Statistical and Machine Learning Models for Diabetes Diagnosis

Sathvik Kommireddy
Rancho Cucamonga High School,
Rancho Cucamonga, CA

Abstract

This study evaluates the use of logistic regression and supervised machine learning models to predict diabetes diagnosis based on demographic factors, medical history, and blood culture results. Logistic regression is employed for feature selection, helping to identify key risk indicators. The performance of several machine learning binary classification algorithms is compared using multiple goodness-of-fit metrics. Results highlight the added value of laboratory data and ensemble methods in improving diagnostic performance.

Keywords: diabetes, diagnosis, binary logistic regression, binary probit regression, binary complementary log-log regression, random forest, gradient boosting, support vector machine, k-nearest neighbor classifier, naive Bayes classifier, artificial neural network

1 Introduction

1.1 Diabetes Mellitus: Mechanisms, Types, and Risk Factors

Diabetes mellitus is a chronic metabolic disorder characterized by elevated levels of blood glucose (hyperglycemia), which result from defects in insulin production, insulin action, or both. Glucose, the body's primary energy source, comes from the food we eat and is absorbed into the bloodstream during digestion. Under normal physiological conditions, insulin, a hormone secreted by the beta cells of the pancreas, facilitates the uptake of glucose into muscle, fat, and liver cells for energy or storage. In people with diabetes, this process is impaired, causing glucose to accumulate in the blood rather than being properly utilized by the body's cells (American Diabetes Association, 2014; National Institute of Diabetes and Digestive and Kidney Diseases, 2022).

There are several types of diabetes, but the two most common forms are type 1 diabetes and type 2 diabetes. Type 1 diabetes is an autoimmune condition in which the immune system mistakenly destroys the insulin-producing beta cells in the pancreas. As a result, the body produces little or no insulin. This form of diabetes is usually diagnosed in children, adolescents, or young adults and requires lifelong insulin therapy for survival (Atkinson et al., 2014). Patients with type 1 diabetes must monitor their blood glucose regularly and administer insulin either via injection or an insulin pump to maintain glycemic control.

Type 2 diabetes, on the other hand, is characterized by insulin resistance, a condition in which the body's cells do not respond properly to insulin. Over time, the pancreas may also produce less insulin, exacerbating the problem. Unlike type 1 diabetes, type 2 diabetes is more prevalent in adults, though it is increasingly being diagnosed in children due to rising rates of obesity and sedentary lifestyles. Risk factors for type 2 diabetes include obesity, physical inactivity, poor diet, family history of diabetes, and advancing age. This form of diabetes can often be managed with lifestyle modifications (diet and exercise), oral medications, and sometimes insulin (DeFronzo et al., 2015).

If left untreated or poorly managed, both types of diabetes can lead to serious complications such as cardiovascular disease, nerve damage (neuropathy), kidney failure (nephropathy), vision problems (retinopathy), and lower-limb amputations. Therefore, early diagnosis and ongoing management are crucial for preventing complications and maintaining quality of life.

1.2 A Brief History of Diabetes: From Ancient Despair to Scientific Breakthroughs

Diabetes has been a persistent and often deadly disease for centuries. The earliest known description of a diabetes-like condition dates back to ancient Egypt around 1500 BCE, where documents such as the Ebers Papyrus mentioned excessive urination and weight loss—classic symptoms of the disease. In antiquity, diabetes was considered a fatal condition. Without an understanding of its physiological causes, treatments were largely ineffective. One common approach in the 19th century involved placing patients on strict starvation diets to reduce sugar intake and extend life slightly. Unfortunately, this often resulted in death from malnutrition or starvation rather than diabetes itself (Bliss, 1982).

A major turning point in understanding the disease came in 1889, when German physicians Joseph von Mering and Oskar Minkowski surgically removed the pancreas from a dog. The animal subsequently developed symptoms now known to be indicative of diabetes mellitus, including polyuria and glycosuria, and ultimately died. This experiment provided the first strong evidence that the pancreas played a critical role in blood sugar regulation (von

Mering and Minkowski, 1890).

Building on this foundation, Canadian surgeon Frederick Banting and medical student Charles Best conducted groundbreaking research in 1921 at the University of Toronto. They replicated Mering and Minkowski's procedure and went further, isolating a pancreatic extract that could lower blood sugar levels in diabetic dogs. Working alongside James Collip and John Macleod, they successfully purified insulin from the pancreas of a cow, which they used in 1922 to treat a 14-year-old patient, Leonard Thompson—the first human to receive insulin therapy. The results were dramatic and lifesaving. This discovery marked the beginning of effective treatment for diabetes, transforming it from a fatal disease into a manageable chronic condition (Bliss, 1993).

Further advancements in diabetes research continued into the 20th century. In 1936, British physician Sir Harold Percival (Harry) Himsworth published a seminal paper distinguishing between two types of diabetes based on insulin sensitivity. He classified type 1 diabetes as insulin-sensitive (insulin-deficient) and type 2 diabetes as insulin-insensitive (insulin-resistant), laying the foundation for our modern classification system (Himsworth, 1936).

In recent decades, technological innovations have dramatically improved the quality of life for people living with diabetes. Continuous Glucose Monitors (CGMs) allow real-time tracking of glucose levels, reducing the need for finger-prick tests and enabling more precise glucose control. When paired with insulin pumps, which deliver insulin automatically in response to glucose levels, these tools form an integrated system that mimics some functions of a healthy pancreas. This closed-loop system—often referred to as an "artificial pancreas"—is particularly transformative for individuals with type 1 diabetes, who now commonly live well into their 70s and beyond with proper care and management (Heinemann et al., 2018).

Looking forward, artificial intelligence (AI) is poised to further revolutionize diabetes care. AI algorithms are being developed to predict blood glucose trends, optimize insulin dosing, and personalize dietary and activity recommendations. These tools promise to enhance both self-management and clinical decision-making, making life with diabetes more manageable and sustainable than ever before (Contreras and Vehi, 2018).

1.3 Literature Review: Machine Learning for Binary Prediction of Diabetes

The application of machine learning (ML) in binary classification tasks such as diabetes prediction has gained significant momentum, offering powerful alternatives to traditional statistical methods. Various supervised ML models—including logistic regression, random forests, and artificial neural networks (ANNs)—have demonstrated excellent performance on clinical and demographic datasets.

Together, these studies suggest that ensemble methods, particularly random forests, consistently outperform other models in binary diabetes prediction tasks. The incorporation of interpretability tools and clinical relevance further supports the integration of these models into real-world healthcare decision-making.

Analyzing similar work, a 2022 study employed many different forms of supervised machine learning, including: logistic regression, KNN, random forest, decision tree, bagging, AdaBoost, XGBoost, Voting, and SVM. XGBoost and Bagging algorithms performed the best, with F1 scores of 0.81 and 0.79 respectively, and accuracies of 81% and 79% respectively, confirming that ensemble methods outperform other models (Tasin et al., 2022). In a similar study, Iparraguirre-Villanueva et al. (2023) evaluated KNN, decision tree, logistic regression, SVM, and BNB, a model not discussed in this study. KNN performed the best with the highest accuracy of 79.6%, with BNB just under performing at 77.2%. Other models, however, did not perform as well. Febrian et al. (2022) built on existing work, specifically comparing KNN and Naive Bayes models. They found that Naive Bayes seemed to consistently outperform KNN models.

In a more detailed analysis, Kaviyaadharshani et al. (2024) utilized many datasets, comparing many of the models listed in the previous study. Most of the datasets proved to be very successful, with accuracies as high as 99.41%.

1.4 Data Description

This dataset, titled Diabetes Prediction Dataset, was obtained from Kaggle. It contains approximately 100,000 entries, and eight variables: patient's gender, age, BMI, blood glucose level, presence of hypertension status, presence of heart disease, smoking history, and lycohemoglobin (HbA1c) level. The outcome variable is a binary diabetes diagnosis (positive or negative), making this dataset suitable for classification tasks.

Table 2 in the appendix contains the full list of predictor variables along with histograms or bar graphs illustrating their distributions, separated by diabetes status.

The dataset is highly imbalanced, with only 8.5% of cases being positive. While such imbalance is common in medical datasets, it did not significantly affect model performance in this study, as the overall fit remained strong across evaluation metrics.

2 Binary Classifiers: Theoretical Framework

Here, we provide a brief theoretical introduction to the statistical and machine learning techniques used to model the data.

2.1 Logistic Regression

Logistic regression extends the linear regression framework by applying the sigmoid (logistic) function, which maps predicted values to the [0,1] interval, allowing interpretation as probabilities (Korosteleva, 2018). The model is based on the log-odds (logit) transformation:

$$\ln \frac{\pi}{1-\pi} \;=\; \beta_0 \,+\, \beta_1 \, x_1 \,+\, \cdots \,+\, \beta_k \, x_k$$

where π denotes the probability that the response variable is equal to 1. Here $x_1, \ldots, x_k$ is a set of predictors, β_0 is the intercept, and the parameters $\beta_0, \ldots, \beta_k$ are the regression slopes.

Once the model is fitted, predicted probabilities are generated and then typically thresholded at 0.5 to classify observations into binary outcomes.

2.2 Probit Regression

Probit regression is conceptually similar to logistic regression but uses the cumulative distribution function (cdf) of the standard normal distribution instead of the logistic function (Korosteleva, 2018). The model takes the form:

$$\pi \;=\; \Phi\big(\beta_0 \,+\, \beta_1 \, x_1 \,+\, \cdots \,+\, \beta_k \, x_k\big)$$

where $\Phi(\cdot)$ denotes the standard normal cdf. Compared to logistic regression, probit regression is less sensitive to outliers.

2.3 Complementary Log-Log Regression

Complementary log-log (cloglog) regression is another alternative for modeling binary outcomes, particularly useful when the event of interest has a low probability of occurring (Korosteleva, 2018). Unlike the symmetric sigmoid curve of logistic regression, the cloglog link function produces an asymmetric curve, making it more appropriate for skewed outcome distributions. The model is defined as:

$$\ln(-\ln(1-\pi)) \;=\; \beta_0 \,+\, \beta_1 \, x_1 \,+\, \cdots \,+\, \beta_k \, x_k.$$

2.4 Random Forest

The random forest algorithm is an ensemble learning method that builds a large number of decision trees and combines their predictions to improve accuracy and reduce overfitting (James et al., 2021). It operates by generating multiple bootstrap samples from the original dataset, that is, each sample is drawn with replacement, and training a separate decision tree on each sample.

For classification tasks, about two-thirds of the data is used in each bootstrap sample to train a tree, while the remaining one-third (called the out-of-bag sample) can be used for internal validation. Each tree is constructed by recursively splitting the data, but at each split, only a random subset of features is considered. This process of combining bagging (bootstrap aggregation) with feature randomness introduces diversity among the trees, which strengthens the ensemble.

Once all trees are trained, the final classification prediction is made by majority voting across the trees.

2.5 Gradient Boosting

Unlike bagging, which builds multiple decision trees in parallel and aggregates their predictions to reduce variance, boosting is a sequential ensemble method that aims to reduce bias by building models iteratively (Cerulli, 2023). Each new model in the sequence attempts to correct the errors made by the previous ones.

Boosting begins with a weak initial model—often a constant predictor—and then fits subsequent models to the residuals (i.e., the difference between the observed and predicted values). At each step, a new weak learner is trained to approximate these residuals, and the overall model is updated by adding the new learner's contribution.

The iterative update rule can be written as:

$$\hat{f}_{new}(x) \;=\; \hat{f}_{old}(x) \;+\; \lambda\, \hat{f}^{b}(x)$$

where $\hat{f}^b$ is the bth weak learner, and λ is the learning rate, a small constant that controls how much each learner contributes.

After B iterations, the final boosted model is the sum of all weak learners:

$$\hat{f}(x) \;=\; \sum_{b=1}^{B} \lambda\, \hat{f}^{b}(x).$$

This approach allows for building strong predictive models by focusing sequentially on hard-to-predict examples, leading to higher accuracy and better generalization in many settings.

2.6 Support Vector Machine

Support Vector Machines are supervised learning models used for binary classification tasks. The main idea behind SVM is to find the optimal hyperplane that best separates data points from two different classes (James et al., 2021). This hyperplane is chosen to maximize the margin, which is the distance between the hyperplane and the nearest data points from each class. These nearest points are called support vectors, and they are the only points that directly influence the position and orientation of the hyperplane.

In the case of linearly separable data, the optimization problem is:

$$\begin{aligned}
&\text{Find } \mathbf{w}, b \text{ such that:}\\
&\min_{\mathbf{w},b} \ \frac{1}{2}\|\mathbf{w}\|^2\\
&\text{subject to: } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \quad \text{for all } i
\end{aligned}$$

where $\mathbf{w}$ is the weight vector perpendicular to the hyperplane, b is the bias term, $\mathbf{x}_i$ are the input vectors, and $y_i \in \{-1, 1\}$ are the class labels.

The decision function is:

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b).$$

However, many real-world datasets are not linearly separable. To handle this, SVMs can use the kernel trick, which maps the original input features into a higher-dimensional space where a linear separator may exist. Common kernels include:

- **Polynomial kernel:**

  $$K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}' + c)^d$$

  where c is a constant and d is the degree of the polynomial.

- **Radial basis function (RBF) kernel / Gaussian kernel:**

  $$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2\right)$$

  where γ controls the width of the Gaussian.

- **Sigmoid kernel (related to neural networks):**

  $$K(\mathbf{x}, \mathbf{x}') = \tanh(\alpha \mathbf{x}^\top \mathbf{x}' + c).$$

By using these kernels, SVMs can create non-linear decision boundaries in the original feature space, making them highly flexible and effective in many complex classification tasks.

2.7 k-Nearest Neighbor

K-Nearest Neighbors is a simple yet powerful non-parametric classification method. It works by classifying a new data point based on the majority class of its closest neighbors in the training set (James et al., 2021). The user specifies a value of k, which determines how many nearby data points are considered when making a prediction.

To classify a new point, the algorithm measures the distance between that point and all others in the training data—most commonly using Euclidean distance—and identifies the k nearest neighbors. The class that appears most frequently among these neighbors is then assigned to the new point.

KNN requires no training phase and is intuitive to understand and implement. However, it can be computationally intensive for large datasets and may perform poorly in high-dimensional spaces. It's also sensitive to the choice of k and to the scale of the input features, making preprocessing like normalization important.

2.8 Naive Bayes

As the name suggests, Naive Bayes classification is based on *Bayes' Theorem*, which provides a way to calculate the probability of a class Y given a set of input features X (Cerulli, 2023). The general form of Bayes' Theorem is:

$$\mathbb{P}(Y \mid X) = \frac{\mathbb{P}(X \mid Y)\mathbb{P}(Y)}{\mathbb{P}(X)}.$$

The algorithm computes the *posterior probability* of each class given the observed features and assigns the label with the highest probability.

The method is considered "naive" because it assumes that all predictors (features) are *conditionally independent* given the class label. That is, the presence or value of one feature does not influence any other, given the class. While this assumption is often violated in practice, the model still performs surprisingly well in many applications.

2.9 Artificial Neural Network

Artificial Neural Networks (ANNs) are among the most important and widely used machine learning methods powering today's advances in artificial intelligence. Inspired by the structure of the human brain, ANNs consist of interconnected layers of nodes, called neurons, which transform input data into meaningful outputs.

An ANN typically takes raw input features and passes them through one or more hidden layers, each containing multiple neurons (Cerulli, 2023). These hidden layers perform complex transformations, extracting hierarchical features and patterns from the data. The structure and number of hidden layers and nodes are usually determined through experimentation and model tuning, as they are not directly interpretable.

The final layer, called the output layer, produces the prediction or classification result based on the processed information. Through a process called training, the ANN adjusts the weights of connections between neurons to minimize prediction error, often using algorithms like backpropagation and gradient descent.

This ability to learn complex, non-linear relationships makes ANNs a highly effective classification tool.

2.10 Performance Measures

To evaluate the effectiveness of classification models, several performance measures are commonly used. Below are brief definitions of key metrics:

- **Accuracy**: The proportion of all correct predictions (both true positives and true negatives) among the total number of cases.
- **Sensitivity (Recall)**: The ability of the model to correctly identify positive cases (true positives) out of all actual positives.
- **Specificity**: The ability of the model to correctly identify negative cases (true negatives) out of all actual negatives.
- **Precision**: The proportion of true positive predictions out of all positive predictions made by the model.
- **F1-Score**: The harmonic mean of precision and sensitivity, providing a balance between the two metrics.
- **ROC Curve (Receiver Operating Characteristic Curve)**: A graphical plot illustrating the trade-off between sensitivity (true positive rate) and 1-specificity (false positive rate) across different classification thresholds.
- **AUC (Area Under the ROC Curve)**: A scalar summary of the ROC curve representing the model's ability to discriminate between classes across all thresholds; values closer to 1 indicate better performance.

3 Applications and Results

3.1 Determining Feature Importance

Random Forest and Gradient Boosting are machine learning algorithms that have built-in methods for calculating feature importance, which help identify the most influential predictors in the model. In addition to these tree-based models, supplementary regression analyses were also conducted to further evaluate the primary contributing factors to diabetes diagnosis.

Figures 1 and 2 in the appendix display the feature importance scores generated by the models. These scores indicate the relative impact each variable has on the prediction outcome, allowing us to better understand which factors play the most significant role in classifying diabetes status.

Across all figures, HbA1c level consistently emerges as the most significant predictor for diabetes diagnosis. HbA1c, or glycated hemoglobin, reflects the average blood glucose level over the past few months, making it a direct and reliable indicator of long-term glucose regulation. As expected, blood glucose level itself also shows strong predictive value.

Body Mass Index (BMI) and age are two additional features that demonstrate moderate yet consistent importance across different algorithms. Given that BMI is a measure of body fat—a known risk factor for metabolic disorders—its relevance in predicting diabetes is well supported. Similarly, the increased susceptibility of older individuals to type 2 diabetes is well documented, and the models reflect this trend.

In contrast, some other health-related features, such as smoking history and heart disease, surprisingly show relatively low importance in this dataset. These results suggest that while such factors may contribute to overall health, their direct impact on diabetes prediction in this population may be limited.

3.2 Comparing Performance Measures

The results of eleven machine learning algorithms designed to predict diabetes diagnosis based on selected predictors are presented below in tabular form (see Table 1).

Table 1. Performance Measures Across Eleven Machine Learning Models.

Algorithm	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1-score**	**AUC**
Random Forest	0.9703	0.6729	1	1	0.8045	0.8643
Gradient Boost	0.9681	0.6799	0.9945	0.9188	0.7815	0.9638
SVM (linear)	0.9609	0.5883	0.9953	0.9209	0.7179	0.9561
SVM (polynomial)	0.9638	0.5825	0.9991	0.9833	0.7316	0.9439
SVM (radial)	0.9632	0.5796	0.9987	0.9765	0.7274	0.9251
SVM (sigmoid)	0.9143	0.4866	0.9538	0.4932	0.4899	0.8497
KNN	0.9558	0.5165	0.9966	0.9340	0.6652	0.9153
Naive Bayes	0.9066	0.6386	0.9312	0.4601	0.5349	0.8984
ANN (1,3)	0.9212	0.7134	0.9746	0.8773	0.7806	0.9279
ANN (2,3)	0.9209	0.6978	0.9781	0.8941	0.7827	0.9559
ANN (tanh)	0.6933	0.2370	0.8102	0.2425	0.2397	0.8363

Analyzing the results, it is unsurprising that the Random Forest algorithm performs the best. This aligns with the findings of the Khan et al. (2024) study, which suggests that the algorithms are functioning properly. In fact, the Random Forest algorithm achieved a slightly higher accuracy than reported in that study. With an accuracy of 0.9703, it outperformed all other algorithms in accuracy, specificity, false positive rate (FPR), precision, and F1 score. Its sensitivity is second only to the Gradient Boost algorithm, with only a minimal difference. However, Gradient Boosting exhibits a much higher area under the ROC curve (AUC), demonstrating its strong predictive capability as well.

All other algorithms achieved accuracy rates above 90%, except for the ANN with the tanh activation function. These findings are supported by both the confusion matrix and ROC curve analyses. Overall, the results indicate that the developed algorithms are effective for diabetes diagnosis.

4 Conclusion

4.1 Summary of Work Completed

This study applied a variety of machine learning algorithms to predict diabetes diagnosis, with Random Forest and Gradient Boosting emerging as the top performers. These findings are consistent with numerous existing studies that highlight the strength of ensemble methods in classification tasks. Notably, nearly all algorithms achieved an accuracy exceeding 90%, demonstrating the robustness and reliability of the models developed.

