Data Description

The dataset used in this project has the following attributes:

- The size of the dataset is 133 columns and 4920 rows. The dataset has been bifurcated into training and testing data. The model will be trained through training dataset and the result will be tested on testing dataset. In this dataset, the 133 columns are split as follows:
  - 132 columns have continuous discrete binary values (Symptoms wherein 0 stands for no symptom and 1 stands for presence of symptom in the patient.)
  - 1 column is qualitative discrete categorical (Prognosis of disease consisting of 41 diseases such as Dengue, Diabetes, Heart attack etc.)

The rows consist of continuous variables i.e. dummy variable taking the value of 0 (false) and 1 (true). The diseases listed in this dataset range from communicable diseases (spread from one person to another) to non communicable diseases (do not spread to others).

Table 1: Attribute summary

Attribute	**Type**
Symptoms (all 133 columns)	Continuous (binary: 0, 1)
Prognosis	Categorical

Below is a list of all 41 diseases:

Table 2: List of Prognosis

Prognosis	**Prognosis**	**Prognosis**
Pneumonia	Acne	Dimorphichemorrhoids (piles)
Heart attack	Psoriasis	Varicose veins
Migraine	Drug Reaction	Hyperthyroidism
Paralysis (brain hemorrhage)	AIDS	Osteoarthritis
Malaria	Gastroenteritis	(vertigo) Paroymsal Positional Vertigo
Dengue	Hypertension	Urinary tract infection
Impetigo	Cervical spondylosis	Hepatitis D
Hepatitis B	Jaundice	Allergy
hepatitis A	Chicken pox	GERD
Chronic cholestasis	Typhoid	Bronchial Asthma
Peptic ulcer disease	Diabetes	Hepatitis C
Fungal infection	Hypothyroidism	Alcoholic hepatitis
Hepatitis E	Hypoglycemia	Common Cold
Tuberculosis	Arthritis	