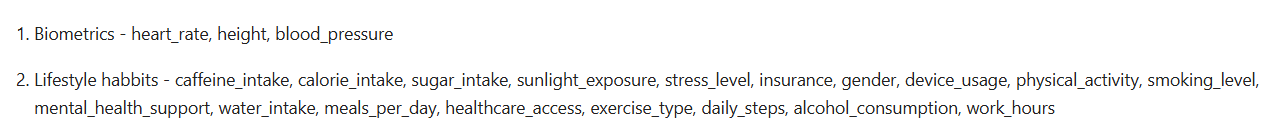
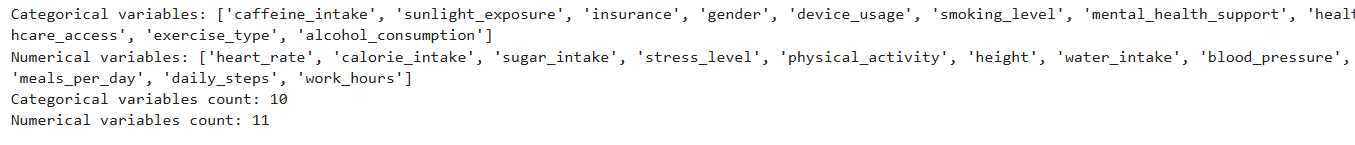
**Insights**

* There are 21 out of 44 variables which are most related to the target variable.
* Above 21 variables can be classified into biometrics and lifestyle habits as follows.



* Categorical and numerical variables among above 21 variables as follows.



* Analyzing categorical variables.
* Number of males and females are approximately same.
* Number of respondents who occasionally consume alcohol is more than twice the number of respondents who consume alcohol regularly.
* Number of respondents whose caffeine intake is high is nearly half of the ones whose caffeine intake is moderate.
* Device usage of respondents is divided equally between 3 categories: high, moderate, and low.
* Approximately half of respondents’ exercise type is strength, and other half is approximately divided among exercise types of cardio and mixed.
* Ease of access to healthcare of respondents equally divided among moderate and poor and the respondents’ access is good is slightly greater than other two.
* Number of respondents who have insurance and not are approximately equal.
* Number of respondents who have access to mental health support and who do not have are approximately equal.
* Number of light smokers and non-smokers are equal, and number of heavy smokers is slightly greater than other two.
* Number of respondents whose daily sunlight exposure is low and moderate are equal, and number of respondents whose daily sunlight exposure high is slightly lower than other two.
* Analyzing numerical variables.
* There are no outliers in any of the variables.

|  |  |
| --- | --- |
| Variable | Mean |
| blood\_pressure | 119.98 |
| calorie\_intake | 2201.46 |
| daily\_steps | 7005.91 |
| heart\_rate | 74.98 |
| height | 170.02 |
| meals\_per\_day | 3.00 |
| physical\_activity | 3.03 |
| stree\_level | 4.99 |
| sugar\_intake | 60.05 |
| water\_intake | 2.00 |
| work\_hours | 8.00 |

* Target variable has two categories: healthy and diseased.

Majority class – healthy (70.10%)

Minority class – diseased (29.90%)

Imbalance ratio = 2.34:1

Gini impurity = 0.4192

Therefore, dataset is imbalanced. (Majority is 2.34 times larger than minority.)

* Analyzing each categorical variable with target variable.
* Many diseased and healthy respondents have low sunlight exposure.
* Many diseased and healthy respondents have an insurance.
* Most of the diseased and healthy respondents and male.
* Most of the diseased and healthy respondents’ device usage is high.
* Most of the diseased and healthy respondents are non-smokers.
* Number of diseased and healthy respondents having access to mental resources and not having are approximately same.
* Most of the diseased respondents have good healthcare access and most of the healthy respondents have poor healthcare access.
* Most of the diseased and healthy respondents exercise types are strength.
* Most diseased and healthy respondents consume alcohol occasionally.
* Analyzing each numerical variable with target variable.

|  |  |  |
| --- | --- | --- |
| Numerical variable | Mean (diseased) | Mean(healthy) |
| blood\_pressure | 120.00 | 119.92 |
| calorie\_intake | 2201.50 | 2201.35 |
| daily\_steps | 7013.38 | 6988.41 |
| heart\_rate | 74.98 | 74.98 |
| height | 170.03 | 170.00 |
| measl\_per\_day | 3.00 | 2.99 |
| physical\_activity | 3.03 | 3.03 |
| stree\_level | 4.99 | 4.99 |
| sugar\_intake | 60.05 | 60.05 |
| water\_intake | 2.00 | 2.00 |
| work\_hours | 8.02 | 7.97 |