# THE ROLE OF MACHINE LEARNING IN PREDICTING DIABETES

**SALEH BABAEI**
**SENECA POLYTECHNIC**

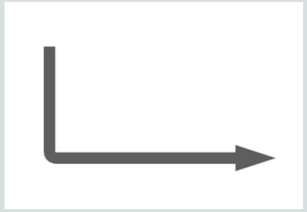
This research evaluates the effectiveness of machine learning models—Logistic Regression, Decision Tree, and XGBoost—in predicting diabetes. By analyzing patient data, including glucose, BMI, and age, the study identifies key risk factors and highlights how ML can improve early detection and intervention.

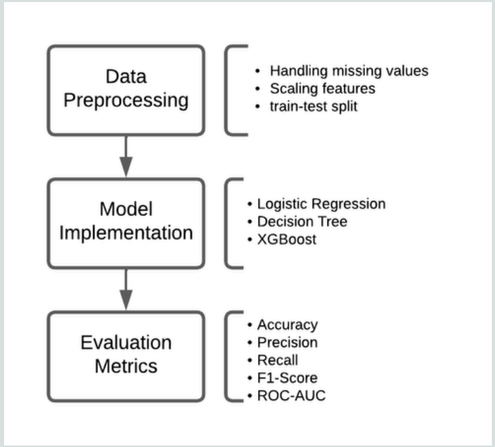
## INTRODUCTION

This study explores how machine learning can help predict diabetes early and identify key factors like glucose levels, BMI, and age. By comparing three models—Logistic Regression, Decision Tree, and XGBoost—this research shows that XGBoost is the most accurate, achieving a **98% success rate**. The findings highlight the potential of machine learning to support healthcare professionals in early detection and personalized interventions, offering a practical solution to tackle one of the world's most prevalent chronic diseases.

## METHODOLOGY

Three models were evaluated using key performance metrics on a dataset of ~3,000 records. Preprocessing included imputing missing values and standardizing features.

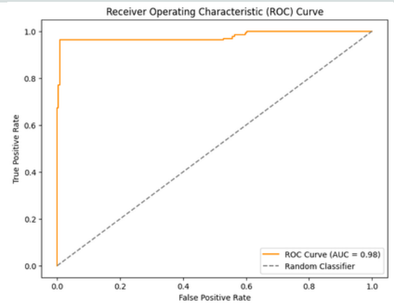





## RESULTS

**XGBoost** achieved the highest accuracy (**98.19%**) and demonstrated superior performance in all metrics. Glucose, BMI, and Age were identified as the top predictors of diabetes.

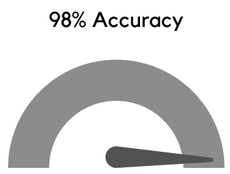
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Non-Diabetic | 0.98 | 0.99 | 0.99 | 367 |
| Diabetic | 0.98 | 0.96 | 0.97 | 187 |
| accuracy | | | 0.98 | 554 |
| macro avg | 0.98 | 0.98 | 0.98 | 554 |
| weighted avg | 0.98 | 0.98 | 0.98 | 554 |










## FINDINGS

- Glucose levels and BMI are critical, modifiable predictors of diabetes.
- Dietary changes like ketogenic diets and regular exercise can mitigate risks.
- XGBoost is a reliable tool for predictive healthcare applications.
- Future work should focus on more diverse datasets and explainable AI techniques.

## CONCLUSION

- Glucose, BMI, and age are key predictors, highlighting the importance of lifestyle interventions like low-carb diets and exercise.
- XGBoost proved to be the most reliable model for accurate diabetes predictions.
- Future work should focus on diverse datasets and explainable AI for better clinical adoption.
- These findings can guide healthcare professionals in early detection and personalized treatment strategies.

### AUTHORS

Saleh Babaei - 121183206
Supervisor: Behshid Behlamal

### AFFILIATIONS

Seneca Polytechnic
Class of BTM 710

### RELATED LITERATURE/REFERENCE

- 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 in older adults with diabetes. PLoS Medicine, 21(4). https://doi.org/10.1371/journal.pmed.1004369
- 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
- Jahan Kakoly, I., Hoque, M. R., & Hasan, N. (2023). Data-driven diabetes risk factor prediction using machine learning algorithms with feature selection techniques. Sustainability, 15(6), 4930. https://doi.org/10.3390/su15064930
- Pore, N. (2023). Healthcare-Diabetes Dataset. Kaggle. https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes