Prediction of Type 2 Diabetes Mellitus Using Machine Learning Methods

A.P. Loria

1. INTRODUCTION

Diabetes mellitus (DM) is a disorder affecting blood sugar regulation that impacts over 500 million people globally, contributing to nearly seven million deaths in 2021 alone [1, 2]. Ranked as the eighth leading cause of death worldwide by the World Health Organization, DM is a serious health concern with complex pathophysiology and significant impacts on quality of life. The two most common forms are type 1 diabetes mellitus (T1DM), usually diagnosed in childhood, and type 2 diabetes mellitus (T2DM), which typically develops in adulthood and is associated with lifestyle factors. Though this report doesn't delve into the biological mechanisms of DM, it's important to recognize that untreated DM can lead to chronic complications like retinopathy, nephropathy, and neuropathy, often resulting in blindness, dialysis, or amputation [3]. As early detection is crucial to preventing these severe outcomes, efficient prediction methods for identifying individuals at risk of developing DM could provide significant clinical benefits.

This report describes the machine learning process and model evaluation for predicting the likelihood of T2DM onset in a dataset of adult women from Pima County, Arizona, aged 21 to 81. The dataset, originally described in the *Proceedings of the Annual Symposium on Computer Application in Medical Care* [4] and accessed via Kaggle [5], includes eight diabetes-related characteristics, including age and number of pregnancies.

Several machine learning models were tested, with the goal of identifying individuals at high risk of developing T2DM. Among the models, two tree-based algorithms, random forest and XGBoost, stood out due to their strong performance across key metrics, including AUC (mid-80% range), AP (above 70%), and recall (upper 80% range). These metrics are particularly important in medical prediction, as they reflect the models' ability to accurately identify at-risk individuals and minimize missed cases. Furthermore, both models demonstrated robustness against overfitting, making them well-suited for capturing the complex patterns in this dataset.

While this study's dataset consists exclusively of female participants and includes pregnancy count as a feature, the recommended model, XGBoost, could be adapted to include both sexes if applied to a more generalized dataset. The model's adaptability and superior performance on imbalanced data make it a promising tool for diabetes risk prediction.

2. METHODS

2.1 Exploratory Data Analysis (EDA)

The dataset contains 768 observations and 9 features for Pima county women. The features are listed in table 1.

Table 1

list of features

Feature	Description	Unit of measurement
pregnancies	number of pregnancies	integer values (no unit)
glucose	plasma glucose concentration at two hours in an Oral glucose tolerance test (OGTT)	milligrams per deciliter (mg/dL)
blood pressure	diastolic	millimeters of mercury (mmHg)
skin thickness	triceps skin fold thickness	millimeters (mm)
insulin	two-hour serum insulin	micro units per milliliter (µU/ml)
BMI (body mass index)	weight and height	kilograms per height in meters squared (kg/m^2)
diabetes pedigree function	a formula described in the publication that provides a numeric value for family history and genetics of DM per person	float values (no unit)
age	age of the person (observation)	integer values (no unit)
outcome	the target variable for absence or presence of DM	0 - no DM 1 - DM

The dataset has no NaN ("not a number") values but does contain values of 0 for many observations of several features, except diabetes pedigree function and age. A value of 0 is reasonable for some features, including (number of) pregnancies and outcome (where 0 means "no diabetes"). However, a value of 0 for other features does not make sense biologically. For example, one cannot have a blood pressure of 0. Therefore, values of 0 for features listed in table 2 are assumed to be missing data and will be treated as such. The percent of observations with values of 0 are also provided for each feature. Since the dataset is small in terms of number of observations, and that observations with 0 values for one or more feature still contain valuable information for other features, the 0 values for these features were imputed.

Table 2

Percent of 0 values per feature

Feature	Percent of dataset
glucose	0.65
blood pressure	4.56
skin thickness	29.56
insulin	48.70
BMI	1.43

Data Distribution

Some features have an approximately normal distribution, including: glucose, blood pressure, and BMI (excluding outliers). The other features are right skewed (pregnancies, skin thickness, insulin, diabetes pedigree function, and age). The target variable is "Outcome" and is not a feature, per se. The histogram below shows that the distribution of the target variable is moderately imbalanced with 500 observations having a target value of 0 (no diabetes) and 268 observations are listed with a target value of 1 (diabetes).

