Analysis and Evaluation of Machine Learning Models for Diabetes Risk Prediction

Sonam R. Gurung
Department of Computer Science
University of South Dakota
Vermillion, SD, USA
sonam.gurung@coyotes.usd.edu

Siddhartha Phuyal
Department of Computer Science
University of South Dakota
Vermillion, SD, USA
siddhartha.phuyal@coyotes.usd.edu

Sabin Adhikari
Department of Computer Science
Univesity of South Dakota
Vermillion, SD, USA
sabin.adhikari@coyotes.usd.edu

Venkat Karthik Poreddy
Department of Computer Science
University of South Dakota
Vermillion, SD, USA
venkatkarthik.poredd@coyotes.usd.edu

***Abstract*— This research investigates the application of machine learning models for predicting diabetes risk using a dataset sourced from Kaggle. The dataset includes demographic, lifestyle, and medical features like age, gender, BMI, hypertension, smoking history, and blood glucose levels. The study focuses on developing models that are not only accurate but also interpretable for clinical use, balancing prediction accuracy with computational efficiency. Five classification models were evaluated: Decision Tree, Random Forest, XGBoost, LightGBM, and CatBoost. The performance of each model was assessed using metrics such as precision, recall, F1-score, ROC AUC, cross-validation accuracy, and standard deviation. LightGBM emerged as the best-performing model based on its high F1-score and ROC AUC, effectively balancing precision, and recall. CatBoost also demonstrated strong performance, particularly in precision and cross-validation accuracy, though its recall was slightly lower compared to other models. Random Forest and XGBoost consistently performed well across all metrics. Decision Tree, while having the highest recall, underperformed overall. Feature importance analysis revealed that HbA1c levels and blood glucose levels were crucial factors for most models. These findings suggest that LightGBM is a promising model for diabetes prediction, offering valuable insights for clinical decision-making. The study highlights the potential of machine learning to improve early diagnosis and intervention strategies for diabetes prevention.**

Keywords—Data mining, Machine Learning, Diabetes Prediction, Classification models, Model Evaluation

I. INTRODUCTION

Diabetes is a chronic condition that affects millions of individuals worldwide. Early diagnosis and prediction of diabetes can lead to better disease management and a reduction in healthcare costs. Classification models have gained popularity in predicting diabetes by leveraging health data. These models can identify complex patterns in 1datasets, making them superior to traditional statistical techniques in many cases. [1]

This project aims to develop machine learning models for predicting diabetes risk using a dataset containing demographic, lifestyle, and medical features. The project also focuses on understanding the influence of key features and balancing accuracy with computational efficiency, ensuring the interpretability of models for clinical use.

The dataset used for this project is sourced from Kaggle and includes features such as: Age, Gender, BMI, Hypertension, Heart disease, Smoking history, HbA1c level, and Blood glucose level.

The project will evaluate five different classification models. Logistic Regression, a simple and interpretable baseline model. Decision Trees can handle nonlinear relationships and offer interpretability but may overfit. Random Forest which is an ensemble method that combines multiple decision trees to reduce overfitting and improve accuracy. Gradient Boosting is known for high accuracy and effectiveness in handling imbalanced datasets. Implementations include XGBoost, LightGBM, and CatBoost.

The project aims to improve prediction accuracy by using techniques like feature engineering, hyperparameter tuning, and cross-validation. Additionally, the project will explore the creation of a novel ensemble method that combines the strengths of the top-performing models. The goal is to develop a scalable and interpretable solution that can effectively predict diabetes risk using real-world health data. This project will contribute to improving early diagnosis and intervention strategies for diabetes prevention.

II. LITERATURE REVIEW

A. Role of Machine Learning in Diabetes Prediction

Machine learning offers the ability to predict diabetes based on structured datasets containing features such as glucose levels, age, BMI, and family history. These methods go beyond simple rule-based systems by learning intricate relationships in the data. For example, [2] highlighted how machine learning could detect

patterns not immediately obvious to human practitioners, enhancing diagnostic accuracy.

B. Classification Models: An Overview

Logistic Regression is a common starting point for predictive modeling due to its simplicity and ease of interpretation. Researchers such as [3] have successfully used logistic regression to achieve decent accuracy on widely used datasets like the PIMA Indian Diabetes Dataset (PIDD). However, its linear nature often limits performance when dealing with complex data.

Decision Trees are another popular choice due to their interpretability. They split data into branches, helping to capture simple decision-making paths. Studies like those by [4] show that while decision trees perform well on smaller datasets, they may overfit unless carefully pruned or combined with other methods.

