
Diabetes Prediction Project Report

Title Page

Project Title: Diabetes Prediction Using Machine Learning
Course: Machine Learning
Team Members:

- Ali Shahat
- Kareem Omar
- Abdelrahman Kadry
- Mahmoud Refaat

---

1. Introduction

The goal of this project is to build a simple machine learning model that predicts whether a person has diabetes based on several medical features such as glucose level, BMI, blood pressure, and age.
This problem is important because early prediction can help reduce health risks and allow early medical intervention.

We tested two basic machine learning algorithms:

- **K-Nearest Neighbors (KNN)**
- **Logistic Regression**

These algorithms were selected because they are easy to implement, simple to understand, and commonly used for classification tasks.

1.1 Define Task

The task is a **binary classification problem**, where the model predicts:

- **0 → No diabetes**
- **1 → Has diabetes**

1.2 Dataset Description

The dataset contains **1000 rows** and **9 columns**.
Features include:

- Pregnancies
- Glucose
- BloodPressure
- SkinThickness
- Insulin
- BMI
- DiabetesPedigreeFunction
- Age
- Target label: Diagnosis (0 or 1)

The dataset had **no missing values**, but it contained some duplicates which were removed.

---

2. Methodology

The main steps in the project included:

- Loading and cleaning the data
- Scaling the features
- Splitting the dataset
- Training two models
- Evaluating accuracy and performance using confusion matrices

2.1 Algorithms Used

K-Nearest Neighbors (KNN)

- Works by checking the nearest data points.
- We used **K = 7**.
- Sensitive to feature scaling.

Logistic Regression

- A simple linear classifier.
- Good for binary prediction.
- `max_iter` increased to ensure convergence.

2.2 Data Preprocessing

- Checked missing values
- Filled missing data (if existed) using median
- Removed duplicate rows
- Scaled features using **StandardScaler**

2.3 Data Splitting

The dataset was split into:

- **80% training**
- **20% testing**

---

3. Results and Technical Discussion

3.1 Results

Model	Accuracy
KNN Classifier	**62%**
Logistic Regression	**69.5%**

3.2 Result Analysis

- Logistic Regression performed better than KNN.
- Logistic Regression predicted class 0 well but failed to predict class 1 due to imbalance in the dataset.

3.3 Error Analysis

- Class 1 (diabetes) accuracy was low.
- The dataset is imbalanced: more people without diabetes than with it.
- This caused models to favor predicting class 0.

---

4. Conclusions

In this project, I learned how to clean data, scale it, split it, and apply two basic machine learning algorithms.
The results were not very high because of class imbalance and simple models, but the project helped me understand how preprocessing and algorithm choice affect performance.

---

5. Appendix

5.1 Source Code

5.2 GitHub Repository Link

The full project is uploaded to GitHub:

🔗 **https://github.com/alycoder19/Diabetes-Prediction.git**
