“Smoking and Drinking Dataset with Body Signal”

Unveiling Patterns and Correlations in Health Behaviors Through Comprehensive Body Signal Analysis

Project Members

- Sarvagya Porwal (21117112)
- Shivam Kumar (21117119)
- Shivam Singh (21117120)
- Shyam Jayakannan (21117126)

CONTENTS

1. MOTIVATION

Often when patients are diagnosed with a respiratory problem or health condition, they are reluctant to reveal information regarding their smoking and drinking status, that is, whether they smoke or drink and if so, how frequently. There may be several reasons ranging from fear of dismissal or appearing as someone who is careless about personal health. In any case, this leads to misinformation and difficulty for medical practitioners in providing proper treatment. Therefore, there is a need for accurate knowledge of a patient's smoking and drinking status without having to rely on the patient's words for the same.

It has been shown that this very information can be predicted based on several measurable body signals such as Blood Pressure, Cholesterol, Urine Proteins, and a few enzymes. Practitioners can use Machine Learning Models trained on this data available for several patients to predict the status of future patients.

This report summarizes an attempt at the problem discussed above. Details have been provided for the smoking status only because the analysis and methodology are the same for the drinking status.

2. DATASET

The dataset for the project is the Smoking and Drinking Dataset with body signal and has been sourced from Kaggle. This dataset was collected from the National Health Insurance Service in Korea and all personal information and sensitive data were excluded.

The data contains 9,91,346 rows and 24 columns.

	sex	age	height	weight	waistline	sight_left	sight_right	hear_left	hear_right	SBP	DBP	BLDS	tot_chole	HDL_chole	LDL_chole	triglyceride	hemoglobin	urine_protein	serum_creatinine	SGOT_AST	SGOT_ALT	gamma_GTP	SMK_stat_type_cd	DRK_YN
0	Male	35	170	75	90.0	1.0	1.0	1.0	1.0	120.0	80.0	99.0	193.0	48.0	126.0	92.0	17.1	1.0	1.0	21.0	35.0	40.0	1.0	1
1	Male	30	180	80	89.0	0.9	1.2	1.0	1.0	130.0	82.0	106.0	228.0	55.0	148.0	121.0	15.8	1.0	0.9	20.0	36.0	27.0	3.0	0
2	Male	40	165	75	91.0	1.2	1.5	1.0	1.0	120.0	70.0	98.0	136.0	41.0	74.0	104.0	15.8	1.0	0.9	47.0	32.0	68.0	1.0	0
3	Male	50	175	80	91.0	1.5	1.2	1.0	1.0	145.0	87.0	95.0	201.0	76.0	104.0	106.0	17.6	1.0	1.1	29.0	34.0	18.0	1.0	0
4	Male	50	165	60	80.0	1.0	1.2	1.0	1.0	138.0	82.0	101.0	199.0	61.0	117.0	104.0	13.8	1.0	0.8	19.0	12.0	25.0	1.0	0
5	Male	50	165	55	75.0	1.2	1.5	1.0	1.0	142.0	92.0	99.0	218.0	77.0	95.0	232.0	13.8	3.0	0.8	29.0	40.0	37.0	3.0	1
6	Female	45	150	55	69.0	0.5	0.4	1.0	1.0	101.0	58.0	89.0	196.0	66.0	115.0	75.0	12.3	1.0	0.8	19.0	12.0	12.0	1.0	0
7	Male	35	175	65	84.2	1.2	1.0	1.0	1.0	132.0	80.0	94.0	185.0	58.0	107.0	101.0	14.4	1.0	0.8	18.0	18.0	35.0	3.0	1
8	Male	55	170	75	84.0	1.2	0.9	1.0	1.0	145.0	85.0	104.0	217.0	56.0	141.0	100.0	15.1	1.0	0.8	32.0	23.0	26.0	1.0	1

FIGURE 1 FIRST 8 ROWS

Given below is a table describing the columns present in the dataset:

