**Diabetes Classification Using Logistic Regression**

**📌 Overview**

This project implements a modular and production-ready pipeline for predicting diabetes outcomes using logistic regression. It reads patient data, explores the dataset, preprocesses features, trains a model, and evaluates its performance.

Developed by **Samata Kalyan Kadam** on **10/08/2025**.

**File Structure**

| **File Name** | **Description** |
| --- | --- |
| diabetes.csv | Input dataset containing patient health metrics and diabetes outcome. |
| main.py | Main executable script containing modular functions for each step. |
| README.md | Documentation and usage instructions. |

**Requirements**

Ensure the following Python packages are installed:

pip install pandas numpy scikit-learn

**How to Run**

1. Place your dataset file named diabetes.csv in the same directory.
2. Run the script using:

python main.py

**Function Breakdown**

Each function in the script is modular and documented with headers:

* read\_data(path)  
  Reads the dataset from a CSV file.
* explore\_data(data)  
  Displays columns, head, and shape of the dataset.
* drop\_column(data)  
  Separates features and target (Outcome column).
* split\_dataset(dataset, train\_percentage, feature\_header, target\_header)  
  Splits data into training and testing sets.
* train\_model(X\_train, Y\_train)  
  Trains a logistic regression model.
* evaluate\_model(model, X\_train, Y\_train, X\_test, Y\_test)  
  Prints training and testing accuracy.

**Output**

After execution, the script prints:

* Dataset structure and sample rows
* Feature and target shapes
* Training and testing accuracy of the model

**Notes**

* Ensure the Outcome column exists in your dataset.
* You can modify train\_percentage or random\_state in split\_dataset() for experimentation.