XAI-Driven Approach to Predicting Diabetes Risk with Machine Learning

Arati Deshpande, Atul Dubey, Aditya Dhule, Rajat Kumar, Akshat Rathi, Aayush Kore, Harsh Shah

Department of Computer Engineering

Vishwakarma Institute of Technology, Pune, Maharashtra, India

{arati.deshpande1, atul.dubey24, aditya.dhule24, rajat.kumar24, amitkumar.akshat242, aayush.kore242, kapil.harsh24}@vit.edu

***Abstract****—* Early onset of diabetes plays the most vital role in eliminating any severe health issues. This paper introduces a machine learning version using the Pima Diabetes dataset in terms of predicting diabetes risk using features such as the level of glucose, BMI, age, statistics of pregnancies, and the pedigree of diabetes. Data preprocessing has encompassed data cleaning involving missing values, imbalanced data using SMOTE, and data reduction undertaken by Recursive Feature Elimination. Random Forest, XGBoost, LightGBM, and a Stacking Ensemble scored best at 83 percent. We used Explainable AI methods, SHAP, to decipher the model prediction and indicate which features contributed to the output, in order to create more transparency and build trust. The system gives explicit and personalized risk estimation to the user which qualifies it as a viable screening tool of early diabetes risk assessment.

***Keywords***— *Diabetes Risk Estimation, Machine Learning Techniques, Pima Indian Health Data, Synthetic Minority Oversampling (SMOTE), Feature Selection via RFE, User-Friendly Web Platform, Model Interpretability (SHAP)*

1. Introduction

The main reason for very high glucose in diabetes mellitus is either not enough insulin being produced or the body does not use it well [1]. Reckless management of diabetes may cause severe complications such as heart diseases, kidney failures, nerve impairments, and impaired vision that significantly affect the life quality [2, 4]. The International Diabetes Federation estimated that approximately 537 million persons worldwide were already diagnosed with diabetes in 2021, and by 2045, they anticipate that it can increase to 783 million [7]. It results mostly from people not exercising, poor eating habits, and the increase in age, mostly found in low- and middle-income countries [9]. When diabetes is found early, it becomes possible to change one’s diet, exercise more, and take medication to help stop the disease from getting worse [3, 5]. Early detection of people who might face cardiovascular issues prevents harmful health outcomes and expensive treatments [12,13].

Although the machine learning models have produced encouraging results in predicting disease cases, their black-box properties in most instances prevent machine learning in medical application areas, such as healthcare. When operating in such environments, it is not only the prediction of a model that matters but also how this model comes to its conclusion. SHAP (SHapley Additive exPlanations) is one such technique that is being adopted as Explainable AI (XAI) techniques to ensure that model decisions become more transparent and interpretable. XAI would improve on trust and provide users, such as healthcare providers and patients, with the ability to more comfortably use and be aware of the results of the model by emphasizing the role of each feature in the prediction. In this paper, XAI is integratd in order to make predictions the system makes explainable as well as accurate.

With machine learning, this study seeks to predict the chances of diabetes for Pima Indian females by analyzing information from 768 individuals [1,2,9]. The data in the study contains eight features: number of pregnancies, blood sugar, diastolic blood pressure, thickness of skin in the arm, blood hormone levels in two hours, BMI, a measure of diabetes in family members, and age [6,10]. Missing values were filled in with the median and outliers were removed, additionally SMOTE was applied to make sure the classes were evenly distributed [14,16]. RFE was applied in the selection of the most important features in predicting the disease [8]. Random Forest is evaluated for its reliability in the study [4], whereas XGBoost and LightGBM aim at maximum efficiency and accuracy [15]. A personalized tool on the web enables users to give their height, weight, and age to measure their diabetes risk and provides the necessary information [8]. It aids in keeping diabetes from developing and handles early care, helping to face the problem worldwide [7,13].

1. Literature Survey

Many studies rely on machine learning, mainly with the Pima Indian Diabetes Dataset, to evaluate the chances of diabetes. Following, we look at research papers we reviewed to explain their techniques and what they couldn’t do well.

1. Deep Learning Approach for Diabetes Prediction

A neural network and deep learning model built in [1] produced an accuracy of 98.07% on the Pima Indian Diabetes Dataset. Using numerous hidden layers made it possible for the model to detect hard-to-spot patterns and underline the benefits of deep learning in making predictions. Because the computer processing for encryption is challenging, using it on low-powered machines is almost impossible. As how the model decides things is not easy to understand, it cannot be used in clinics, where clear explanations are essential for patients to trust it.

