DiaBETic: Predicting Diabetes Risk from Health and Demographic Factors

Tyler G. Rose
Department of Chemistry, Brown University
https://github.com/tylergeorgerose/diabetic

I. Introduction

Purpose

Despite our understanding of risk factors, diabetes poses a significant burden on the American population, with an estimated 136 million diabetic or prediabetic adults [1]. Predicting diabetes risk from accessible personal data is therefore of interest to healthcare providers and individuals, facilitating informed diabetes testing decisions.

Data Source & Prior Work

The Behavioral Risk Factor Surveillance Survey (BRFSS) is an annual CDC telephone survey, collecting self-reported demographic and healthcare data [2]. Its comprehensive feature set and large sample size makes it an attractive dataset for a wide range of biomedical machine learning tasks, including diabetes prediction.

Previously, researchers used a subset of the 2014 BRFSS with 138,146 individuals and 27 features to predict type II diabetes risk [3]. More recently, researchers used a different subset of the 2015 BRFSS for prediction, highlighting the success of tree-based models [4]. The latter approach is attractive, including more points and more recent data; however, the lack of available code and methodological detail limits reproducibility while extraordinarily high performance (94% recall and 99% precision for the best model) warrant serious data leakage concerns. This work provides a transparent, robust machine learning pipeline for assessing diabetes risk using the latter dataset.

Dataset

The dataset consists of 253,680 individuals and 21 features. Most features are binary indicators (14, e.g. high cholesterol), with some continuous (3, e.g. body mass index) and ordinal-encoded (4, e.g. age category) features (summarized at README.md). The target variable is a binary indicator of an individual being nondiabetic (negative class) or prediabetic or diabetic (positive class). The dataset contains no missing values.

II. Exploratory Data Analysis

Imbalance

Exploratory data analysis revealed imbalance, skew, and outliers within the dataset, potentially impacting model performance. The target variable is slightly imbalanced, with 13.9% of points belonging to the positive class (Figure 1). Furthermore, various feature values are imbalanced or contain significant outliers (Figure 2).

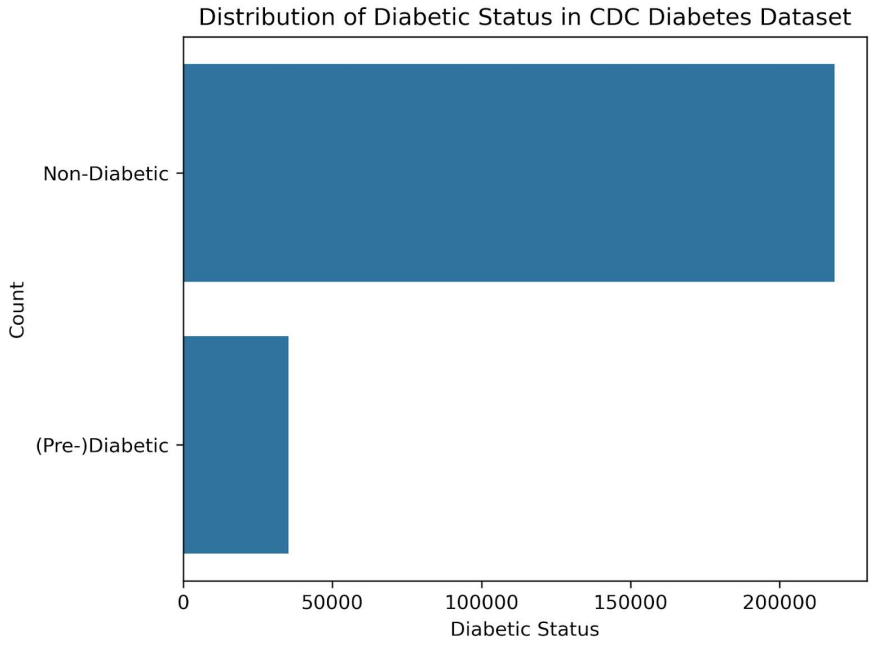

Figure 1. Imbalance in the target variable.

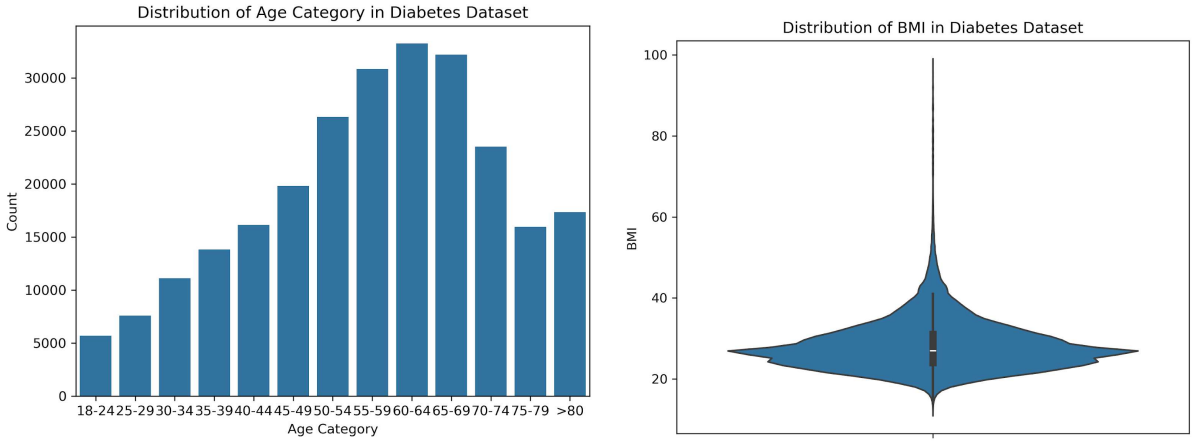

Figure 2. Imbalance and outliers in demonstrative features.

