Problem Statement

This assignment focuses on a multi-class diabetes dataset with three classes: N (Non-diabetic), Y (Diabetic), and P (Predict-diabetic). Data features and their clinical significance are described in the provided table.

URL Link: https://data.mendeley.com/datasets/wj9rwkp9c2/1

Data Description

Feature Name	Description	Unit / Range	Notes / Medical Significance
Gender	Biological sex of the subject (F = Female, M = Male)	M or F	May influence metabolic parameters such as cholesterol and BMI
AGE	Age in years	number (20 to 79)	Age is a major risk factor for Type 2 Diabetes
Urea	Urea level in blood	mg/dL (Normal: ~7–20)	Elevated levels may indicate kidney issues, common in diabetics
Cr (Creatinine)	Blood creatinine	mg/dL (Normal: ~0.6–1.3)	High values can suggest impaired kidney function
HbA1c	Glycated Hemoglobin percentage	% (Normal: <5.7%, Prediabetes: 5.7–6.4%, Diabetes: ≥6.5%)	Key diagnostic test for diabetes

Chol	Total cholesterol	mg/dL (Desirable: <200)	High cholesterol can indicate dyslipidemia, linked to diabetes
TG	Triglycerides	mg/dL (Normal: <150)	High TG often indicates insulin resistance
HDL	High-Density Lipoprotein	mg/dL (Optimal: >40 for men, >50 for women)	Low HDL is a risk marker for cardiovascular disease
LDL	Low-Density Lipoprotein	mg/dL (Optimal: <100)	Elevated LDL increases heart disease risk
VLDL	Very-Low-Density Lipoprotein	mg/dL (Normal: 2–30)	High levels may indicate metabolic syndrome
BMI	Body Mass Index	kg/m² (Normal: 18.5–24.9; Overweight: 25–29.9; Obese: ≥30)	Strongly correlated with diabetes risk
CLASS	Target variable: N = Non-diabetic, Y = Diabetic, P = Predict-diabetic	Categorical	Labels for classification task

Tasks

Build a classification and clustering algorithm to predict the diabetic class/cluster for a given dataset.

Deliverables

- Exploratory Data Analysis (EDA)
- Model Building
- Classification and Clustering (compare both along with evaluation)
- Analysis/Interpretation
- Error Analysis
- Future Enhancements

Submission

- Python Notebook along with output, dependencies, and formatted notebook
- PDF Report of your solution

Note: There are no right or wrong answers, be as creative as possible.