Dataset Preparation

This dataset consists of 55692 observations, 26 independent variables and a target variable being smoking. Smoking value is either 0 or 1, with 0 being non-smoking and 1 being smoking. Note that the smoking status did not include e-cigarettes. All observations and variables will be included for the next step and the descriptions of all variables are presented in Table 1.

Variable Name	Description	Data Type
ID	Index	Numerical
Gender	Gender of the individual (F=female, M=Male)	Categorical
Age	Age of individual	Numerical
Height.cm.	Height of individual in cm	Numerical
Weight.kg.	Weight of individual in kg	Numerical
Waist.cm.	Waist circumference in cm	Numerical
Eyesight.left.	Visual acuity measure for left eye	Numerical
Eyesight.right.	Visual acuity measure for right eye	Numerical
Hearing.left.	Hearing ability for left ear (1=can hear, 0=cannot hear)	Numerical
Hearing.right.	Hearing ability for right ear (1=can hear, 0=cannot hear)	Numerical
Systolic	Systolic blood pressure	Numerical
Relaxation	Diastolic blood pressure	Numerical
Fasting.blood.sugar	Fasting blood sugar	Numerical
cholesterol	Total Cholesterol (measurement of lipid panel)	Numerical
Triglyceride	Triglyceride (measurement of lipid panel)	Numerical
HDL	High-density lipoprotein (measurement of lipid panel)	Numerical
LDL	Low-density lipoprotein (measurement of lipid panel)	Numerical
Hemoglobin	Heamoglobin	Numerical
Urine.protein	Urine protein (measurement of kidney function	Numerical
Serum.creatinine	Serum creatinine (measurement of kidney function)	Numerical
AST	Alanine transaminase (measurement of liver function)	Numerical
ALT	Aspartate transaminase (measurement of liver function)	Numerical
Gtp	Gamma-glutamyl transpeptidase (measurement of liver function)	Numerical
Oral	Oral examination status	Categorical
Dental.caries	Dental care	Numerical
tartar	Tartar status	Categorical
smoking	Smoking status	Numerical

Table 1 Features Description of Dataset