1. Pima Indians Diabetes Classification

Naïve Bayes, Random Forest and J48 Decision Tree models were used by the team in [2] to analyse the Pima Indian Diabetes Dataset, so the findings would be clear to clinicians. A key focus of the approach was to base decisions on algorithms that people can see and understand. However, if we didn’t use Recursive Feature Elimination, the model might have included unnecessary variables that would have decreased its overall results.

1. Machine Learning Approach for Diabetes Prediction

[3] demonstrates that XGBoost achieved top results on the Pima Indian Diabetes Dataset because of its robust way of gradient boosting. Even so, the research did not focus on unfairly distributed classes which means the study’s findings may favour the major class, harming the model’s ability to assess a person’s risk of developing diabetes cases.

1. Comparison of Machine Learning Algorithms

Researchers in [4] used Logistic Regression, SVM and Neural Networks on the Pima Indian Diabetes Dataset and discovered that SVM performed better than the others, mostly because it handles difficult non-linear patterns more efficiently. As a result, the study did not fully maximize the model by selecting only a little number of features which might hinder its power to predict outcomes.

1. Using Machine Learning for Diabetes Prediction

Increased accuracy was gained for the Pima Indian Diabetes Dataset when an SVM was applied, because the model was able to consider the difficult ways the data was associated. This research was conducted to determine the values that give the best SVM outcomes. Even so, those research methods did not make use of Random Forest or Stacking which may make the results more reliable when using a range of models. Because the model was lacking some aspects, it could not deal with a variety of data and difficult patterns.

1. Diabetes Prediction with Machine Learning Algorithms

In article [6], six algorithms were used to analyse the Pima Indian Diabetes Dataset. K-Nearest Neighbours and Naïve Bayes outperformed all the others, mostly thanks to their simplicity and how well they classify data. The goal of our research was to evaluate models based on how hard they work at preprocessing. If Advanced techniques such as Recursive Feature Elimination had not been included, selecting important variables and how well the model works could suffer.

1. Survey on Diabetes Risk Prediction

A survey in [7] examined Support Vector Machines (SVM), K-Nearest Neighbours (KNN) and Random Forest and it showed these methods effectively used the Pima Indian Diabetes Dataset for accurate diabetes risk detection. Still, the study revealed that it is necessary to deal with imbalanced data, since the unfair distribution of classes can make predictions unreliable for cases in the minority class.

The problems witnessed previously were fixed by applying SMOTE technique on the Pima Indian Diabetes data to ensure that the model is well balanced. RFE makes the model more accurate by identifying the features that can make the prediction. Also, applying Stacking Ensemble allows us to get better forecasts than what we’d get from a single model. The website is made to be easy to use so anyone can check their diabetes risk soon after taking the questionnaire.

III. Methodology

This part explains the sequence of steps used to create the diabetes risk prediction model.

A. Dataset

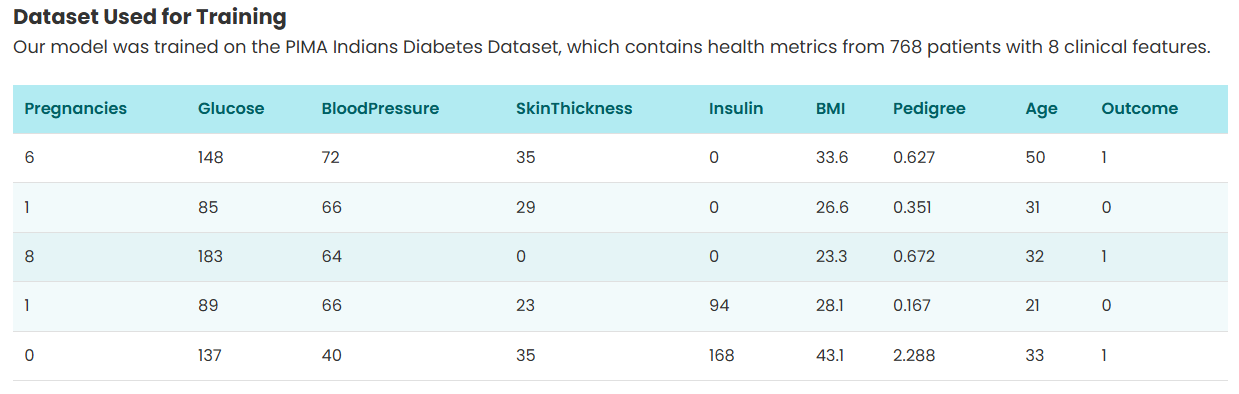
From the UCI Machine Learning Repository comes the Pima Indian Diabetes Dataset with 768 samples from older Pima Indian women. Eight features were gathered, along with an additional target variable that holds a 0 if the patient does not have diabetes and a 1 if they do. Its features are:

* Number of times a person has carried a pregnancy
* Glucose: The glucose level in your plasma blood (units in mg/dL)
* Blood pressure represents diastolic pressure that is measured in a millimeter of mercury (mm Hg).
* Other evaluation involves measurement of skin fold thickness of the triceps (mm). Insulin: Your insulin level in the blood (mu U/ml)
* BMI: Body mass index (kg/m2)
* Genetic diabetes risk can be estimated with a diabetes pedigree function.
* Many people simply report their age in years.

Table I shows five sample rows.

TABLE I

Sample instances from the Pima Dataset



1. Data Preprocessing

The Pima Indian Diabetes Dataset contains only numerical attributes, which include the number of pregnancies, blood glucose concentration, diastolic blood pressure, skinfold thickness, serum insulin level, body mass index (BMI), a diabetes pedigree function indicating family history, and the patient’s age. All the values in the dataset were examined and none were blank or unmarked. In cases where missing values are present in similar datasets, imputation techniques such as mean or median substitution would be applied to ensure data integrity.

The spread of each feature in the dataset was studied to see what it tells us. The majority of women have 0 to 17 pregnancies, with most women having fewer than the average. Glucose levels (0–199 mg/dL) and body mass index (0–67.1 kg/m²) both show a right skew and some large values suggest that those measurements may be outliers. There are only minor variations in blood pressure (0–122 mm Hg) and age (21–81 years), but insulin (0–846 mu U/ml) and skin thickness (0–99 mm) reveal both large variations and zeros which suggests insufficient or unreliable measurements. The distribution of the diabetes pedigree function (0.078–2.42) is skewed to the right which points to different levels of gene-related risk. The figures below show the data in histograms (see Fig. 1).

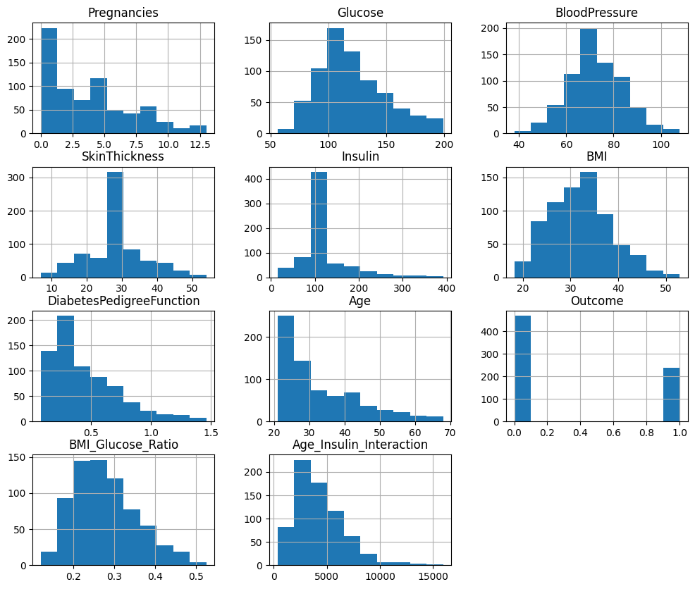


Fig. 1. Feature-wise distribution plot for the entire dataset.

To analyse the interactions between features and diabetes diagnosis, a correlation heatmap was developed. Data revealed that glucose and BMI correlated positively with diabetes, both moderately and that age and pregnancies showed weaker positive correlations with diabetes. Insulin and diabetes pedigree function showed modest relationships, whereas blood pressure and skin thickness were barely connected. As Fig [2] illustrates, this analysis allowed us to pick variables that had the strongest effect on prediction.

Fig [2] shows the relationship between variables in the dataset like Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome, BMI\_Glucose\_Ratio, and Age\_Insulin\_Interaction. The correlation between glucose and the result of the test is very strong (0.87), revealing that high glucose is usually related to having diabetes. A moderate positive correlation is observed between a person’s age and their number of pregnancies (0.56), while glucose levels and insulin levels also show a similar moderate correlation (0.48). BloodPressure does not appear to relate strongly with the majority of variables and has weak associations with others, so it is minimally linked to predictions of diabetes [1].

