Diabetes Diagnosis

Zhirui Li
Data Science Institute @ Brown
https://github.com/ZhiruiLi1/Diabetes_Diagnosis.git

1. Introduction:

The prevalence of diabetes mellitus is a significant global health concern, and its early and accurate diagnosis is critical for effective management and reducing the risk of complications. This report explores machine learning techniques to enhance diabetes diagnosis, focusing on models such as Logistic Regression with L1 and L2 regularization, Random Forest, Support Vector Machine (SVM), and XGBoost. Among these, the optimal XGBoost model exhibited an f2-beta score of 0.724.

Previous works in this field have laid a solid foundation for using machine learning in diabetes diagnosis. For example, Vyas, Ranjan, Singh, and Mathur (2019) compared various machine-learning algorithms to build predictive models for diabetes diagnosis, with the optimal model having a 91.7 precision [2]. Another study (Agliata et al., 2023) employed artificial neural networks for Type 2 diabetes prediction, achieving an accuracy of up to 86%, underscoring the effectiveness of deep learning in this area [1].

However, previous studies primarily focus on precision and accuracy, and while these metrics are necessary, they may not emphasize recall sufficiently. In medical diagnostics, recall is crucial, as it relates to a model's ability to minimize false negatives. Missing a diagnosis of diabetes can pose more significant risks than false alarms. Thus, the F2-beta score of our XGBoost model, standing at 0.724, is particularly relevant. This score strongly emphasizes recall, balancing precision with the critical need to reduce false negatives in diabetes diagnosis.

2. EDA:

The dataset used in our study, sourced from Kaggle, comprises 100,000 records, four continuous variables, and four categorical variables. The continuous variables in the dataset include age, Body Mass Index (BMI), Hemoglobin A1c (HbA1c) level, and blood glucose level. These variables are critical as they are commonly used indicators in diabetes screening and monitoring. The categorical variables include gender, hypertension status, heart disease presence, and smoking history. These factors are crucial because of their known association with the risk of developing diabetes. The target variable of our analysis is binary, indicating the presence or absence of diabetes, where '0' denotes the absence and '1' signifies the presence of the condition.

After conducting a thorough exploratory data analysis, I have identified several vital visualizations that offer valuable insights into our dataset. These include distributions and correlations among various variables, which are critical for understanding the underlying patterns and relationships pertinent to diabetes diagnosis.

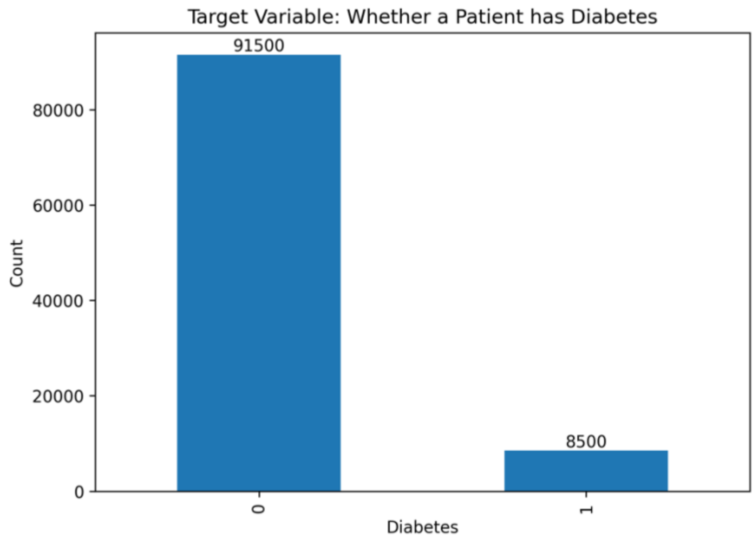


Figure 1 illustrates the uneven distribution of the 'Diabetes' target variable in the dataset, indicating a significant imbalance. To address this, a stratified split strategy will be employed for more balanced data representation in model training and testing.

