Diabetes Prediction Using Machine Learning

Naseefah Chowdhury [1]

[1]Morehouse College | Center for Broadening Participation in Computing

Abstract

Diabetes often develops quietly, and many people are not diagnosed until serious problems arise. I used a public dataset of about 100,000 patient records with age, weight, blood sugar, and lifestyle factors. The goal was to build a model that predicts who is at higher risk for diabetes using this basic health information. My final model correctly identified over 80% of diabetic patients while keeping overall accuracy high. This shows that simple health data can power early-warning tools to support doctors and help patients take action sooner.

Introduction

Diabetes is one of the most common chronic diseases worldwide, yet many people remain undiagnosed until complications appear. Early detection is critical, since lifestyle changes and timely treatment can prevent long-term damage. With millions at risk, doctors need simple tools to quickly flag patients who may need closer monitoring. My project set out to answer: **Can I use basic health and lifestyle information to predict diabetes risk before it is formally diagnosed?**

Dataset and Features

- **Source:** Kaggle Diabetes Prediction dataset.
- **Size:** ~100,000 patient records after cleaning.
- **Contents:** demographic info (age, gender), medical history (hypertension, heart disease), lifestyle (smoking history), and lab results (BMI, HbA1c, blood glucose).
- **Cleaning:** removed duplicates, unrealistic ages, and extreme lab values; encoded categories like gender and smoking history as numbers.
- **Tools:** Python, pandas = cleaning, scikit-learn = preprocessing, and XGBoost = modeling.

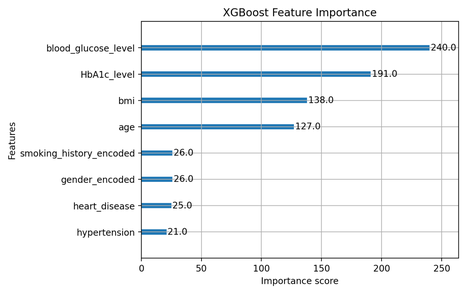

Figure 1. Feature Importances – Glucose, HbA1c, and BMI were the strongest predictors.

Methods

1. **Algorithm:** Used XGBoost, a tree-based model balancing accuracy and interpretability.
2. **Training/Testing:** Data split 80/20 with stratification to keep class balance.
3. **Validation:** Tuned hyperparameters for ROC-AUC; adjusted threshold to prioritize recall.
4. **Metrics:** Evaluated with recall, precision, F1-score, and AUC.
5. **Key Choice:** I lowered the classification threshold (0.21) so the model caught more true diabetes cases, even if that meant a few more false alarms.

Results: Confusion Matrix

- Recall = 0.81, Precision = 0.74 at threshold 0.21.
- **Recall means:** of all true diabetic patients, the model correctly identified 81%.
- Most diabetic cases detected, though some were missed.
- Reflects healthcare tradeoff: prioritize recall to reduce false negatives.

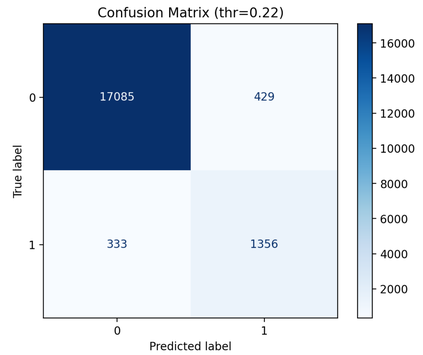

Figure 2. Confusion Matrix – Recall (81%) means most true diabetic cases were caught; Precision (74%) shows that most flagged cases were correct.

Results: ROC Curve

The ROC curve demonstrates excellent discrimination with AUC = 0.98. This means the model is highly effective at separating diabetic and non-diabetic patients, showing strong potential for early intervention.

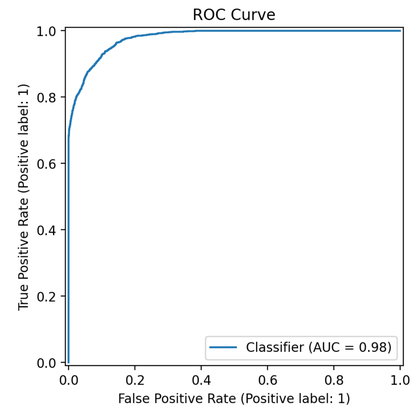

Figure 3. ROC Curve – AUC of 0.98 shows excellent ability to separate diabetic vs. non-diabetic patients across thresholds.

Ethical Considerations

Machine learning in healthcare carries risks if not applied carefully.

- **Bias:** Underrepresented groups (e.g., women, younger patients) could face reduced accuracy.
- **Recall priority:** In this context, recall means the ability to catch as many true diabetes cases as possible. Missing fewer patients was considered more important than reducing false alarms.
- **Privacy:** All data were de-identified, but real-world use would require strong protections.

These considerations shaped my evaluation strategy and threshold tuning.

Future Work

Next steps to strengthen this project:

- **More data:** Add family history, medications, and lifestyle factors to boost accuracy.
- **Stronger models:** Explore deeper architectures or ensembles beyond XGBoost.
- **Limits:** Current dataset is noisy and compute was limited — future work should scale up.
- **Opportunities:** Refining the model opens the door for integration into routine screenings and preventative care.

Conclusion

The goal of this project was to predict diabetes risk using simple health and lifestyle data. I built and tuned an XGBoost model, carefully adjusting the threshold to prioritize recall so fewer true cases were missed. The final model achieved strong performance, identifying over 80% of diabetic patients while maintaining high overall accuracy.

Takeaway: Predictive models are not a replacement for doctors, but they can serve as an early-warning tool that flags hidden risks and helps providers act sooner.

Acknowledgments

I am deeply grateful to my mentors at Morehouse College and the Center for Broadening Participation in Computing for their guidance, patience, and encouragement throughout this project. Their support helped me refine my technical skills while also growing in confidence as a professional.

I also want to recognize my peers in the program — your encouragement, insights, and shared energy made this journey meaningful.

This experience has not only advanced my knowledge of machine learning, but has also reinforced my commitment to using technology for positive impact in communities.

References

1. Mohammed Mustafa. *Diabetes Prediction Dataset*. Kaggle. `https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset`
2. Morehouse College, Center for Broadening Participation in Computing.