4.2 Proposed Future Directions

Given the promising potential of machine learning to revolutionize disease diagnosis, as evidenced by this study and the broader literature, there are countless opportunities to enhance diabetes healthcare through advanced analytics. Future work might focus on applying similar machine learning approaches to predict the length of hospital stay for diabetic patients, incorporating relevant clinical and demographic predictors. Additionally, a more detailed classification of diabetes types and severity levels could be undertaken to better understand their impact on patient outcomes and tailor predictive models accordingly.

Supplemental Materials

The dataset, R code, and all relevant outputs (including the ROC curves) are available in the project's GitHub repository (https://github.com/sathvik-kommireddy/Diabetes-Machine-Learning.git) for reproducibility and further exploration.

Acknowledgments

I would like to sincerely thank Dr. Olga Korosteleva, a professor at California State University, Long Beach, for her invaluable guidance and support throughout this project. I am also grateful to my dedicated math teachers, Mr. Jared Derksen and Mr. Dave Oberhauser, for their encouragement and assistance. Finally, I wish to express my deepest appreciation to my family members for their wholehearted support and motivation.

References

American Diabetes Association. (2014). Diagnosis and classification of diabetes mellitus. *Diabetes Care*, 37(Supplement 1), S81-S90. https://doi.org/10.2337/dc14-S081.

Atkinson, M. A., Eisenbarth, G. S., and A. W. Michels. (2014). Type 1 diabetes. *The Lancet*, 383(9911), 69–82. https://doi.org/10.1016/S0140-6736(13)60591-7.

Bliss, M. (1982). *The Discovery of Insulin.* University of Chicago Press.

Bliss, M. (1993). *Banting: A Biography.* University of Toronto Press, 2nd edition.

Cerulli, G. (2023). *Fundamentals of Supervised Machine Learning With Applications in Python, R, and Stata.* Springer.

Contreras, I. and J. Vehi. (2018). Artificial intelligence for diabetes management and decision support: Literature review. *Journal of Medical Internet Research*, 20(5), e10775. https://doi.org/10.2196/10775.

DeFronzo, R. A., et al. (2015). Type 2 diabetes mellitus. *Nature Reviews Disease Primers*, 1(1), e15019. https://doi.org/10.1038/nrdp.2015.19.

Febrian, M.E., Ferdinan, F.X., Sendani, G.P., Suryanigrum, K.M., and R. Yunanda. (2023). Diabetes prediction using supervised machine learning, *Procedia Computer Science*, 216, 21-30, https://doi.org/10.1016/j.procs.2022.12.107.

Heinemann, L. et al. (2018). Real-time continuous glucose monitoring in adults with type 1 diabetes and impaired hypoglycaemia awareness or severe hypoglycaemia treated with multiple daily insulin injections (HypoDE): a multicentre, randomised controlled trial. *Lancet*,

Lancet, 391(10128):1367-1377. https://doi.org/10.1016/S0140-6736(18)30297-6.

Himsworth, H.P. (1936). Diabetes mellitus: its differentiation into insulin-sensitive and insulin-insensitive types. *Lancet*, 1:127-130.

Iparraguirre-Villanueva, O., Espinola-Linares, K., Flores Castañeda, R.O., and M. Cabanillas-Carbonell. (2023). Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes. *Diagnostics (Basel)*, 13(14):2383.
https://doi.org/10.3390/diagnostics13142383.

James, G., Witten, D., Hastie, T., and R. Tibshirani. (2021). *An Introduction to Statistical Learning with Applications in R*. Springer, 2nd edition.

Kaviyaadharshani, D. et al. (2024). Diagnosing Diabetes using Machine Learning-based Predictive Models. *Procedia Computer Science*, 233, 288-294.
https://doi.org/10.1016/j.procs.2024.03.218.

Korosteleva, O. (2018). *Advanced Regression Models with SAS and R*, Chapman and Hall/CRC.

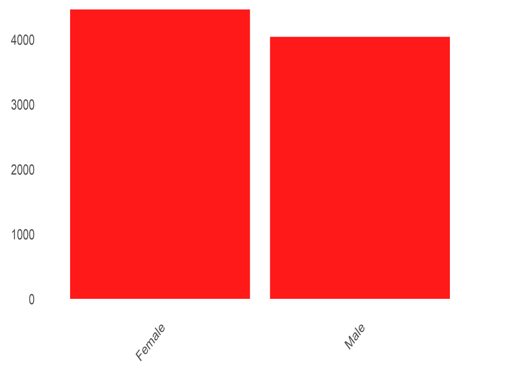
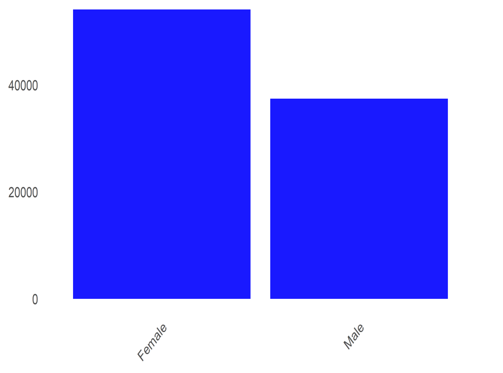
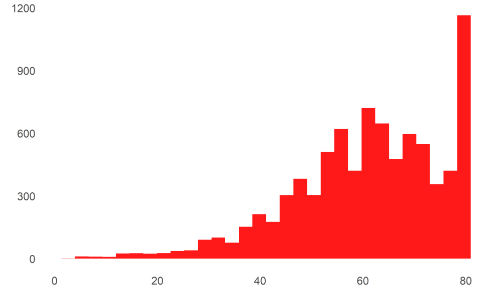
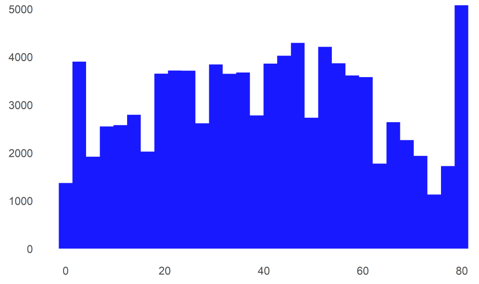
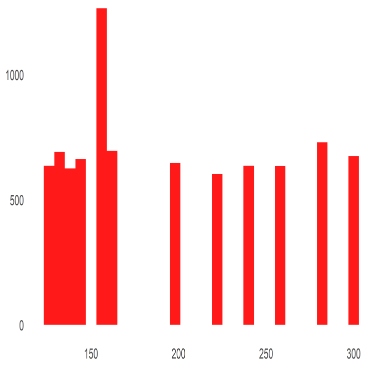
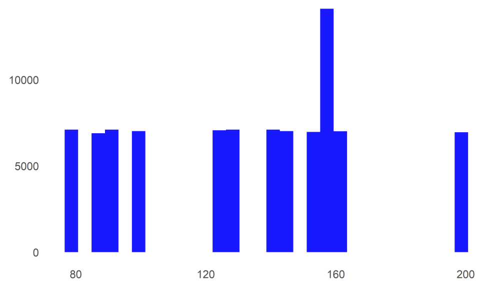
National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). (2022). What is Diabetes? https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes

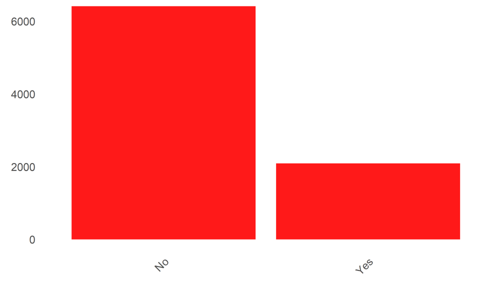
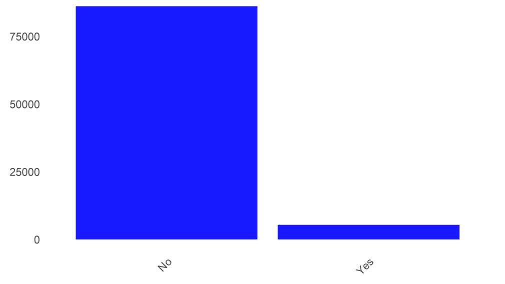
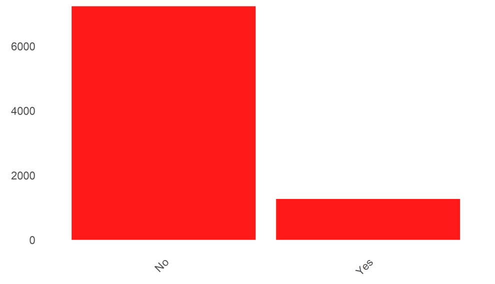
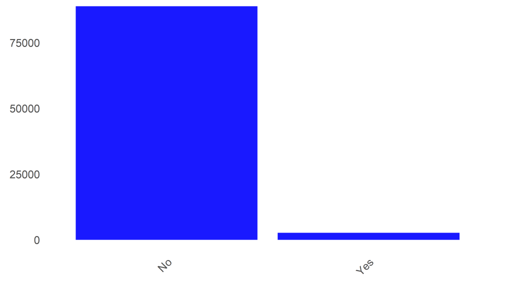
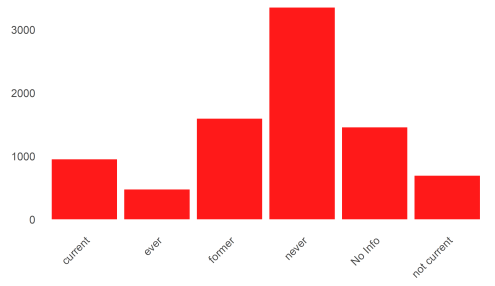
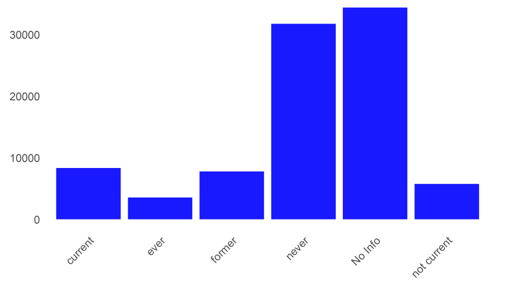
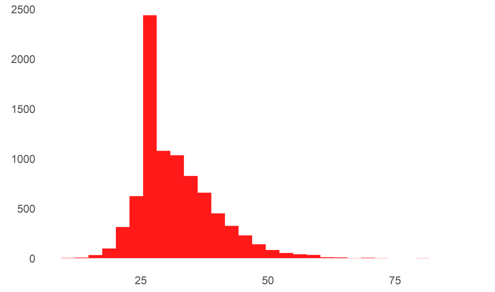
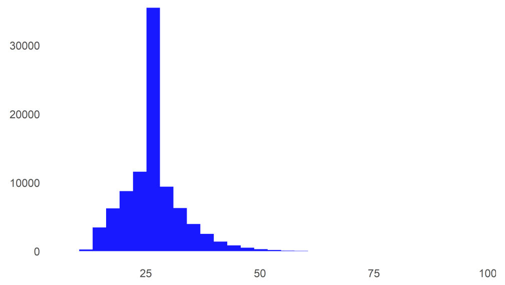
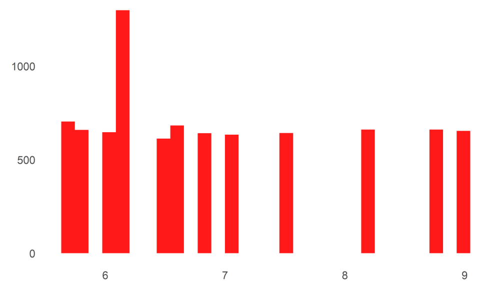
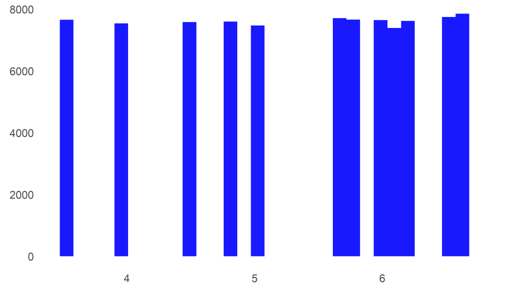
Tasin, I., Nabil, T.U., Islam, S., and R. Khan. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, 10 (1-2), 1-10.
https://doi.org/10.1049/htl2.12039.

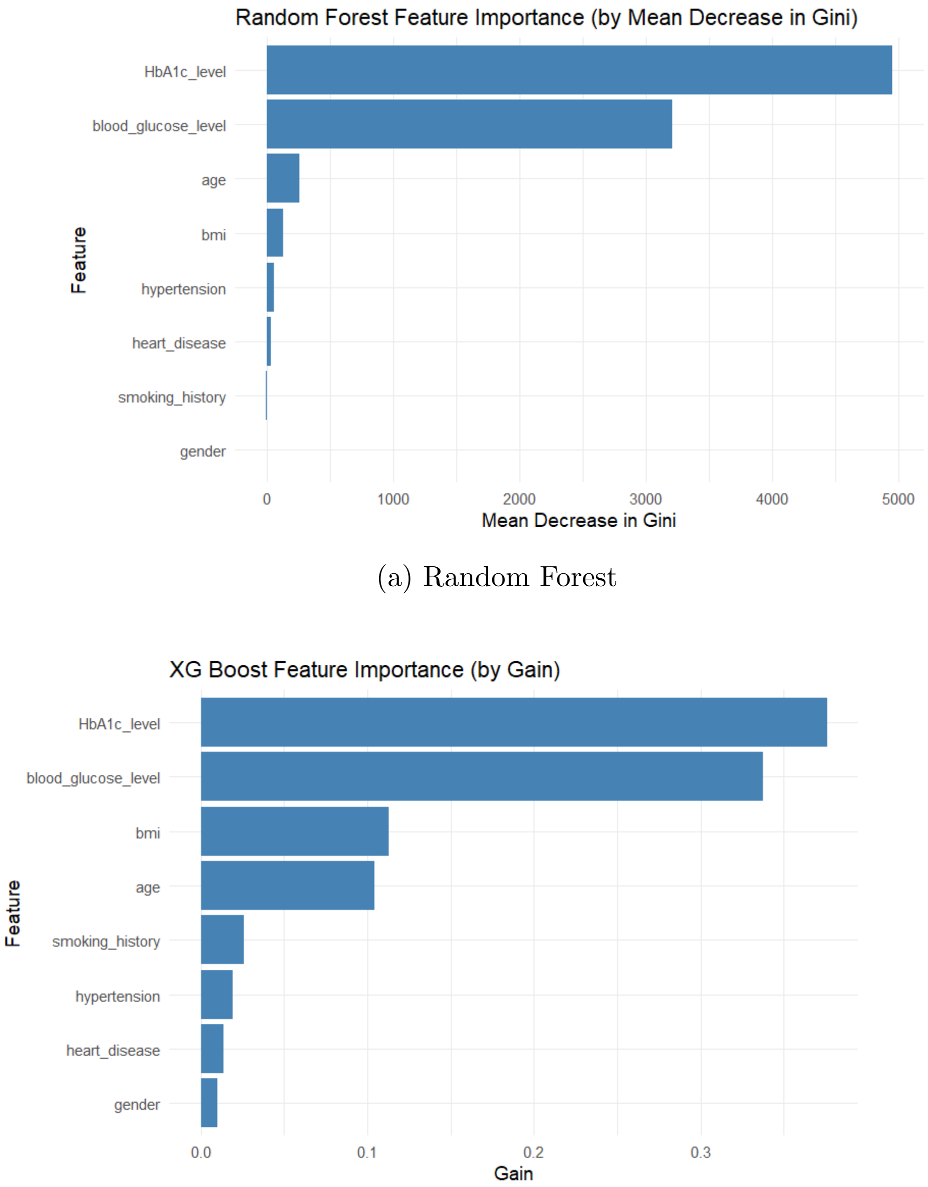
von Mering, J., and O. Minkowski. (1890). Diabetes mellitus nach Pankreasextirpation. *Archiv für experimentelle Pathologie und Pharmakologie*, 26, 371–387.
https://doi.org/10.1007/BF01831214.

