Diabetes Prediction

Submitted by:

Aarohi Keshari(102217165)
Ashutosh Kumar Swarnakar(102217263)

BE Third Year

CSE

Submitted to:

Dr. Anjula Mehto

Assistant Professor

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering

Department Thapar Institute of Engineering

and Technology, Patiala

November 2024

TABLE OF CONTENTS

Introduction

Diabetes is a common health problem that happens when the body cannot control blood sugar levels properly. If left untreated, it can lead to serious issues like heart disease, kidney problems, and even blindness. Detecting diabetes early is very important because it helps people manage the condition better and avoid complications.

With the growing use of technology in healthcare, machine learning (ML) has become a powerful tool for solving medical problems. ML uses computers to find patterns in data and make predictions. In this project, we use machine learning to predict whether a person has diabetes based on their health information, such as blood sugar levels, age, and body weight.

The goal of this project is to create a simple and accurate system that can help doctors identify people who are at risk of diabetes. By doing this, we hope to make early diagnosis easier and more effective.

Problem Statement

Traditional methods of diagnosing diabetes often require extensive medical testing, which may not always be accessible or affordable for everyone. There is a need for a quick, accurate, and cost-effective way to predict the likelihood of diabetes using readily available health information.

This project aims to address this challenge by developing a machine learning-based model that can predict whether an individual is diabetic or not based on key health indicators. The system will analyze features such as glucose levels, BMI, age, and family history to make predictions. The ultimate goal is to create a reliable tool that can assist healthcare providers in identifying high-risk individuals, enabling timely diagnosis and treatment to improve health outcomes.

Overview of the Dataset used

1. Dataset Description

Objective: To classify individuals as diabetic (1) or non-diabetic (0) based on specific health metrics.

Type: Tabular data with numerical and categorical attributes.

Outcome: Binary classification label indicating whether a person is diabetic (1) or non-diabetic (0).

2. Features

The dataset includes the following health-related features:

Features	Description
Pregnancies	Number of pregnancies (relevant for females).
Glucose	Plasma glucose concentration(mg/dL) measured during an oral glucose tolerance test.

BloodPressure	Diastolic blood pressure(mm Hg).
SkinThickness	Triceps skinfold thickness(mm).
Insulin	2-hour serum insulin
BMI	Body Mass Index(weight in Kg/height in m^2)
DiabetesPdeigreeFunction	A score indicating the likelihood of diabetes based on family history.
Age	Age of individual(years)

3. Dataset Characteristics

Balanced or Imbalanced: The dataset may have imbalanced classes, where the number of diabetic cases is lower than non-diabetic cases.

Data Distribution: The data may require normalization or standardization to improve model performance.

4. Dataset Preprocessing Steps

Feature Scaling: Normalize or standardize features like Glucose and BMI for algorithms sensitive to scaling.

Data Splitting: Divide the dataset into training and testing sets (e.g., 70-30 split) to evaluate model performance.

5. Usage

This dataset is well-suited for binary classification problems and helps evaluate the performance of machine learning models such as logistic regression, decision trees, random forests, and neural networks.

Project Workflow

1.Problem Understanding

Define the project goal: Predict whether a person is diabetic or not based on health metrics.

Identify the target audience: Healthcare professionals or patients.

Understand the dataset and its features.

2.Data Collection

Source: Obtain the dataset (e.g., Kaggle).

Format: Ensure the dataset is in a structured format, such as a CSV file.

Attributes: Understand the features (e.g., glucose levels, BMI, pregnancies, etc.) and the target variable (diabetic or not).

3.Data Preprocessing

Handle missing or incorrect data:

Replace zeros in features like Glucose, BMI, and Insulin with appropriate values (mean/median).

Perform exploratory data analysis (EDA):

Visualize distributions, correlations, and trends using histograms, pair plots, or heatmaps.

Encode categorical data (if applicable).

Normalize or standardize numerical features for consistent scaling.

Split the data:

Training set (70-80%) and testing set (20-30%).

Results

The results of the diabetes prediction project are summarized based on the performance of the machine learning model, insights from the dataset, and the achievement of the project's goals. Below is an outline of the key outcomes:

1. Model Performance

Algorithm Used: Logistic Regression

Accuracy: The model achieved an accuracy of 78% .

2. Key Insights

Important Features:

Features like Glucose, BMI, and Age were identified as the most influential factors in predicting diabetes.

Pregnancies and Diabetes Pedigree Function also contributed significantly.

Data Trends:

Higher glucose levels and BMI were strongly correlated with the likelihood of diabetes.

Age showed a positive association with diabetes risk.

3. Achievements

Successfully built a machine learning model capable of predicting diabetes with high accuracy and reliability.

Provided insights into the health metrics most associated with diabetes risk.

Demonstrated the use of machine learning in healthcare for early detection.

4. Limitations

Model performance might vary with different datasets or populations.

5. Future Scope

Improve the model by collecting larger and more diverse datasets.

Incorporate additional features, such as lifestyle habits, diet, or physical activity levels.

Conclusion

This project underscores the potential of machine learning in transforming healthcare by enabling early detection and personalized care. With further development, the model could be deployed as a practical tool to assist healthcare professionals in improving patient outcomes and combating the growing prevalence of diabetes.

Github link: https://github.com/swarnakar06/ML_project