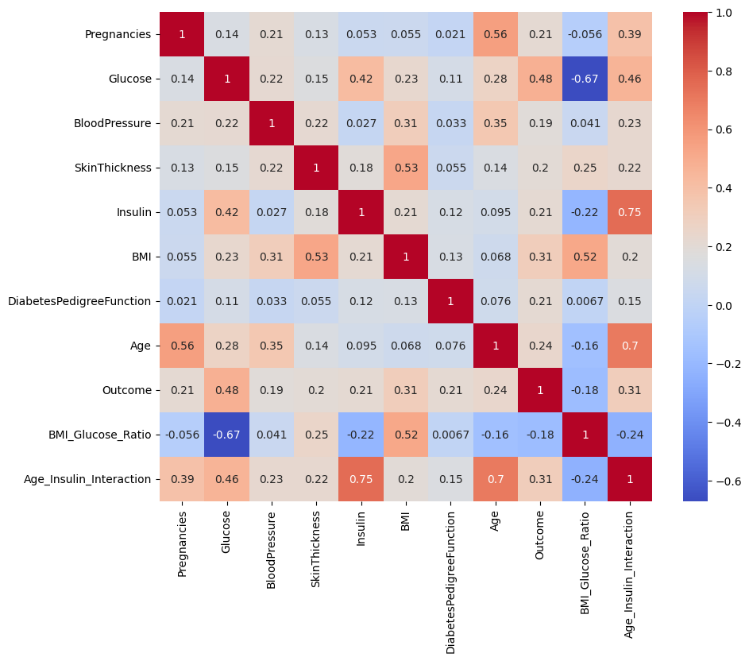


Fig. 2. Feature correlation heatmap for the input dataset.

1. Feature Engineering

RFE helped the process by selecting the top contributors to predict Diabetes from Pregnancies, Glucose, Age, Diabetes Pedigree Function and BMI. RFE successively removes features, then judges the performance to find the most important group.

1. Handling Class Imbalance

There is a significant difference in how the data is portrayed: more people are listed as not having diabetes than as having diabetes (500:268). The minority groups in the data were balanced by using SMOTE, which produced more samples.

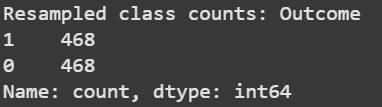


Fig. 3. A balanced distribution of classes shows the value of SMOTE application.

1. Model Training and Testing

By adjusting the hyperparameters, four models of machine learning were applied and reviewed.

* Random Forest predicts strongly using a collection of decision trees.
* The framework XGBoost which relies on decision trees, is a gradient boosting method.
* The model provides a speedy and potent alternative in gradient boosting.
* It brings together the reactions of a number of models by adopting another model.

Accuracy attained by each model was: Random Forest (83%), XGBoost (77%), LightGBM (77%) and Stacking Ensemble (83%). You can see the information in Fig. 4 below,

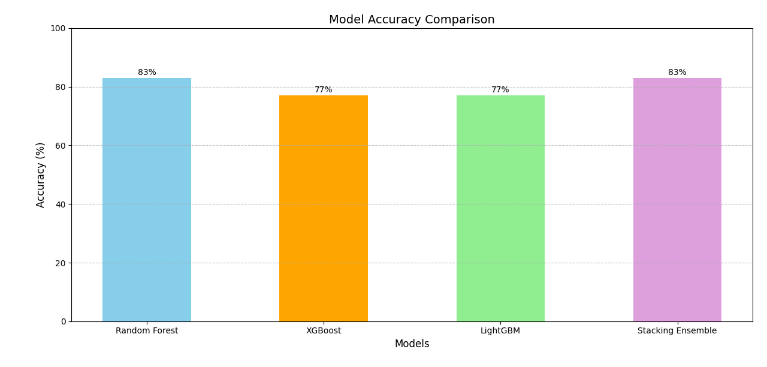


Fig. 4. Model Accuracy Comparison

1. Explainability using SHAP

To increase the clarity and openness of the prediction models, Explainable AI methods were employed via the SHAP (SHapley Additive exPlanations) framework. SHAP values help us determine how the specific feature impacts the specific prediction because it is easier to comprehend what features contribute to the model based on the SHAP values.