Ensemble Models like Random Forest and Gradient Boosting have revolutionized predictive analytics. Random Forest, as demonstrated by [5]. (2021), reduces overfitting by combining predictions from multiple trees. Gradient Boosting methods, including XGBoost and LightGBM, optimize sequentially to minimize errors. As reported by [6] their superior performance in capturing nonlinear interactions, making them ideal for health datasets.

Support Vector Machines (SVM), which rely on optimal hyperplanes to separate data, are effective for high-dimensional problems. [7] utilized SVM with kernel functions to handle nonlinearity, achieving robust results in diabetes prediction.

Neural Networks, though resource-intensive, are becoming increasingly important. [8] demonstrated their exceptional accuracy when applied to large datasets. Neural networks can uncover subtle relationships but require careful tuning and large volumes of labeled data to perform well.

C. Data and Challenges

The performance of these models heavily depends on the quality and representation of data. Diabetes datasets often suffer from class imbalance, where the number of positive cases is significantly smaller than negative cases. [9] emphasized the importance of addressing this issue using techniques like Synthetic Minority Oversampling Technique (SMOTE).

Feature selection also plays a vital role. Variables such as glucose levels, BMI, and family history are strong predictors, but including redundant features can reduce model efficiency. Feature engineering techniques like Principal Component Analysis (PCA) are often used to enhance performance.

Researchers evaluate models using metrics like accuracy, precision, recall, F1-score, and AUC-ROC. However, relying solely on accuracy can be misleading in imbalanced datasets. Metrics like recall and F1-score are more informative in such cases, as they better reflect the model's ability to capture positive instances.

III. Methodology

A. Problem Statement

Diabetes prediction is a critical task in medical diagnosis, requiring accurate and interpretable models to assist healthcare providers in early detection and intervention. Using the provided dataset, which includes demographic, lifestyle, and medical features such as age, gender, BMI, hypertension, smoking history, and blood glucose levels, this project aims to develop machine learning models capable of predicting diabetes risk. The project also focuses on understanding the influence of key features, balancing accuracy, and computational efficiency, and ensuring interpretability for clinical use.

B. Dataset Description

The dataset diabetes_prediction_dataset.csv was used, containing features like age, gender, BMI, hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. The target variable is diabetes status (positive/negative)

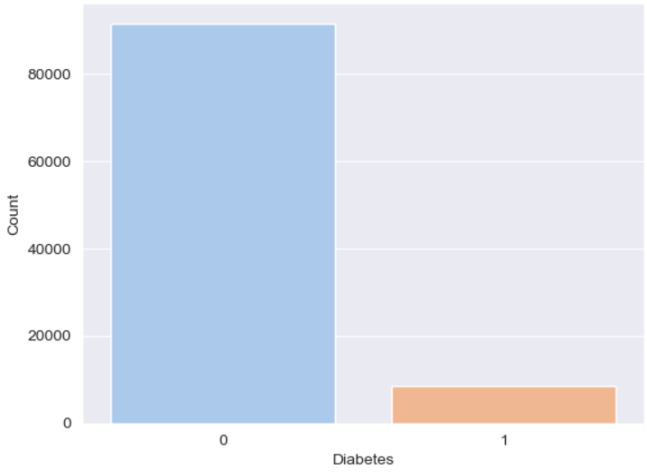

Fig. 1. Class imbalance between people with diabetes (1) and people without (0)

C. Data Preprocessing

- Data Cleaning: Removed entries with missing values and “Other” gender.
- Categorical Encoding: One-hot encoded “gen der” and “smoking history”.
- Feature Scaling: Standardized continuous variables (age, BMI, HbA1c, blood glucose level).

D. Model Implementation

The implemented models include Decision Tree, Random Forest, XGBoost, LightGBM, and CatBoost. The dataset was split into 80% training and 20% testing sets using stratified sampling to ensure the class distribution remained consistent across both subsets.

E. Model Evaluation

The performance metrics employed for the experimental results are precision, recall, f1-score, ROC-AUC, confusion matrix, cross validation accuracy and cross validation standard deviation.

Precision measures the proportion of correctly identified positive value out of all positive classification made. Recall, also called sensitivity measures the proportion of actual positive that were correctly classified. F1-score is the harmonic means of precision and recall. The ROC-AUC evaluates the model's performance across various thresholds, a higher value indicates better classification. A confusion matrix consists of a table which represents the classification performance of the models by counting the true positives, false positives, true negatives, and false negatives as shown in Table I. Cross validation accuracy measures the average accuracy of the model across multiple folds during cross-validation. It evaluates the generalizability of the model. Similarly, cross validation standard deviation measures the standard deviation of cross validation accuracy where lower values indicate better consistency of the model [10], [11].

