

Diabetes Risk Prediction using Machine Learning

Shubham Aggarwal Atharva Sharma Bhavya Punj

Milestone 1 Project

February 28, 2026

1 Introduction

Early detection of diabetes is critical in preventive healthcare. This project aims to develop a machine learning-based system to predict diabetes risk using health indicators. The objective is to build a reliable classification model that balances predictive performance while minimizing both false positives and false negatives.

2 Dataset Description

The dataset used is the **Diabetes Health Indicators Dataset**.

It is derived from the **BRFSS 2015 survey** conducted by the CDC and contains approximately **253,680 records** with 21 features and one binary target variable.

Target Variable:

- 0: Non-diabetic
- 1: Diabetic or pre-diabetic

3 Feature Encoding

Several features in the dataset are encoded categorical variables:

Age: Categorized into 13 groups:

- 1: 18–24
- 2: 25–29
- 3: 30–34
- 4: 35–39
- 5: 40–44
- 6: 45–49
- 7: 50–54

- 8: 55–59
- 9: 60–64
- 10: 65–69
- 11: 70–74
- 12: 75–79
- 13: 80+

Education (Grade Level):

- 1: No formal education (Kindergarten or below)
- 2: Grades 1–8 (Elementary)
- 3: Grades 9–11 (Some high school)
- 4: Grade 12 / GED (High school graduate)
- 5: Grades 13–15 (Undergraduate / College 1–3 years)
- 6: Grade 16+ (College graduate or higher)

Income:

- 1: Less than \$10,000
- 2: \$10,000 – \$15,000
- 3: \$15,000 – \$20,000
- 4: \$20,000 – \$25,000
- 5: \$25,000 – \$35,000
- 6: \$35,000 – \$50,000
- 7: \$50,000 – \$75,000
- 8: \$75,000 or more

General Health (GenHlth):

- 1: Excellent
- 2: Very good
- 3: Good
- 4: Fair
- 5: Poor

These encodings preserve ordinal relationships and improve model learning.

4 Data Preprocessing

The following preprocessing steps were applied:

- Removal of redundant features: Stroke, HeartDiseaseorAttack
- Removal of duplicate records
- Feature scaling using StandardScaler
- Handling class imbalance using SMOTE

5 Model Development

Four models were implemented:

- Logistic Regression (LR)
- Random Forest (RF)
- XGBoost (XGB)
- Artificial Neural Network (ANN)

All models were trained under consistent conditions using SMOTE, hyperparameter tuning, and threshold optimization.

6 Model Evaluation

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.765	0.356	0.654	0.461
Random Forest	0.779	0.369	0.613	0.460
XGBoost	0.786	0.378	0.609	0.466
ANN	0.782	0.374	0.622	0.467

Table 1: Performance Comparison of Models

All models show comparable performance. ANN achieved the highest F1-score, while XGBoost achieved the highest AUC.

7 Confusion Matrix Analysis

Observations:

- Logistic Regression produces fewer false negatives
- ANN achieves slightly better balance but misses more diabetic cases
- Logistic Regression is more reliable for healthcare screening

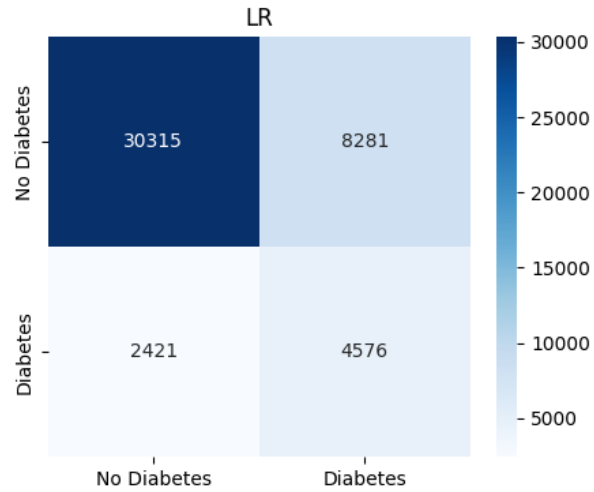


Figure 1: Confusion Matrix for Logistic Regression

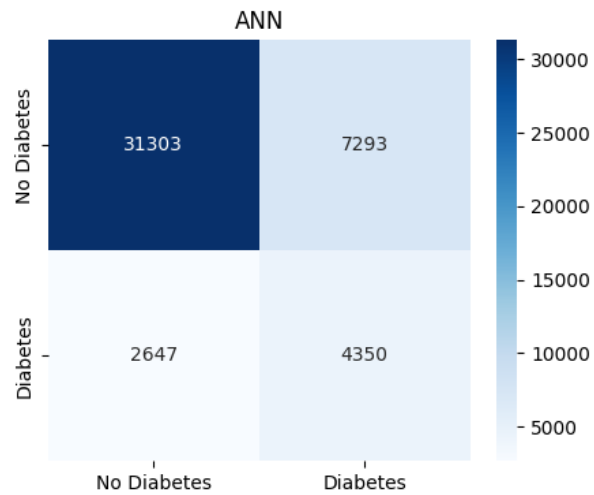


Figure 2: Confusion Matrix for ANN

8 Final Model Selection

Logistic Regression was selected as the final model due to:

- Higher Recall (0.654)
- Lower False Negatives
- Better suitability for healthcare applications

In healthcare, minimizing false negatives is critical as missing a diabetic patient can have serious consequences.

9 Final Model Pipeline

- StandardScaler

- SMOTE
- Logistic Regression ($C = 0.01$)
- Threshold = 0.58

The model was retrained on the full dataset for deployment.

10 Deployment

The model is deployed using a Streamlit application:

Live Application Link

The application:

- Collects user health data
- Predicts diabetes probability
- Displays risk classification

11 Project Resources

- **Stage 1:** Notebook
- **Stage 2:** Notebook
- **Final Model:** Notebook

12 Conclusion

This project demonstrates a complete machine learning workflow for diabetes prediction. Logistic Regression was selected due to its strong recall and lower false negative rate, making it more suitable for healthcare applications.

13 Future Work

- Integration with AI assistants
- Model explainability (SHAP/LIME)
- Real-time monitoring systems

References

- [1] A. Teboul, *Diabetes Health Indicators Dataset*, Kaggle. Link