Title	Description
Sex	male, female
Age	round up to 5 years
Height	round up to 5 cm[cm]
Weight	[kg]
Sight_left	eyesight(left)
Sight_right	eyesight(right)
Hear_left	hearing left, 1(normal), 2(abnormal)
Hear_right	hearing right, 1(normal), 2(abnormal)
SBP	Systolic blood pressure[mmHg]
DBP	Diastolic blood pressure[mmHg]
BLDS	BLDS or FSG(fasting blood glucose)[mg/dL]
Tot_chole	total cholesterol[mg/dL]
HDL_chole	HDL cholesterol[mg/dL]
LDL_chole	LDL cholesterol[mg/dL]
Triglyceride	triglyceride[mg/dL]
Hemoglobin	hemoglobin[g/dL]
Urine_protein	protein in urine, 1(-), 2(+/-), 3(+1), 4(+2), 5(+3), 6(+4)
Serum_creatinine	serum(blood) creatinine[mg/dL]
SGOT_AST	SGOT(Glutamate-oxaloacetate transaminase) AST(Aspartate transaminase)[IU/L]
SGOT_ALT	ALT(Alanine transaminase)[IU/L]
Gamma_GTP	y-glutamyl transpeptidase[IU/L]
SMS_stat_type_cd	Smoking state, 1(never), 2(used to smoke but quit), 3(still smoke)
DRK_YN	Drinker or Not

3. PRELIMINARY ANALYSIS

3.1. Under sampling

It was observed that out of the 9,91,346 patients, around 6,00,0000 were non – smokers, which means that the proportion of patients who smoke or used to smoke is relatively small. To remove the oversampling, the number of non–smokers was reduced to 2,50,000 by random selection.

3.2. Invariable Features

Plotting values in each column as a parameter vs. the smoking status of a patient, it was observed in the case of columns urine_protein, hear_left, and hear_right that the observed values lie in the normal range for smokers and non–smokers alike. Hence, these three parameters were not taken into consideration in making predictions.

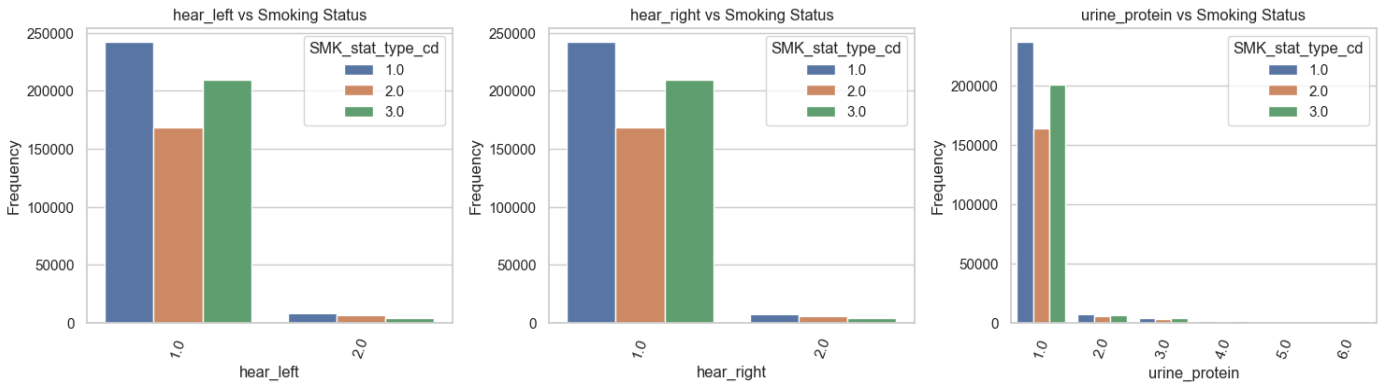

FIGURE 2 PLOTTING HEAR_LEFT, HEAR_RIGHT AND URINE_PROTEIN

3.3. Measures of Central Tendency

The mean, standard deviation and median of the data points for each parameter have been plotted in the figure below. The horizontal axis distinguishes smokers as 'Never Smoked,' 'Former Smoker' and 'Current Smoker.' We can see that all three values are more – or – less similar for all types of smokers.

- **Blue:** Mean
- **Orange:** Standard Deviation
- **Green:** Median

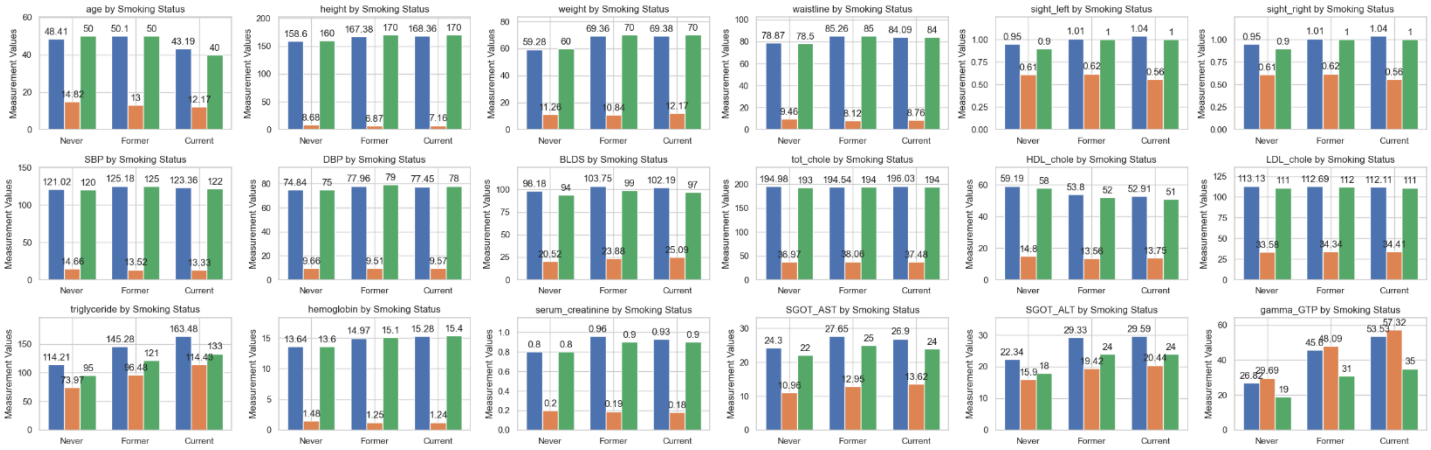

FIGURE 3 MEASURES OF CENTRAL TENDENCY

3.4. Removing Outliers

Outliers are observations that differ considerably from other observations. Such data were removed from the dataset by deleting entries whose parameters had data points lying in the top 0.1% of the complete data. 8,433 such points were found and their removal brought down the number of entries to 6,30,472.

3.5. Count plot

Count plots were made for all the remaining parameters for a random sample of 1,00,000 patients to understand their distribution better. The plots for all three types of smokers emerged similar for most of the parameters, but there were slight exceptions like in the case of hemoglobin.