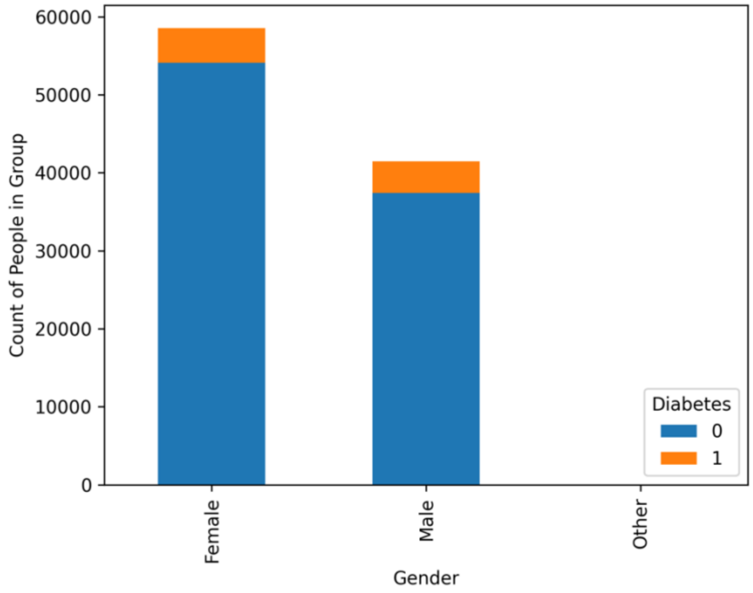


Figure 2 displays the gender distribution in our dataset, highlighting a higher proportion of female patients, approximately one-third more than male patients. Both male and female categories have around 5000 cases each.

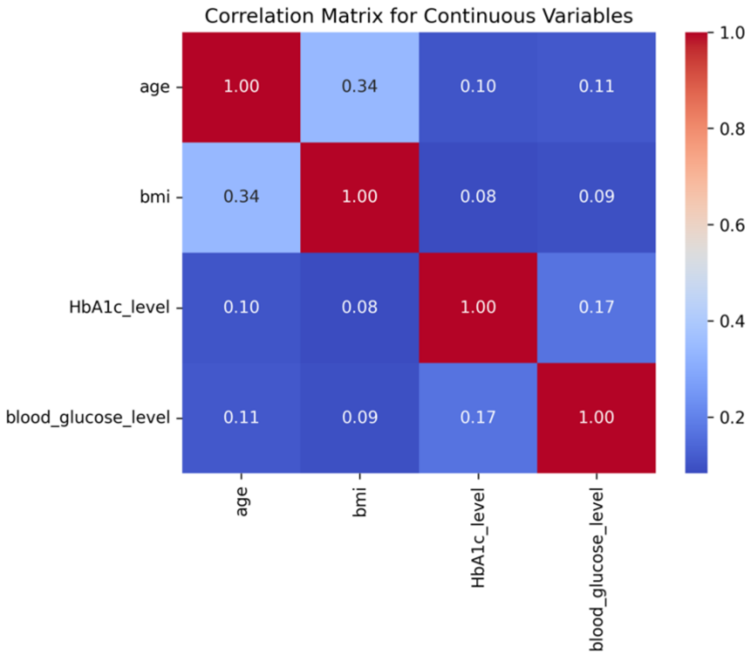


Figure 3 presents the correlation matrix for all continuous variables in our dataset, indicating no strong correlation.

3. Methods:

To address the imbalance in the dataset, I implemented a stratified split approach for dividing the data into training, validation, and testing sets. Additionally, I utilized the standard scaler to normalize the four continuous variables and one-hot encoding for the four categorical variables, resulting in a dataset of 100,000 rows and 15 features. The cross-validation pipeline for the study was structured as follows:

1. Random State Selection: Five distinct random states were chosen to ensure the reproducibility and robustness of the model's performance and to account for uncertainties due to splitting and non-deterministic ML methods.
2. Data Splitting: The dataset was stratified and split into training (60%), validation (20%), and testing (20%) sets. This approach ensured a proportional representation of the target classes across each subset.
3. Preprocessing: Standard scaling was applied to normalize the continuous variables, and one-hot encoding was used to convert categorical variables into a format suitable for model training.
4. Hyperparameter Tuning: An exhaustive grid search was conducted across various hyperparameter combinations to identify the configuration that best performs the validation set.
5. Model Evaluation: Finalize the optimal model based on grid search results and evaluate its effectiveness on the previously unseen test set to assess its generalization capability.

This comprehensive process was designed to rigorously develop and validate the predictive models, ensuring accurate and reliable results in our diabetes diagnosis study.