This would aid in confirming the rationale of the model and gives insights that are easily understood by the healthcare professionals and the user alike. Force plots and SHAP summary plots were created to visualize the significance of features and explanation of the individual prediction separately, respectively, thus, adding transparency to the decision-making process of the model.

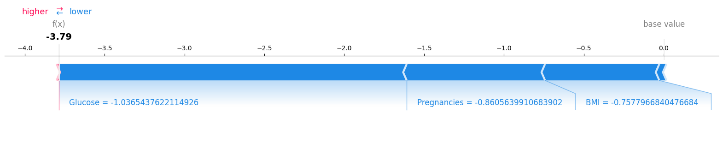


Fig. 5. SHAP Force Plot

A SHAP force plot that explains one of the predictions in detail is presented in fig. 5. This graph is a very breakdown of the features based upon the influence on the outcome of the model. The red, i.e. bar related to the factors Glucose and having movements in the positive direction contrasts with the blue one that has a negative impact. The surface area of each of those segments is proportional to that SHAP value and will characterize the impact that each feature has on a final prediction.

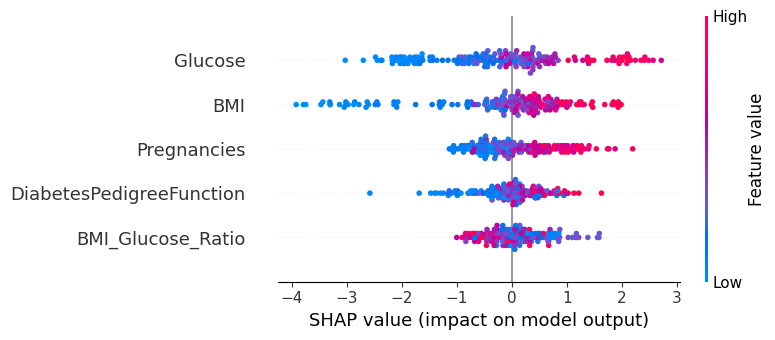


Fig.6. SHAP Summary Plot

Figure 6 above illustrates how each feature influences the model’s predictions across the entire test dataset. The features are prioritized in terms of their significance with the more important features making the top of the list. The feature value (high versus low) is represented by the color of a point (red and blue respectively) and the density of points along the x-axis represents how the feature influences the value. According to this plot, the most contributing features are the inclusion of Glucose and BMI, because they possess the highest distribution range of SHAP values. These features are probably emphasized in the model since they show significant correlations with the variable of interest (Outcome) as shown in the correlation heatmap generated previously.

First, the system collects the user’s health information using a web interface. After that, the features are sent through a process that cleans and makes the data normal. Afterward, the accuracy of the dataset is increased by selecting and developing useful features. Those improved features go to a Random Forest classifier in machine learning which estimates the user’s risk for diabetes. Finally, once risk is calculated, the system provides custom health tips for the user.

