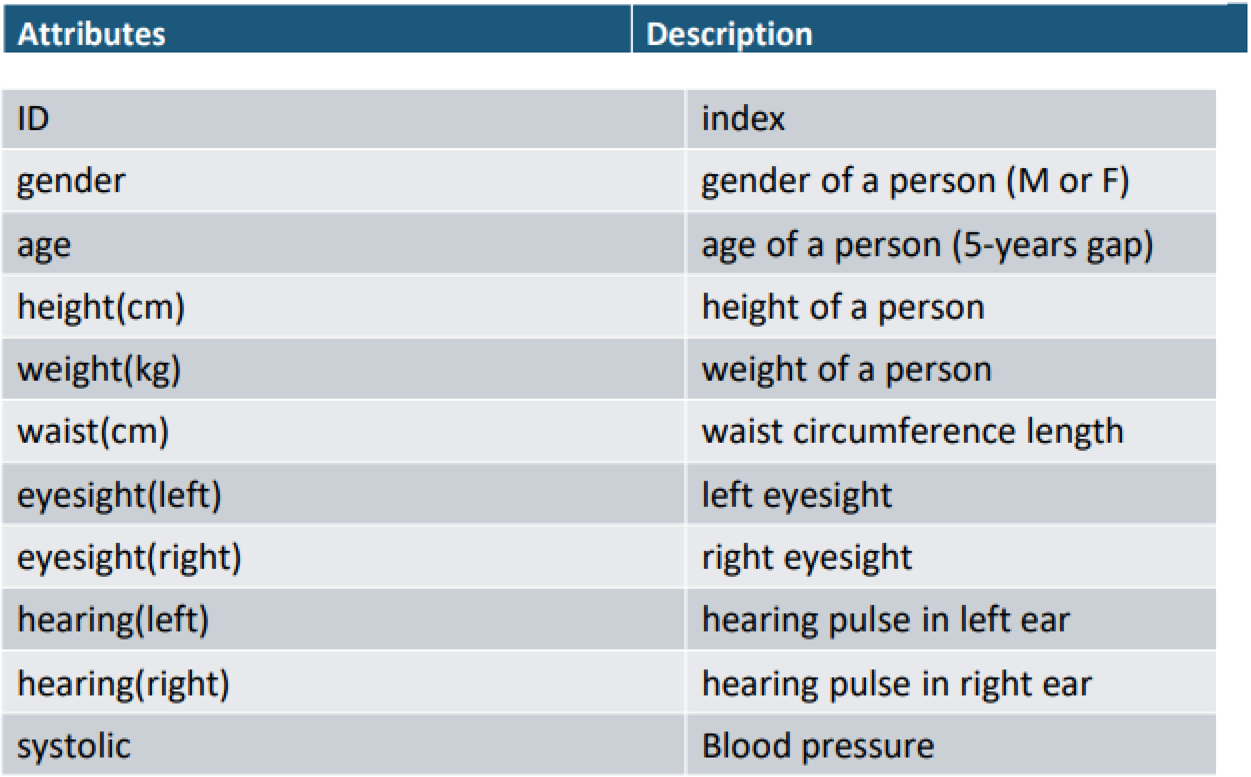
C-DAC Kharghar

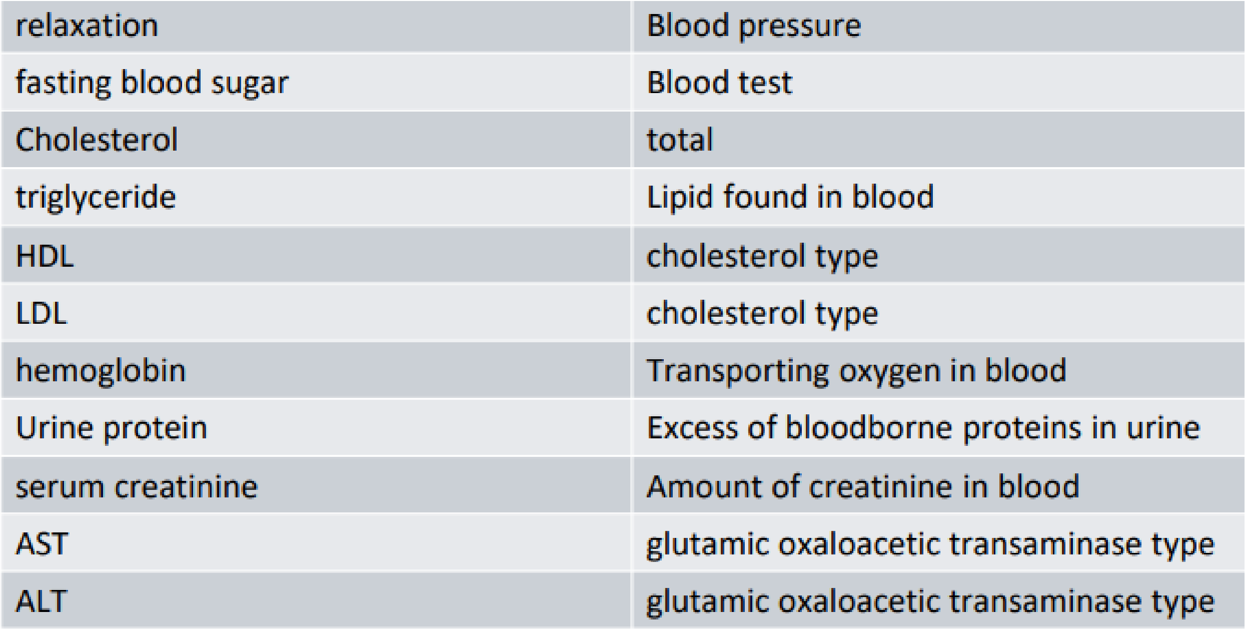
Bio-Signal Analysis for Smoking (ML case study)

As a data scientist at a global organization, you are entrusted with utilizing a vast amount of individual-level information collected over the years. The company's leadership aims to develop a smart system capable of identifying whether a person is a smoker or not, based on their bio-signals. Your task is to build a machine learning model that can accurately classify the smoking status of individuals using this data. Demonstrate Logistic regression, Decision tree & Random forest and compare accuracy also do feature selection. Give conclusion which ML model is perfect for this case study.

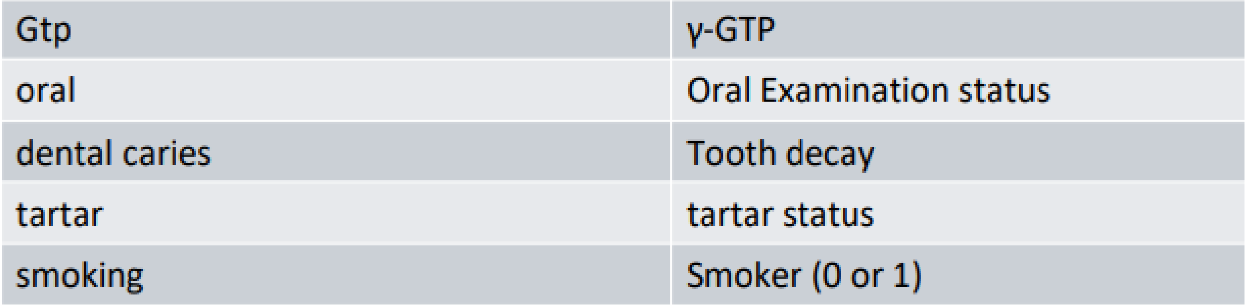
This dataset is a collection of basic health biological signal data which contains around 55K record with 27 attributes.



Attributes	Description
ID	index
gender	gender of a person (M or F)
age	age of a person (5-years gap)
height(cm)	height of a person
weight(kg)	weight of a person
waist(cm)	waist circumference length
eyesight(left)	left eyesight
eyesight(right)	right eyesight
hearing(left)	hearing pulse in left ear
hearing(right)	hearing pulse in right ear
systolic	Blood pressure



relaxation	Blood pressure
fasting blood sugar	Blood test
Cholesterol	total
triglyceride	Lipid found in blood
HDL	cholesterol type
LDL	cholesterol type
hemoglobin	Transporting oxygen in blood
Urine protein	Excess of bloodborne proteins in urine
serum creatinine	Amount of creatinine in blood
AST	glutamic oxaloacetic transaminase type
ALT	glutamic oxaloacetic transaminase type



Gtp	γ-GTP
oral	Oral Examination status
dental caries	Tooth decay
tartar	tartar status
smoking	Smoker (0 or 1)

****