EduLinkUp Capstone Project Report

Submitted By: Tanuja Nalage

Date: 4th February 2026

Diabetes Risk Assessment using Machine Learning

⚠ Medical Disclaimer

This project is developed strictly for **educational purposes only**. It is not intended to diagnose, treat, cure, or prevent diabetes. The predictions and insights generated by this system should not be used as a substitute for professional medical advice. Always consult a qualified healthcare professional for medical decisions.

1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels. According to global health studies, the prevalence of diabetes is increasing rapidly due to sedentary lifestyles, unhealthy diets, and genetic factors. Early detection of diabetes is critical to prevent severe complications such as cardiovascular disease, kidney failure, nerve damage, and vision loss.

Machine learning provides effective tools to analyze medical data and identify hidden risk patterns. This project aims to build a **diabetes risk assessment system** using machine learning models that can assist in early-stage screening by analyzing patient health indicators.

2. Problem Statement

Traditional diabetes diagnosis requires laboratory tests and clinical evaluations, which may not always be accessible for early screening. The challenge is to design a data-driven system that:

- Analyzes patient health metrics
- Predicts diabetes risk accurately
- Minimizes false negatives
- Supports preventive healthcare decisions

3. Dataset Description

The project uses the **Pima Indians Diabetes Dataset**, a widely used benchmark dataset for diabetes prediction tasks.

- Total Records: 768
- Number of Features: 8 input features + 1 target variable
- Target Variable: Outcome
    - 1 → Diabetic
    - 0 → Non-Diabetic

Key Attributes

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI
- Diabetes Pedigree Function
- Age

4. Data Preprocessing

Medical datasets often contain missing or invalid values. In this dataset, zero values in certain features are medically impossible and represent missing data.

Steps Performed:

- Identified zero values in Glucose, Blood Pressure, Insulin, BMI, and Skin Thickness
- Treated zero values as missing values
- Applied **median imputation**, which is suitable for healthcare data as it is robust to outliers

This step improves data quality and model reliability.

5. Feature Engineering

Feature engineering was performed to convert raw numerical data into medically meaningful representations.

Engineered Features:

- **BMI Category**: Underweight, Normal, Overweight, Obese
- **Age Group**: Young, Adult, Middle-Aged, Senior
- **Glucose Level Category**: Normal, Prediabetes, Diabetes

Categorical features were encoded using **one-hot encoding** to make them suitable for machine learning algorithms.

6. Machine Learning Models

Two classification algorithms were implemented and compared:

6.1 Logistic Regression

Logistic Regression is a widely used linear classification algorithm in medical applications. It provides interpretable results and balanced performance across evaluation metrics.

6.2 Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem. Despite its simplicity, it performs well on structured medical datasets and is particularly effective in identifying high-risk cases.

7. Model Evaluation Strategy

The models were evaluated using the following metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

Special emphasis was placed on **recall**, as missing a diabetic patient (false negative) can have serious medical consequences.

8. Results and Comparison

- Logistic Regression achieved balanced accuracy and stable predictions
- Naive Bayes achieved very high recall for diabetic patients

Although Naive Bayes showed lower overall accuracy, it minimized false negatives, making it suitable for early-stage diabetes risk screening.

9. Importance of Recall in Medical AI

In healthcare applications, recall is often more important than accuracy. A false negative means a patient with diabetes is classified as healthy, delaying diagnosis and treatment.

Therefore, models with higher recall are preferred for screening purposes, even if they have slightly lower accuracy.

10. Health Insights Derived

- High glucose levels are the strongest predictor of diabetes
- Obesity and high BMI significantly increase diabetes risk
- Age above 35 increases susceptibility, especially when combined with high BMI
- Insulin abnormalities strongly correlate with diabetic outcomes
- Multiple moderate risk factors together can result in high overall risk

These insights highlight the value of machine learning in preventive healthcare.

11. Ethical and Legal Considerations

- The model is not a replacement for medical professionals
- Predictions should be used only for educational and screening purposes
- Patient data privacy and responsible AI use are critical
- Clear medical disclaimers are essential when deploying healthcare AI systems

12. Limitations

- Dataset size is limited
- Model trained on a specific population group
- Real-world clinical deployment requires validation with diverse datasets

13. Conclusion

This project demonstrates how machine learning can be applied to healthcare for diabetes risk assessment. By focusing on recall and medically meaningful features, the system supports early-stage screening and preventive care.

Naive Bayes emerges as a suitable choice for early risk detection due to its high recall, while Logistic Regression provides balanced and interpretable predictions.

14. Future Scope

- Inclusion of additional clinical parameters
- Testing on larger and more diverse datasets
- Deployment as a web-based screening tool
- Integration with electronic health record systems