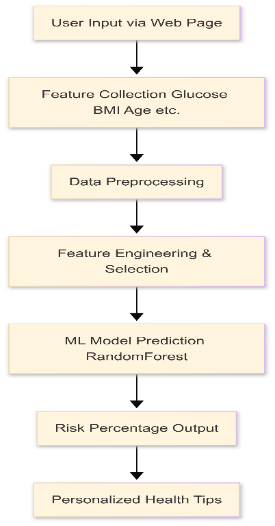
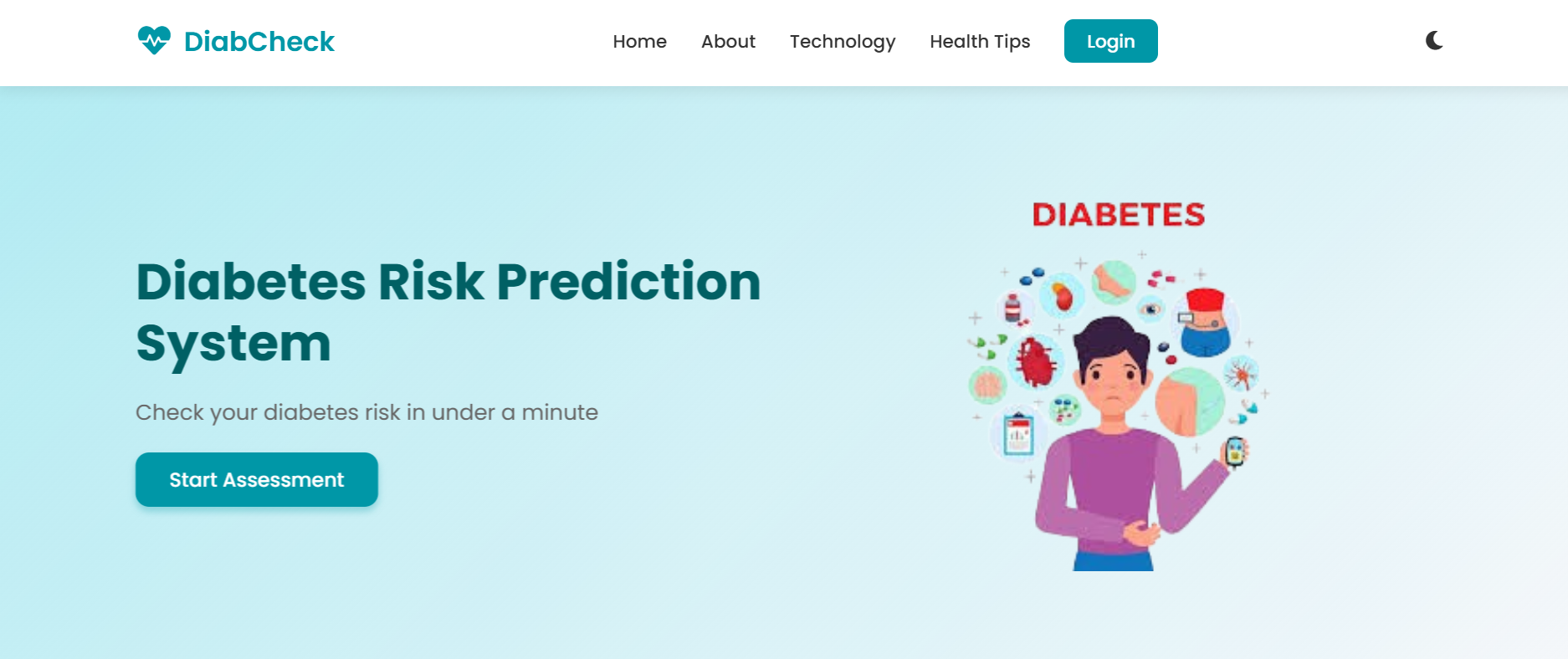


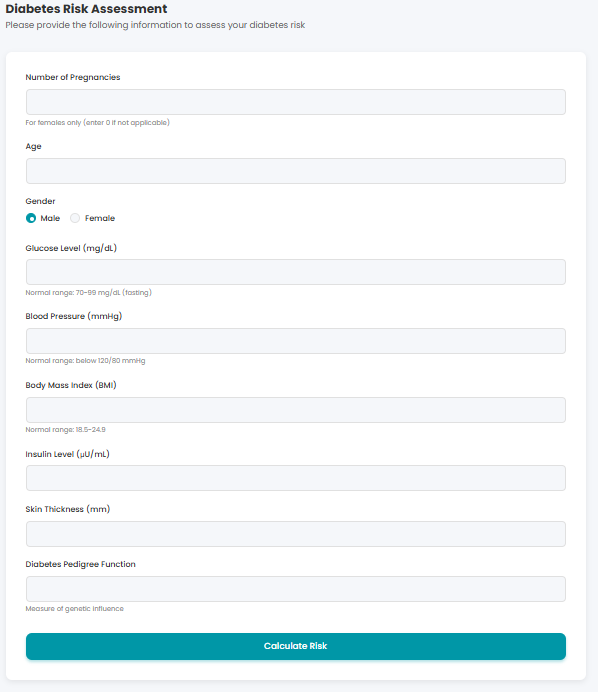
Fig. 5. Process Flow Diagram

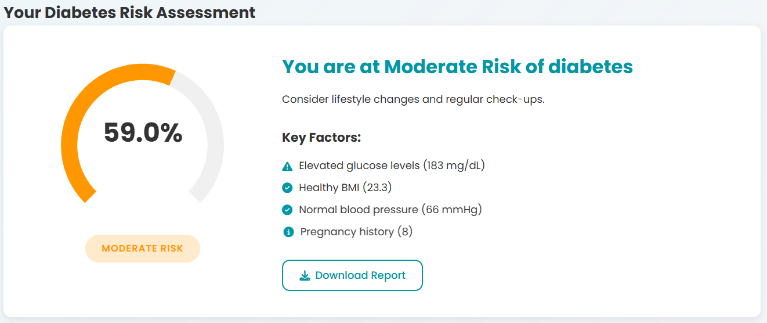
IV. Webpage Interface (User Module)

HTML, CSS and were used to develop the webpage. JavaScript takes charge of the user interface, supported by a Python Flask backend to run the prediction tasks. A form allows users to provide their values for Pregnancies, Glucose, Age, Diabetes Pedigree Function and BMI. Following processing with the model, the backend shows the prediction, for instance, “High risk” or “Low risk.” Photos of the Homepage, Input Form and Output Results are shown below.

