

Problem Statement: Predictive Modeling for Disease Diagnosis

Background:

Healthcare professionals often rely on various diagnostic tests and biomarkers to assess an individual's health status and diagnose diseases. In this scenario, we have access to a dataset containing multiple health-related attributes such as cholesterol levels, blood cell counts, hormone levels, and other physiological measurements. The dataset also includes information on whether the individual has been diagnosed with a specific disease or not.

Objective:

The objective of this project is to develop a predictive model that can accurately classify individuals into diseased or non-diseased categories based on their health attributes. By leveraging machine learning algorithms, we aim to create a reliable tool that healthcare providers can use to assist in disease diagnosis and prognosis.

Data Description:

The dataset consists of the following attributes:

- Cholesterol: Level of cholesterol in the blood (mg/dL)
- Hemoglobin: Protein in red blood cells carrying oxygen
- Platelets: Blood cells aiding in clotting
- White Blood Cells (WBC): Immune system cells fighting infections
- Red Blood Cells (RBC): Cells carrying oxygen
- Hematocrit: Percentage of blood volume occupied by RBC
- Mean Corpuscular Volume (MCV): Average volume of RBC
- Mean Corpuscular Hemoglobin (MCH): Average amount of hemoglobin in RBC
- Mean Corpuscular Hemoglobin Concentration (MCHC): Average concentration of hemoglobin in RBC
- Insulin: Hormone regulating blood sugar levels
- BMI (Body Mass Index): Measure of body fat based on height and weight
- Systolic Blood Pressure (SBP): Pressure in arteries during heartbeats
- Diastolic Blood Pressure (DBP): Pressure in arteries at rest between beats
- Triglycerides: Type of fat found in blood (mg/dL)
- HbA1c (Glycated Hemoglobin): Measure of average blood sugar levels over past 2-3 months
- LDL (Low-Density Lipoprotein) Cholesterol: "Bad" cholesterol
- HDL (High-Density Lipoprotein) Cholesterol: "Good" cholesterol
- ALT (Alanine Aminotransferase): Liver enzyme
- AST (Aspartate Aminotransferase): Enzyme found in liver and heart
- Heart Rate: Number of heartbeats per minute (bpm)

- Creatinine: Waste product produced by muscles and filtered by kidneys
- Troponin: Protein released into bloodstream during heart muscle damage
- C-reactive Protein (CRP): Marker of inflammation in the body
- Disease: Binary indicator (1: Diseased, 0: Non-diseased)

Task:

Given this dataset, the task is to build a machine learning model capable of predicting the likelihood of an individual having a specific disease based on their health attributes. The model's performance will be evaluated using appropriate evaluation metrics such as **accuracy, precision, recall, and F1-score.**

Outcome:

The ultimate goal is to develop a robust predictive model that can assist healthcare professionals in early disease detection and patient management. By accurately identifying individuals at risk of certain diseases, interventions can be initiated promptly, potentially improving patient outcomes and reducing healthcare costs.