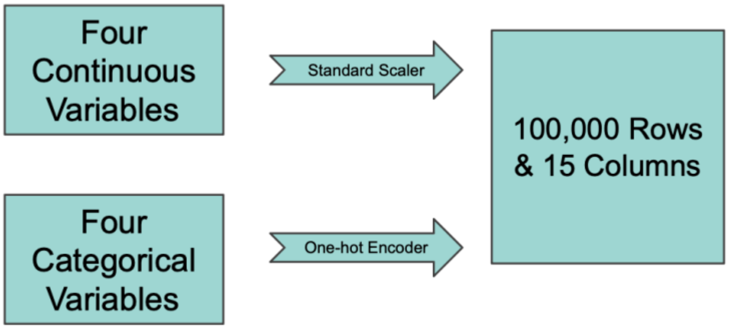
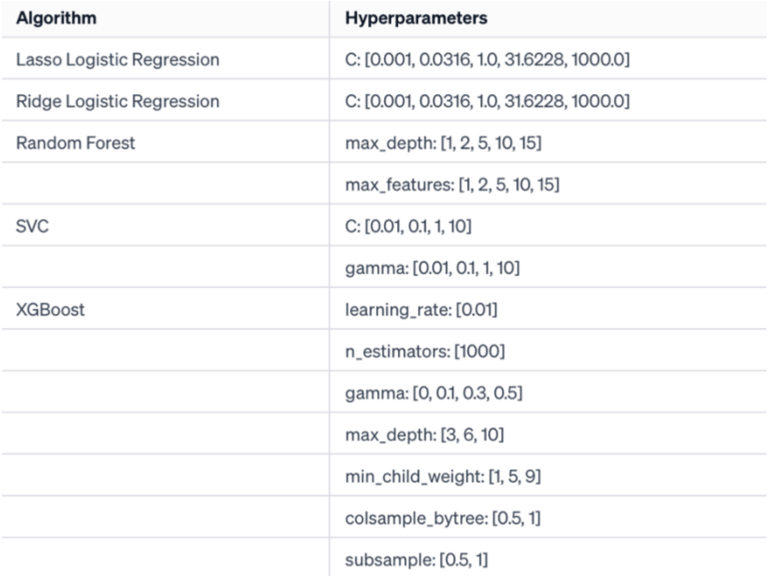

Four Continuous Variables
Standard Scaler
Four Categorical Variables
One-hot Encoder
100,000 Rows & 15 Columns

Figure 4 represents the preprocessing procedures for our continuous and categorical features.



Algorithm	Hyperparameters
Lasso Logistic Regression	C: [0.001, 0.0316, 1.0, 31.6228, 1000.0]
Ridge Logistic Regression	C: [0.001, 0.0316, 1.0, 31.6228, 1000.0]
Random Forest	max_depth: [1, 2, 5, 10, 15]
	max_features: [1, 2, 5, 10, 15]
SVC	C: [0.01, 0.1, 1, 10]
	gamma: [0.01, 0.1, 1, 10]
XGBoost	learning_rate: [0.01]
	n_estimators: [1000]
	gamma: [0, 0.1, 0.3, 0.5]
	max_depth: [3, 6, 10]
	min_child_weight: [1, 5, 9]
	colsample_bytree: [0.5, 1]
	subsample: [0.5, 1]

Figure 5 showcases the five models we trained on: Lasso Logistic Regression, Ridge Logistic Regression, Random Forest, Support Vector Classifier, and Extreme Gradient Boosting, as well as all the combinations of hyperparameters to tune for.

We chose the F2-beta score as the evaluation method because of the reason discussed earlier and prioritized recall due to the significant impact of false negatives.

4. Results:

The plot and table below illustrate the performance of five machine learning models, with error bars indicating the variability due to data splitting and the non-deterministic nature of these methods. A red line in the plot marks the performance of the baseline model, which naively predicts a positive diabetes diagnosis for every patient. This comparison shows that all five models significantly outperform the baseline. Notably, the Random Forest and XGBoost models emerge as top performers, with XGBoost slightly surpassing Random Forest by about 0.002 in the mean F2-beta score, and both are approximately 67 standard deviations away from the Baseline.

