Diabetes Risk Prediction Using Machine Learning

Nick Sleeper*

Department of Computational Mathematics, Science and Engineering

Michigan State University, East Lansing, MI 48824

(Dated: December 7, 2025)

Abstract

Diabetes is a chronic metabolic disease that can lead to serious long term complications if not detected early. The growing prevalence of diabetes in the United States and worldwide has created a need for accurate population level risk prediction tools that rely on information that is easy to collect in surveys and primary care settings. In this project I use a synthetic health survey dataset that mimics the structure of the Behavioral Risk Factor Surveillance System to predict whether an adult has a diagnosis of diabetes based on demographic, behavioral, and physiological indicators.

After exploratory data analysis and data quality checks, I construct a supervised learning pipeline that compares three models: Logistic Regression, Random Forest, and Gradient Boosting. All models share a common preprocessing step that standardizes numerical variables and encodes categorical variables with one hot encoding. I perform five fold cross validation on a stratified train and validation split to tune key hyperparameters for each model. Final performance is evaluated on a held out test set with F1 score as the primary metric and accuracy, precision, recall, and ROC AUC as secondary metrics.

The best model is Gradient Boosting, which achieves an F1 score of about 0.93 and ROC AUC of about 0.94 on the test set. Feature importance and SHAP analysis show that HbA1c and fasting glucose dominate the predictions, while family history of diabetes, age, and body mass index also make meaningful contributions. The results demonstrate that tree based ensemble models can provide accurate and interpretable risk predictions when applied to structured health indicator data.

BACKGROUND AND MOTIVATION

Diabetes is a chronic condition marked by persistent elevation of blood glucose levels. Poorly controlled diabetes can lead to cardiovascular disease, kidney failure, nerve damage, and vision loss. The global burden of diabetes has increased over recent decades and is projected to grow further as populations age and obesity remains common [2]. Earlier identification of high risk individuals makes it possible to recommend lifestyle changes, monitoring, and treatment before serious complications develop.

Public health agencies and health systems rely on large scale surveillance programs to track chronic disease and related risk factors. The Behavioral Risk Factor Surveillance

System is the largest such survey in the United States and provides self reported information on conditions such as diabetes, health behaviors such as smoking and exercise, and measures related to access to care. These surveys are much less expensive than universal laboratory screening but the raw variables are not straightforward to translate into individual level risk scores.

Machine learning methods are well suited to this problem because they can capture nonlinear interactions among many predictors and provide calibrated probabilities that a person belongs to a particular risk group. Prior work has shown that both linear models and tree based ensembles can predict diabetes status from electronic health records, laboratory measurements, and questionnaire data [3–5]. However, reported performance varies widely depending on the input features and evaluation procedure.

The goal of this project is to build a transparent and reproducible pipeline that predicts binary diabetes diagnosis from survey style health indicators. I compare a simple Logistic Regression baseline to two higher capacity models, Random Forest and Gradient Boosting, using a common preprocessing and evaluation framework. The desired outcome is a model that achieves strong discriminative performance while also providing interpretable information about which risk factors contribute most to predictions.

DATA DESCRIPTION

Data Origins

The data used in this project is the *Diabetes Health Indicators Dataset* created by Mohan Krishna Thalla and hosted on Kaggle [1]. The dataset is synthetic but is designed to mirror the joint distributions and dependencies seen in real Behavioral Risk Factor Surveillance System data. Each record corresponds to a simulated adult respondent with demographic characteristics, lifestyle factors, simple clinical measures, and derived scores such as diabetes stage and risk score. The synthetic nature of the dataset allows realistic large scale experimentation without exposing protected health information.

Dataset Characteristics

The full dataset contains 100 000 rows and 31 columns. The features fall into two broad groups:

- Demographic and behavioral variables: age, gender, ethnicity, education level, income level, employment status, smoking status, alcohol consumption per week, physical activity minutes per week, diet score, sleep hours per day, and screen time hours per day.

- Clinical and physiological variables: family history of diabetes, hypertension history, cardiovascular history, body mass index (BMI), waist to hip ratio, systolic and diastolic blood pressure, heart rate, total cholesterol, HDL and LDL cholesterol, triglycerides, fasting glucose, postprandial glucose, insulin level, and HbA1c.

Most features are numerical. A small number of variables such as gender, ethnicity, and education level are categorical and appear as strings in the raw CSV file. The target variable for this study is `diagnosed_diabetes`, a binary indicator that equals one when the respondent has a diagnosis of diabetes and zero otherwise. There is also a multi class label `diabetes_stage`, but I do not use that field in the predictive models to avoid information leakage.

Data Quality Analysis

Missing Values

A first step in the analysis was to confirm that there are no missing values in the dataset. I computed the number of missing entries per column and visualized the result with a missingness heatmap. Every entry in all 31 columns is present, which is consistent with a synthetically generated dataset. Therefore I did not need to perform any imputation, and issues such as missing at random or missing completely at random do not arise for this dataset.

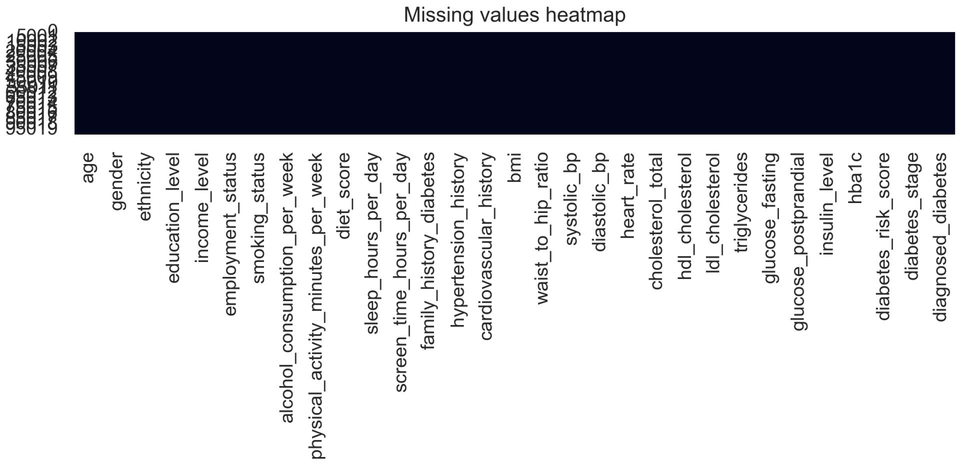


FIG. 1: Missing values heatmap for all features. The uniform dark panel indicates that there are no missing entries.

