Exploring Machine Learning Methods for Diabetes Prediction

Samarth Kumar
STAT 6000
szk0187@auburn.edu

***Abstract*—Predicting cases of diabetes is a vital challenge in the field of healthcare and enables the creation of early intervention strategies. This project utilizes the following supervised machine learning algorithms: Naive Bayes, Logistic Regression and Random Forest. Through a dataset containing medical and demographic information for 100,000 patients, this approach evaluates the influence of various parameters, such as age, gender, BMI, HbA1c levels, blood glucose levels, smoking history, and heart disease, on the likelihood of a patient having diabetes. The model achieved accuracy scores of 92%, 96% and 97%, respectively, proving that machine learning can be used to effectively predict diabetes across several combinations of the parameters. Additionally, HbA1c and Blood Glucose Levels were the most significant predictors of diabetes. The models from this project provide early-stage tools for predicting cases of diabetes.**

***Index Terms*—Healthcare, Diabetes, Machine Learning, Classification, Gaussian, Naive Bayes, Logistic Regression, Decision Tree, Random Forest.**

I. Introduction

A. Background

Diabetes is a chronic disease that impacts millions worldwide. Due to its various causes and risk factors, early prediction and intervention remain critical to preventing severe effects on health despite advances in the field of medicine. Doctors currently predict cases of diabetes by examining a patient's medical history, family history, and administering blood tests. However, traditional methods can often be challenging, therefore, supervised machine learning quickly became a popular and transformative method for disease prediction and management [1]. In this project, the following supervised machine learning algorithms were selected: Gaussian Naive Bayes, Logistic Regression, and Random Forest Classifier. These algorithms are known for their efficiency in handling large datasets as well as their computational speeds.

B. Purpose of Analysis

Through this project, I hope to evaluate and compare the effectiveness of multiple machine learning algorithms in predicting diabetes. Various demographic and lifestyle factors along with related medical conditions influence the occurrence of diabetes. By exploring relationships between these features, I aim to note the most relevant predictors of diabetes and the most efficient algorithm for classifying individuals at the highest risk of the disease. The analysis leverages various performance metrics, such as precision-recall curves and ROC curves to ensure a thorough evaluation. The goal is to provide a transformative tool for prevention strategies and clinical decision-making.

II. Methodology

A. Data Description

1) Dataset: The dataset consists of various parameters based on the medical and demographic information from 100,000 patients [2]. Below is a list of each parameter:

- Gender
- Age
- Hypertension
- Heart Disease
- Smoking History
- Body Mass Index (BMI)
- Hemoglobin A1c (HbA1c) Level
- Blood Glucose Level

2) Data Preprocessing: First, I began with the discrete features, which are gender, smoking history, hypertension, and heart disease. For consistency, each feature was set to use numerical categories. For gender, the values "Male", "Female", and "Other" were replaced with integers 0, 1, and 2, respectively. Similarly, the six categories for smoking history

were replaced with integers ranging from 0 through 6, where 0 indicates that the patient has never smoked and 4 indicates that the patient currently smokes. However, there were cases where a patient's smoking history was denoted as "No Info", so this category was denoted by the integer value of 5. For Hypertension and Heart Disease, I kept the original categories of 0 and 1, as these features indicate other health conditions to specify their presence or absence.

Next, I found that the distributions for the continuous features, BMI, HbA1c, and Blood Glucose Level, were skewed. Therefore, I applied a logarithmic transformation to these parameters so that they would follow more of a uniform distribution in order to improve model performance.

The figure below illustrates the class distribution from the dataset. Since the negative class, or patients without diabetes, accounts for 90% of the samples, it is clear that the dataset is imbalanced. To account for this, I initially implemented a random oversampling technique to the dataset. However, randomly oversampling the minority class ultimately harmed overall model performance.

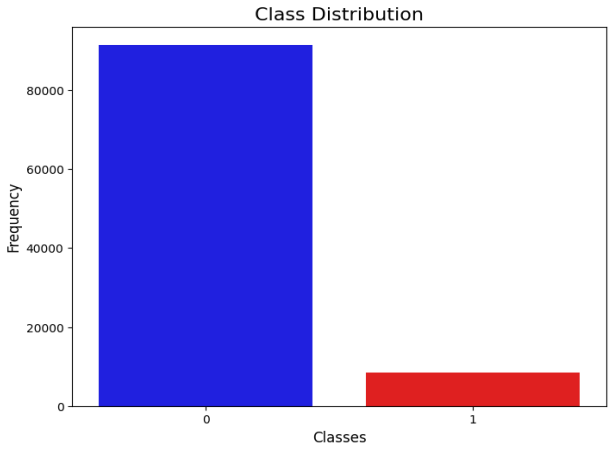

Fig. 1. Class Distribution

B. Model Selection

The dataset consists of predefined labels, therefore supervised machine learning models were used. For this project, I selected Naive Bayes, Logistic Regression, and Random Forest Classifier.

1) Naive Bayes: When the Gaussian Naive Bayes algorithm is used, it is assumed that each feature is conditionally independent, which simplifies the calculation of the posterior probabilities. The conditional probability formula for events X, and Y, is given below:

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)} \tag{1}$$

In the equation, $P(Y|X)$ represents the likelihood of event Y given the class X, P(X) represents the prior probability of class X while P(Y) is the evidence. For continuous features, the probability is modeled by a Gaussian Distribution, with the following equation:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2}$$

2) Logistic Regression: Logistic Regression is another widely used algorithm for binary classification. Rather than predicting continuous values, Logistic Regression predicts the probability of belonging to a given class. The algorithm relies on the Sigmoid Function to force the values to be between 0 and 1. The equation is given below, where z represents a linear combination of features modeled by $z = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b$.

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

3) Random Forest Classifier: Before discussing the Random Forest Classifier, it is important to begin with the Decision Tree algorithm, as it serves as the building block for the Random Forest. A decision tree is a structure consisting of nodes and branches and the structure can be used for making decisions by examining relationships between different variables [3].

Random Forest starts by using the bootstrap sampling technique, which creates random samples from the dataset with replacement, ensuring each tree receives a unique training set [4]. Next, the model splits the trees via random feature subsets. Then, the trees grow using the bootstrap samples and feature subsets and make splits until they reach a stopping point, based on criteria like maximum tree depth or minimum sample size. The final decision depends on the resulting votes from each tree. This project is a classification problem, so the

final decision is simply the majority vote across the trees instead of taking the average of the tree votes for a regression.

Random Forest is advantageous over using Decision Tree alone because the algorithm is better equipped for handling larger datasets and more efficiently reduces overfitting as multiple, unique trees are used.

III. Analysis

A. Data Description

When evaluating the performance of each algorithm, several metrics were used to provide greater insight. Accuracy describes the rate of correct predictions compared to the total number of predictions made. In the equation below, TP and TN refer to true positive and true negative cases while FP and FN refer to the false positive and false negative cases.

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

However, relying on accuracy alone can sometimes be misleading, therefore using additional metrics is more beneficial. Next, precision measures the proportion of true positive cases and all predicted positive cases using the formula below. In other words, precision measures the accuracy of positive predictions.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Third, recall is used to measure the prediction of true positive cases and all cases that are actually positive, thus also accounting for false negative cases.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Lastly, the F1 score combines precision and recall in order to measure the balance between both metrics, accounting for false positive and false negative cases at once. The F1 score, also a weighted average of precision and recall, is represented by the equation below:

$$\text{F1-Score} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

All three models were able to achieve high overall accuracies, thus demonstrating an exceptional performance for predicting cases of diabetes. Tables I, II, and III represent the overall results for Naive Bayes, Logistic Regression, and Random Forest Classifier, respectively. The Random Forest Classifier had the highest accuracy, of 97%, while Naive Bayes had the lowest accuracy of 92%. For all three models, the precision, recall, and f1 scores were consistently high for the negative class. However, due to the dataset imbalance, focusing on the minority class, or positive class in this case, is crucial. The Naive Bayes model had the lowest f1 score for this class while Random Forest had the highest and Logistic Regression was in between. The same was true for precision, but Logistic Regression had a precision value significantly higher than that of Naive Bayes and was closer to that of the Random Forest Classifier. For the recall, however, Naive Bayes outperformed the other models with Random Forest only having a lower recall by 0.05. Overall, Random Forest Classifier was the most efficient model due to accuracy, recall, and f1-score.

TABLE I
NAIVE BAYES MODEL PERFORMANCE

Class	Precision	Recall	F1-Score	Support
0	0.97	0.94	0.96	18300
1	0.52	0.74	0.61	1700
Overall Scores				
			0.92	20000
Macro Avg	0.75	0.84	0.78	20000
Weighted Avg	0.94	0.92	0.93	20000

TABLE II
LOGISTIC REGRESSION MODEL PERFORMANCE

Class	Precision	Recall	F1-Score	Support
0	0.97	0.99	0.98	18300
1	0.87	0.64	0.74	1700
Overall Scores				
			0.96	20000
Macro Avg	0.92	0.81	0.86	20000
Weighted Avg	0.96	0.96	0.96	20000

B. Results

The numerical analysis of the model performance is presented through a confusion matrix, ROC curve,

TABLE III
RANDOM FOREST MODEL PERFORMANCE

Class	Precision	Recall	F1-Score	Support
0	0.97	1.00	0.98	18300
1	0.94	0.69	0.80	1700
Overall Scores				
			0.97	20000
Macro Avg	0.96	0.84	0.89	20000
Weighted Avg	0.97	0.97	0.97	20000

and precision-recall curve. The confusion matrix reveals that the model correctly identified 17,159 instances of Class 0 but misclassified 1,141 as Class 1, indicating a relatively high rate of false positives for this class. Conversely, the model identified 1,256 instances of Class 1 correctly, but misclassified 444 as Class 0, suggesting challenges in recognizing true Class 1 instances effectively.

The precision-recall curve shows a decline in precision as recall increases, which is typical for models handling imbalanced datasets. The curve starts with a high precision at low recall levels, reflecting effectiveness in identifying a subset of positive cases accurately. However, as the recall increases, precision drops significantly, indicating a compromise in precision to capture more true positives. Figure 1 shows the curves for all three models.

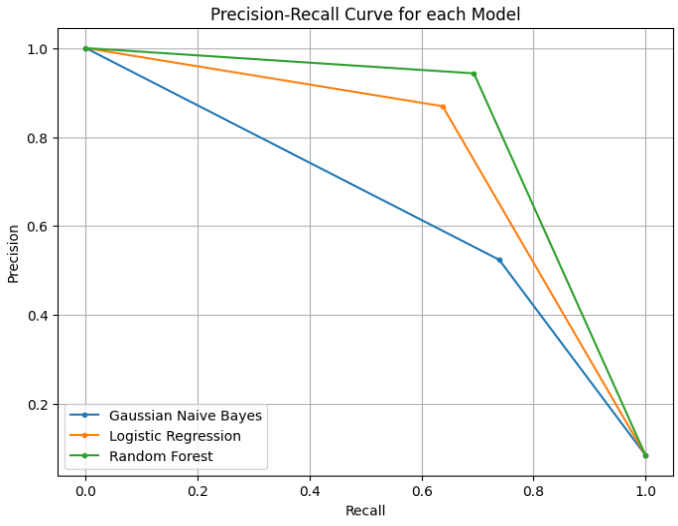

Fig. 2. Precision-Recall Curves

The ROC curve demonstrates a strong capability of each model in distinguishing between the classes across various thresholds. The curve's steep initial ascent and plateau before reaching a false positive rate of 0.2 is indicative of the model's effectiveness at achieving high true positive rates while maintaining a low rate of false positives.