Correlations and Feature Selection

In general, low inter-feature Pearson correlations are observed (Figure 3), as well as low correlations with the target variable (Figure 4), indicating all features likely contribute independent information to the model and should be used.

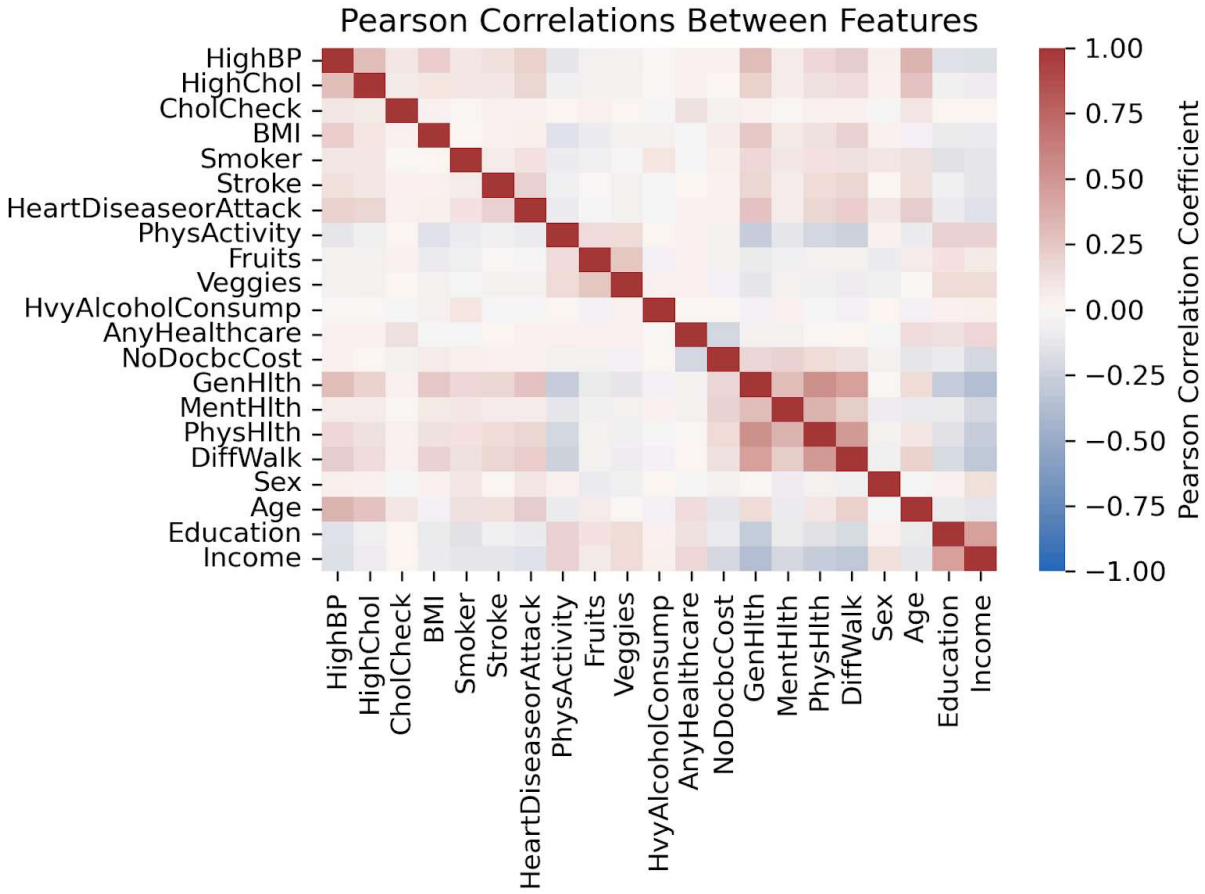

Figure 3. Intra-feature Pearson correlations.

