Diabetes Prediction using Classical Machine Learning

Project Type: Individual Capstone Project

Module: **Module III**

Submitted by: **Sushmitha S**

Submission Date: 10 January 2026

TABLE OF CONTENTS

PROJECT DOCUMENT

Diabetes Prediction using Classical Machine Learning

---

1. Problem Statement

Diabetes is a long-term health condition that can lead to serious complications if it is not identified and managed at an early stage. Many individuals remain unaware of their condition until symptoms become severe. Early prediction can support timely medical consultation and lifestyle adjustments.

The goal of this project is to build a machine learning–based system that predicts whether a person is likely to have diabetes using commonly available medical and demographic data. The project focuses on applying classical machine learning techniques and understanding model behavior rather than using overly complex methods.

---

2. Approach

This project follows a structured, data-driven approach to predict diabetes outcomes. The approach involves using a publicly available medical dataset, performing careful data preprocessing and exploratory analysis, and applying classical machine learning models. The focus is on building interpretable models and evaluating them using healthcare-relevant metrics rather than maximizing complexity.

3. Dataset Description

This project uses the PIMA Indians Diabetes Dataset, a widely used dataset in healthcare-related machine learning studies. It contains medical measurements collected from female patients along with a diabetes outcome label.

Dataset Overview:

- Total records: 768
- Number of input features: 8
- Target variable: Outcome
    - 0 indicates non-diabetic
    - 1 indicates diabetic

Features Included:

- Number of pregnancies
- Glucose level
- Blood pressure
- Skin thickness
- Insulin level
- Body Mass Index (BMI)
- Diabetes pedigree function
- Age

All features are numerical, making the dataset suitable for classical machine learning classification models.

4. Methodology

The project follows a clear and structured machine learning workflow to ensure clarity and reproducibility. The main steps include:

- Loading and inspecting the dataset
- Identifying and handling missing or invalid values
- Performing exploratory data analysis to understand patterns
- Scaling features for fair model training
- Splitting data into training and testing sets
- Building and evaluating machine learning models
- Comparing model performance and selecting the most suitable one

This approach ensures that each stage of the pipeline is well understood and logically connected.

---

5. Data Preprocessing

Although the dataset does not explicitly contain missing values, several medical features include zero values that are not realistic in a real-world medical context. For example, glucose level or BMI cannot be zero.

Preprocessing Steps Applied:

- Zero values in selected medical features were treated as missing data.
- Median imputation was used to replace missing values, as it is less affected by extreme values.
- Feature scaling was applied using standardization so that all features contribute equally during model training.

These preprocessing steps improved data quality and helped the models perform more reliably.

---

6. Exploratory Data Analysis (EDA)

Exploratory data analysis was carried out to better understand the data distribution and relationships between features and the target variable.

Key Observations:

- Glucose levels showed a clear difference between diabetic and non-diabetic cases.
- BMI and age also showed noticeable trends related to diabetes risk.
- The dataset showed a slight imbalance, with more non-diabetic cases.
- Correlation analysis indicated no strong multicollinearity among features.

EDA helped in identifying important predictors and validating assumptions before model building.

Feature Relationship Analysis

• To understand the relationships between different input features and their influence on diabetes prediction, a correlation analysis was performed. The correlation heatmap helps in identifying positively and negatively correlated features, which is useful for feature selection and model understanding.

