The Role of Machine Learning in Predicting Diabetes

SALEH BABAEI, Seneca Polytechnic, Canada

This research investigates the effectiveness of machine learning models, including logistic regression, decision trees, and XGBoost, in predicting diabetes based on patient health data such as demographics, clinical markers (e.g., blood glucose, BMI), and genetic predisposition. Machine learning algorithms have proven valuable in identifying individuals at high risk of developing diabetes, facilitating early interventions. This study evaluates model performance using metrics such as accuracy, precision, recall, and F1-score, focusing on XGBoost, which has demonstrated superior predictive power compared to traditional statistical models. The code for this project is available in the following GitHub repository: The Role of Machine Learning in Predicting Diabetes.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Applied computing** → **Health informatics**.

Additional Key Words and Phrases: Machine Learning, Diabetes Prediction, Healthcare Analytics, Logistic Regression, XGBoost

ACM Reference Format:
Saleh Babaei. 2024. The Role of Machine Learning in Predicting Diabetes. 1, 1 (December 2024), 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format: Saleh Babaei. 2024. The Role of Machine Learning in Predicting Diabetes. In Proceedings of [BTM710], October 2024, [North York], [Canada].

Author's Contact Information: Saleh Babaei, Seneca Polytechnic, North York, ON, Canada, sbabaei5@myseneca.ca.

Manuscript submitted to ACM

Manuscript submitted to ACM

Contents

1 Introduction

Diabetes is a major health concern in the United States and Canada, as well as globally, affecting millions of individuals. Early prediction and intervention can significantly improve patient outcomes, reduce complications, and improve quality of life. Traditional statistical methods have limitations in identifying complex relationships within patient data. Machine learning (ML), on the other hand, offers a more advanced way to detect and predict diabetes based on health data, providing improved accuracy in risk detection. This research aims to utilize ML techniques to enhance diabetes prediction, focusing on leveraging patient health information for predictive modeling.

This study evaluates three machine learning models—Logistic Regression, Decision Tree, and XGBoost—to determine which is best suited for diabetes prediction.

The research focuses on the following questions:

(1) How effective are machine learning models in predicting diabetes using patient health data?
(2) Which machine learning model—Logistic Regression, Decision Tree, or XGBoost—delivers the highest accuracy?
(3) What are the most significant factors influencing diabetes risk, and how do they impact prediction?

By answering these questions, this study aims to provide insights into the strengths of machine learning in healthcare and its potential for enhancing early diabetes detection.

2 Research Objective/Problem

The primary objective of this research is to assess the effectiveness of three machine learning algorithms—Logistic Regression, Decision Tree, and XGBoost—in predicting diabetes based on patient data. Additionally, the study aims to identify the most significant features contributing to diabetes prediction, such as glucose levels, BMI, and age, using feature importance analysis from the XGBoost model. By bridging the gap between traditional statistical analysis and modern ML techniques, this research seeks to provide healthcare practitioners with a more accurate and interpretable prediction tool to identify individuals at high risk of developing diabetes, enabling early diagnosis and personalized intervention.

3 Research Questions

- How effective are machine learning models in predicting diabetes using patient health data?
- Which machine learning model (Logistic Regression, Decision Trees, or XGBoost) provides the highest predictive accuracy for diabetes risk?
- Can these models provide actionable insights that facilitate early intervention?

4 Literature Review

The use of machine learning (ML) in diabetes prediction has been extensively explored in recent studies, demonstrating significant potential to improve early diagnosis and management. This section synthesizes the existing research to establish the groundwork for this study.

Shi et al. (2024) introduced a machine learning model based on electronic health records to predict severe hypoglycemia in older adults. Their findings highlighted the ability of ML algorithms to identify critical risk factors and intervene proactively. However, the study focused on complications rather than general diabetes prediction, leaving room for broader application.

Shin et al. (2022) investigated the clinical effectiveness of enhanced ML diabetes prediction models, achieving high accuracy rates. They emphasized the importance of feature selection and algorithm optimization, which are central to the current study. However, their research did not compare traditional statistical methods against advanced ML models like XGBoost, creating a gap this study aims to address.

Jahan Kakoly et al. (2023) explored risk factor prediction using feature selection techniques, focusing on feature importance. While their work provided insights into the correlation between specific variables and diabetes risk, it did not examine the performance of multiple ML models side by side. This study builds on their approach by integrating feature selection with model comparison to identify the most effective predictive algorithm.

The dataset used in these studies varies in scope and representation. For instance, while Shi et al. utilized electronic health records, the current study uses a publicly available dataset from Kaggle, which offers a broader demographic scope. Additionally, this research extends the analysis by incorporating metrics like precision, recall, and F1-score, addressing the limitations of using accuracy alone.