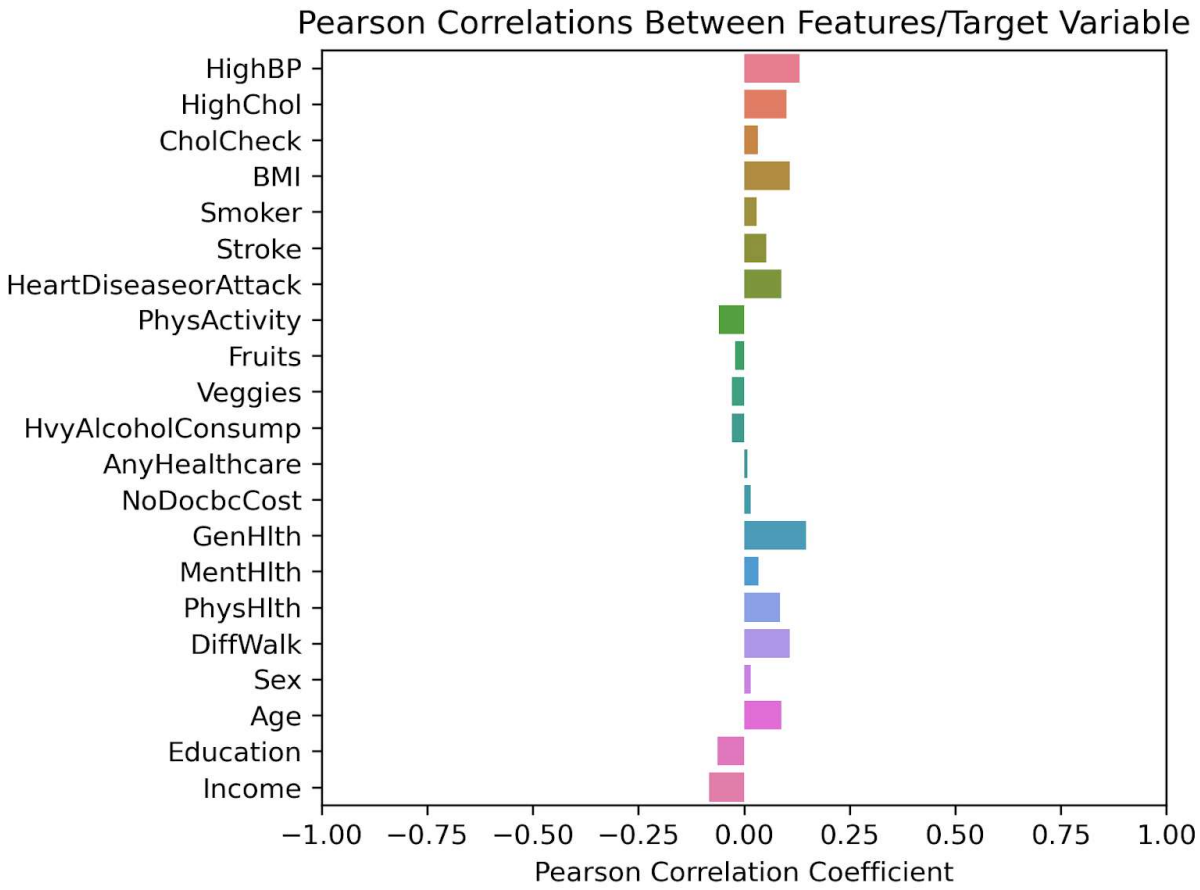

Figure 4. Pearson correlations between features and target variable.

Further exploratory data analysis, including visualizing distributions and summary statistics for each feature, is available at src/src.ipynb.

III. Methods

a. Model Training and Evaluation

20% of the data was held out for testing. For four classifiers (logistic regression, random forest, decision tree, and XGBoost), hyperparameter tuning was performed through a grid search over values specified in Table 1 using 4-fold cross-validation on the remaining 80%. Stratification was used to ensure proportional representation among unbalanced classes at all splitting steps. Standard scaling was fit and applied to the training data to ensure all features were compared on the same scale, impacting regularization and interpretability. All values are numeric in the original dataset with binary/ordinal encodings applied *a priori*. The corresponding scaler was applied to transform validation and testing data. F1-score was used as the evaluation metric to select hyperparameters and evaluate test set performance, balancing the healthcare system burden of false positives with the individual risk of false negatives. This was repeated for five random seeds to estimate uncertainty due to random splitting and nondeterministic models. To account for class imbalance, balanced class weights were used for all models.

Model	Hyperparameter Ranges	Non-Default Fixed Parameters
LogisticRegression	C: 10^{-3}, 10^{-2}, …, 10^{2}, 10^{3} l1_ratio: 0.1, 0.3, 0.5, 0.7, 0.9	max_iterations: 10,000 penalty: Elastic Net solver: SAGA
RandomForestClassifier	max_features: 0.1, 0.3, 0.5, 0.7, 0.9 max_depth: 3, 5, 10, 30	–
DecisionTreeClassifier	max_features: 0.1, 0.3, 0.5, 0.7, 0.9 max_depth: 3, 5, 10, 30	–
XGBClassifier	reg_lambda: 10^{-3}, 10^{-2}, …, 10^{2}, 10^{3} max_depth: 3, 5, 10, 30, 50, 100, 300	subsample: 0.66 colsample_bytree: 0.9 learning_rate: 0.03 early_stopping_rounds: 50

Table 1. Tuned and fixed hyperparameters by model.

b. Interpretation

For the best-performing XGBoost model, global feature importances were calculated using model-specific importance metrics (cover, gain, total cover, total gain, and weight). For the same model and a 1% sample of the corresponding test set, SHapley Additive exPlanation (SHAP) values were calculated as a complementary measure.

Local feature importances were investigated for two randomly sampled individuals from each of the predicted positive and negative classes (for a total of four individuals).

IV. Results

a. Model Training and Evaluation

Upon the initial evaluation round, each model's mean F1-score exceeded the baseline, calculated as $F1=2p_1/(1+p_1)$ for a dummy model predicting all positive points given the true positive class proportion p_1. XGBoost boasted the highest mean test F1-score, approximately 100 standard deviations above the baseline mean. All models showed comparable uncertainty due to random splitting on the order of a tenth of a percent.

Each model demonstrated good agreement between means and standard deviations of test scores versus highest observed cross-validation scores. Optimized hyperparameter values were distributed within the tested range, indicating effective exploration of the parameter space. The decision tree and random forest models showed remarkable consistency in optimal depth, although the optimal feature subset varied. XGBoost showed consistency across max depth and L2 regularization. Logistic regression performance depended less on regularization parameters, which may have been unnecessary for model performance.