Appendix

Table 2. Comparison of Patient Characteristics by Diabetes Status.

Variable Name	Plot for Diabetic Patients	Plot for Non-diabetic Patients
Gender		
Age		
Blood glucose level		

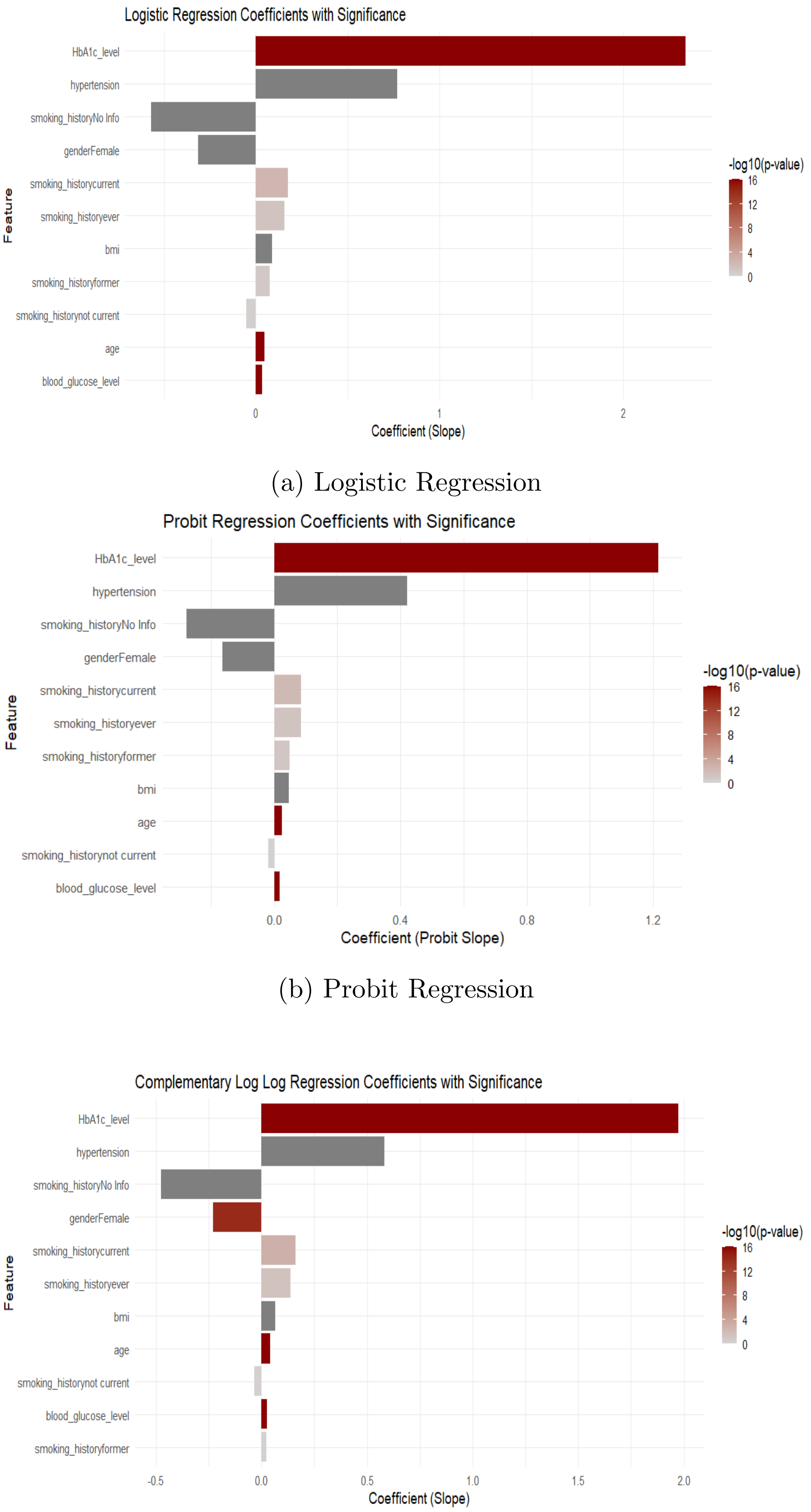
Variable Name	Plot for Diabetic Patients	Plot for Non-diabetic Patients
Hypertension		
Heart disease		
Smoking history		
BMI		
HbA1c level		



(a) Random Forest

(b) Gradient Boosting

Figure 1. Feature Importance Graphs for Random Forest and Gradient Boosting.



(a) Logistic Regression

(b) Probit Regression

(c) Cloglog Regression

Figure 2. Feature Importance Graphs for Binary Regressions.