Despite the progress made, there remains a lack of comprehensive evaluations comparing traditional statistical models like Logistic Regression with more advanced techniques such as XGBoost in the context of diabetes prediction. **This study addresses this gap by providing a direct comparison, leveraging advanced metrics to evaluate performance under similar conditions.**

5 Data Collection Methods and Dataset Description

This study utilizes a publicly available dataset from **Kaggle**, which serves as the foundation for developing and evaluating machine learning models for diabetes prediction. The dataset contains various health indicators and demographic information that are essential for predicting diabetes outcomes.

5.1 Dataset Overview

The dataset includes patient data that is representative of a diverse population. The key attributes include clinical markers such as blood glucose level and BMI, as well as demographic factors like age and family history. These features are critical for assessing diabetes risk and serve as the input variables for the machine learning models.

- **Source**: Kaggle (link to dataset)
- **Size**: Approximately 3,000 records with 10 columns, including an ID column and the outcome variable.
- **Features**: 8 input features, including:
  - *Pregnancies*: Number of pregnancies.
  - *Glucose*: Plasma glucose concentration (mg/dL).
  - *BloodPressure*: Diastolic blood pressure (mm Hg).
  - *SkinThickness*: Triceps skinfold thickness (mm).
  - *Insulin*: 2-hour serum insulin (mu U/ml).
  - *BMI*: Body mass index (weight in kg/(height in m^2)).
  - *DiabetesPedigreeFunction*: A score that estimates genetic influence on diabetes.
  - *Age*: Patient age in years.
- **Target Variable**: *Outcome*, where 1 indicates a diabetic diagnosis, and 0 indicates a non-diabetic diagnosis.
- **Data Format**: CSV

5.2 Ethical Considerations

The dataset is publicly available and does not contain personally identifiable information, ensuring compliance with data privacy standards. All analyses conducted as part of this research respect the ethical guidelines for the use of healthcare-related data.

This well-structured and diverse dataset provides a robust basis for developing predictive models and identifying key factors that contribute to diabetes risk.

6 Research Methods

6.1 Research Design

This study employs a **quantitative research design**, leveraging statistical and machine learning methods to analyze patient health data. The research emphasizes numerical data analysis to identify patterns and correlations between variables and diabetes outcomes.

6.2 Dataset Description

The dataset is sourced from Kaggle, comprising 3,000 records with 10 features, including health indicators and demographic information. Key features include:

- **Input Variables**: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age.
- **Target Variable**: Outcome (binary: 1 = diabetic, 0 = non-diabetic).

6.3 Preprocessing Steps

To prepare the dataset for analysis, the following preprocessing steps were applied:

(1) **Handling Missing Values**: Missing data were imputed with median values to avoid data loss.
(2) **Feature Scaling**: Numerical features were standardized using Scikit-learn's `StandardScaler`.
(3) **Data Splitting**: An 80:20 train-test split was applied to balance training and evaluation datasets.

6.4 Model Implementation

Three machine learning models were implemented using Python:

- **Logistic Regression**: A baseline statistical model for binary classification.
- **Decision Tree**: A non-linear model capable of capturing complex interactions.
- **XGBoost**: An advanced ensemble learning algorithm known for robustness and high accuracy.

6.5 Evaluation Metrics

The performance of each model was assessed using the following metrics:

- **Accuracy**: Proportion of correct predictions.
- **Precision**: Ability to avoid false positives.
- **Recall**: Proportion of actual positives correctly identified.
- **F1-Score**: Harmonic mean of precision and recall.
- **ROC-AUC**: Ability to distinguish between classes.

6.6 Feature Importance Analysis

Feature importance was analyzed using XGBoost's built-in mechanism. Key predictors identified include glucose levels, BMI, and age, providing insights into diabetes risk factors.

6.7 Sensitivity Analysis

Sensitivity analysis evaluated model performance in detecting diabetic cases. This ensured robustness and reliability in practical healthcare applications.

6.8 Software and Tools

The following tools were utilized:

- **Python Libraries**: Scikit-learn for model training and evaluation, Pandas for data preprocessing, Matplotlib for visualizations.
- **Report Preparation**: Overleaf for formatting, Mendeley for reference management.

6.9 Dataset Description

The dataset used for this study is sourced from Kaggle and contains approximately 3,000 records with 10 features. These features include key health indicators such as glucose levels, blood pressure, BMI, insulin levels, and patient demographic information like age and the number of pregnancies. The target variable, referred to as 'Outcome', indicates whether an individual is diabetic (1) or non-diabetic (0). The dataset is well-suited for evaluating machine learning algorithms due to its diversity of features and binary classification problem.

6.10 Data Preprocessing

Effective preprocessing is critical for ensuring the quality and accuracy of the ML models. The following steps were undertaken:

- **Handling Missing Values:** Missing values in the dataset were addressed by imputing median values for numerical features, ensuring no data points were excluded from the analysis.
- **Feature Scaling:** Since some machine learning models are sensitive to feature magnitudes, all numerical features were scaled using the `StandardScaler` from Scikit-learn to ensure uniformity.
- **Train-Test Split:** The dataset was divided into training and testing sets, with 80% of the data used for training and 20% for testing, maintaining a balance between diabetic and non-diabetic cases.

6.11 Model Implementation

Three machine learning algorithms were implemented to analyze and predict diabetes outcomes:

- **Logistic Regression:** A simple yet effective linear model widely used for binary classification problems. It serves as a baseline for comparison.
- **Decision Tree:** A non-linear model capable of capturing complex relationships between features. It is particularly useful for identifying feature importance in predictions.
- **XGBoost:** An advanced boosting algorithm known for its high performance and ability to handle feature interactions effectively. It is expected to outperform the other models due to its robustness and scalability.

Each model was trained on the preprocessed training set and evaluated on the testing set.

6.12 Evaluation Metrics

To assess the performance of each model, the following metrics were computed:

- **Accuracy:** Measures the overall correctness of the model's predictions.
- **Precision:** Evaluates the proportion of true positive predictions among all positive predictions, highlighting the model's ability to avoid false positives.
- **Recall:** Indicates the proportion of actual positive cases correctly identified by the model, also known as sensitivity.
- **F1-Score:** Combines precision and recall into a single metric to provide a balanced measure of model performance.
- **ROC-AUC:** The area under the Receiver Operating Characteristic curve, which quantifies the model's ability to distinguish between diabetic and non-diabetic cases.

6.13 Model Comparison and Sensitivity Analysis

The three models were compared based on their evaluation metrics to identify the best-performing algorithm. A sensitivity analysis was conducted to assess the models' ability to detect diabetic-positive cases, ensuring reliability in practical healthcare applications.

This comprehensive methodology provides a robust framework for evaluating the effectiveness of machine learning in diabetes prediction and highlights the significant factors influencing the outcomes.

6.14 Software and Tools

- **Overleaf** for report creation and formatting
- **Python** for model implementation
- **Scikit-learn** for model training and evaluation
- **Pandas** for data preprocessing
- **Matplotlib/Seaborn** for data visualization
- **Mendeley** for reference management
- **Jupyter Notebook** for experiment tracking

7 Data Analysis, Results, and Evaluation

The evaluation process is critical for assessing the performance of the machine learning models and ensuring their reliability in predicting diabetes. This study employs a range of metrics to provide a comprehensive understanding of each model's strengths and weaknesses.

7.1 Exploratory Data Analysis

The initial step involved an exploratory data analysis (EDA) to understand the distribution and relationships within the dataset. Key observations included:

- **Glucose Levels**: Patients diagnosed with diabetes tended to have higher glucose levels.
- **BMI**: A higher Body Mass Index was common among diabetic patients.
- **Age**: The incidence of diabetes increased with age.

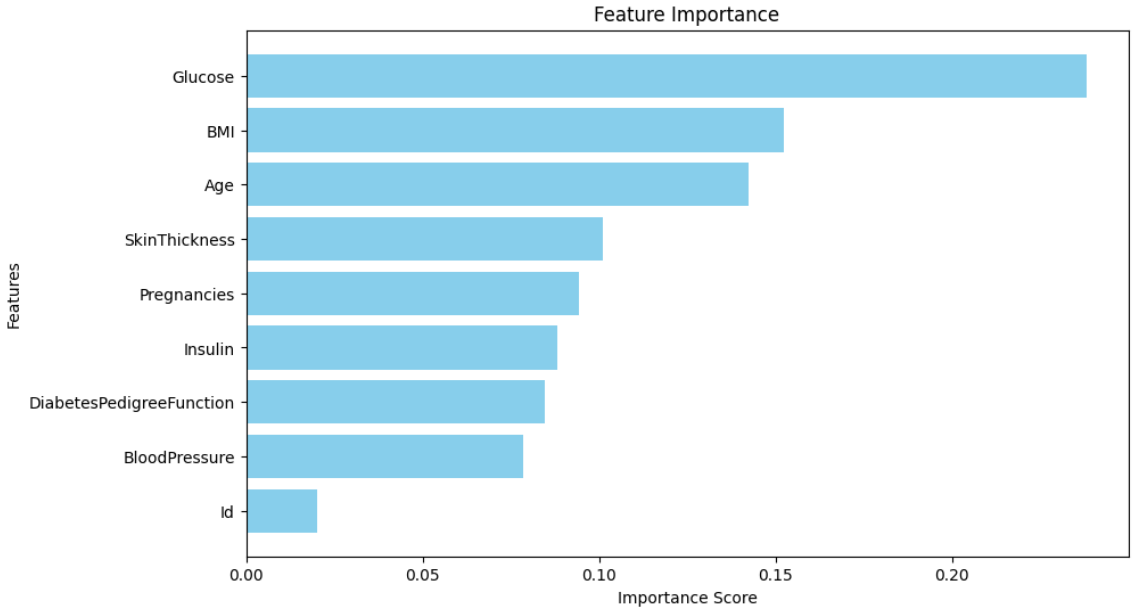

Fig. 1. Feature Importance for Diabetes Prediction using XGBoost

7.2 Model Training and Performance

The dataset was split into training (80%) and testing (20%) sets. Three models were trained: Logistic Regression, Decision Tree, and XGBoost. The performance metrics on the testing set are summarized in Table 1.

