G H Patel College of Engineering & Technology

(The Charutar Vidya Mandal (CVM) University)

Vallabh Vidyanagar

COMPUTER ENGINEERING DEPARTMENT

Mini Project Report

On

Diabetes Prediction

Submitted By:

Name of Student: Mo.Jisan Shaikh(12202040501042)
Nainam Hodiwala(12202040501044)

Artificial Intelligence and Machine Learning
(202046702)
A.Y. 2024-25 EVEN TERM

Diabetes Prediction Using Logical Regression

Objective

The aim of this project is to predict whether a patient is diabetic or not based on various health-related attributes. Early prediction can lead to timely medical intervention and improved patient outcomes.

Dataset Used

The dataset contains multiple health indicators relevant to diabetes diagnosis. Key features include:

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

Target Variable: Outcome (0: Non-Diabetic, 1: Diabetic)

Model Chosen

To perform binary classification on the diabetes dataset, the following model was used:

Logistic Regression – Due to its simplicity, interpretability, and effectiveness for binary outcomes.

Methodology

- Exploratory Data Analysis (EDA): Understand distribution, correlation, and outliers in the dataset.
- Data Preprocessing: Handling missing values, feature scaling (Standardization), train-test split.
- Model Training: Logistic Regression model trained on the preprocessed dataset.
- Model Evaluation: Evaluated using standard classification metrics.

Performance Metrics

- Accuracy – Percentage of correct predictions over the total predictions.

- Precision – Proportion of positive identifications that were actually correct.
- Recall – Proportion of actual positives that were correctly identified.
- F1-Score – Harmonic mean of precision and recall.
- Confusion Matrix – To analyze true vs false classifications.

Model Evaluation Results

For Logistic Regression:

- Accuracy: 0.78
- Precision: 0.74
- Recall: 0.69
- F1-Score: 0.71

Challenges

- Dealing with imbalanced data in the dataset.
- Deciding threshold for classification beyond 0.5.
- Ensuring interpretability of the model for healthcare usage.

Learnings

- Understood logistic regression in-depth and its application to medical datasets.
- Gained practical skills in cleaning and scaling data to improve model performance.
- Learned how to interpret model coefficients and understand their impact on prediction.
- Evaluated model results using multiple metrics beyond accuracy.
- Developed a Streamlit app for easy and real-time diabetes prediction.

Conclusion

This project illustrates how logistic regression can effectively predict diabetes using basic health indicators. With good accuracy and interpretability, the model serves as a valuable tool in early diagnosis. The deployment via Streamlit enhances usability, enabling real-time predictions that can assist both patients and healthcare professionals in making informed decisions.