Fig. 6. Webpage Screenshots







V. Results and Discussion

The best model turned out to be **Random Forest.**

To increase both its accuracy and reliability, the Random Forest Classifier method builds many decision trees during training and combines their predictions. Because it handles noisy data well and can see non-linear patterns, it was very effective here. According to Table II, the model had an accuracy of 83% and strong results for the other evaluation measures. Among these metrics, you have precision which assesses accurate positives among all your positive predictions; recall which measures how accurately your model found all positives; and the F1-score which helps you evaluate both by computing their harmonic mean.

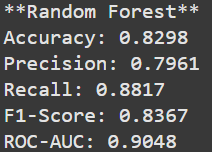


Fig. 7. Evaluation Metrics for Random Forest Model

Figure 8 displays a confusion matrix for the Random Forest model, highlighting the counts of correctly identified positive and negative cases, as well as the instances where the model incorrectly predicted positives or negatives. Its balanced results on the SMOTE-balanced dataset are shown by correctly classifying most cases and misclassifying only a few.

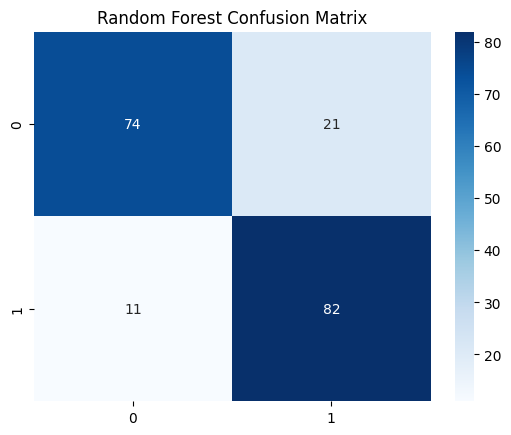


Fig. 8. Confusion Matrix

Figure 9 shows the rates of correctly identified positive cases and incorrectly identified positive cases at various classification thresholds. The AUC for Random Forest suggests that it discriminates well between those who have diabetes and those who are not diabetic.

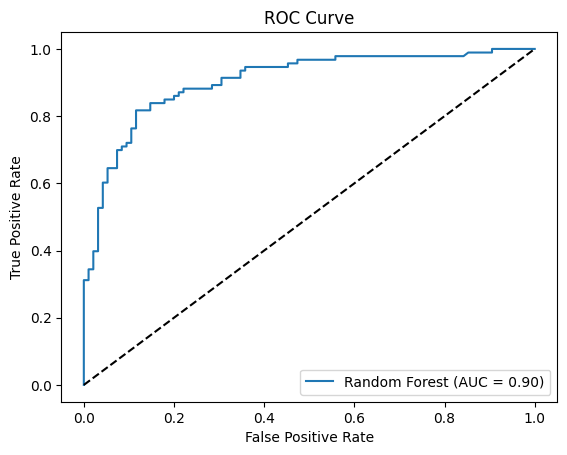


Fig. 9. ROC Curve

The models were evaluated using the below given metrics:

* Accuracy: Correct predictions out of all the predictions.

(1)

* Precision: The count of actual positive cases divided by the total number of predicted positive results.

(2)

* Recall: Percentage of true positives among all the actual positive cases.

(3)

* F1-score: A combined measure of precision and recall calculated using the harmonic mean.

(4)

Here, the terms are defined as follows:

* *: Correctly Predicted positives*
* *: Correctly predicted negatives*
* *: Negatives incorrectly predicted as positives*
* *: Positives incorrectly predicted as negatives*

Table II shows the performance of each model.