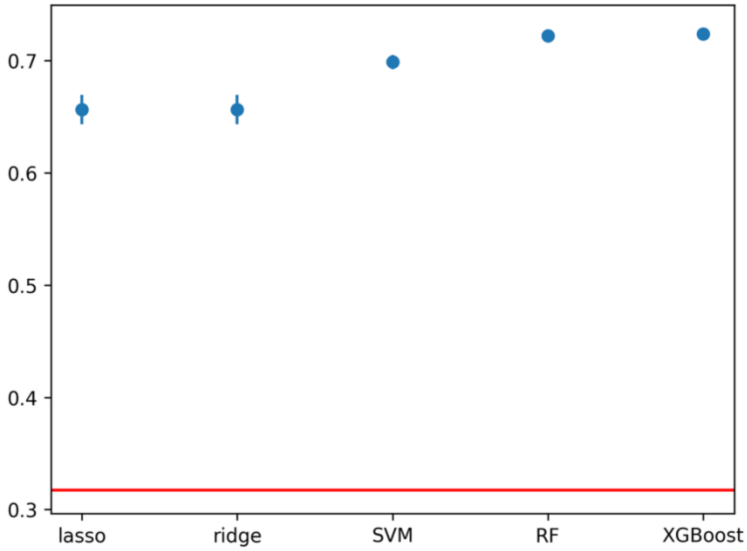
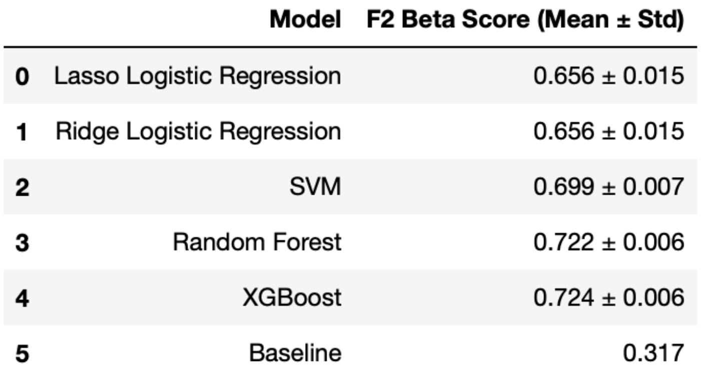


Figure 6 shows the performance of each model with an error bar where the Random Forest and the XGBoost are the optimal ones.



	Model	F2 Beta Score (Mean ± Std)
0	Lasso Logistic Regression	0.656 ± 0.015
1	Ridge Logistic Regression	0.656 ± 0.015
2	SVM	0.699 ± 0.007
3	Random Forest	0.722 ± 0.006
4	XGBoost	0.724 ± 0.006
5	Baseline	0.317

Figure 7 summarizes the mean F2-beta score with standard deviation for each model.

To further analyze the prediction outcomes of the Random Forest and XGBoost models, we can see from subsequent confusion matrices that the Random Forest yields fewer false negatives and false positives compared to the XGBoost model. Specifically, the Random Forest model results in 481 false negatives and 12 false positives, while the XGBoost model has 573 false negatives and 14 false positives. This suggests a slightly better performance in terms of error reduction by the Random Forest model in this dataset.

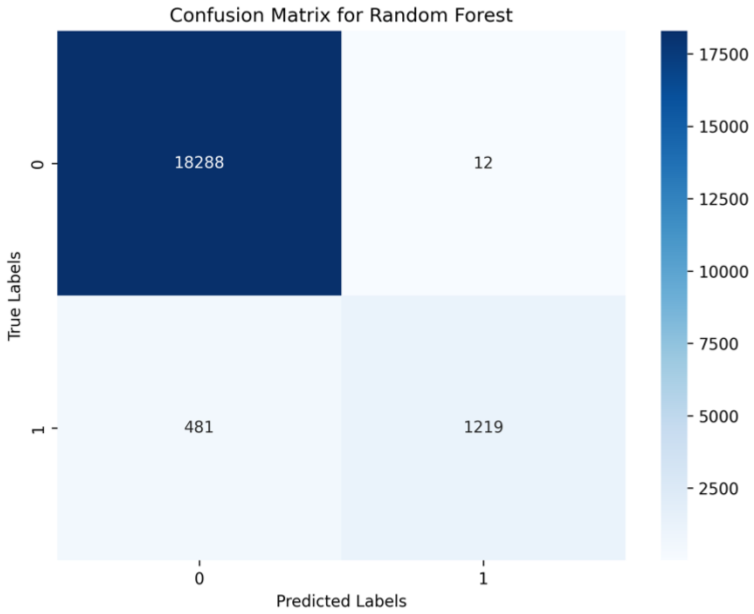


Figure 8 shows the confusion matrix for the random forest model.