Model	Ideal Hyperparameters (By Seed)	Best CV F1-Score (mean ± std)	Best Test F1-Score (mean ± std)
Logistic Regression	C: 10^{1}, 10^{-3}, 10^{-2}, 10^{-3}, 10^{1} L1 Ratio: 0.1, 0.1, 0.1, 0.7, 0.9	0.443 ± 0.001	0.442 ± 0.002
Random Forest	Max Depth: 10, 10, 10, 10, 10 Max Features: 0.3, 0.7, 0.3, 0.3, 0.7	0.447 ± 0.000	0.445 ± 0.002
Decision Tree	Max Depth: 10, 10, 10, 10, 10 Max Features: 0.5, 0.5, 0.5, 0.7, 0.9	0.428 ± 0.001	0.429 ± 0.003
XGBoost	Max Depth: 50, 50, 50, 30, 50 L2 Regularization: 10^{1}, 10^{1}, 10^{1}, 10^{1}, 10^{1}	0.456 ± 0.001	0.455 ± 0.002
Baseline	–	0.245 ± 0.000	0.245 ± 0.000

Table 2. Ideal parameters and cross-validation (CV)/test scores for each model (mean ± std).

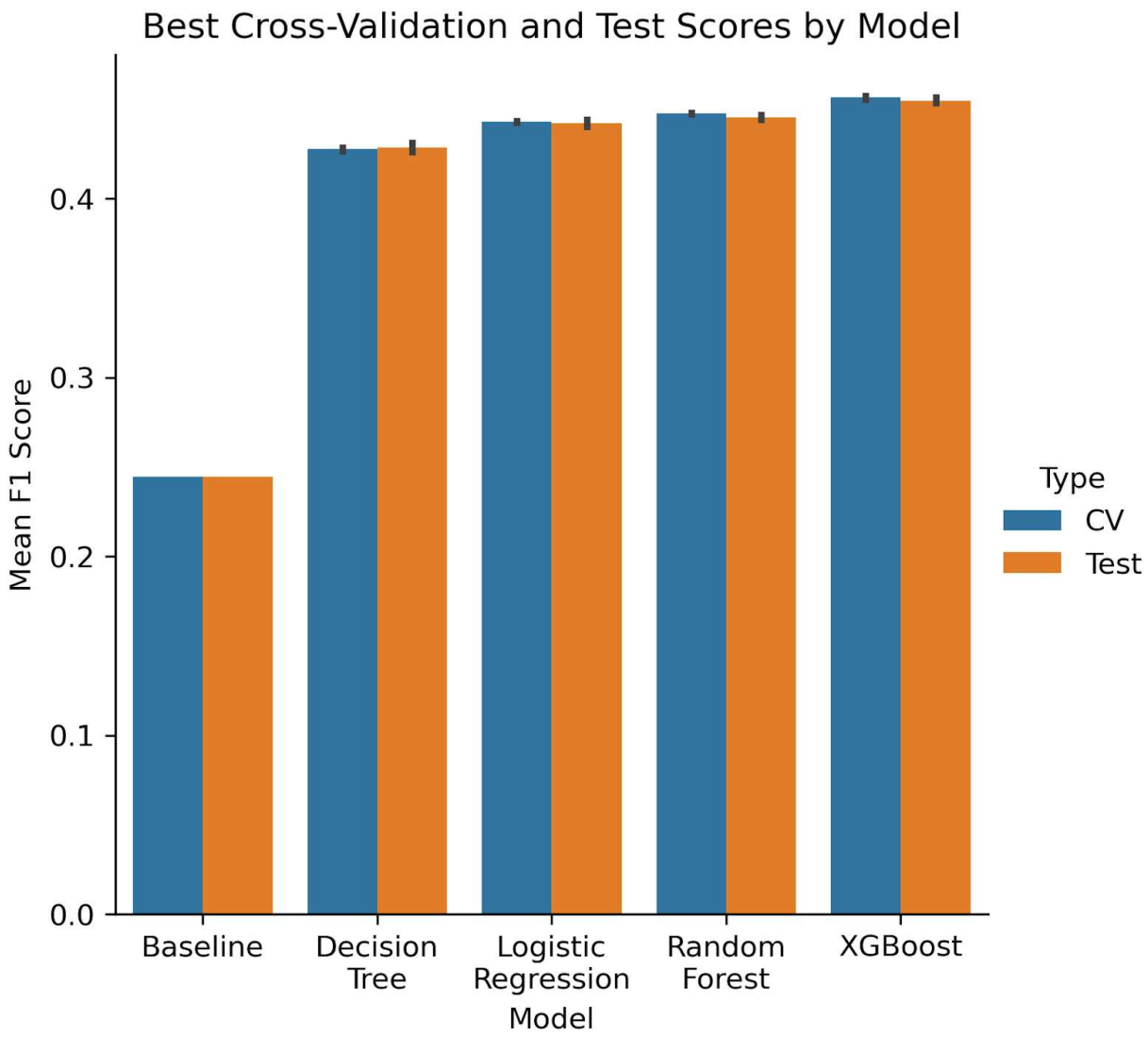

Figure 5. Comparison of cross-validation and test scores.

Analysis of the confusion matrices (Figure 6) indicates that, for most models, correct classification of condition positive points was on average slightly easier than correct classification of condition negative points (22-24% vs. 27-30%). The performance advantage of XGBoost, misclassifying 30% of condition positive points, was its unique ability to correctly classify positive points.

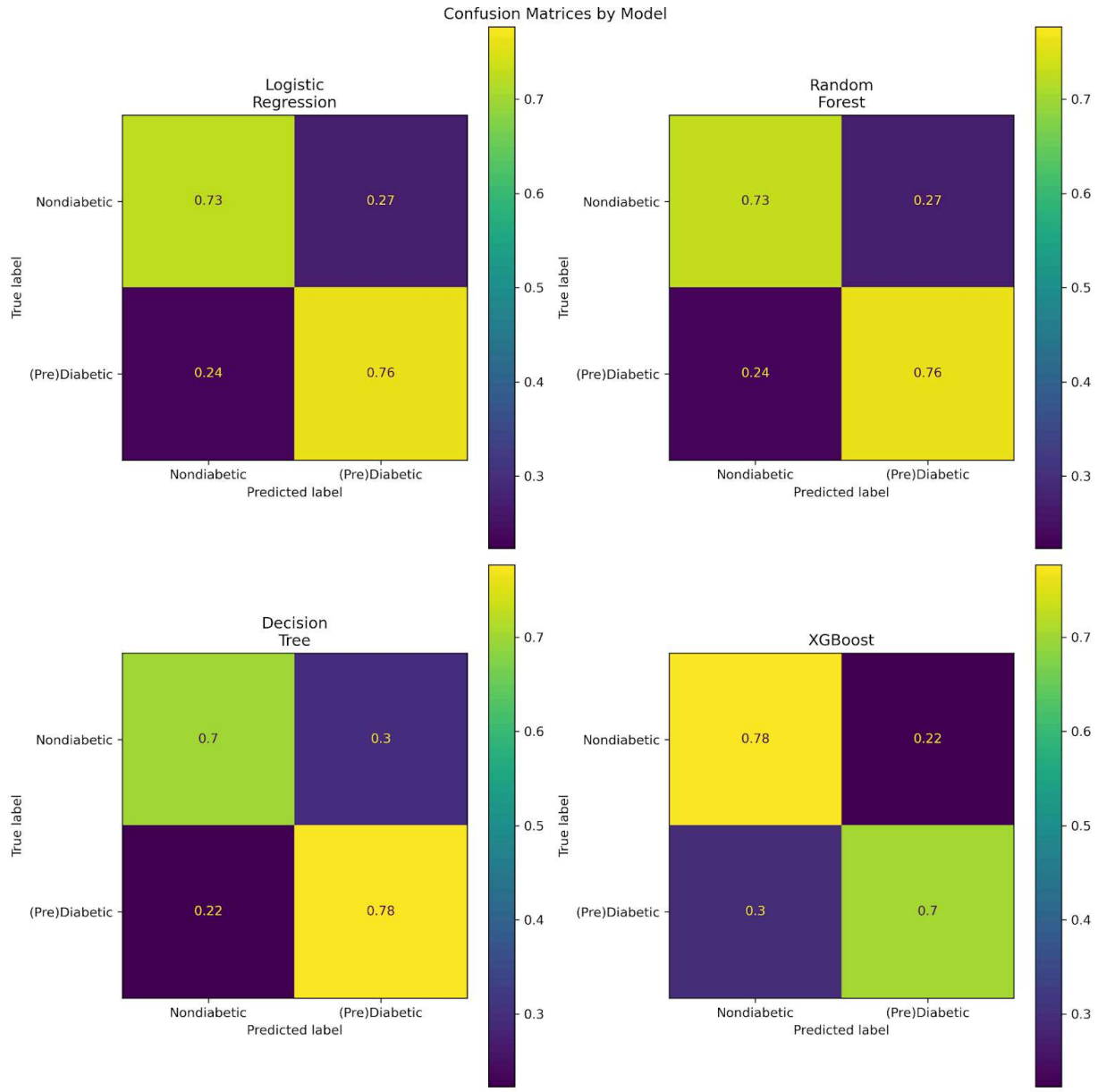

Figure 6. Confusion matrices on test sets averaged across all best models from each algorithm and normalized by true class proportions.

b. Interpretation

The XGBoost feature importances varied slightly by measure, although general trends were consistent (Figure 7). Cover and gain showed close mutual agreement, with high blood pressure as the most predictive feature. The second and third most important features by these metrics were (unordered) self-assessed general health and recent cholesterol checks, closely followed by further physical factors including high cholesterol, age, BMI, and alcohol consumption. This makes sense given high blood

pressure and cholesterol are correlated with heart disease, a known risk factor for type I diabetes, and age and BMI are known risk factors for type II [1]. Demographic and behavioral factors were found to be lowly important.

Total cover, total gain, and weight were also found to be self-consistent, identifying BMI, age, and general health among the top few features. Surprisingly, by weight, high blood pressure was ranked as least important, despite appearing in the top five by total cover and total gain, perhaps biased by its low cardinality or prevalence towards the top of tree splits. These measures weighted demographic factors like income and education higher than cover and gain and produced a more even feature importance distribution. The consensus between all five measures suggests blood pressure, cholesterol levels, general health, age, and BMI are important determinants of diabetic status, in line with intuitive expectations and known risk factors [1].

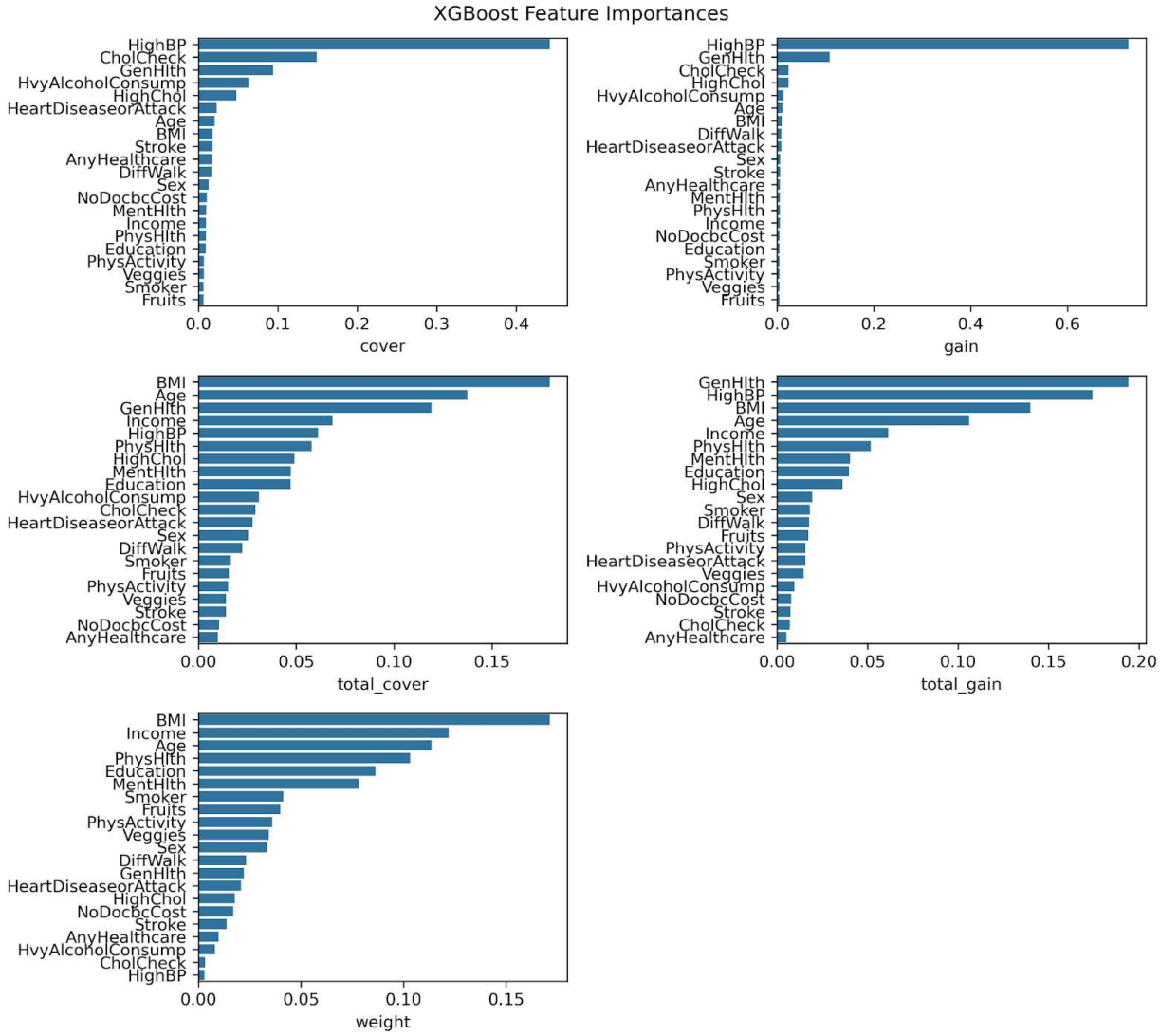

Figure 7. XGBoost feature importances using all model-specific measures.

Global SHAP values produced relative feature importances consistent with the internal XGBoost results, identifying a similar subset of top predictive features (Figure 8). The relationship between each variable and model prediction also align with expectations; poor general health, high blood pressure, high BMI, old age, and high cholesterol are associated with higher predicted diabetic probabilities. This measure also highlights specific relationships between behavioral and demographic factors with diabetes risk. For example, low income, which may burden an individual's access to healthcare or adversely influence dietary and activity habits, is associated with higher predicted diabetes probability. Being male, a known risk factor for type II diabetes [5], is associated positively with predicted probability. More niche health characteristics (e.g.

heart disease or attack history, difficulty walking, stroke status, and alcohol consumption) were globally less important but tended to exert significant influence on the points affected by these.