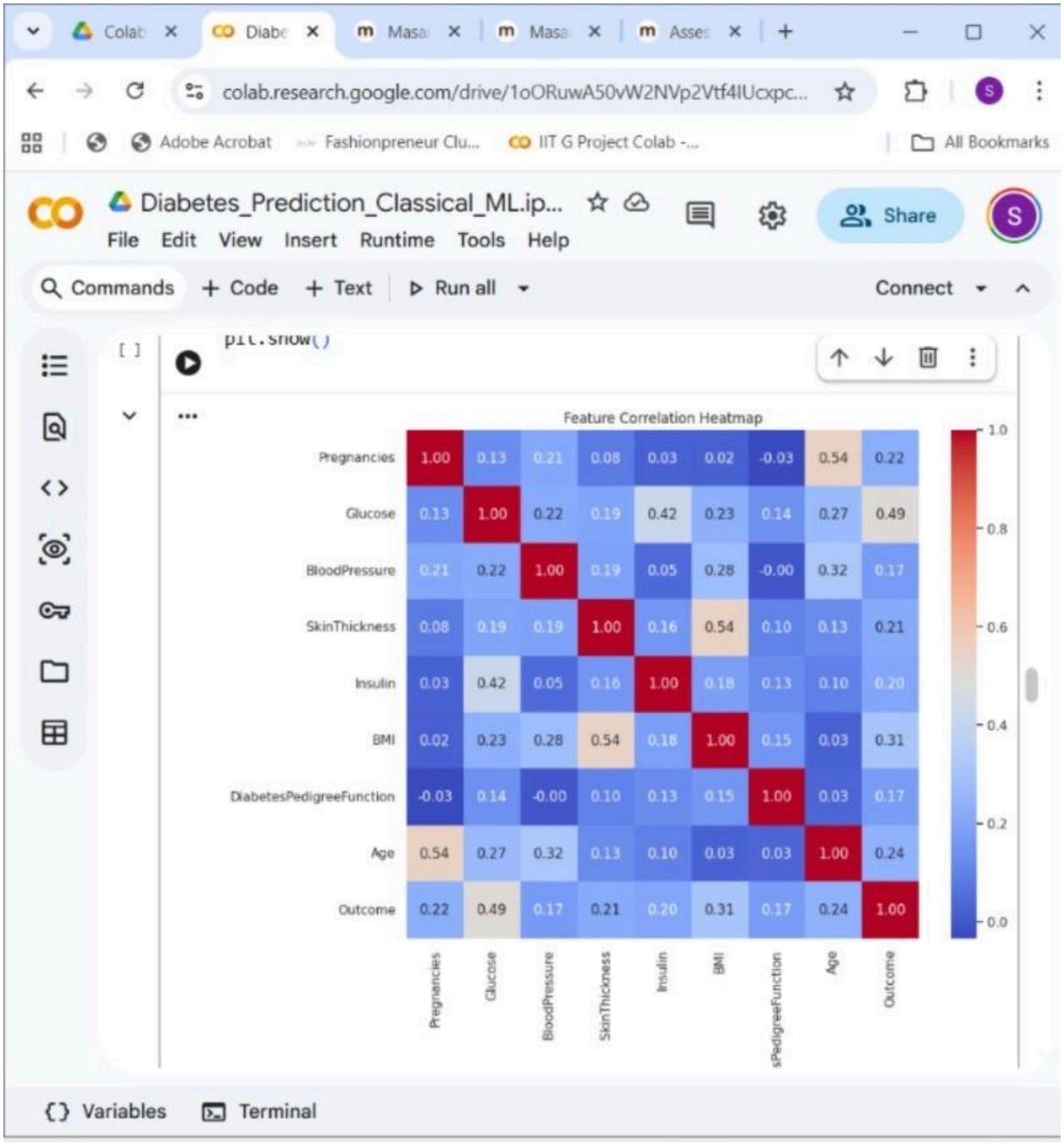

Figure 1: Correlation Heatmap showing relationships among clinical features and diabetes outcome.

7. Model Building

Two classical machine learning models were implemented and evaluated:

Logistic Regression

Logistic Regression was used as a baseline model due to its simplicity and interpretability. It is well-suited for binary classification problems and is commonly used in healthcare-related applications.

K-Nearest Neighbors (KNN)

KNN is a distance-based model that classifies data points based on similarity. It is sensitive to feature scaling, making it a good choice to evaluate the effectiveness of preprocessing steps.

Both models were trained on the same processed dataset to ensure a fair comparison.

Performance Evaluation of KNN Classifier

• The performance of the K-Nearest Neighbors (KNN) classifier was evaluated using a confusion matrix. This matrix provides a clear view of correct and incorrect predictions made by the model, helping assess its classification effectiveness on

unseen test data.