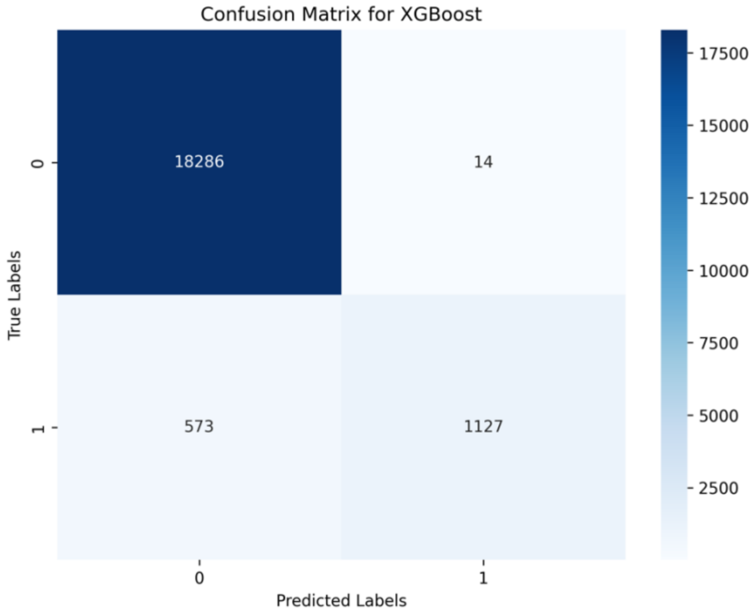


Figure 9 shows the confusion matrix for the XGBoost model.

To decode the XGBoost model's decision-making process, I analyzed its feature importance using 'Gain' as the metric. This metric is crucial for understanding the relative impact of each feature on the model's predictions. The analysis revealed that 'HbA1c_level' emerged as the most significant feature in the XGBoost model for predicting diabetes in patients, ‘blood_glucose_level,’ ‘hypertension,’ and ‘heart_disease’ also play pivotal roles in the model's decision-making process. Sub-categories of smoking history are the least important features.

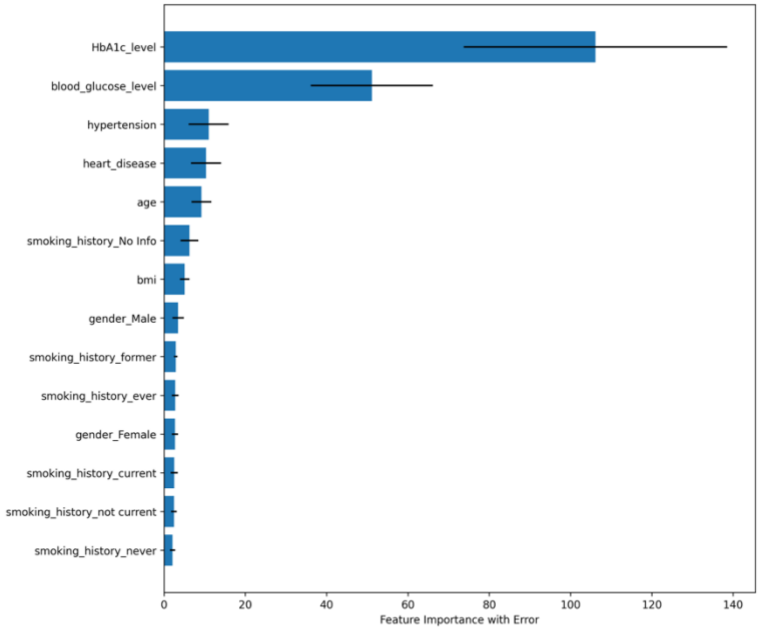


Figure 10 shows the global feature importance of the XGBoost model, where 'HbA1c_level' is the most influential one.

Upon examining the XGBoost model's prediction for a specific case, the model confidently predicts the individual does not have diabetes. The feature 'HbA1c_level' contributes substantially to a lower prediction score, suggesting that for this individual, the 'HbA1c_level' was a decisive factor leading to the model's prediction of a non-diabetic status.

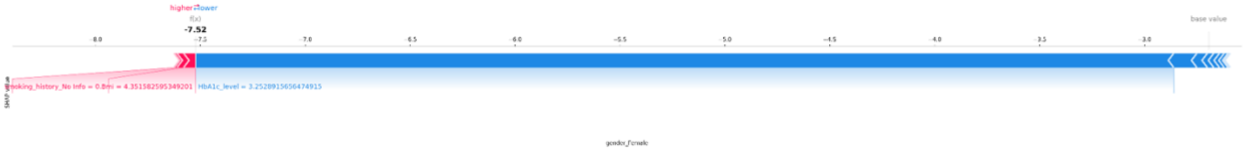


Figure 11 shows the feature importance of a particular case for the XGBoost model.