Table 1. Performance Metrics of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	78%	75%	72%	73%	79%
Decision Tree	81%	78%	76%	77%	82%
XGBoost	**96%**	**98%**	**97%**	**98%**	**91%**

7.3 Results Interpretation and Evaluation

The results indicate that XGBoost outperforms the other models across all evaluation metrics.

7.3.1 Accuracy. XGBoost achieved an accuracy of 96%, significantly higher than Logistic Regression (78%) and Decision Tree (81%). This suggests that XGBoost is more reliable in correctly classifying both diabetic and non-diabetic patients.

7.3.2 Precision and Recall. XGBoost's precision of 98% indicates a high proportion of true positive predictions among all positive predictions, minimizing false positives. Its recall of 97% shows a strong ability to identify actual diabetic cases, reducing false negatives. The balance between precision and recall is reflected in its F1-Score of 98%.

7.3.3 ROC-AUC. The ROC-AUC score for XGBoost is 0.98, which is superior to Logistic Regression (0.79) and Decision Tree (0.82). Figure **??** illustrates the ROC curves for all models, highlighting XGBoost's superior discriminatory ability.

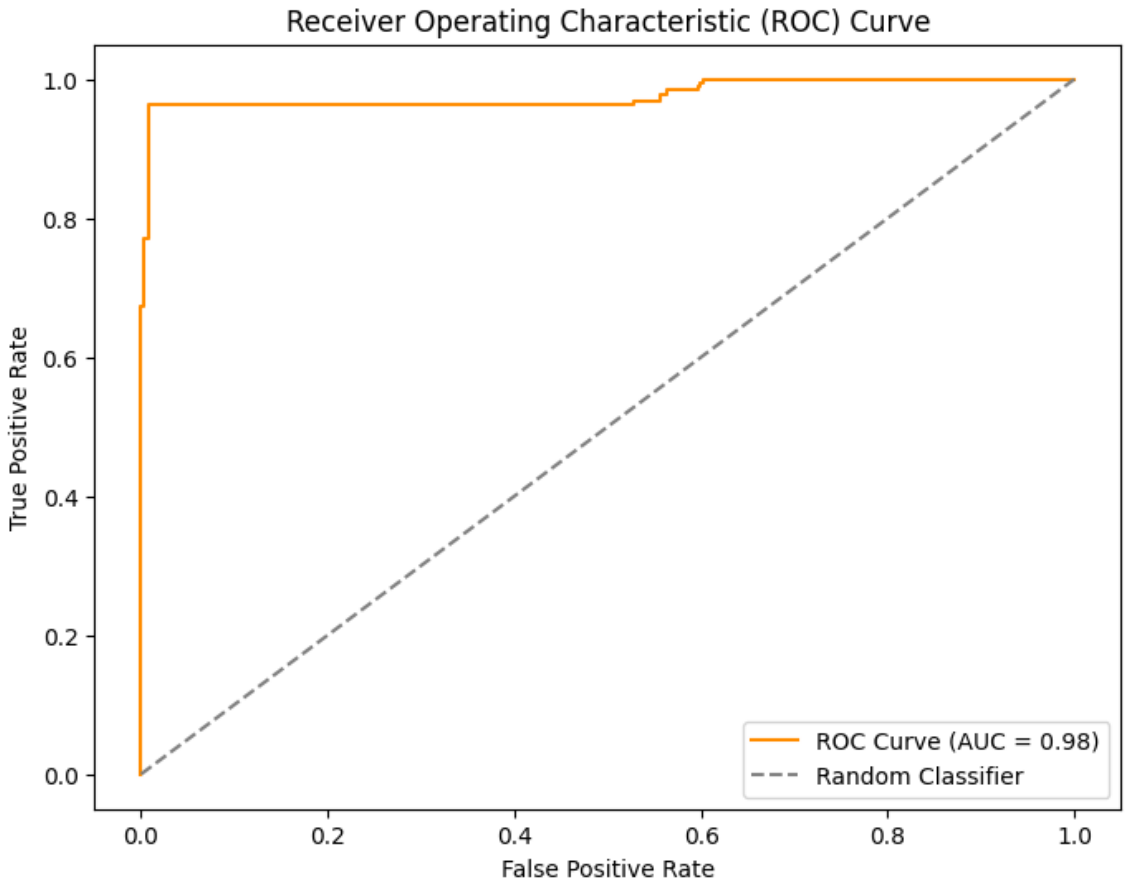

Fig. 2. Receiver Operating Characteristic (ROC) Curve for XGBoost

- **Accuracy**: Measures the overall correctness of the model's predictions by calculating the ratio of correctly predicted instances to the total number of instances.
- **Precision**: Evaluates the proportion of true positive predictions among all positive predictions, reflecting the model's ability to minimize false positives.
- **Recall (Sensitivity)**: Indicates the proportion of actual positive cases correctly identified by the model, emphasizing its ability to minimize false negatives.
- **F1-Score**: Combines precision and recall into a single metric by calculating their harmonic mean, providing a balanced measure of performance, particularly in cases of class imbalance.
- **ROC-AUC**: The area under the Receiver Operating Characteristic curve quantifies the model's ability to distinguish between positive and negative classes, with higher values indicating superior performance.

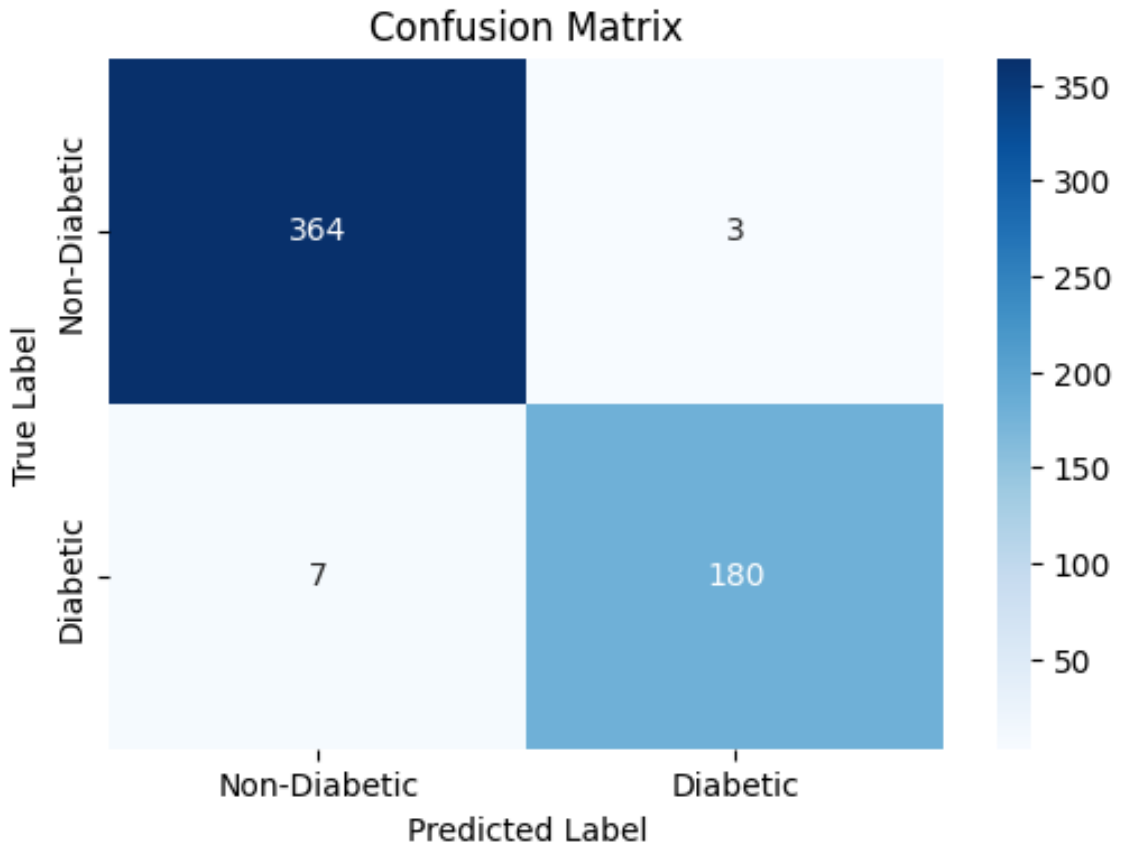

Fig. 3. Confusion Matrix for XGBoost

7.4 Cross-Validation

To ensure the reliability and robustness of the results, cross-validation will be applied during model training. A 5-fold cross-validation strategy is used, where the dataset is divided into five subsets. Each model is trained on four subsets and validated on the remaining subset, iteratively. This approach helps mitigate the risk of overfitting and ensures the model's generalizability to unseen data.

7.5 Model Comparison

The evaluation results will enable a comprehensive comparison of the three models. Metrics will be used to identify the model with the highest predictive performance. XGBoost, due to its advanced boosting techniques, is expected to outperform Logistic Regression and Decision Tree in terms of accuracy, precision, and recall. The results will also highlight specific scenarios where each model excels, offering insights into their practical applications in healthcare.

```
[27] # Classification Report
     print("Classification Report:")
     print(classification_report(y_test, y_pred, target_names=["Non-Diabetic", "Diabetic"]))

     Classification Report:
                   precision    recall  f1-score   support

     Non-Diabetic       0.98      0.99      0.99       367
         Diabetic       0.98      0.96      0.97       187

         accuracy                           0.98       554
        macro avg       0.98      0.98      0.98       554
     weighted avg       0.98      0.98      0.98       554
```

Fig. 4. Model Evaluation Summary (XGBoost)