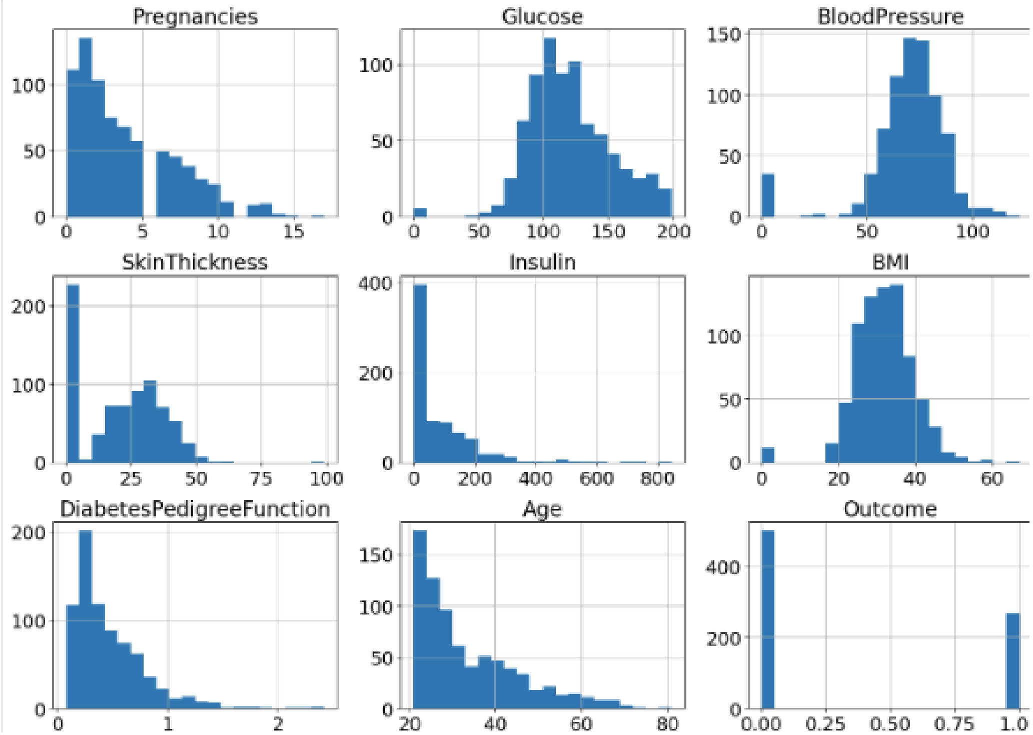


Figure 1

Distribution of Feature Data

All features have outlies based on a standard boxplot. Here, an outlier is defined as a data point that exist below Q1 - 1.5 and above Q3 + 1.5 for a specific feature.

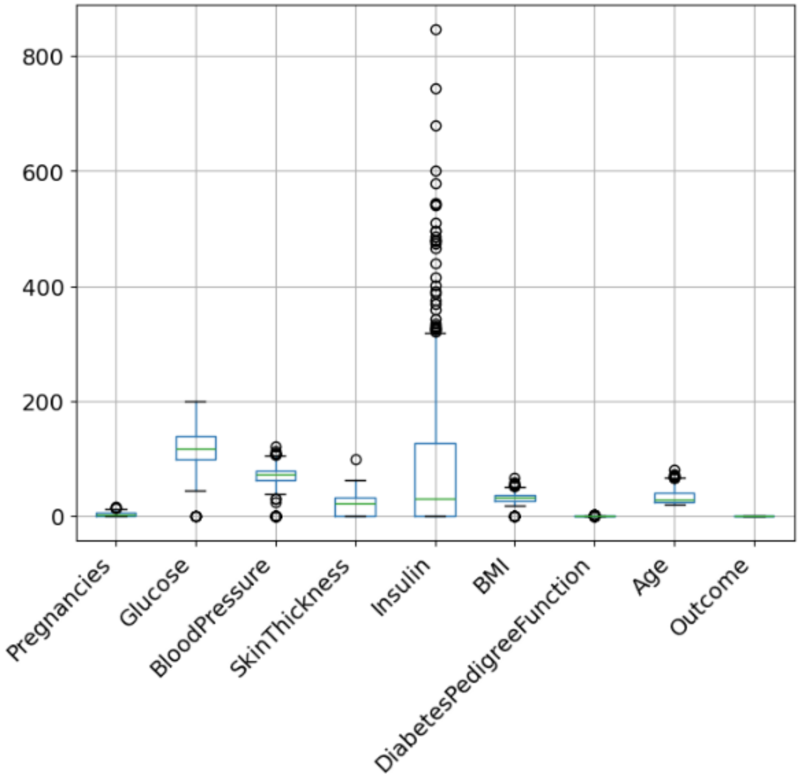


Figure 2
Boxplot of Features

A search of the literature helps to define expected or average values for healthy individuals to corroborate the mathematical definition of an outlier. Of course, there will be variance in the expected values based on age and other factors. But these values provide a biological framework for considering the range of values observed for each feature. Average values for features with substantial outliers are provided in table 3 [6-9].

Table 3
Average values for adult women

Feature	Average value
blood pressure (diastolic)	<80 mmHg [6]
insulin (two-hour serum level)	16-166 µU/ml [7]
BMI	18.5-24.9 kg/m^2 [8]
triceps skin fold thickness	23.6 +/- 7.5 mm [9]

Correlation

Based on a correlation matrix plot, none of the feature combinations show a substantial degree of correlation.

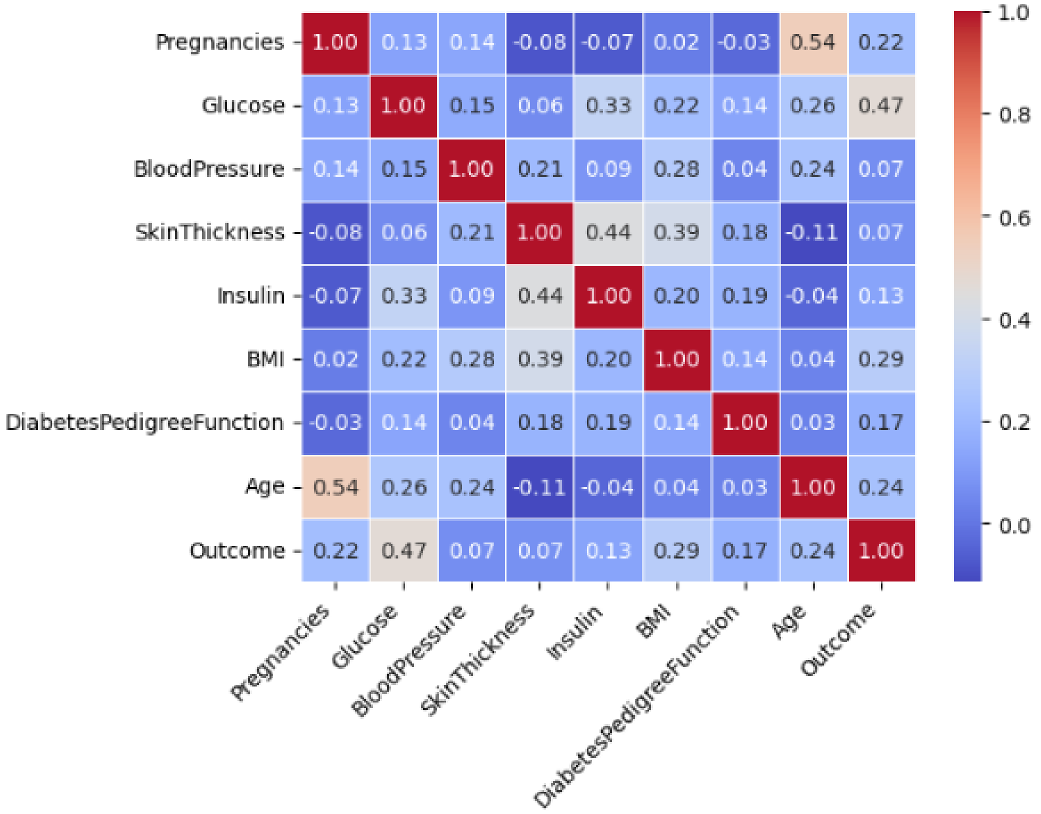


Figure 3

Correlation Matrix Heatmap

A pair plot reveals some target class separation in feature combinations, including glucose vs age, glucose vs diabetes pedigree function, glucose vs blood pressure, glucose vs skin thickness, glucose vs BMI, glucose vs insulin, and to a lesser extent, BMI vs age. Glucose seems to be an important feature for target class separation (see the pairplot in the mL notebook).

Mean values grouped by outcome variable

The mean value for every feature is greater for the those diagnosed with T2DM (outcome variable = 1) compared to those who are diagnosed to not have T2DM (outcome variable = 0). Number of pregnancies and insulin level show the greatest mean difference in the population that has been diagnosed with T2DM.

Table 4

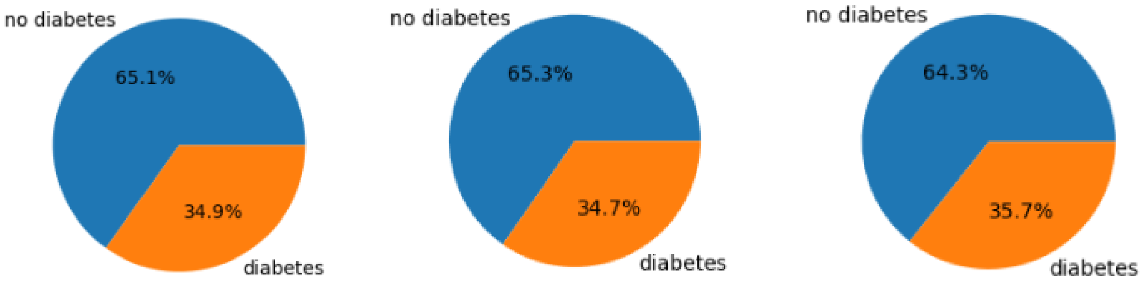
Mean values grouped by the outcome variable