TABLE I. CONFUSION MATRIX

	Predicted	
Actual	TP	FN
	FP	TN

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{F1-score} = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (3)$$

IV. PROPOSED FRAMEWORK

In this section, we present the proposed framework for predicting diabetes using machine learning models. The framework outlines the steps from data collection to model evaluation and comparison.

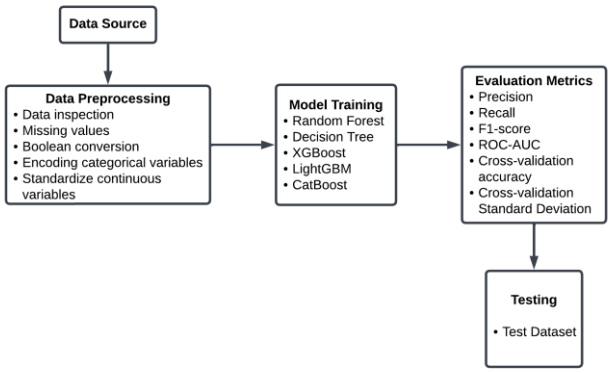

Fig. 2. Proposed Framework

The framework consists of the following main steps:

- Data Source: The dataset containing various health features (age, gender, BMI, HbA1c level, etc.).
- Data Preprocessing: Includes cleaning the data, encoding categorical variables, handling missing values, and scaling continuous features.
- Model Training: Training multiple machine learning models such as Decision Tree, Random Forest, XGBoost, LightGBM, and CatBoost.
- Evaluation Metrics: Metrics like precision, recall, F1-score, ROC-AUC, and cross-validation are used to evaluate the models.

V. RESULTS AND DISCUSSIONS

The performance of all the models were evaluated using the same evaluation metrics, namely, precision, recall, f1-score, ROC-AUC, confusion matrix, cross validation accuracy and cross validation standard deviation. Different evaluation metrics were employed to compare the performance of each model. Along with evaluation, we calculate the importance of the features with respect to the model being used. From our experiments, we got the following evaluation results, shown in Table II.

The results show that Decision tree had the worst performance out of all the other models except for its Recall score, which was the highest. Random forest and XGBoost models performed well with consistent score across all metrics. CatBoost excelled in precision and cv-accuracy but had lower recall which affected its f1-score. The LightGBM model is, according to this evaluation metric, the best model since it has a higher f1-score and ROC AUC value which are crucial metrics in evaluating classification models. LightGBM performed the best in terms of numbers, but CatBoost was also not far behind with its superior recall and cv-accuracy score and similar f1-score to LightGBM.

We also performed experiments to find the most important features within the dataset for each model. From our observations, we found that HbA1c level (average blood sugar over past 2-3 months) and blood glucose level are the most important features for 4 out of 5 models, decision tree, random forest, XGBoost, and CatBoost. For LightGBM model, the most important features were found to be body mass index (BMI), age and glucose level, in that order.

TABLE II. EVALUATION RESULTS

Models	Precision	Recall	F1 Score	ROC AUC	CV Accuracy
Decision Tree	0.7030	0.7435	0.7227	0.8570	0.9519
Random Forest	0.9430	0.6906	0.7973	0.9632	0.9699
XGBoost	0.9562	0.6929	0.8035	0.9783	0.9710
LightGBM	0.9655	0.6912	0.8056	0.9792	0.9717
CatBoost	1.0	0.6729	0.8045	0.9580	0.9718

