

Type II diabetes affects over 10% of adults, with nearly half undiagnosed [1]. Early detection is critical for preventing complications. Even after accounting for health insurance, immigrants and racial/ethnic minority adults have increased odds of undiagnosed diabetes [2].

RO1: Investigates whether demographic and biometric factors can predict HbA1c levels.
RO2: Cross-examines different classification models to determine the best method for accurately attributing Diabetes diagnosis

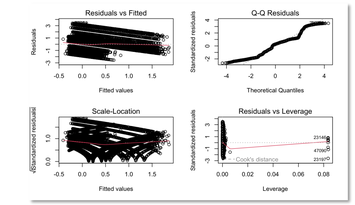
Approach

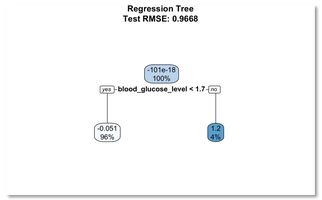
Cleaned and explored a dataset (n = 100,000) with predictors: age, gender, BMI, glucose, hypertension, heart disease, and smoking history.

RO1: The data was split 80-20 train-test to evaluate model performance. We built four regression models to predict HbA1c levels: Linear regression w. interaction term, regression tree, Lasso regression, and Ridge regression. We standardized numeric predictors for Lasso/Ridge, and used 10-fold CV to select the optimal λ. For the linear model, we checked assumptions of linearity, independence, and constant variance, and assessed multicollinearity using VIF.

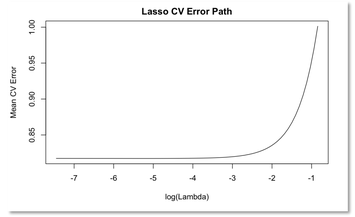
RO2: With "Diabetes" diagnosis as the target variable, we conducted a 90-10 train-test split with 10-fold CV in the training set. This led to Confusion Matrices for the following models: Logistic, Lasso, Ridge, and ElasticNet Regressions, KNN, LDA, and Random Forest.

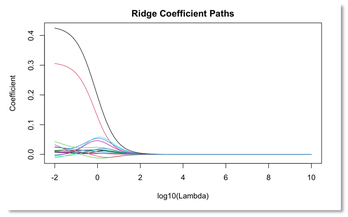
RO1: Demographic/Biometric Features vs HbA1c levels





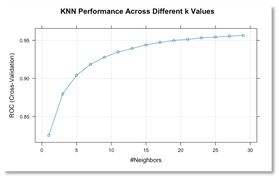
Model <chr>	RMSE <dbl>
Linear (w/ interactions)	0.9678781
Tree	0.9668248
Lasso	0.9050378
Ridge	0.9053900

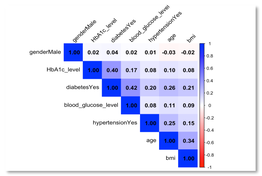




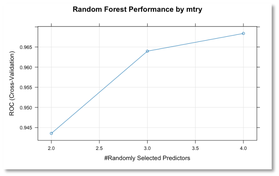
- Diagnostic plots and visuals highlighted the need for interactions and non-linear terms.
- Lasso & Ridge (RMSE = 0.905) captured complex patterns more effectively than the linear model.

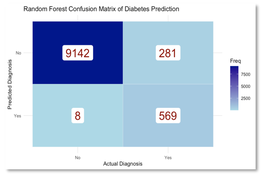
RO2: Finding best Classification Model for Diabetes diagnosis





Model	Accuracy	Precision	Recall	F1	AUC	KNeighbors.Best_MTry.lambda
Logistic Regression	0.9600000	0.8652597	0.6270588	0.7271487	0.9589914	NA
KNN	0.9573000	0.9731544	0.5117647	0.6707787	0.9552005	39
LDA	0.9547000	0.8358714	0.5811765	0.6856350	0.9517544	N/A
Random Forest	0.9711000	0.9861352	0.6694118	0.7974772	0.9680602	4
Lasso Regression	0.9605444	0.8790457	0.6213072	0.7280386	0.9619369	0.000514695191165332
Ridge Regression	0.9573222	0.9269222	0.5405229	0.6828503	0.9616148	0.0113497018771895
ElasticNet Regression	0.9602667	0.8926369	0.6053595	0.7214519	0.9621162	0.00186683648874033





- ANOVA determined all variables relevant in modeling
- Random Forest (m = 4, trees = 500) led to every highest metric

Executive Summary

Type II Diabetes impacts over 10% of adults, with nearly half undiagnosed. Racial and ethnic minorities are disproportionately impacted. Identifying key risk factors may lead to increased diagnoses and longer life expectancy. Utilizing various biometric and demographic features, we found:

Regression models (Lasso & Ridge) returned **RMSE values of 0.905** while predicting HbA1c, a key diabetes indicator

A Classification model (Random Forest) which has **97% accuracy** at correctly predicting diabetes diagnoses on unseen cases

Results/Implications

Even a simple model using routine health data can support early risk detection.

RO1: Lasso Regression had the lowest RMSE at 0.9050. Ridge followed closely behind. LM underperformed, despite including interaction & polynomial terms.

RO2: With our best model (Random Forest), an unseen case has a 97% chance of an accurate diabetes diagnosis.

Assumptions/Limitations/ Challenges or Secondary Results

RO1: RF and Gradient Boosting were not explored, and capturing complex relationships in the data could have provided better accuracy.

RO2: SVM, PCA/PCS not explored, though they may produce more accurate/efficient models.

Both: Dataset limited in scope and may not represent all demographics equally.
Future work should include more predictors (e.g., diet, physical activity). As medical data becomes more robust, more interesting and surprising relationships will be discovered.

Primary References

1. International Diabetes Federation. (2021). *Diabetes facts & figures*. https://idf.org/about-diabetes/diabetes-facts-figures/
2. Loretta Hsueh, Wei Wu, Adam T. Hirsh, Mary de Groot, Kieren J. Mather, Jesse C. Stewart. Undiagnosed diabetes among immigrant and racial/ethnic minority adults in the United States: National Health and Nutrition Examination Survey 2011–2018. *Annals of Epidemiology*, 51: 14-19, 2020.
3. Boye, K. S., Lage, M. J., Shinde, S., Thieu, V., & Bae, J. P. (2021). Trends in HbA1c and body mass index among individuals with type 2 diabetes. *Diabetes Therapy, 12*(7), 2077–2087. https://doi.org/10.1007/s13300-021-01084-0
4. Rohlfing, C. L., Wiedmeyer, H.-M., Little, R. R., England, J. D., Tennill, A., & Goldstein, D. E. (2002). Defining the relationship between plasma glucose and HbA1c: Analysis of glucose profiles and HbA1c in the Diabetes Control and Complications Trial. *Diabetes Care, 25*(2), 275–278. https://doi.org/10.2337/diacare.25.2.275

Data

Mustafa, M. (n.d.). *Diabetes prediction dataset*. Kaggle. Retrieved March, 2025, from https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

Size: 100,000 patient records
Scale: Demographics, biometric info (age, BMI, blood glucose), and medical history (heart disease, hypertension, smoking)

Acknowledgments

- Professor Ahmad Mousavi, STAT 427/627-001 Spring 2025