Supervised ML (Assignment5)

Assignment Project

Problem Statement

A leading biomedical research institute, **NovaGen Research Labs**, is conducting large-scale population health studies to better understand how underlying health conditions influence disease risk and long-term health outcomes. Every year, the institute recruits thousands of volunteers who undergo medical examinations, lifestyle assessments, and clinical tests. However, the researchers currently lack a reliable way to consistently distinguish between individuals with generally **healthy** profiles and those who may be **at higher health risk**, limiting the effectiveness of participant selection and stratified analysis in their studies.

To address this challenge, the institute has compiled a dataset consisting of **health records from 9,800 individuals** collected over multiple observational studies. Each record represents a unique participant to avoid sampling bias and ensure independent observations. The dataset includes a combination of **numerical and categorical health indicators**, such as physiological measurements, lifestyle factors, and medical history attributes, providing a comprehensive view of each individual's health status.

You are engaged as a **Data Scientist** to develop a predictive model that classifies individuals as **"healthy"** or **"unhealthy"** based on the available health data. This classification will support key research decisions, including:

- Selecting eligible participants for clinical trials and longitudinal studies
- Stratifying populations for risk-based analysis and outcome comparison

Dataset Description

Feature Name	Description
Age	Age of the individual (in years)
BMI	Body Mass Index, measuring body fat based on height and weight
Blood_Pressure	Systolic blood pressure (mmHg)
Cholesterol	Cholesterol level (mg/dL)
Glucose_Level	Blood glucose level (mg/dL)
Heart_Rate	Resting heart rate (beats per minute)
Sleep_Hours	Average number of sleep hours per day
Exercise_Hours	Average hours of exercise per day
Water_Intake	Daily water intake (litres)
Stress_Level	Stress level on a predefined scale (e.g., 1–10)
Smoking	Smoking habit (1 = Smoker, 0 = Non-smoker)
Alcohol	Alcohol consumption (1 = Yes, 0 = No)
Diet	General diet category encoded numerically
MentalHealth	Mental health score or condition indicator
PhysicalActivity	Overall physical activity level
MedicalHistory	Presence of prior medical conditions
Allergies	Presence of known allergies
Diet_Type__Vegan	One-hot encoded: 1 if diet is Vegan
Diet_Type__Vegetarian	One-hot encoded: 1 if diet is Vegetarian
Blood_Group_AB	One-hot encoded: Blood group AB
Blood_Group_B	One-hot encoded: Blood group B
Blood_Group_O	One-hot encoded: Blood group O
Target	**Target variable** representing health outcome or risk