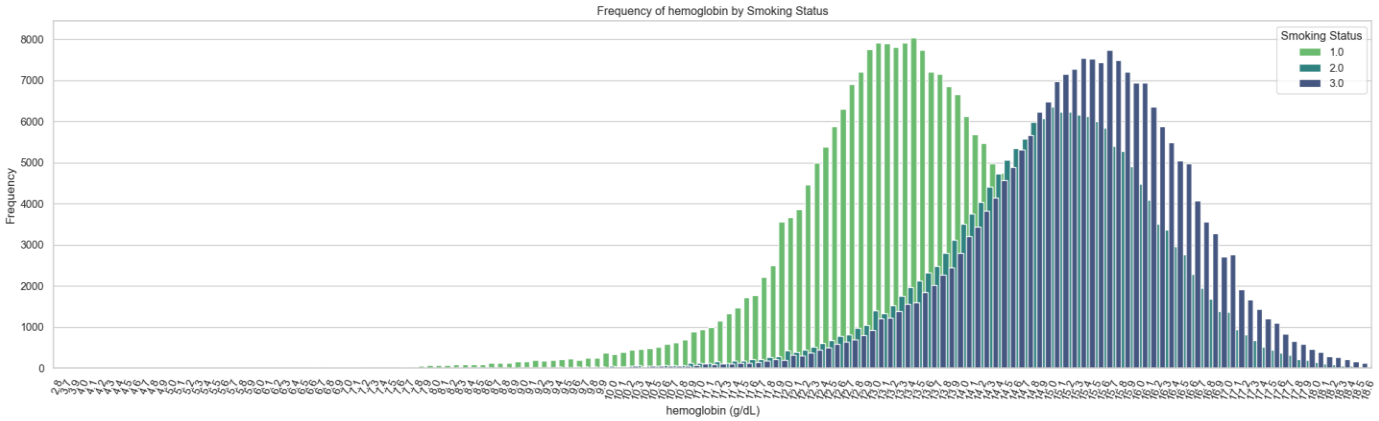

FIGURE 4 HEMOGLOBIN LEVELS IN BLOOD OF PATIENTS

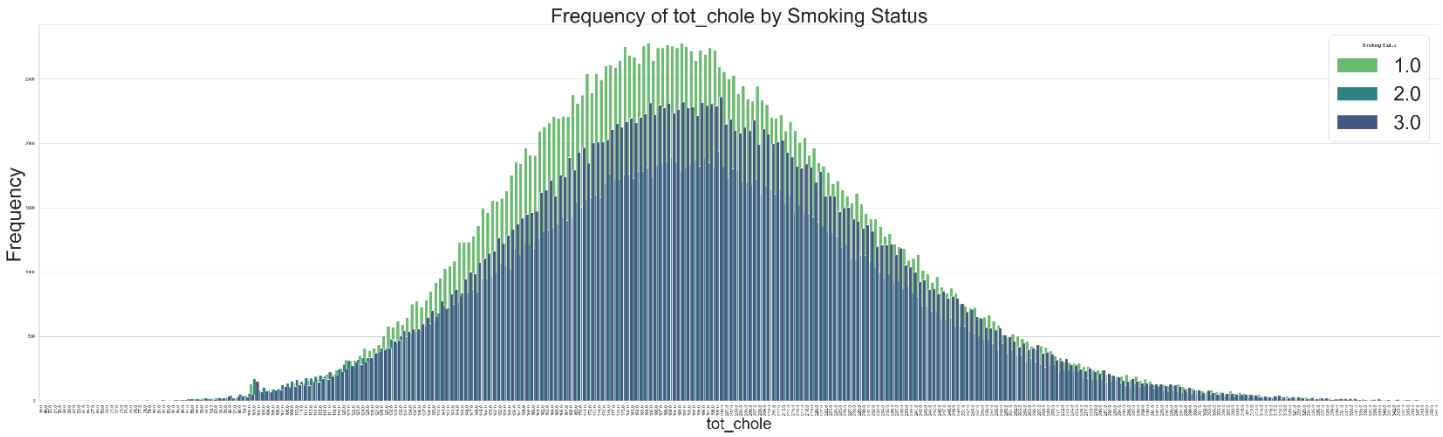

FIGURE 5 TOTAL CHOLESTEROL

3.6. Heatmap

Along with relations to the smoking status, relations between the parameters themselves also provide important insights. A heatmap is a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colors. A heatmap was made to study relations among the parameters.

