APPLIED BUSINESS ANALYTICS

2024-25, Spring Semester

Exercises to Practice – Week 1

1. Load the **heart.csv** data file. This dataset dates from 1988 and includes information about 1025 patients, such as age, cholesterol levels, blood pressure, and whether they have heart disease.

   This dataframe contains 14 columns, including the predicted attribute. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

VARIABLE	DESCRIPTION
age	
sex	
chest pain type	4 values
resting blood pressure	
serum cholesterol in mg/dl	
fasting blood sugar > 120 mg/dl	
resting electrocardiographic results	Values 0, 1, 2
maximum heart rate achieved	
exercise induced angina	
oldpeak	ST depression induced by exercise relative to rest
the slope of the peak exercise ST segment	Number of cylinders (missing for Mazda RX-7, which has a rotary engine)
number of major vessels colored by flourosopy	0-3
thal	0 = normal; 1 = fixed defect; 2 = reversable defect

   a. Load the dataset into a Pandas dataframe and display the first five rows and inspect the column names and their data types.

   b. Check the shape of the dataframe and identify missing values, if any.

   c. Rename the column **rest bp** to **Resting_Blood_Pressure**.

   d. Replace spaces in column names with underscores.

e. Classify all the variables in the dataset. Using **.dtypes method,** check the data type of all columns in the dataframe. Convert **target** to a categorical variable**,** and ensure **cholesterol** is a continuous variable.

f. Select and display the **cholesterol** levels of the first four patients.

g. Select the first four rows of columns **age, cholesterol,** and **max heart rate**.

h. Combine non-consecutive columns **age, sex,** and **max heart rate** into a new dataframe.

i. Generate a random sample of 10 patients from the dataset. Oversample patients older than 60 years for a new sample.