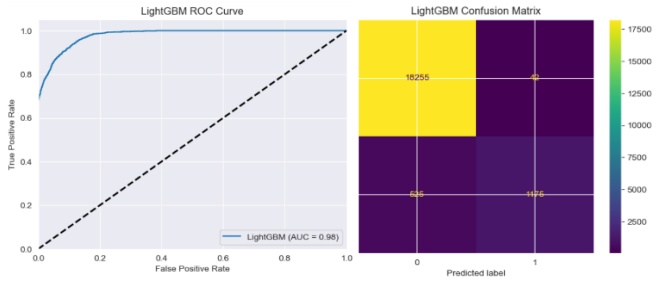

Fig. 3. LightGBM ROC Curve and Confusion Matrix

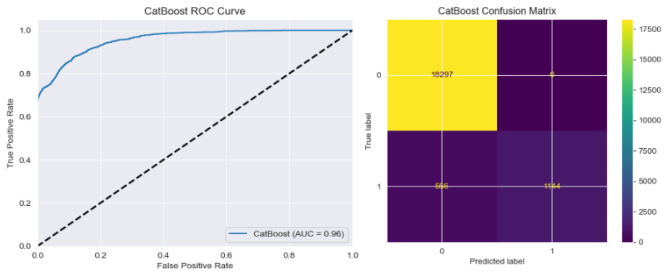

Fig. 4. CatBoost ROC Curve and Confusion Matrix

VI. CONCLUSION AND RECOMMENDATIONS

This study compared five machine learning models, Decision Tree, Random Forest, XGBoost, LightGBM, and CatBoost - to predict diabetes. The evaluation metric used were precision, recall, F1-score, ROC AUC, cross-validation accuracy, and standard deviation. LightGBM emerged as the best model, with the highest F1-score and ROC AUC while balancing precision and recall effectively. CatBoost showed strong performance in precision and cross-validation accuracy but had lower recall compared to other models. Random Forest and XGBoost consistently delivered reliable results across all the metrics. Decision Tree underperformed overall despite having the highest recall. Feature importance analysis identified HbA1c levels and blood glucose as crucial factors for most models.

These results position LightGBM as the most effective model for diabetes prediction, providing valuable clinical insights into the disease. CatBoost was the second-best model with negligible differences with LightGBM.

To improve and continue this project, future work can focus on incorporating a larger and more diverse dataset to train the models on a wider range of patient profiles, potentially improving accuracy and generalizability. Further exploration of feature engineering techniques like Principal Component Analysis (PCA) could lead to the discovery of new predictive variables or more informative features from existing ones. Addressing class imbalance in the dataset with techniques like SMOTE (Synthetic Minority Oversampling Technique) can improve the models' performance, particularly in recall and F1-score. Fine-tuning hyperparameters through techniques like grid search or Bayesian optimization can also lead to significant performance improvements.

REFERENCES

[1] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, Jul. 2013, pp. 1–7. doi: 10.1109/ICCCNT.2013.6726842.

[2] Y. Yang, X. Liu, and Q. Li, "Machine Learning for Predicting Diabetes: Advances and Future Directions," *J. Healthc. Inform. Res.*, vol. 2, no. 3, pp. 45–58, 2018.

[3] A. Alghamdi, S. Alshammari, and R. Khan, "Logistic Regression in Predicting Diabetes Using PIMA Dataset," *Int. J. Artif. Intell. Res.*, vol. 5, no. 1, pp. 23–30, 2020.

[4] P. Kukreja and A. Vashisht, "Decision Trees for Diabetes Prediction: Strengths and Weaknesses," *Data Sci. Appl. J.*, vol. 4, no. 2, pp. 95–102, 2019.

[5] R. Tirthapura, V. Singh, and M. Rao, "A Comparative Analysis of Random Forest and Gradient Boosting in Healthcare Predictions," in *Proceedings of the IEEE Conference on Data Science*, 2021, pp. 457–468.

[6] X. Zhang, H. Wang, and Z. Lin, "Gradient Boosting for Predicting Diabetes: A Review of Current Advances," *Mach. Learn. Healthc.*, vol. 6, no. 1, pp. 89–101, 2022.

[7] R. Gupta, S. Kumar, and M. Shukla, "Support Vector Machines in Medical Diagnoses: The Case of Diabetes," *Biomed. Eng. Insights*, vol. 8, no. 3, pp. 112–121, 2020.

[8] M. Al-Rakhami, S. Al-Amir, and M. Al-Mohanna, "Deep Learning in Diabetes Prediction: Achieving Accuracy and Scalability," *Appl. Artif. Intell.*, vol. 35, no. 4, pp. 178–192, 2021.

[9] N. Mishra, A. Verma, and K. Choudhary, “Addressing Privacy Challenges in Healthcare Machine Learning Models,” *J. Med. Inform. Secur.*, vol. 9, no. 1, pp. 34–49, 2022.
[10] H. Dalianis, “Evaluation Metrics and Evaluation,” in *Clinical Text Mining: Secondary Use of Electronic Patient Records*, H. Dalianis, Ed., Cham: Springer International Publishing, 2018, pp. 45–53. doi: 10.1007/978-3-319-78503-5_6.
[11] H. M and S. M.N, “A Review on Evaluation Metrics for Data Classification Evaluations,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 01–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.