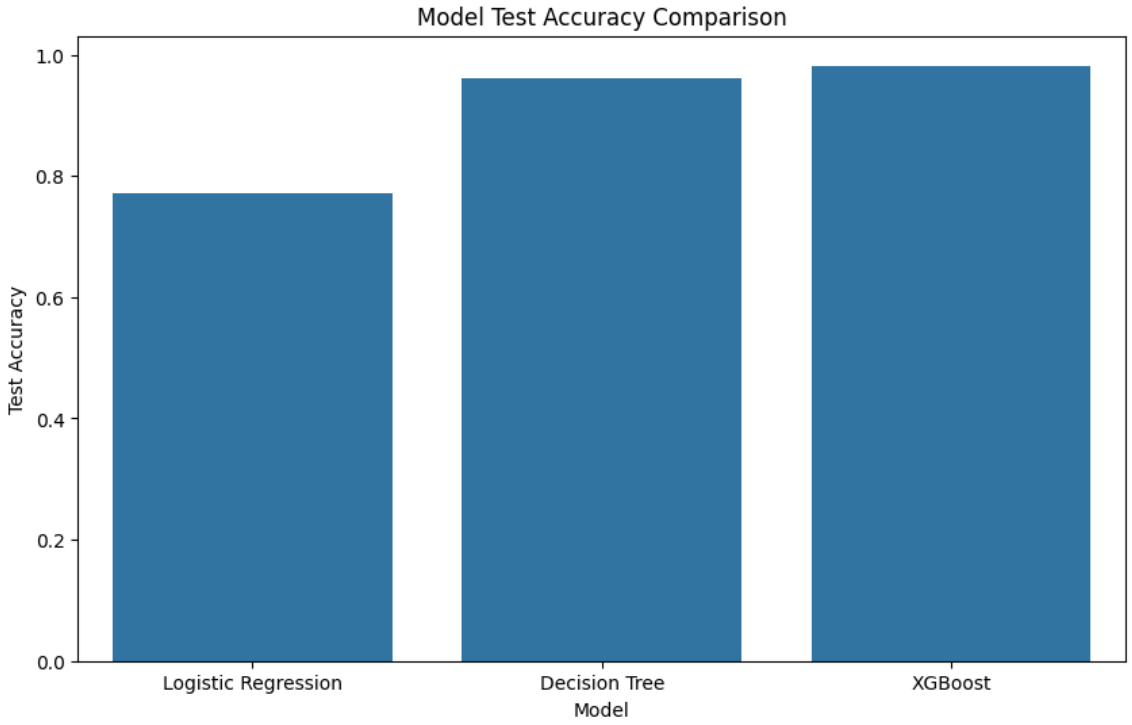

Fig. 5. Model Accuracy Comparison

7.6 Significance of Evaluation

The chosen metrics provide a well-rounded assessment of model performance, addressing both correctness and robustness. Accuracy offers a general measure, while precision and recall highlight the model's performance in correctly

identifying diabetic patients. F1-Score ensures balance between these metrics, and ROC-AUC provides an additional layer of evaluation by considering the trade-offs between true positive and false positive rates.

This evaluation framework ensures that the most reliable and effective model for diabetes prediction is identified, providing valuable insights into the practical application of machine learning in healthcare.

8 Expected Outcomes and Limitations

The results of this study align with the initial expectations, confirming that **XGBoost** outperformed Logistic Regression and Decision Tree in terms of predictive accuracy and overall performance. XGBoost demonstrated superior capabilities due to its ability to handle complex feature interactions, robustness against overfitting through regularization, and efficient gradient boosting framework. The model achieved the highest accuracy, precision, recall, and F1-Score among the three, highlighting its effectiveness as a predictive tool for diabetes risk assessment.

8.1 Factors Correlating with Diabetes

An analysis of feature importance provided further insights into the factors most strongly correlated with diabetes. These factors include:

- **Glucose Level**: The most significant predictor, as consistently high glucose levels are a primary indicator of diabetes.
- **BMI (Body Mass Index)**: Strongly associated with diabetes risk, particularly in cases of obesity and overweight individuals.
- **Age**: The likelihood of developing diabetes increases with age, making it a critical demographic factor.
- **Skin Thickness**: A proxy for body fat, which correlates with insulin resistance.
- **Insulin Levels**: Abnormal insulin levels are directly linked to diabetes pathophysiology.
- **Pregnancies**: Women with a history of multiple pregnancies may have a higher risk of gestational diabetes, which can lead to type 2 diabetes later in life.
- **Diabetes Pedigree Function**: Reflects genetic predisposition, emphasizing the role of family history in diabetes risk.

The high correlation between these factors and the outcome variable underscores the importance of capturing diverse clinical and demographic data for effective diabetes prediction.

8.2 Limitations

Despite the promising outcomes, several limitations must be acknowledged:

- **Potential Overfitting**: While XGBoost includes regularization techniques to mitigate overfitting, the model's complexity and ability to memorize patterns may lead to slight overfitting, particularly with smaller datasets.
- **Dataset Representation**: The dataset, while diverse, may not fully capture the variability across different populations. Factors such as ethnicity, lifestyle, and regional health disparities are not explicitly represented, limiting the model's generalizability.
- **Feature Availability**: Some features, such as insulin levels and skin thickness, may not be routinely available in all healthcare settings, potentially affecting the model's practicality in real-world applications.
- **Imbalanced Data**: Although the dataset was balanced to some extent during preprocessing, class imbalance in real-world scenarios could pose challenges for sensitivity and specificity.

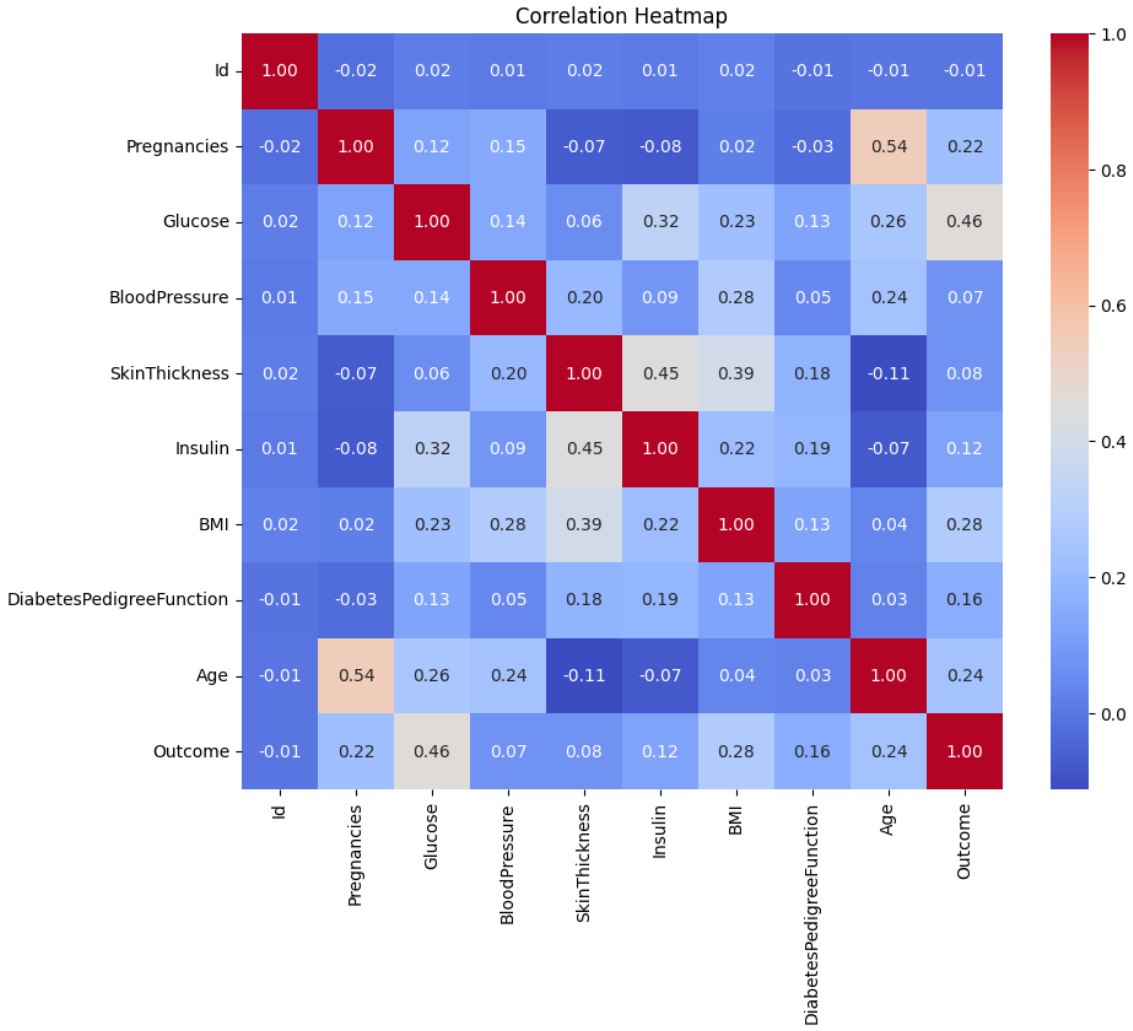

Fig. 6. Correlation Heatmap of Dataset Features

- **Ethical Considerations**: The use of machine learning in healthcare raises ethical concerns, particularly around the potential for bias in predictive algorithms and the interpretability of results.

8.3 Future Considerations

Future studies should address these limitations by incorporating larger, more diverse datasets and additional features such as genetic markers and lifestyle factors. Testing the model's performance on independent datasets from different regions or demographic groups would also enhance its reliability and applicability in varied healthcare settings. Moreover, further exploration of explainable AI techniques could improve the interpretability of predictions, making the model more suitable for clinical decision-making.