Additionally, the scatter plot between SHAP values and features reveals intriguing insights into how each feature influences the XGBoost model's predictions. We can see examples below:

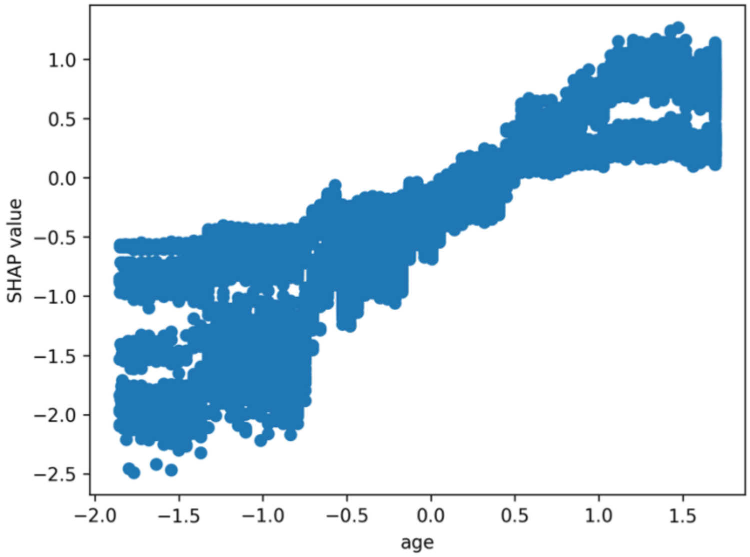


Figure 12 indicates that patients with higher age are more likely to be predicted as having diabetes.

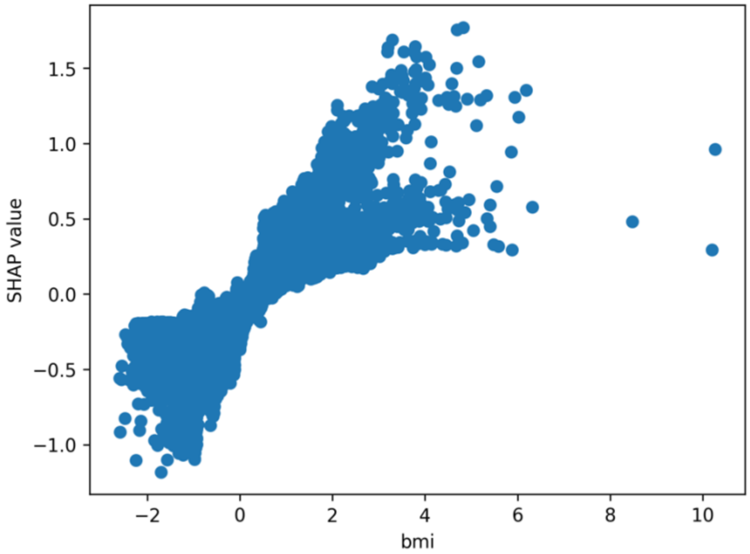


Figure 13 demonstrates a trend where higher BMI values are associated with a greater probability of patients being predicted as diabetic.

5. Outlook:

In conclusion, our analysis unveils positive correlations between diabetes status and features such as age and BMI, alongside negative correlations with features like HbA1c level, a critical factor in diabetes diagnosis. Moreover, our analysis has laid a foundation for future enhancements. Expanding the dataset size is a primary objective to improve our models' accuracy and generalization. A more comprehensive approach to hyperparameter tuning will also be pursued to refine model performance and consider possible interaction effects among features.

Exploring advanced predictive models, such as neural networks, could offer avenues for more sophisticated predictions. Lastly, augmenting our understanding of each feature through extensive literature reviews and consultations with domain experts will enhance our models' interpretability and clinical relevance. These future steps are crucial for evolving our analysis into a more robust and insightful tool in diabetes diagnosis.

6. Reference:

1. Agliata A, Giordano D, Bardozzo F, Bottiglieri S, Facchiano A, Tagliaferri R. Machine Learning as a Support for the Diagnosis of Type 2 Diabetes. *International Journal of Molecular Sciences*. 2023; 24(7):6775. https://doi.org/10.3390/ijms24076775.
2. Sonali Vyas, Rajeev Ranjan, Navdeep Singh, Arohan Mathur, "Review of Predictive Analysis Techniques for Analysis Diabetes Risk," *2019 Amity International Conference on Artificial Intelligence (AICAI)*, pp.626-631, 2019.

7. GitHub:

https://github.com/ZhiruiLi1/Diabetes_Diagnosis.git