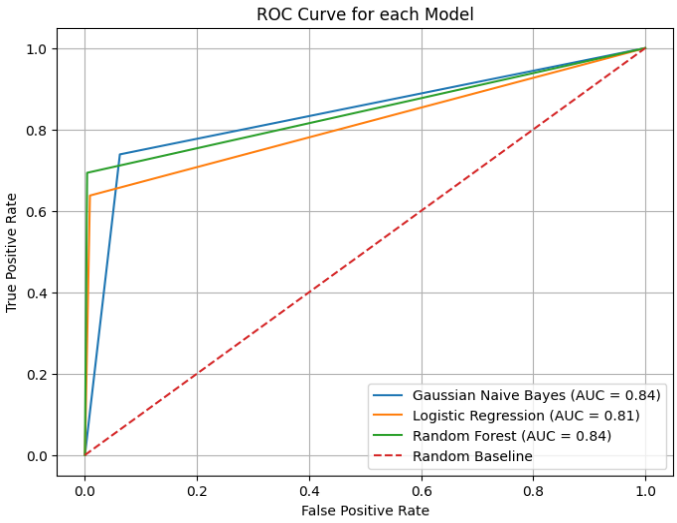

Fig. 3. ROC Curves with the AUC Values

To determine which features were the most significant for predicting diabetes, I used two techniques. First, I used the Logistic Regression model to retrieve the coefficients and found that HbA1c level and Blood Glucose levels had the most significant values. Second, I used feature importance from the Random Forest model and identified the same two features as the most significant. However, there were some inconsistencies for the other features. For example, the Logistic Regression model found that age had little effect on predicting diabetes while Random Forest assigned a higher importance to age, making it the most significant after HbA1c and Blood Glucose levels.

TABLE IV
FEATURE COEFFICIENTS AND IMPORTANCE

Feature	**LR Coefficient**	**RF Importance**
Gender	-0.30	0.010
Age	0.05	0.075
Hypertension	0.67	0.010
Heart Disease	0.67	0.007
Smoking History	-0.10	0.030
BMI	0.33	0.060
HbA1c Level	-6.20	0.190
Blood Glucose Level	-3.80	0.160

IV. Discussion

When comparing the Precision-Recall curves for each model, it is clear that Random Forest Classifier performed the strongest, as the model maintained a higher precision across a wider range of recall values while the Naive Bayes model's precision quickly deteriorated. The Logistic Regression model performed similarly to the Random Forest model, but its precision decreased slightly faster.

While the precision-recall curve favored the Random Forest Classifer over Naive Bayes, these models both had an area under the curve (AUC) value of 0.84 when evaluating the Receiver Operating Characteristic (ROC) curve. When using this metric, the Logistic Regression model had the lowest AUC of 0.81, slightly weaker than the other two models.

In conclusion, while the models perform adequately well in distinguishing between the classes, the observed trade-offs and the challenges in classifying the minority class (Class 1) underscore the need for continued model refinement. Future work should focus on enhancing the model's sensitivity to the minority class without adversely impacting the overall accuracy and precision, ensuring balanced performance across both classes. The Random Oversampling technique should be explored further. The models would also benefit from hyperparameter tuning to help improve recall and f1 scores for the minority class. This balanced approach is crucial, especially in practical applications where the cost of misclassification can be significant.

References

[1] O. Iparraguirre-Villanueva., "Application of machine learning models for early detection and accurate classification of type 2 diabetes," 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10378239/

[2] M. Mustafa, "Diabetes prediction dataset," 2024. [Online]. Available: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data

[3] GeeksforGeeks, "Decision tree," 2024. [Online]. Available: https://www.geeksforgeeks.org/decision-tree/

[4] S. Baladram, "Random forest, explained: A visual guide with code examples." Towards Data Science, 2024. [Online]. Available: https://towardsdatascience.com/random-forest-explained-a-visual-guide-with-code-examples-9f736a6e1b3c