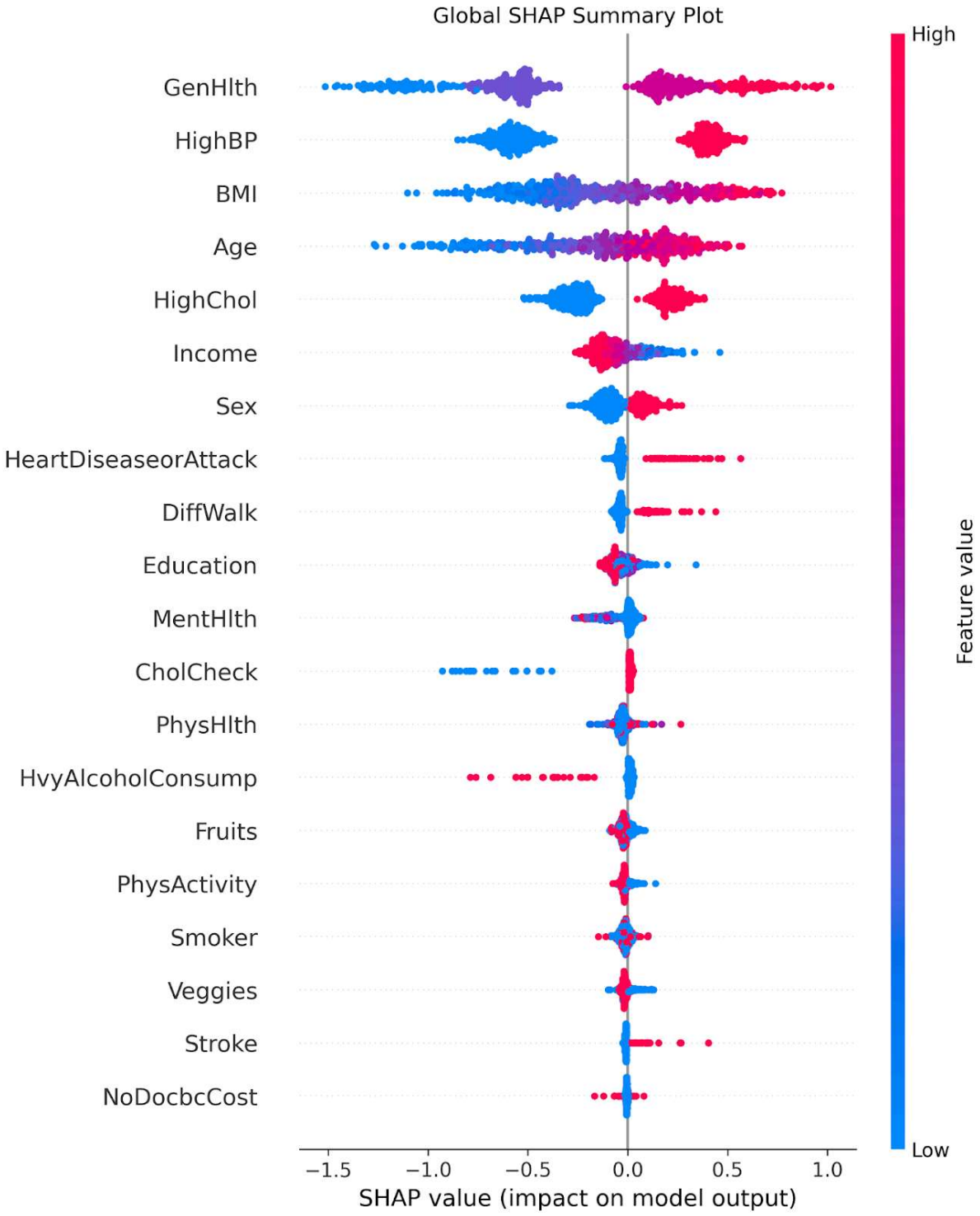

Figure 8. SHAP summary plot measured for 1% of the test set for the best-performing XGBoost model (in logit space).

Local SHAP values reveal local importances aligning sensibly with global importances (Figures 9, 10). Two randomly sampled predicted nondiabetic points (Figure 9) boasted strong general health self-assessments and no elevated cholesterol, impacting their negative prediction. One had high blood pressure but low BMI, respectively increasing and decreasing the predicted diabetic probability, while the other with normal blood pressure but high BMI observed an inverse effect.

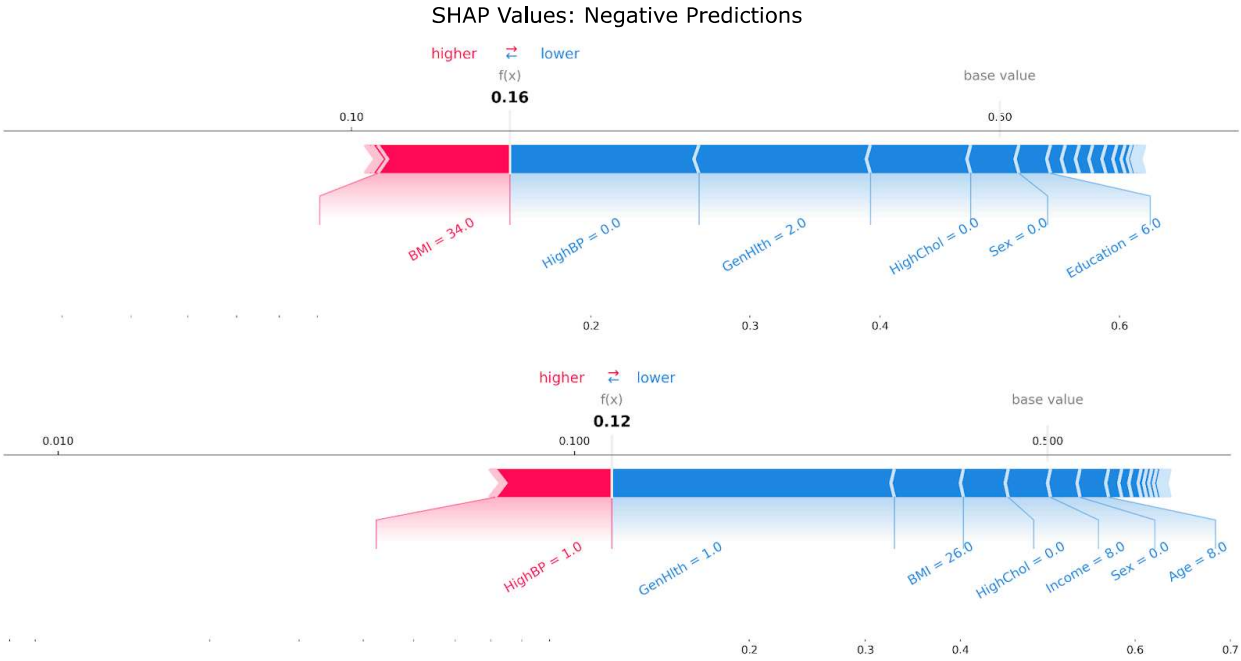

Figure 9. SHAP force plots for randomly sampled negative predicted test points.