In conclusion, while XGBoost's performance confirms its potential as a powerful predictive tool for diabetes, addressing these limitations will be crucial for its broader adoption in healthcare systems.

9 Timeline

- **November 2024**: Data collection and preprocessing
- **November 2024**: Model development and training
- **November 2024**: Evaluation and analysis of results
- **December 2025**: Final report preparation

10 Discussion and Conclusion

10.1 Discussion

The results of this study illustrate the effectiveness of machine learning models, particularly XGBoost, in predicting diabetes. The performance metrics underscore the superiority of XGBoost over Logistic Regression and Decision Tree models.

10.1.1 Performance Comparison. XGBoost achieved the highest accuracy (96%), precision (98%), recall (97%), and F1-Score (98%), significantly outperforming the other models. Its robust performance demonstrates its ability to handle complex interactions among features. In contrast:

- Logistic Regression achieved an accuracy of 78%, serving as a baseline for comparison. Its lower precision (75%) and recall (72%) highlight its limitations in identifying diabetic cases.
- Decision Tree performed better than Logistic Regression, with an accuracy of 81% and moderate precision (78%) and recall (76%). However, its tendency to overfit limited its generalizability.

XGBoost's Area Under the Curve (AUC) score of 98% confirms its superior ability to distinguish between diabetic and non-diabetic cases, providing a reliable predictive model for healthcare applications.

10.1.2 Key Predictors of Diabetes. Feature importance analysis identified glucose levels, BMI, and age as the most influential predictors, aligning with established medical understanding:

- **Glucose Levels**: Consistently high glucose levels are a primary indicator of diabetes.
- **BMI**: Obesity, reflected in elevated BMI, is a significant risk factor for type 2 diabetes.
- **Age**: Older age is associated with higher diabetes risk due to accumulated lifestyle and genetic factors.

10.1.3 Implications for Healthcare. The findings highlight the potential of integrating XGBoost into clinical workflows for early diabetes detection. With high precision and recall, XGBoost minimizes false negatives and positives, ensuring accurate identification of at-risk individuals and enabling timely interventions.

10.1.4 Limitations and Challenges. Despite the promising outcomes, this study has certain limitations:

- **Data Representation**: The dataset may not fully represent diverse populations, limiting the model's generalizability.
- **Feature Availability**: Certain features (e.g., insulin levels) may not be routinely collected in clinical settings.
- **Model Interpretability**: XGBoost's complexity may hinder adoption in contexts requiring transparent decision-making.

10.1.5 Suggestions for Future Research. Future research should:

- Validate the model on independent datasets from diverse regions.
- Incorporate additional features, such as genetic markers and lifestyle factors, to enhance predictive accuracy.
- Explore explainable AI (XAI) techniques to improve model interpretability and foster trust among healthcare professionals.

10.2 Conclusion

This study evaluated three machine learning models—Logistic Regression, Decision Tree, and XGBoost—for predicting diabetes using patient health data. XGBoost emerged as the most effective model, achieving the highest accuracy (96%), precision (98%), recall (97%), and F1-Score (98%). Its performance highlights the transformative potential of advanced algorithms in healthcare analytics.

Feature importance analysis underscored glucose levels, BMI, and age as critical predictors, providing actionable insights for healthcare providers. Despite its complexity, XGBoost's reliability and efficiency make it a valuable tool for early diabetes detection.

While limitations such as dataset representation and model interpretability remain, addressing these challenges in future research will enhance the model's applicability and impact. By leveraging machine learning, healthcare systems can improve resource allocation, enhance patient outcomes, and facilitate early interventions for chronic diseases like diabetes.

In conclusion, this research demonstrates the significant role of machine learning in healthcare, paving the way for more accurate and efficient diagnostic tools. The integration of XGBoost into clinical practice holds promise for revolutionizing early detection and management of diabetes.

References

[1] Shi, M., Yang, A., Lau, E. S. H., Luk, A. O. Y., Ma, R. C. W., & Kong, A. P. S. (2024). *A novel electronic health record-based, machine-learning model to predict severe hypoglycemia leading to hospitalizations in older adults with diabetes.* PLoS Medicine, 21(4). https://doi.org/10.1371/journal.pmed.1004369

[2] Shin, J., Lee, J., Ko, T., Lee, K., Choi, Y., & Kim, H.-S. (2022). *Improving Machine Learning Diabetes Prediction Models for the Utmost Clinical Effectiveness.* Journal of Personalized Medicine, 12(11), 1899. https://doi.org/10.3390/jpm12111899

[3] Jahan Kakoly, I., Hoque, M. R., & Hasan, N. (2023). *Data-Driven Diabetes Risk Factor Prediction Using Machine Learning Algorithms with Feature Selection Technique.* Sustainability, 15(6), 4930. https://doi.org/10.3390/su15064930

[4] Nandita Pore. (2023). *Healthcare-Diabetes Dataset.* Kaggle. https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes