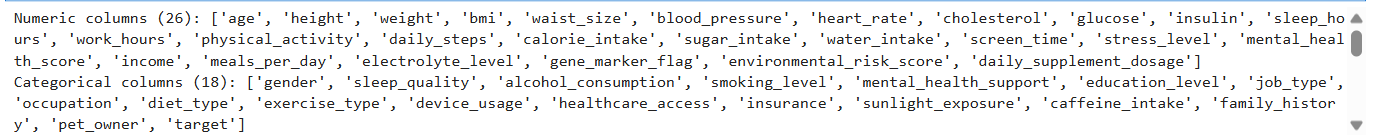
**Description**

1. Used the preprocessed dataset.
2. Preprocessed dataset consists of 44 columns and 10000 rows.
3. There are 21 columns of data type float, 5 columns of data type integer, and 18 columns of data type object.
4. There are 26 numerical variables and 18 categorical variables.



1. Identified how many categories per categorical variable.

A screenshot of a computer

AI-generated content may be incorrect.

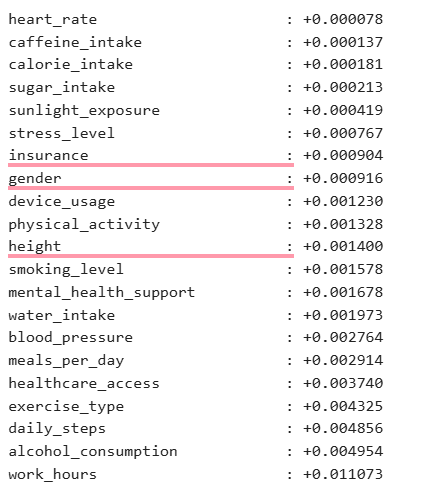
1. Created df\_encoded (encoded categorical variables)

Save enoded dataset as ‘encoded\_dataset.csv’ and mappings as ‘label\_mappings.csv’.

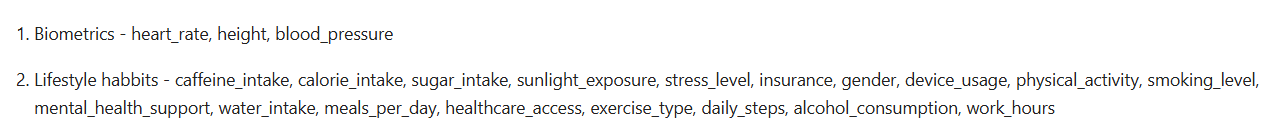
1. Round-off numerical variables to 6 decimal places and save dataset as df\_encoded\_rounded.
2. Generated correlation heat map using df\_encoded\_rounded to identify variables which are most related to the target variable.

There are only 21 variables which have positive correlation (weak positive correlation; 0<r<0.4). Correlation values are rounded upto 3 demical places.

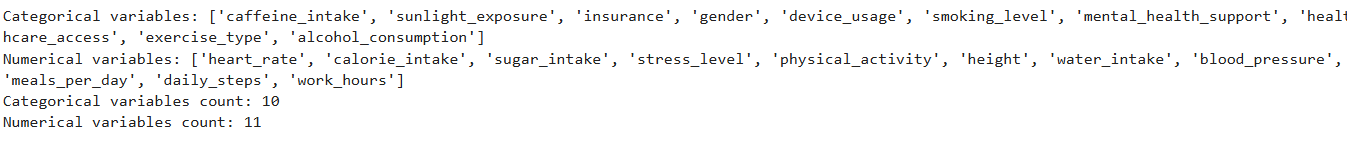
Hereafter, analysis is carried out with these 21 variables as they are the most related to the target variable.



1. Classify variables with positive correlations, biometrics and lifestyle habits.



1. Identify categorical and numerical variables among variables with positive correlation.



1. Draw pie charts for categorical variables which have positive correlations to understand each of them.
2. Draw box plots for numerical variables which have positive correlations to understand each of them.
3. Finding class imbalance using gini impurity.
4. Compare categorical variables with target variable using two-way frequency tables.
5. Comparing numerical variables with target variable using box plots.