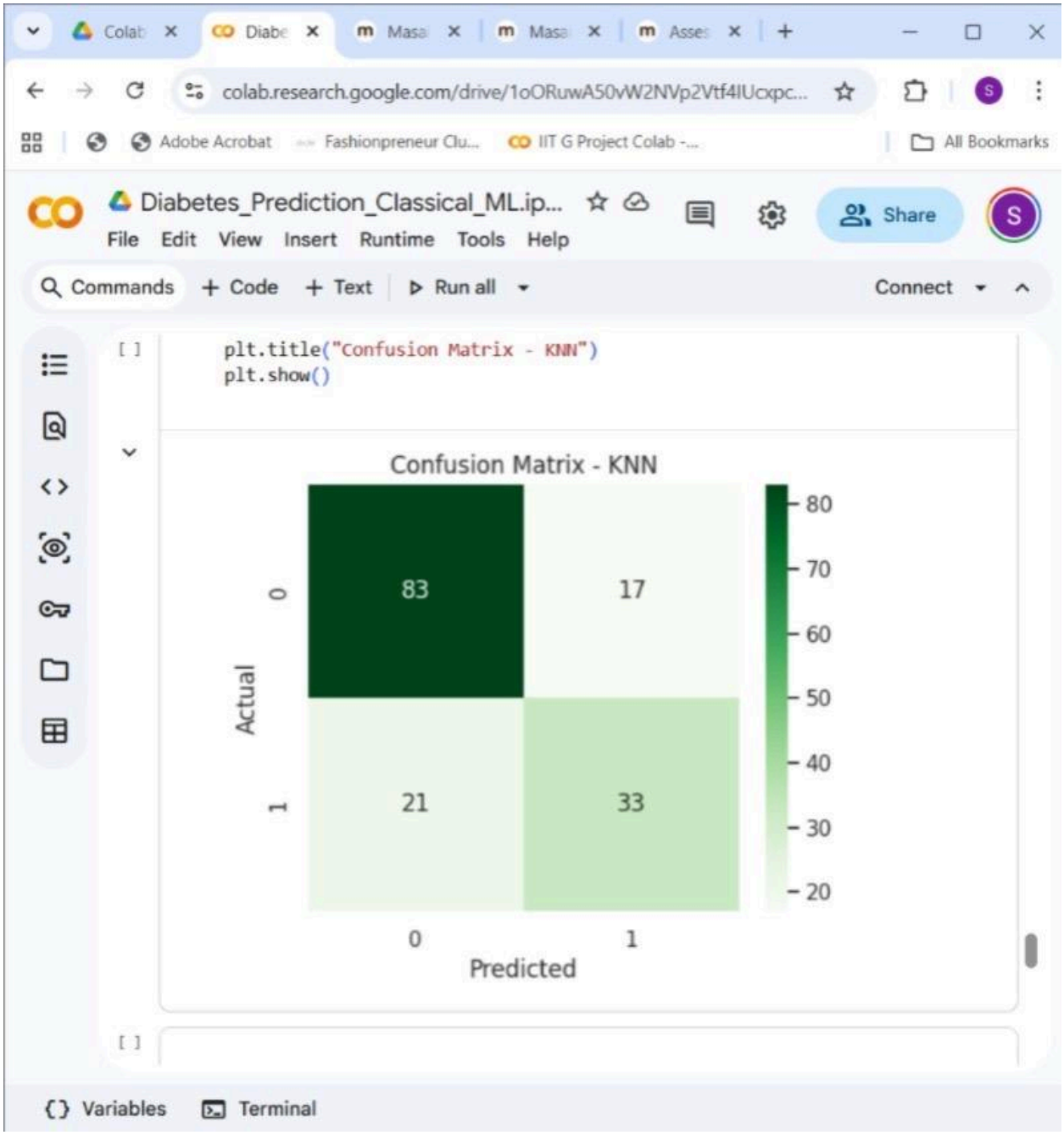

Figure 2: Confusion Matrix for KNN Model showing prediction performance.

8. Model Evaluation

Instead of relying only on accuracy, multiple evaluation metrics were used to assess model performance more meaningfully.

Evaluation Metrics Used:

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

In medical prediction tasks, recall is particularly important because failing to identify a diabetic patient can have serious consequences.

---

9. Results and Insights

Key Findings:

- Logistic Regression showed more stable and balanced performance compared to KNN.
- Logistic Regression achieved better recall, which is desirable in healthcare screening scenarios.
- KNN performance varied more and was influenced by data distribution and neighborhood selection.

Model Selection:

Based on overall performance and interpretability, Logistic Regression was selected as the final model.

10. Limitations

- The dataset is limited in size and represents a specific population group.
- Several features required imputation, which may affect prediction accuracy.
- The model has not been tested on real-time or longitudinal medical data.
- Only classical machine learning models were explored.

---

11. Future Scope

- Experiment with advanced models such as Random Forest or Gradient Boosting.
- Perform hyperparameter tuning to improve performance.
- Apply cross-validation for more reliable evaluation.
- Deploy the model as a simple web application for real-time use.
- Include additional health and lifestyle-related features if available.

---

12. Conclusion

This project demonstrates a complete machine learning workflow for predicting diabetes using classical algorithms. By focusing on proper data preprocessing, exploratory analysis, and meaningful evaluation metrics, the project highlights how interpretable models can support early disease

detection. The results show that even simple machine learning techniques can provide valuable insights when applied carefully.