Outcome/ Features	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
0	3.30	109.98	68.18	19.66	68.79	30.30	0.43	31.19
1	4.87	141.26	70.82	22.16	100.34	35.14	0.55	37.07
% diff	47.53	28.44	3.87	12.71	45.85	15.97	28.10	18.84

Training and Testing Sets

There is reasonable class separation across features and substantial differences in mean values per class, suggesting that a model with good predictive power should be achievable, even with this small dataset.

Before any feature engineering, the dataset was split into training (80%) and testing (20%) sets. Splitting the data before engineering features helps prevent data leakage, which could artificially inflate prediction metrics. Figure 4 compares the data imbalances in the full dataset with those of the training and testing sets. Stratifying the target variable did not substantially alter the degree of data imbalance of the training, testing, and full datasets.



Full Dataset

Training Set

Testing Set

Figure 4

Data Imbalance

Data Imbalance

The outcome variable in both the training and test sets is moderately imbalanced, with a ratio of approximately 2:1 for 0s to 1s. This kind of imbalance is common, particularly in medical datasets where the number of disease cases is often much smaller than the number of non-cases. Since this dataset is only moderately imbalanced, I addressed the imbalance during model fitting using specific techniques tailored to handle imbalanced data.

2.2 Feature Engineering

To maximize feature performance, multiple feature engineering techniques were applied to create new features that could reveal more about potential diabetes risk factors. Features with values of 0 (aside from pregnancies) were imputed with the median to retain valuable observations without further reducing the dataset size. A log transformation was then applied to reduces skewness and approximate a normal distribution, which can help improve performance of linear models. Each feature was then scaled to address differences in units and value ranges, standardizing them for more effective learning across models.

Several additional features were engineered to capture complex interactions and non-linear effects relevant to diabetes prediction, as well as to provide more granular categories. The combination of engineered features increased the feature count from 8 to 24, adding discriminative power but also a risk of overfitting, necessitating the use of models with robust regularization capabilities.

Interaction features

To capture the multiplicative effects of risk factors, interaction terms were created between glucose, BMI, insulin, and age:

- Glucose x Age: Both high glucose levels and age are independent risk factors for the onset of T2DM. As people age, glucose metabolism often declines, making this interaction a potential indicator of diabetes risk.

- Glucose x BMI: This interaction explores the compounding effect of glucose levels with obesity, as individuals with high BMI and high glucose are at an increased risk.

- Insulin x BMI: Insulin resistance is often associated with high BMI and is a key factor in T2DM. This feature captures the combined impact of obesity and insulin production on diabetes risk.

- Glucose-to-Insulin Ratio: This ratio is critical for assessing insulin sensitivity. A high glucose-to-insulin ratio may indicate insulin resistance or impaired insulin production, differentiating metabolic states with predictive relevance.

Polynomial Features

- BMI^2, Glucose^2, Age^2: Non-linear effects were introduced by squaring features including BMI, glucose, and age. For example, BMI and diabetes may not have a linear relationship, with elevated BMI levels increasingly impacting diabetes risk. Squared terms help capture these accelerating risk patterns.

Binned Features

To reflect clinical standards and add more granularity, certain continuous features were categorized:

- Glucose (mg/dl): Aligned with clinical diagnostic criteria as defined by the American Diabetes Association [10].
  - Normal: < 140
  - prediabetic: ≥ 140 < 200
  - diabetic: ≥ 200

- BMI (kg/m^2): Stratified according to the U.S. Centers for Disease Control BMI categories [11].
    - underweight: < 18.5
    - healthy weight: ≥ 18.5 < 25
    - overweight: ≥ 25 < 30
    - obesity (class 1): ≥ 30 < 35
    - obesity (class 2): ≥ 35 < 40
    - obesity (class 3): ≥ 40

- Insulin (µU/ml): Categories based on typical clinical ranges [12].
    - low: < 16
    - normal: ≥ 16 < 166
    - high: > 166

Considerations and Limitations

The expanded feature set provides additional predictive power by capturing non-linear effects, interactions, and clinically significant categories. However, this increase in features also risks overfitting due to complexity. To mitigate this, models with built-in regularization, such as lasso and ridge or ensemble methods, such as random forests and XG Boost, were employed.

2.3 Models

Five classification models were used for analyzing the Pima Indian Diabetes dataset: logistic regression, K-Nearest Neighbors (KNN), random forest, XGBoost, and Multi-Layer Perceptron (MLP) classifier. These models were selected for their strengths in handling data with multiple features, especially since feature engineering increased the feature count from 8 to 24.

Logistic Regression

Logistic regression is an efficient, linear model suitable for binary classification tasks, such as diabetes prediction. It adjusts coefficients to reflect the influence of each feature on the target outcome. To prevent overfitting, Lasso (L1) and Ridge (L2) regularization are applied, which penalize large coefficients and can eliminate or minimize unimportant features.

K-Nearest Neighbors (KNN)

KNN is a non-parametric, instance-based model that captures complex relationships through distance calculations. It is especially useful in capturing local patterns in data where classes are clustered. However, as a distance-based model, KNN can be computationally intensive with many features, so PCA (Principal Component Analysis) was used to reduce dimensionality and improve efficiency.

Random Forest

Random forest is an ensemble model that aggregates multiple decision trees to improve robustness and reduce overfitting. It uses bootstrap sampling (with replacement) and random feature selection for each tree, which enhances its generalization performance in high-dimensional datasets. The model also provides feature importance scores, offering insights into which features most influence predictions. This ensemble approach makes random forest highly resilient to overfitting, particularly in cases with many features.

XGBoost

XGBoost is a boosting model based on decision trees, optimized for speed and performance on complex data. Unlike random forest, which builds trees independently, XGBoost builds trees sequentially, with each tree learning from the residual errors of previous ones. This method captures subtle patterns in data. XGBoost includes three forms of regularization - L1 (alpha), L2 (lambda), and gamma - to control model complexity. Gamma regularization, in particular, limits overly complex splits, helping to prevent overfitting.

Multi-Layer Perceptron (MLP) Classifier

The MLP classifier, a type of artificial neural network (ANN), is adept at modeling non-linear relationships by learning complex patterns across multiple layers. This model is beneficial for data with a high number of features and interactions. MLPs require scaling of features to improve convergence and benefit from careful tuning of hyperparameters to avoid overfitting.

3. METRICS and RESULTS

The Pima Indian Diabetes dataset is designed for binary classification, with a target variable of 0 or 1, indicating whether an individual is predicted to develop diabetes within five years. To evaluate model performance effectively, focus is placed on metrics that capture the model's ability to discriminate between classes and handle the imbalanced nature of the dataset.

Importance of Sensitivity and True Positives in Disease Prediction

In disease prediction models, particularly with imbalanced datasets, true positives and recall (sensitivity) are prioritized. Correctly identifying individuals at risk of developing a disease is crucial, as it can prompt early intervention. While false positives may lead to further testing, they are less detrimental than false negatives, which could miss an opportunity for timely medical intervention. By using model-agnostic metrics like Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves, the performance of different models can be compared objectively to select one with high recall and reasonable balance between precision and recall.

ROC for Classification Performance

Initially, models are evaluated using receiver operating characteristic (ROC) curves and Area Under the Curve (AUC) scores. ROC curves plot the true positive rate (also known as sensitivity or recall) against the false positive rate across various decision thresholds. The true positive rate (TPR) is the same as recall and tells us how many of the actual positive instances the model correctly identified as positive. The false positive rate (FPR) is focused on the negative class (no disease) and is the ratio of false positive predictions to the actual negative instances in the dataset.

Table 5

Definitions of true positive rate and false positive rate

$TPR = \frac{true\ positives}{true\ positives + false\ negatives}$	*A high TPR is best and indicates the fraction of people likely to develop the disease that have been identified.*
$FPR = \frac{false\ positives}{false\ positives +\ true\ negatives}$	*A low FPR is best since it indicates the fraction of people incorrectly predicted to develop the disease*

The Area Under the Curve (AUC) metric informs about how the model ranks a randomly chosen observation (patient). As an example from the data in figure 5, the logistic regression model with an AUC score of 0.83 indicates that there is an 83% chance that the model will rank a randomly chosen patient who is likely to develop the disease higher than a randomly chosen healthy patient. A high ROC-AUC score indicates a strong ability to distinguish between individuals likely to develop diabetes and those who are not.

The ROC curves and AUC scores for all models are shown in Figure 5. Overall, all models demonstrate similar ROC-AUC performance, with the random forest and XGBoost models achieving the highest scores, indicating the strongest predictive performance among the models evaluated.

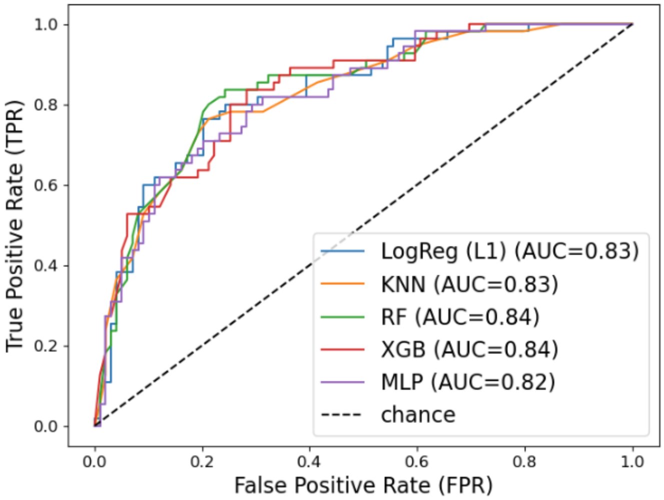


Figure 5

ROC Curves and AUC Scores for All Models

ROC-AUC plots are informative for evaluating a model's ability to discriminate between two classes in binary classification problems. However, in datasets with class imbalance - such as those involving rare events like fraud or disease - the dominant negative class reduces the impact of False Positive Rate (FPR) on overall model evaluation, making it less indicative of the model's ability to predict the minority positive class. This can lead ROC-AUC to overestimate model performance in identifying minority class instances. The diabetes dataset, with a 2:1 class ratio, does not exhibit severe imbalance, so these metrics remain reasonably reliable. To gain a more nuanced understanding of the model's performance on the minority class, Precision-Recall (PR) curves were also plotted. PR curves focus specifically on the trade-off between precision and recall, making them more informative for imbalanced datasets.

Precision-Recall curves for Imbalanced Datasets

Precision-Recall (PR) curves visualize the trade-off between precision and recall across all possible classification thresholds. Precision measures the proportion of correctly identified positive predictions out of all positive predictions made by the model, while recall (also called sensitivity or True Positive Rate) measures the proportion of actual positive cases correctly identified (Table 6). Together, these metrics focus exclusively on the performance of the positive class (e.g., individuals likely to develop diabetes).

Table 6

Definitions of precision and recall

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \qquad recall = \frac{true\ positives}{true\ positives + \ false\ negatives}$$

A commonly used metric to evaluate PR curves is the Average Precision (AP) score, which represents the area under the PR curve. The AP score summarizes the model's performance across all classification thresholds, weighting precision according to increases in recall. PR curves and AP scores are particularly valuable for evaluating imbalanced datasets because they focus solely on the model's ability to predict the positive class, without being skewed by the abundance of negative cases. Unlike metrics like accuracy or even the ROC curve, which incorporate true negatives, PR curves provide a clearer picture of the model's effectiveness at identifying true positives, especially in datasets with significant class imbalance. By using PR curves and AP scores, we can directly assess the model's ability to detect individuals at risk of diabetes, even in scenarios where positive cases are less common.

The PR curves along with AP scores are shown in figure 6. Similar to the ROC-AUC metrics, all models demonstrate similar PR performance, with the random forest and XGBoost models achieving the highest AP scores. Here XGBoost provides the strongest predictive performance among the models evaluated.

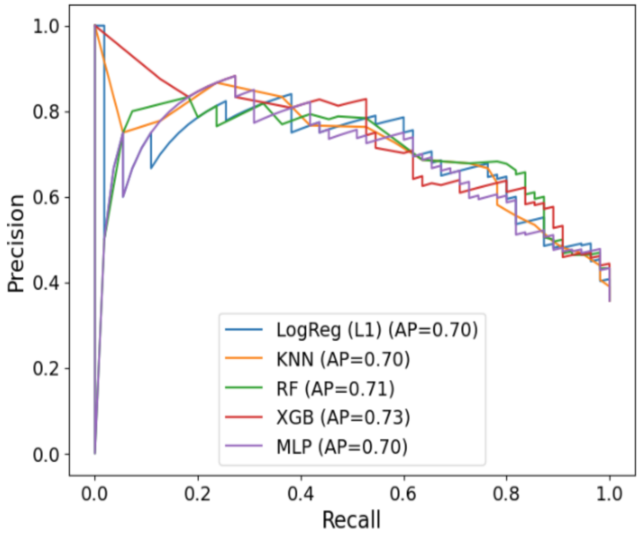


Figure 6
Precision-Recall and AP Scores for All Models

Confusion Matrix and Classification Report Metrics

In addition to ROC-AUC and PR metrics, confusion matrices offer more granular insights into model performance at specific thresholds. The confusion matrix shows counts for true positives, true negatives, false positives, and false negatives. From these counts, additional metrics were calculated, including accuracy, precision, recall, and F1 score, which are summarized in classification reports.

The F1 score, a harmonic mean of precision and recall, provides a single measure of a model's accuracy on the positive class, balancing false positives and false negatives — a useful metric for imbalanced datasets where false negatives may carry a higher cost.

These metrics are provided and compared for all models in Figure 7. The KNN model achieved the highest accuracy, but this metric is not the best for evaluating the performance of models on imbalanced datasets. Accuracy is the ratio of the sum of all positive predictions and negative predictions to the total number of predictions (true positives + true negatives + false positives + false negatives). If 95% of instances are negative and 5% are positive, a model that always predicts negative would have an accuracy of 95% without identifying any positive instances. Therefore, recall and F1 are preferred metrics for imbalanced datasets. KNN, random forest, and XGBoost models tied for highest F1 score, meaning that they performed equally well at identifying the positive class. However, random forest and XGBoost excelled at minimizing the false negatives as indicated by the recall scores.

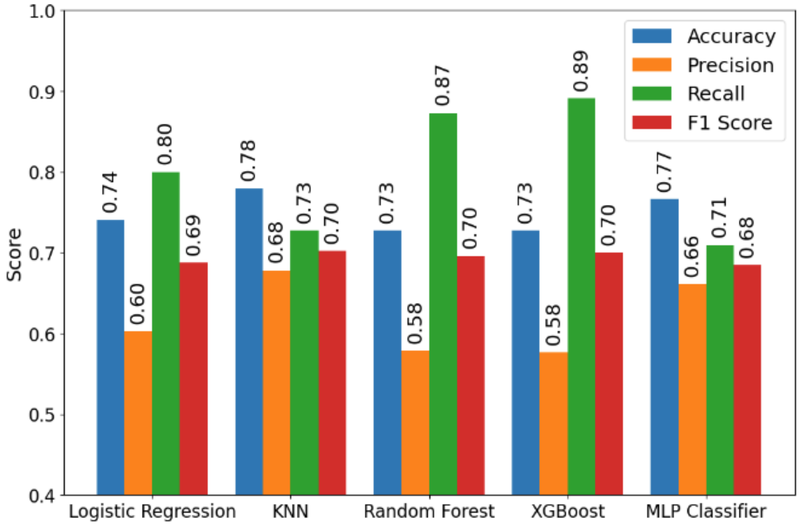


Figure 7
Plot of Classification Report Metrics for All Models

Additionally, the random forest and XGBoost models demonstrated the least overfitting on the training data, indicating their robustness for further feature engineering and potential application to other datasets (Figure 8). Finally, XGBoost is the recommended model based on the balance of all metrics. This model provides the highest scores for AUC, AP, and recall, and the lowest degree of overfitting the data.

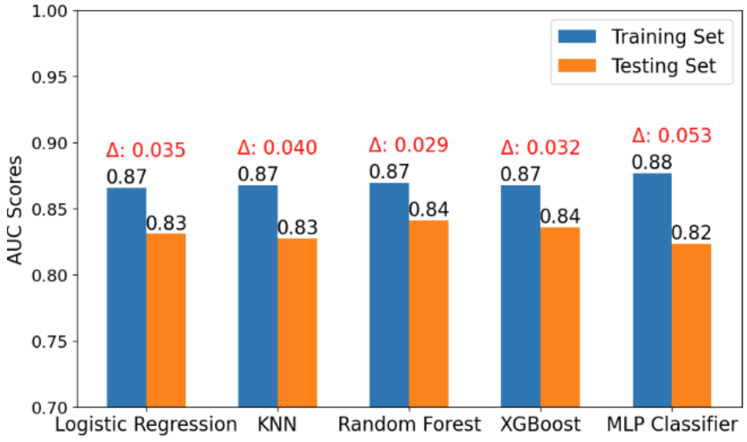


Figure 8
Training and Testing AUC scores for Model Comparison

4. DECISION THRESHOLD

The ROC and PR curves summarize model performance across all possible decision thresholds. A threshold is the value that determines whether an observation is classified into class 0 (no diabetes) or class 1 (likely to develop diabetes). Classification models generate a predicted probability for each observation, and if the probability exceeds the threshold, the observation is assigned to class 1; otherwise, it is assigned to class 0.

Threshold values range from 0 to 1, where 0 is the least strict and 1 is most strict. The default threshold for most models is 0.5. Adjusting the threshold allows for a trade-off between precision and recall:

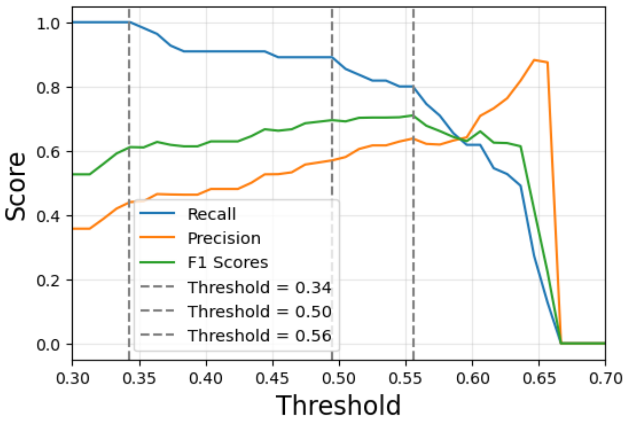
- A higher threshold classifies only the observations with the highest probabilities into class 1, reducing false positives but increasing false negatives (higher precision, lower recall).
- A lower threshold assigns more observations to class 1, which reduces false negatives but increases false positives (higher recall, lower precision).

In imbalanced datasets focused on critical outcomes, such as disease prediction, recall is often prioritized over precision to minimize false negatives. As a result, a lower threshold may be preferred to ensure that most positive cases are identified, even at the cost of some false positives.

The decision of which threshold to use is guided by analyzing the trade-off between precision, recall, and F1-score across different thresholds. Figure 9 visualizes these metrics across a range of thresholds. Three thresholds have been highlighted based on their significance:

1. Threshold 0.34: Recall is at its highest before beginning to decrease. This threshold prioritizes identifying as many positive cases as possible, even at the expense of some precision.
2. Threshold 0.50: The model's default threshold provides a balanced set of metrics for general predictions and serves as a baseline.
3. Threshold 0.56: This threshold corresponds to the maximum F1-score, representing the optimal balance between precision and recall.

By considering these thresholds and the associated trade-offs, stakeholders can select the value that best aligns with the model's purpose, such as prioritizing recall for disease prediction or balancing precision and recall for general classification tasks.



Metric/ Threshold	0.34	0.50	0.56
Precision	0.44	0.57	0.64
Recall	1.00	0.89	0.80
F1	0.61	0.70	0.71

Figure 9

Precision, Recall, and F1 Scores vs. Threshold Values

5. CONCLUSIONS

The K-Nearest Neighbors (KNN) model achieved the highest accuracy and tied for the top F1 score. However, given the dataset's imbalanced target variable, accuracy is not the most reliable metric. In predictive modeling for disease diagnosis, recall is the priority metric, as it ensures individuals at risk for T2DM are accurately identified, minimizing false negatives.

Both the random forest and XGBoost models matched the KNN model's F1 score but outperformed it in recall, AUC (Area Under the Curve), and AP (Average Precision) metrics, albeit with a slight decrease in precision. XGBoost, in particular, excels in handling imbalanced data and capturing intricate feature interactions. Its wide range of hyperparameters for optimization enables further improvements in performance, making it the preferred choice for this dataset.

The threshold selection also plays a crucial role in model performance. At the recommended threshold of 0.34, XGBoost achieves high recall while maintaining reasonable precision, aligning with the objective of reducing false negatives in disease prediction. This threshold ensures the model performs effectively in identifying at-risk individuals without significant overprediction.

Random forest remains a strong alternative, offering advantages in computational efficiency and robustness to feature correlations. If the dataset were expanded or simplified, random forest could become a viable option due to its reduced processing demands.

In summary, XGBoost is the recommended model for this dataset, as it provides the best balance of recall, precision, and flexibility. Its ability to handle imbalanced data and reduce false negatives makes it well-suited to predictive modeling for diabetes risk.

REFERENCES

1. International Diabetes Federation. (2024). *Diabetes facts & figures*. https://idf.org/about-diabetes/diabetes-facts-figures/

2. World Health Organization. (2024). *The top 10 causes of death*. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

3. Unger, R. H., & Orci, L. (2010). Paracrinology of islets and the paracrinopathy of diabetes. *Proceedings of the National Academy of Sciences, 107*(37), 16009–16012. https://doi.org/10.1073/pnas.1006639107

4. Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261–265). IEEE Computer Society Press.

5. Kaggle. (2024). *Diabetes dataset – Pima Indians*. https://www.kaggle.com/datasets/nancyalaswad90/review/data

6. National Institute on Aging. (2024). *High blood pressure and older adults*. https://www.nia.nih.gov/health/high-blood-pressure/high-blood-pressure-and-older-adults

7. Medscape. (2024). *2-hour serum insulin levels*. https://emedicine.medscape.com/article/2089224-overview

8. World Health Organization. (2024). *BMI classification*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8984168

9. National Institutes of Health. (2024). *Triceps skinfold thickness*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9127233

10. American Diabetes Association. (2024). *Glucose levels*. https://diabetes.org/about-diabetes/diagnosis

11. Centers for Disease Control and Prevention. (2024). *BMI categories*. https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html

12. Medscape. (2024). *Insulin levels*. https://emedicine.medscape.com/article/2089224-overview