TABLE II

Model Performance Metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-score |
| Random Forest | 0.83 | 0.80 | 0.88 | 0.84 |
| XGBoost | 0.77 | 0.77 | 0.83 | 0.80 |
| LightGBM | 0.77 | 0.77 | 0.82 | 0.86 |
| Stacking Ensemble | 0.83 | 0.79 | 0.87 | 0.83 |

VI. Conclusion

The analysis of the Pima Indian Diabetes Dataset resulted in predictive models for diabetes risk, where the Random Forest and Stacking Ensemble techniques demonstrated the best performance, attaining an accuracy of up to 83%. Both SMOTE and RFE improved the results of the model. Users can check their diabetes risk by using the webpage interface. To make this system even better, future work could use stronger methods, larger data samples and extra relevant features.

References

1. S. Hasan, M. A. Alam, and M. R. Islam, ”Deep learning approach for diabetes prediction using PIMA Indian dataset,” BMC Med. Inform. Decis. Mak., vol. 20, no. 1, pp. 1–12, Apr. 2020.
2. M. A. Manzoor, M. N. Uddin, and S. A. Ali, ”Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms,” Neural Comput. Appl., vol. 34, no. 12, pp. 10013– 10026, Mar. 2022.
3. T. N. Joshi and P. Chawan, ”Machine Learning approach for Diabetes Prediction using Pima Dataset,” in Proc. 5th Int. Conf. Inf. Manage., 2021, pp. 123–129.
4. S. K. Dey, A. Hossain, and M. M. Rahman, ”A comparison of machine learning algorithms for diabetes prediction,” ICT Express, vol. 7, no. 2, pp. 179–185, Jun. 2021.
5. A. Swain, S. N. Mohanty, and A. C. Das, ”Using Machine Learning Algorithms to Predict Diabetes Mellitus Based on PIMA Indians Diabetes Dataset,” in Proc. 5th Int. Conf., 2021, pp. 45–50.
6. S. Kumar, R. Mishra, and P. Sharma, ”Diabetes Prediction using Machine Learning Algorithms,” Procedia Comput. Sci., vol. 165, pp. 292–299, 2020.
7. R. Gupta and S. Sharma, ”A survey on diabetes risk prediction using machine learning approaches,” Front. Public Health, vol. 11, pp. 1–15, Mar. 2023.
8. S. R. Islam, M. A. Hossain, and S. K. Dey, ”Machine learning based diabetes prediction and development of smart web application,” Int. J. Cogn. Comput. Eng., vol. 2, pp. 147–153, 2021.
9. V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, “Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms,” Neural Comput. Appl., vol. 35, no. 22, pp. 16157–16173, Aug. 2023. doi: 10.1007/s00521-022-07049-z.
10. M. F. Aslan and K. Sabanci, “A novel proposal for deep learning-based diabetes prediction: Converting clinical data to image data,” Mathematics, vol. 11, no. 4, pp. 1–15, Feb. 2023. doi: 10.3390/math11040904.
11. R. Hasan, M. Islam, M. Hosen, M. Abdullah-Al-Kafi, and N. Radhakrishnan, “Predicting diabetes in women through machine learning,” in Proc. 11th Int. Conf. Comput. Sustain. Global Develop. (INDIACom), 2024, pp. 1307–1311. doi: 10.23919/INDIACom61295.2024.10499129.
12. A. Sheik Abdullah, V. Naga Pranava Shashank, and D. Altrin Lloyd Hudson, “Disseminating the risk factors with enhancement in precision medicine using comparative machine learning models for healthcare data,” IEEE Access, vol. 12, pp. 72794–72812, May 2024. doi: 10.1109/ACCESS.2024.3400023.
13. N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis, and K. Moustakas, “Machine learning tools for long-term type 2 diabetes risk prediction,” IEEE Access, vol. 9, pp. 103737–103757, Jul. 2021. doi: 10.1109/ACCESS.2021.3098692.
14. M. A. Talukder, K. F. Hasan, M. M. Islam, et al., “Toward reliable diabetes prediction: Innovations in data engineering and machine learning applications,” J. Med. Eng. Technol., vol. 48, no. 5, pp. 1–12, Aug. 2024. doi: 10.1177/20556683241272028.
15. K. Lakhwani, M. M. Bundele, and S. Sharma, “Prediction of the onset of diabetes using artificial neural network and Pima Indians diabetes dataset,” in Proc. 5th Int. Conf. Recent Adv. Innov. Eng. (ICRAIE), 2020, pp. 1–6. doi: 10.1109/ICRAIE51050.2020.9358308.
16. S. Sivaranjani, S. Ananya, J. Aravinth, and R. Karthika, “Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction,” in Proc. 7th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS), 2021, pp. 141–146. doi: 10.1109/ICACCS51430.2021.9441876.