Conversely, two randomly sampled points predicted as prediabetic or diabetic (Figure 10) overwhelmingly demonstrated risk factors, producing highly confident diabetic predictions. Each point had high blood pressure and cholesterol, a “good” to “fair” health self assessment, elevated BMI, and difficulty walking. High age (75-79) and low income ($15k-20k) made confidence slightly higher for the more confidently predicted point. Overall, the impact of the features on these predicted probabilities align expectedly with global feature importances and known risk factors.

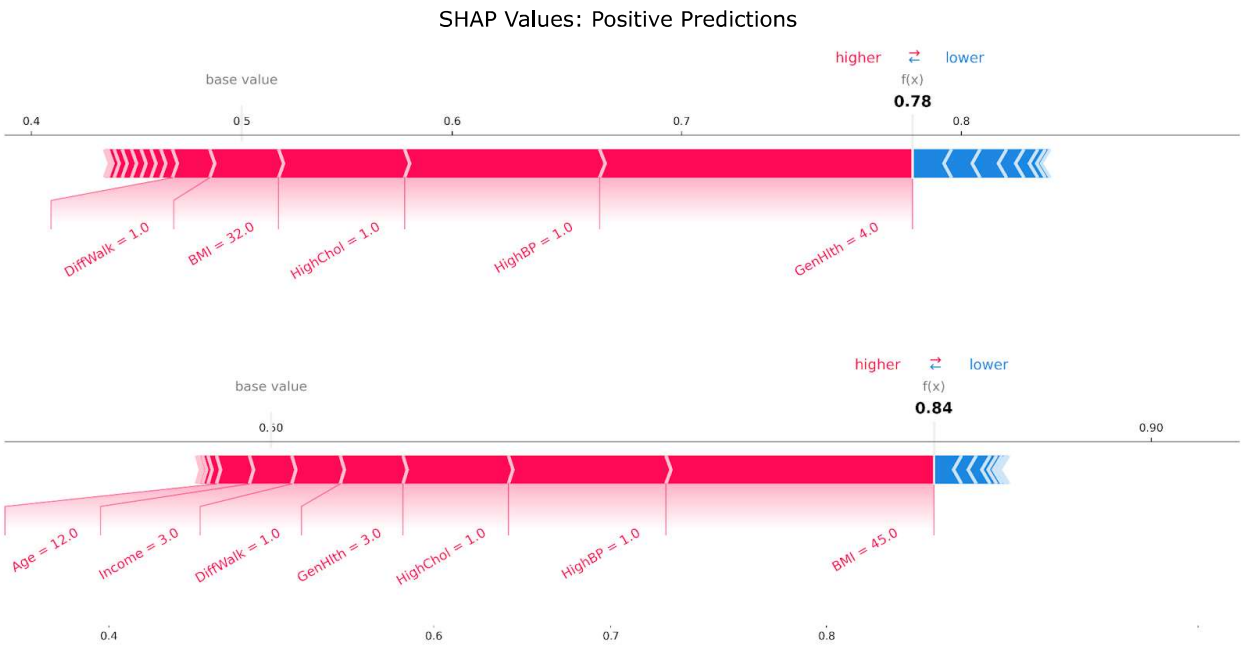

Figure 10. SHAP force plots for randomly sampled positive and negative predicted test points.

V. Outlook

A challenge in this work was computational complexity. Preliminary studies on 10% subsampled data showed a sizable decrease in performance for all models, indicating all data should be used, but making models like k-nearest neighbors or Gaussian support vector machines, whose prediction time scales training size, inaccessible during deployment. While these results also suggested such models less suitable for training and predicting on large datasets were unlikely to significantly improve performance, this stands to be proven for the full dataset.

The choice of the F1 evaluation metric depended on a conscious choice to balance healthcare system and individual burden, warranting room for exploration of other metrics. The best-performing model, XGBoost, showed poor ability to classify diabetic points, which may be preferable when healthcare resources are limited and complementary approaches are used to assess risk but undesirable otherwise.

The dataset suffers from coarse-grained detail. For example, age is binned into categories when the raw number would be preferable, and blood cholesterol is designated “high” or “low” without any laboratory measurement details. Finer resolution in feature values would likely improve model robustness. The subset of features selected, lack of feature engineering, and inclusion of all data points may have

worsened performance [3]. Model performance was underwhelming, with a maximum test F1-score of 0.455, reflecting the challenges of diabetes diagnosis without laboratory testing.

Regardless, this work demonstrates the modest ability of machine learning models to predict diabetes from limited personal data, which could be used to inform healthcare decisions.

VI. References

(1) CDC. *A Report Card: Diabetes in the United States Infographic.* Diabetes. https://www.cdc.gov/diabetes/communication-resources/diabetes-statistics.html (accessed 2025-12-10).

(2) *Behavioral Risk Factor Surveillance System*. https://www.cdc.gov/brfss/index.html (accessed 2025-12-13).

(3) Xie, Z.; Nikolayeva, O.; Luo, J.; Li, D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Prev Chronic Dis* **2019**, *16*, E130. https://doi.org/10.5888/pcd16.190109.

(4) Majcherek, D.; Ciesielski, A.; Sobczak, P. AI-Driven Analysis of Diabetes Risk Determinants in U.S. Adults: Exploring Disease Prevalence and Health Factors. *PLOS ONE* **2025**, *20* (9), e0328655. https://doi.org/10.1371/journal.pone.0328655.

(5) CDC. *Diabetes and Men.* Diabetes. https://www.cdc.gov/diabetes/risk-factors/diabetes-and-men.html (accessed 2025-12-12).