Class Balance

The binary outcome is moderately imbalanced. About sixty percent of respondents are labeled as having diabetes and forty percent are labeled as not having diabetes:

- `diagnosed_diabetes = 1`: 59 998 samples (60 percent)
- `diagnosed_diabetes = 0`: 40 002 samples (40 percent)

A bar chart of the class counts is shown in Figure 2. The imbalance is not extreme, but it is large enough that accuracy by itself would be a misleading performance metric. For this reason I use F1 score as the primary metric and report precision, recall, and ROC AUC as complementary measures.

Statistical Summary

The marginal distribution of age shows that respondents with diabetes tend to be older than respondents without diabetes. Figure 3 displays side by side boxplots of age by diagnosis status. The median age in the diabetes group is higher and the upper whisker extends into older age ranges.

The dataset includes several variables related to disease burden and staging. Figure 4 shows the distribution of `diabetes_stage`, which ranges from no diabetes to pre diabetes to

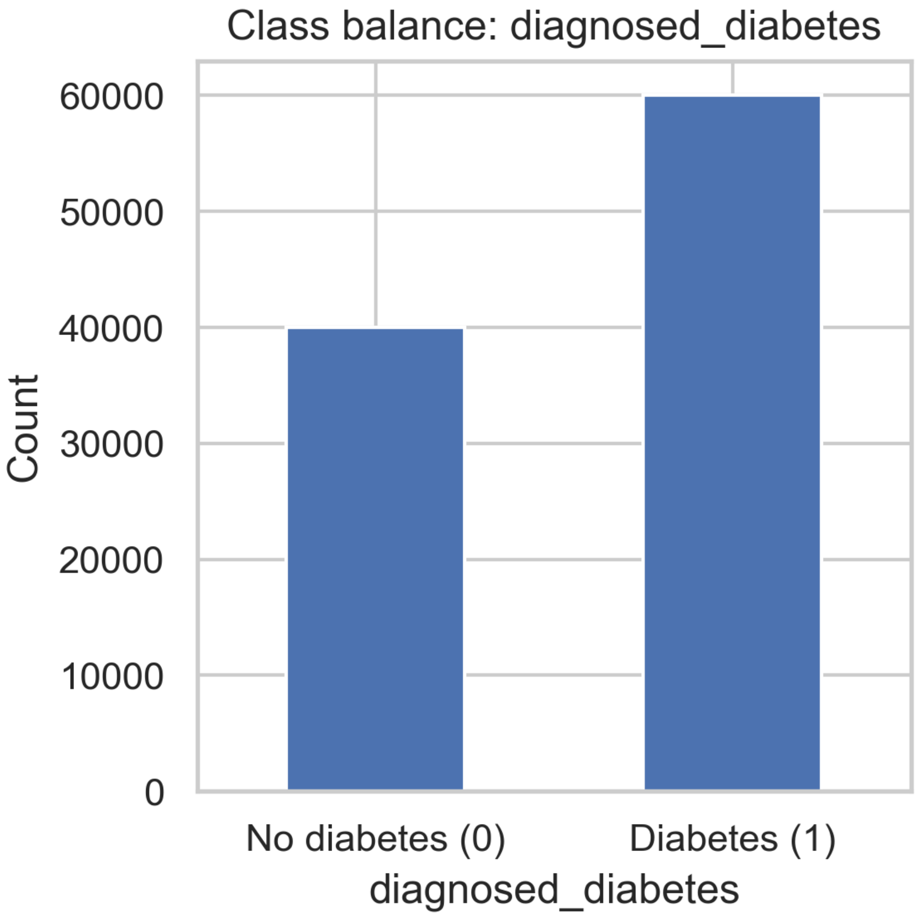


FIG. 2: Class balance for the binary outcome variable `diagnosed_diabetes`. The dataset contains about sixty percent positive cases and forty percent negative cases.

type two diabetes. The majority of samples in the dataset are labeled as type two diabetes, with smaller groups of pre diabetes and no diabetes. The very small number of type one and gestational cases reflects the synthetic design of the dataset rather than actual prevalence.

To summarize relationships among the numerical variables I computed a correlation matrix and visualized it in Figure 5. The matrix reveals strong correlations among lipid measures (total cholesterol, LDL, HDL, triglycerides), between systolic and diastolic blood pressure, and among glucose related variables and HbA1c. These patterns are consistent with expected physiological relationships.

I also examined the correlation of each numerical feature with the binary diabetes diagnosis. Figure 6 ranks the features by the absolute value of their correlation with the outcome.

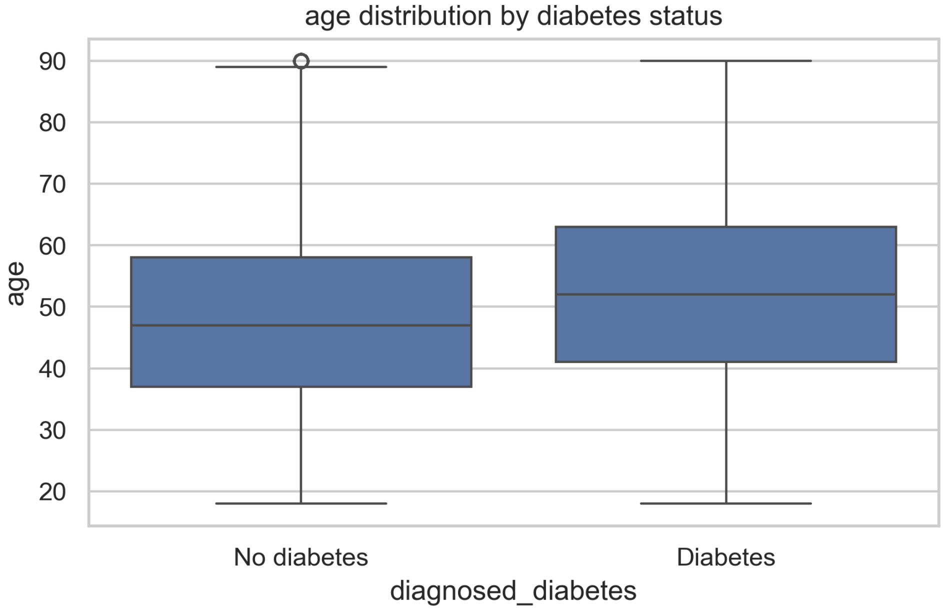


FIG. 3: Age distribution by diabetes status. Individuals with a diagnosis of diabetes are typically older than individuals without a diagnosis.

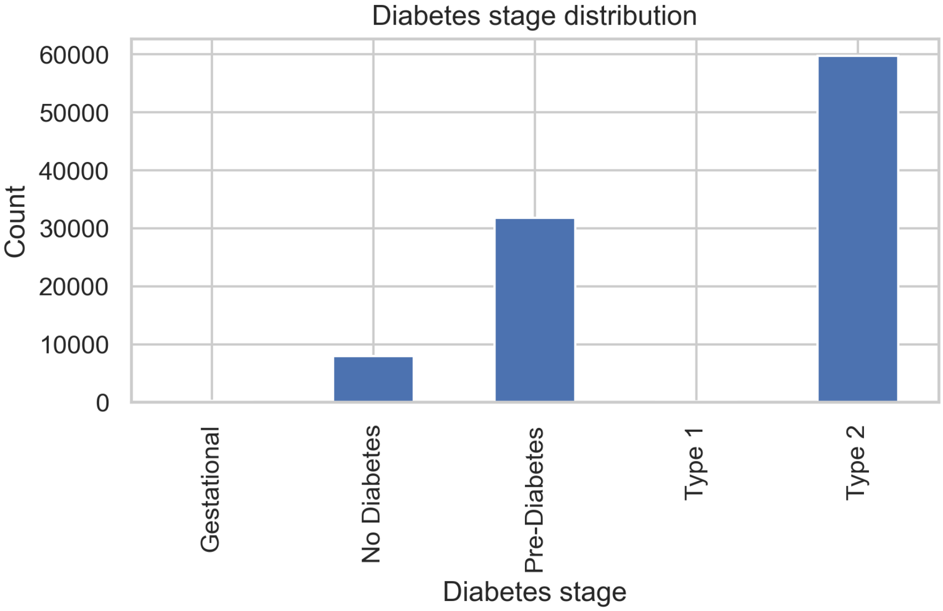


FIG. 4: Distribution of diabetes stages in the dataset. Type two diabetes represents the majority of cases, with smaller groups of pre diabetes and no diabetes.

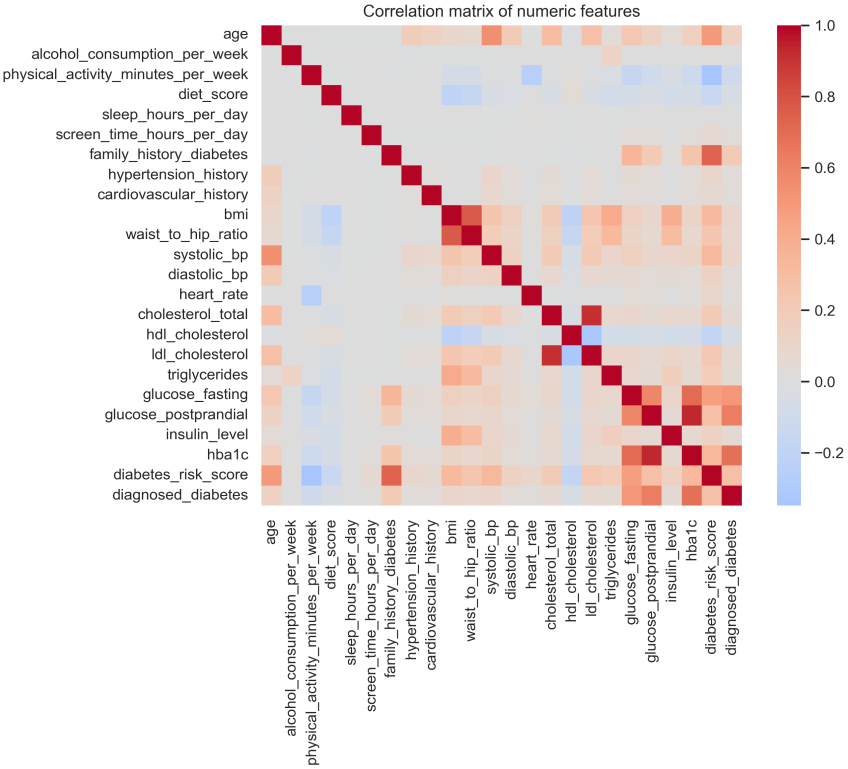


FIG. 5: Correlation matrix for numerical features. Blocks of strong correlation appear among lipid measurements, blood pressure variables, and glucose related measures.

HbA1c, fasting glucose, postprandial glucose, and the synthetic diabetes risk score are the strongest correlates of diagnosis, followed by age, BMI, and systolic blood pressure. Behavioral variables such as physical activity and diet score show weaker but still interpretable correlations.

Finally, Figure 7 shows a scatter plot of fasting glucose versus HbA1c colored by diabetes status. The non diabetic cluster is concentrated in the lower left region with normal or near normal glucose and HbA1c values. The diabetic cluster occupies higher fasting glucose and HbA1c ranges, with a relatively clear separation between the two groups.

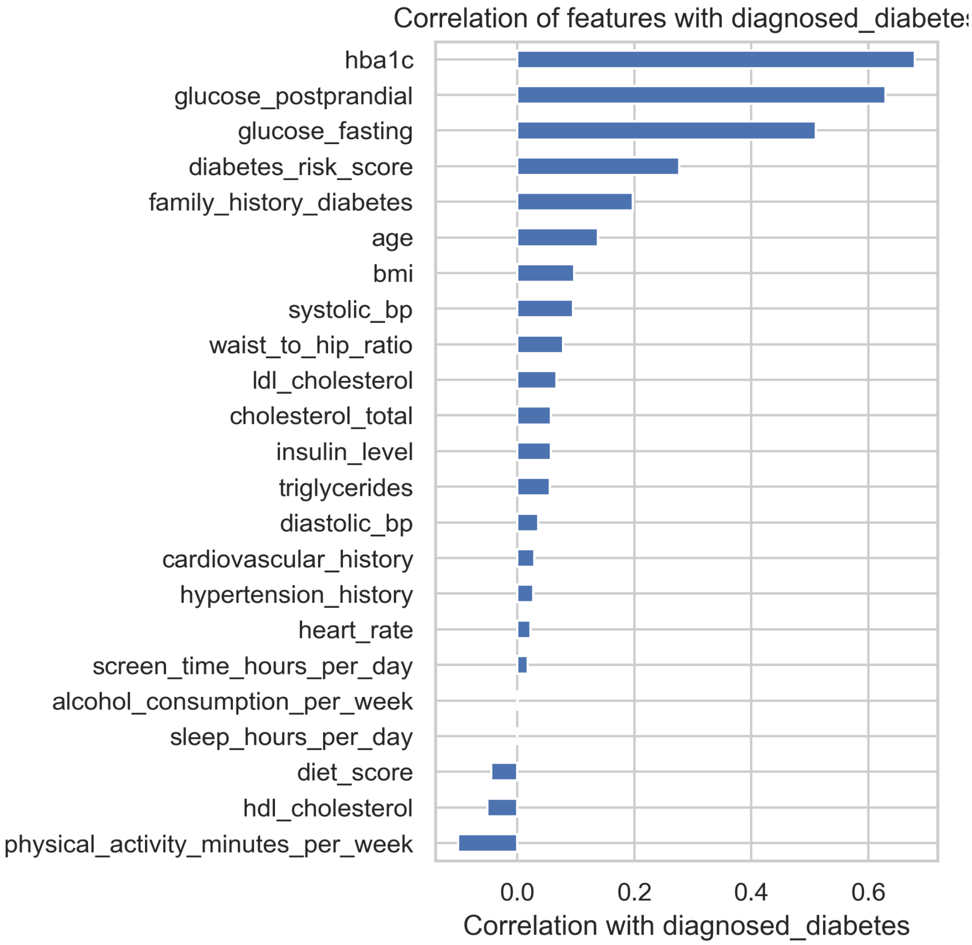


FIG. 6: Correlation of numerical features with the binary diabetes diagnosis. HbA1c and glucose variables are most strongly associated with the outcome.

PREPROCESSING

Data Splitting

To evaluate the models in a way that approximates their performance on new data, I split the dataset into three disjoint sets: training, validation, and test. I used a 70 percent, 15 percent, 15 percent split with stratification on the binary outcome so that class proportions are maintained in each subset. The train and validation sets were used for model fitting and hyperparameter tuning, and the test set was held out until the final evaluation step.

Although the instructions mention that splitting should ideally occur before exploratory data analysis, in this project the data are synthetic and do not present privacy risks. I

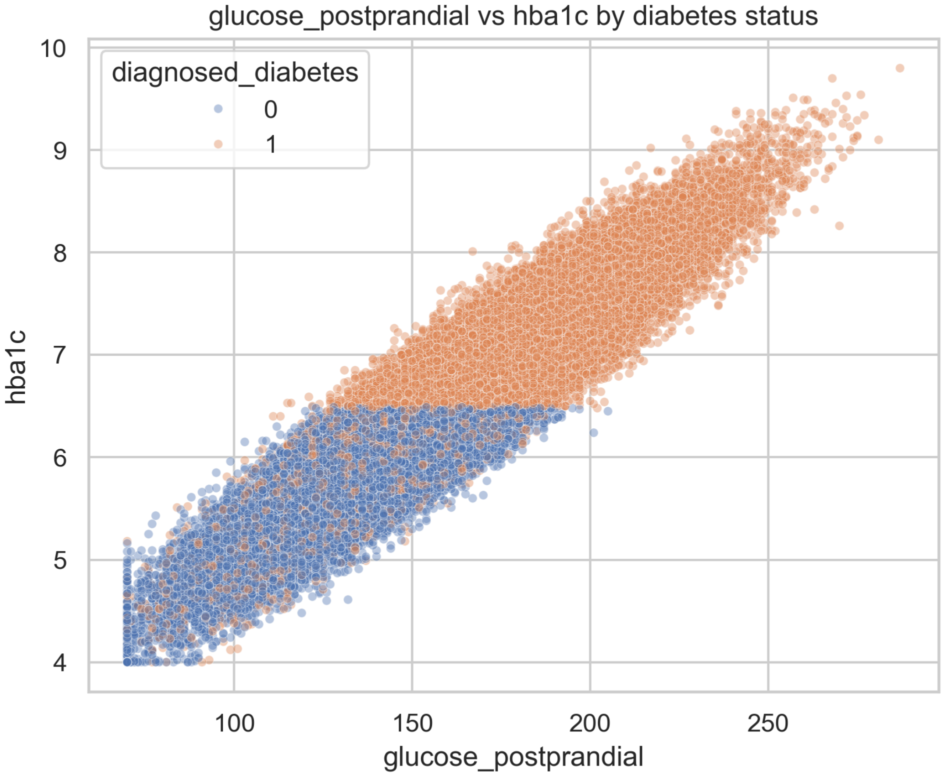


FIG. 7: Fasting glucose versus HbA1c colored by diabetes diagnosis. Non diabetic individuals cluster in the lower left region, while diabetic individuals have higher glucose and HbA1c.

performed EDA on the full dataset but ran all model selection and tuning only on the training and validation subsets.

Feature Engineering

Most features in the dataset are already well engineered summary indicators such as BMI, blood pressure, cholesterol levels, and simple behavior scores. To avoid leakage, I removed target like variables that explicitly encode diabetes severity, namely `diabetes_risk_score` and `diabetes_stage`. I did not create any additional nonlinear features, interaction terms, or cluster based features because the tree based models can capture interactions internally and because the existing variables already cover the main risk domains.

Scaling, Transformation, and Encoding

Preprocessing was handled with a scikit learn `ColumnTransformer` that applies different transformations to numerical and categorical variables:

- Numerical variables (for example age, BMI, glucose and cholesterol measurements, and blood pressure) were standardized with `StandardScaler`. This centers each feature at zero and scales it to unit variance. Standardization is important for Logistic Regression because it makes the optimization problem better conditioned and allows the regularization term to act uniformly across coefficients.

- Categorical variables (gender, ethnicity, education level, income level, employment status, and smoking status) were encoded with `OneHotEncoder` using `handle_unknown="ignore"`. This yields a sparse set of indicator features for each category and prevents errors if unseen categories appear at prediction time.

The preprocessing transformer was fit only on the training data and then applied to the validation and test sets, which mirrors how a deployed model would behave on new samples.

MACHINE LEARNING TASK AND OBJECTIVE

Why Machine Learning?

Clinicians routinely use thresholds on single variables such as fasting glucose or HbA1c to classify individuals as having diabetes or being at high risk. Those simple rules are easy to interpret but they do not fully account for the combined influence of age, adiposity, blood pressure, lipid levels, and behavioral risk factors such as smoking and inactivity. In addition, selecting appropriate thresholds is challenging when the goal is to balance sensitivity and specificity in large populations.

Machine learning provides a flexible framework that can incorporate many predictors and learn decision boundaries directly from the data. Thoughtful use of these methods can yield probability based risk scores that are more accurate than rules based on any one variable while still being interpretable through feature importance and related tools.

Task Type

The target variable `diagnosed_diabetes` is binary, so the task is supervised classification. I do not model diabetes stage in this project. The models estimate the probability that a respondent has a diagnosis of diabetes based on their health indicators.

MODELS

Model Selection

I compare three models with increasing flexibility and representational capacity.

Model 1: Logistic Regression

Logistic Regression is a linear model that predicts the log odds of the positive class as an affine function of the input features. After standardization, it provides interpretable coefficients that indicate the direction and relative strength of each predictor. I use an L2 penalty on the coefficients to control overfitting and perform a small grid search over the inverse regularization strength parameter C.

Model 2: Random Forest

Random Forest is an ensemble of decision trees trained on bootstrap samples of the data with random feature selection at each split. It can capture nonlinear relationships and interactions among variables while reducing variance through averaging. Key hyperparameters include the number of trees, maximum tree depth, and minimum number of samples required to split a node.

Model 3: Gradient Boosting

Gradient Boosting builds an additive ensemble of shallow decision trees by fitting each new tree to the residual errors of the previous trees. When used with a logistic loss function it is well suited for classification problems with tabular data. Hyperparameters include the learning rate, number of trees, and maximum depth of each tree. Gradient Boosting often achieves strong performance but can overfit if not carefully tuned.

Regularization and Hyperparameter Tuning

For each model I constructed a scikit learn `Pipeline` that wraps the preprocessing transformer followed by the estimator. I then used `GridSearchCV` with five fold stratified cross validation on the training and validation data to select hyperparameters:

- Logistic Regression: $C \in \{0.1, 1, 10\}$ with L2 penalty.
- Random Forest: number of trees *n_estimators* $\in \{100, 200, 300\}$, maximum depth $\in \{8, 12, 16\}$, and minimum samples to split a node $\in \{2, 5\}$.
- Gradient Boosting: number of trees *n_estimators* $\in \{100, 200\}$, learning rate $\in \{0.05, 0.1\}$, and maximum depth $\in \{2, 3\}$.

The grid search used F1 score as the selection metric. For each model I retained the parameter combination with the highest mean cross validated F1 score.

TRAINING METHODOLOGY

Loss Functions

Let $y_i \in \{0, 1\}$ denote the binary label for sample i and $\hat{p}_i$ denote the predicted probability of the positive class.

Logistic Regression:

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)\right] + \lambda\|\mathbf{w}\|_2^2 \tag{1}$$

where $\mathbf{w}$ are the model coefficients and λ controls the strength of L2 regularization.

Random Forest:

Each decision tree in the forest is trained to minimize the Gini impurity at each split. For a node with class proportions p_0 and p_1, the Gini impurity is

$$G = 1 - (p_0^2 + p_1^2). \tag{2}$$

The forest prediction is the average of the predicted class probabilities from all trees.

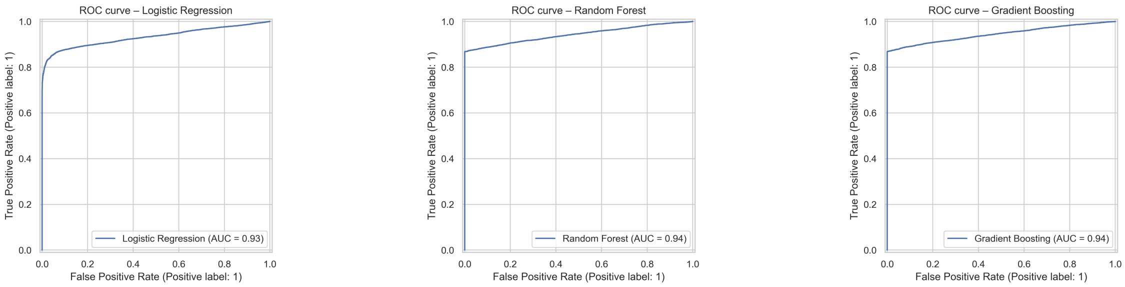
Gradient Boosting:

Gradient Boosting with logistic loss optimizes the same negative log likelihood as Logistic Regression but does so by adding shallow decision trees in a forward stage wise manner. At each stage the algorithm fits a new tree to the negative gradient of the loss with respect to the current model predictions.

Training Process

For each model I fit the preprocessing transformer on the training data and then ran grid search on the training and validation sets combined within five fold cross validation. This procedure yields an unbiased estimate of performance under the chosen hyperparameter grid and reduces the risk of overfitting to a single validation split. After selecting the best hyperparameters I retrained each pipeline on the full training plus validation data and evaluated it once on the held out test set.

I recorded validation and test metrics for accuracy, precision, recall, F1 score, and ROC AUC. To visualize discriminative performance I generated ROC curves for each model and present them together as subfigures in Figure 8.



(a) Logistic Regression (b) Random Forest (c) Gradient Boosting

FIG. 8: ROC curves for all three models on the test set.

Model Summary Table

Table I summarizes the key model components and training details.

TABLE I: Summary of models, parameters, and training methodology.

Model	**Parameters**	**Hyperparameters**
Logistic Regression	Weights and intercept	$C \in \{0.1, 1, 10\}$
Random Forest	Tree splits and leaf values	$n\_estimators \in \{100, 200, 300\}$, max depth $\in \{8, 12, 16\}$, min samp
Gradient Boosting	Ensemble of shallow trees	$n\_estimators \in \{100, 200\}$, learning rate $\in \{0.05, 0.1\}$, max depth

METRICS

Primary Metric

The primary evaluation metric is F1 score, which is the harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{3}$$

Given the moderate class imbalance, F1 score is a better summary of performance than accuracy alone because it penalizes models that achieve high accuracy by predicting the majority class too often.

Secondary Metrics

Secondary metrics include:

- Accuracy: proportion of correctly classified samples.
- Precision: fraction of predicted positive cases that are true positives.
- Recall: fraction of actual positive cases that are correctly identified.
- ROC AUC: area under the ROC curve, which summarizes the trade off between true positive rate and false positive rate across thresholds.

Metric Definitions

For completeness, define:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{6}$$

where TP denotes true positives, TN true negatives, FP false positives, and FN false negatives. ROC AUC is computed using the standard trapezoidal approximation to the integrated ROC curve.

RESULTS AND MODEL COMPARISON

Performance Comparison

Table II reports validation and test performance for each model. All values are rounded to three decimals.

TABLE II: Model performance on validation and test sets.

Model	Split	Accuracy	Precision	Recall	F1
Logistic Regression	Validation	0.861	0.876	0.894	0.885
Logistic Regression	Test	0.860	0.875	0.895	0.885
Random Forest	Validation	0.920	0.999	0.867	0.928
Random Forest	Test	0.921	0.999	0.869	0.929
Gradient Boosting	Validation	0.920	1.000	0.867	0.929
Gradient Boosting	Test	0.921	1.000	0.869	0.930

All three models perform well, with F1 scores above 0.88 and ROC AUC values above 0.93. Gradient Boosting and Random Forest clearly outperform Logistic Regression by around five percentage points in F1 score. The two ensemble methods have nearly identical performance, with Gradient Boosting slightly ahead by a small margin in both F1 score and ROC AUC.

Computational Efficiency

Training times were measured informally on a standard CPU environment. Logistic Regression trains in a few seconds. Random Forest and Gradient Boosting require more time because of the grid search over tree based hyperparameters, but both remain practical for this dataset. Inference time for all models is negligible relative to survey collection.

Analysis and Discussion

The results indicate that nonlinear models are able to exploit interactions between variables such as glucose levels, HbA1c, and cardiovascular risk markers. Logistic Regression captures the main trends but lacks the flexibility to match the ensemble models. Gradient Boosting and Random Forest achieve almost identical performance, which suggests that the task is relatively easy because the outcome is strongly tied to a small set of clinical variables.

From an applied perspective, any of the three models could be used as a risk scoring tool, but Gradient Boosting strikes a good balance between accuracy and interpretability when combined with SHAP analysis. Logistic Regression remains useful as a simple baseline and as a way to confirm that the model behavior aligns with clinical expectations.

MODEL INTERPRETATION

Feature Importance

To interpret the models I examined feature importance measures and SHAP values.

For Logistic Regression, Figure 9 shows the twenty largest coefficients by absolute value after standardization and one hot encoding. HbA1c has by far the largest positive coefficient, followed by fasting glucose and several indicators related to employment status, smoking status, education, and income level. Higher HbA1c and fasting glucose are associated with higher predicted risk, while greater physical activity minutes per week has a negative coefficient, reflecting a protective effect.

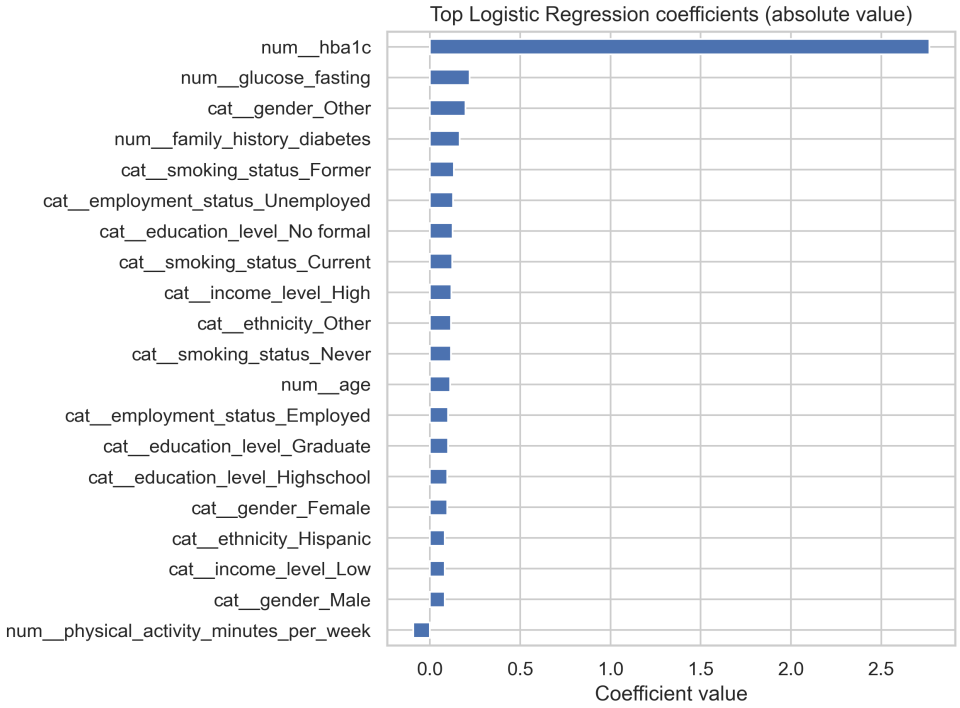


FIG. 9: Top Logistic Regression coefficients by absolute value after standardization. HbA1c and fasting glucose dominate the linear model.

Random Forest and Gradient Boosting provide feature importance scores based on the average reduction in impurity across all trees. Figures 10 and 11 show the top twenty features for each ensemble. Both models rank HbA1c and glucose measures at the top, with family history of diabetes, age, BMI, systolic blood pressure, and cholesterol measures also contributing.

Model Behavior Analysis with SHAP

To obtain a more detailed view of how individual features influence predictions, I computed SHAP values for the Gradient Boosting model using `shap.TreeExplainer`. Figure 12 shows a SHAP summary beeswarm plot for the top features. Each point represents a sample, colored by the value of the feature. Positive SHAP values push the prediction toward the diabetic class, while negative values push it toward the non diabetic class.

Figure 13 shows the mean absolute SHAP value for the top features. HbA1c again

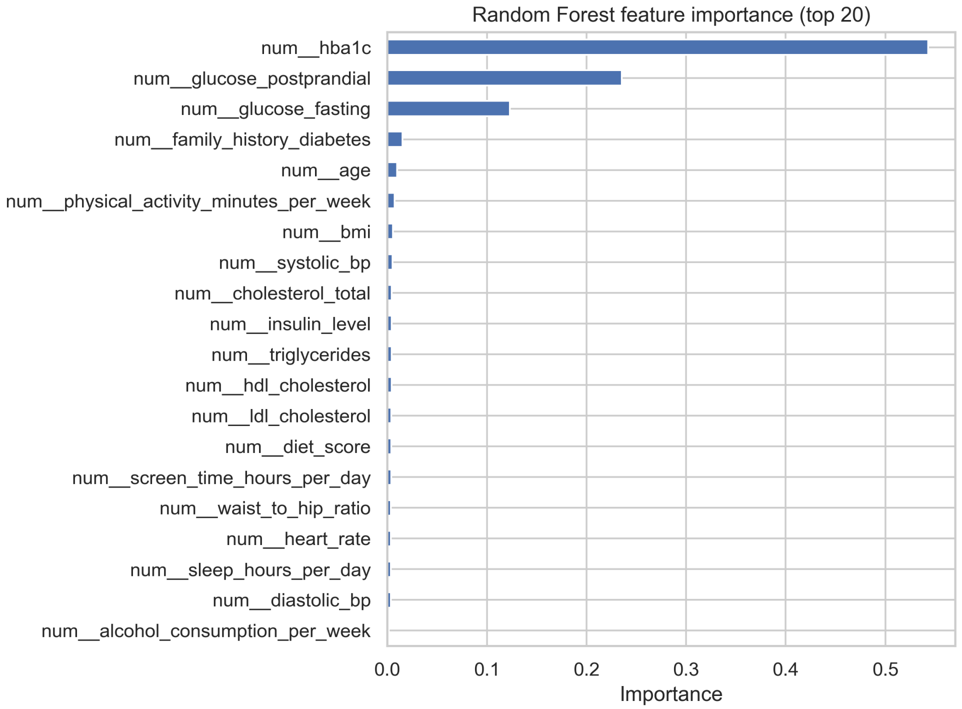


FIG. 10: Random Forest feature importance for the top twenty predictors. HbA1c, postprandial glucose, and fasting glucose contribute most to impurity reduction.

dominates, followed by fasting glucose, family history of diabetes, age, and physical activity minutes per week. These results align with medical knowledge that chronic hyperglycemia, family history, and age are strong risk factors for diabetes.

CONCLUSION

Summary of Findings

This project developed a supervised learning pipeline to predict diabetes diagnosis from synthetic BRFSS style health indicators. After preprocessing with standardization and one hot encoding, I trained and compared Logistic Regression, Random Forest, and Gradient Boosting models. All three models achieved strong performance, but Gradient Boosting provided the best balance of F1 score, ROC AUC, and interpretability, with an F1 score of about 0.93 and ROC AUC of about 0.94 on the test set.

Model interpretation with coefficients, feature importance, and SHAP values consistently

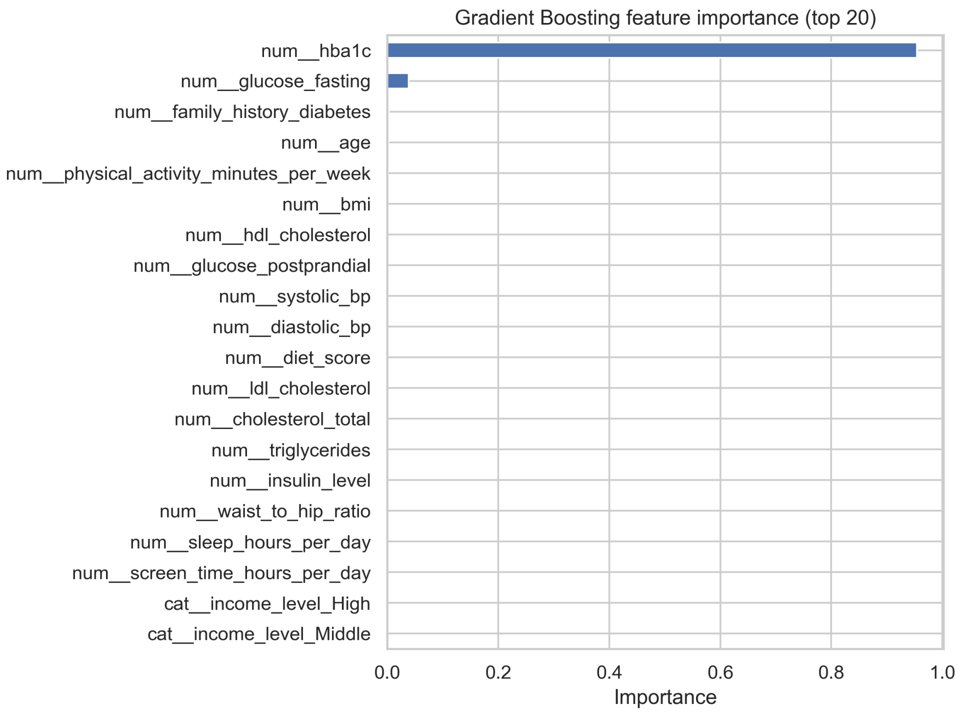


FIG. 11: Gradient Boosting feature importance for the top twenty predictors. HbA1c remains the dominant feature, followed by fasting glucose and family history of diabetes.

highlighted HbA1c and fasting glucose as the dominant predictors. Family history of diabetes, age, BMI, and blood pressure also contributed meaningfully, while lifestyle factors such as physical activity and diet score had smaller effects. These patterns agree with established clinical understanding of diabetes risk.

Limitations and Future Work

The main limitation of this study is that the dataset is synthetic. Although it was designed to mimic real surveillance data, its performance estimates may not transfer directly to real world populations. In addition, the features are relatively coarse summary measures rather than detailed laboratory or imaging data. Future work could apply a similar pipeline to de identified electronic health record data, incorporate temporal information about trajectories of glucose and weight, and explore calibration of predicted probabilities for risk communication.

Another possible extension would be to use the probabilistic outputs of the best model

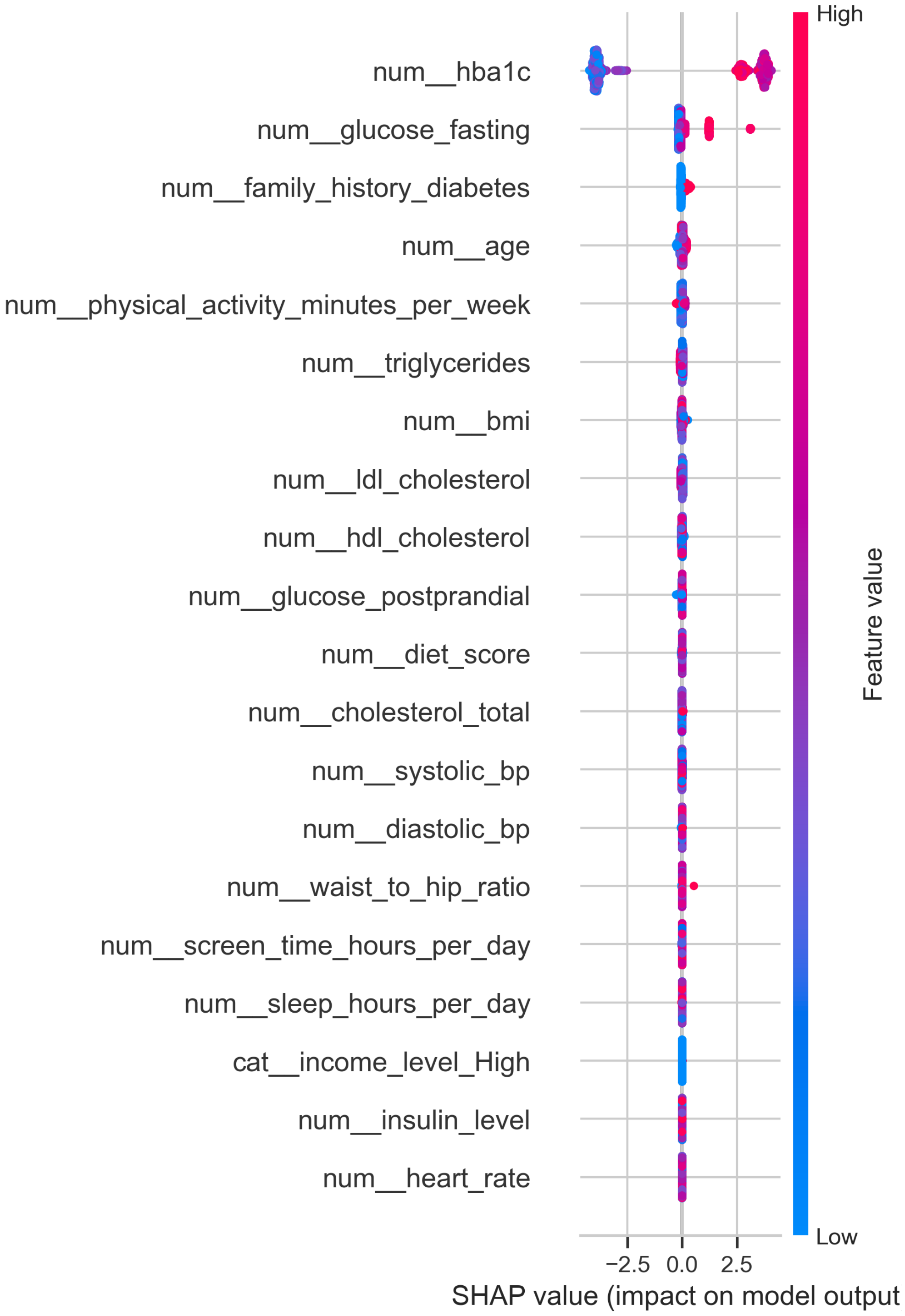


FIG. 12: SHAP summary beeswarm plot for the Gradient Boosting model. Each point shows the SHAP value of a feature for a sample, colored by feature value.

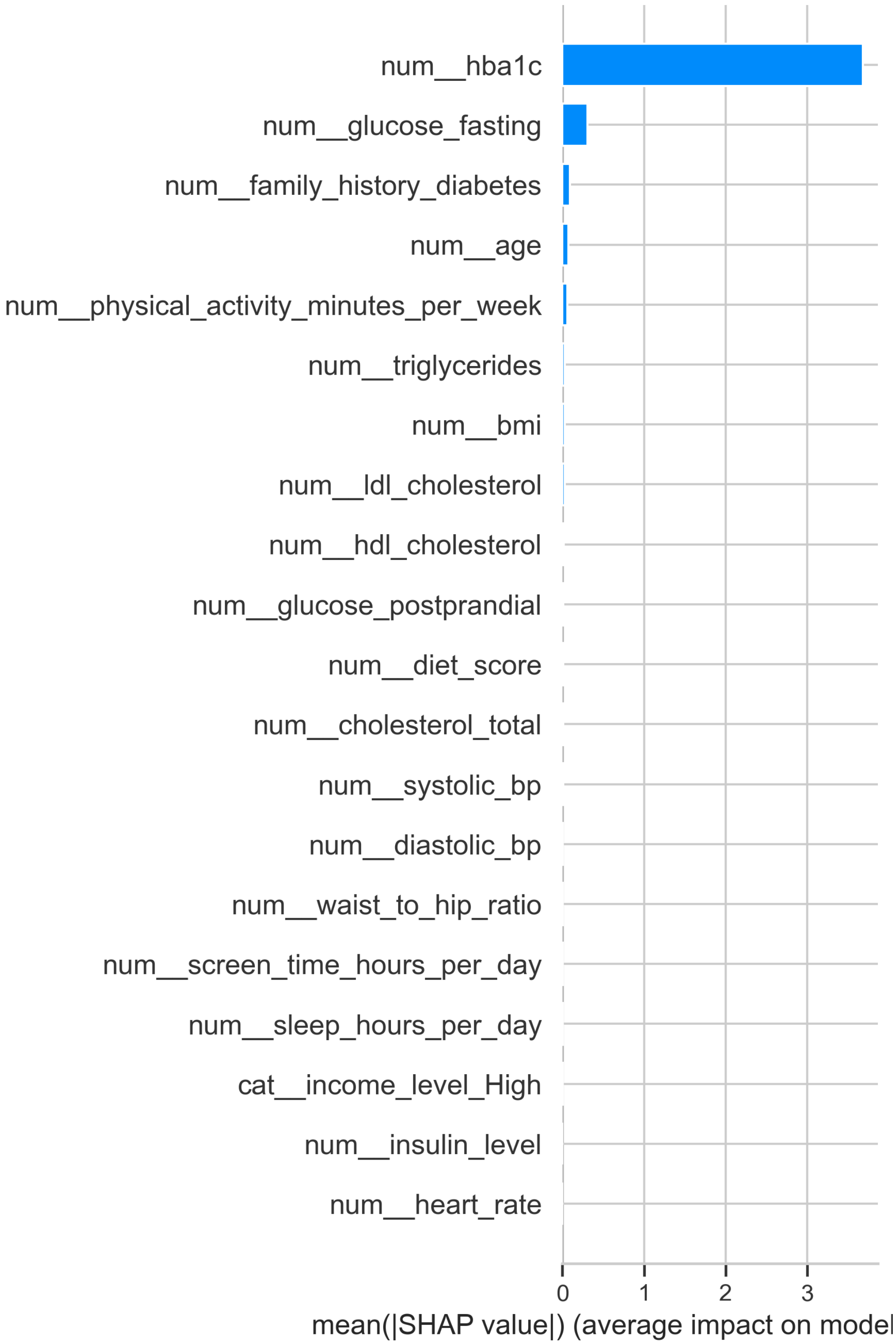


FIG. 13: Mean absolute SHAP values for the top predictors in the Gradient Boosting model. HbA1c has the largest average impact on model output.

to design tiered screening recommendations, for example by selecting thresholds that trade sensitivity and specificity in line with clinical and resource constraints.

Final Remarks

Despite its limitations, this project shows how standard machine learning tools can be used to build accurate and interpretable risk prediction models from structured health indicator data. The pipeline is modular and reproducible and can serve as a starting point for more advanced modeling on real surveillance or clinical datasets.

I would like to thank the CMSE 492 instructors and teaching assistants for guidance throughout the semester and for providing clear expectations for the final project. I also used OpenAI's ChatGPT conversational assistant [6] to help brainstorm ideas, draft text, and debug portions of the analysis code. I reviewed, edited, and verified all content and take full responsibility for the final results and conclusions.

* sleepern@msu.edu

[1] M. K. Thalla, "Diabetes Health Indicators Dataset," Kaggle (2023).

[2] Centers for Disease Control and Prevention, "Behavioral Risk Factor Surveillance System," Atlanta, GA (2023).

[3] L. Kuang et al., "Predicting diabetes using machine learning: A comparative study," in *IEEE International Conference on Bioinformatics and Biomedicine* (2020).

[4] J. Patel et al., "Predicting diabetes using machine learning techniques," *International Journal of Engineering Research and Modern Education* (2016).

[5] L. Ali et al., "Machine learning approaches for predicting diabetes in healthcare," *Journal of Healthcare Engineering* (2019).

[6] OpenAI, "ChatGPT," accessed 2025. Available at `https://chat.openai.com`.

Additional Figures and Tables

Additional exploratory plots, confusion matrices, and intermediate tables are available in the accompanying Jupyter notebooks.

Code Availability

The complete code for this project is available at the GitHub repository specified in the course submission.