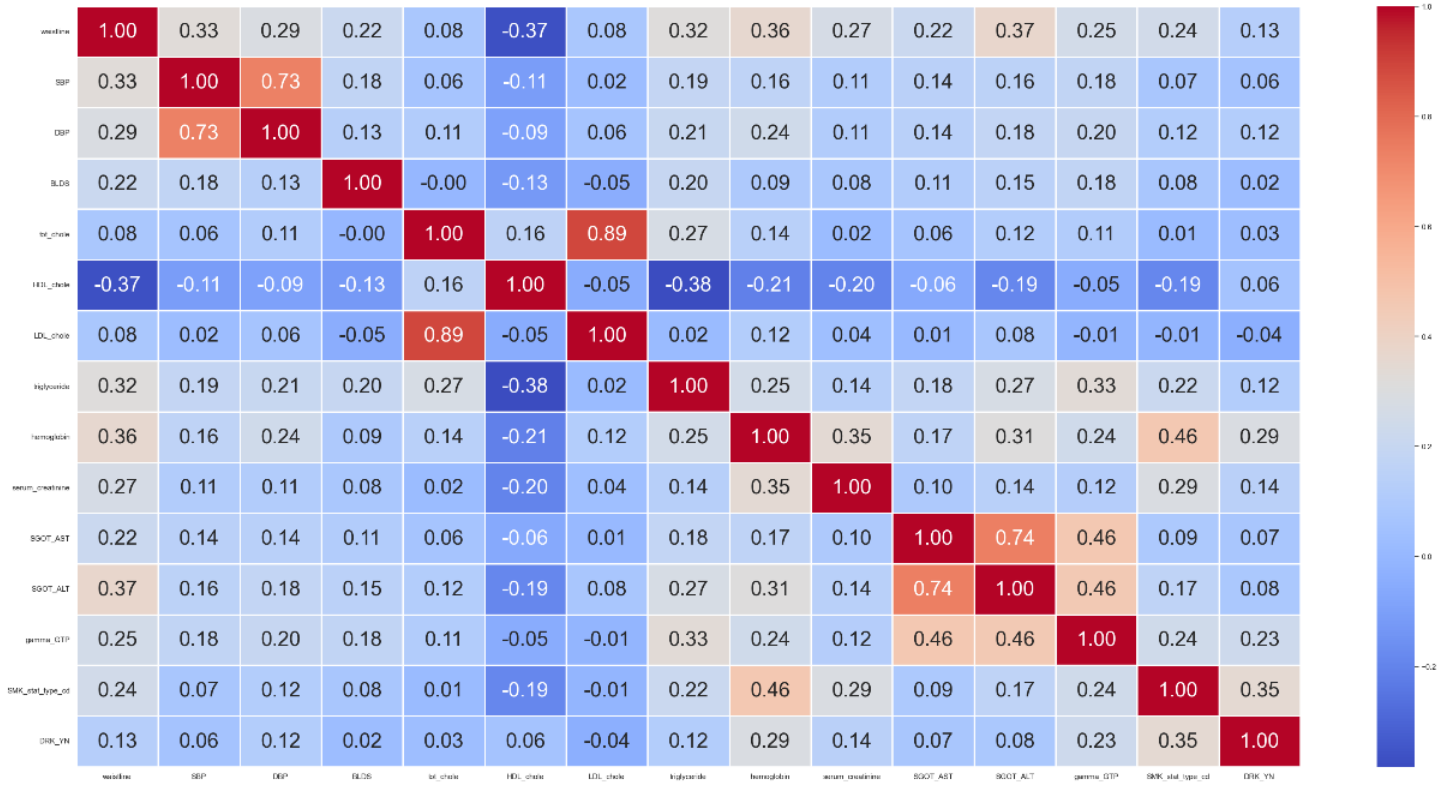

FIGURE 6 HEATMAP

From the heatmap, we can see strong correlations between SBP – DBP, LDL_chole – tot_chole and SGOT_ALT – SGOT_AST. There was also a considerable correlation between sight_left – sight_right, which, being categorical variables (a variable that can take on only one of a small fixed list of values) were not plotted on the heatmap. So, we may consider reducing these parameters from 8 to 4 using methods such as PCA, SVD and LDA to reduce the number of dimensions.

At the same time, we must be careful not to reduce the number of parameters too much because there already are only 22 parameters remaining to describe around 1,00,000 data – points. Keeping this in mind, parameters were reduced only in situations where one of the three methods (PCA, SVD, LDA) resulted in a data loss of less than 0.5%.

Data loss in each case was calculated using the formula:

$$\text{data\_loss} = 1 - \text{explained\_variance}[0]$$

Finally, LDL_chole – tot_chole, SGOT_ALT – SGOT_AST and sight_left – sight_right were reduced to a single parameter each using LDA.

4. METHODOLOGY

After the initial analysis and reduction of parameters, we proceeded to the Machine Learning Models.

4.1. Logistic Regression

Logistic Regression is justified for predicting "isDrinker" in the dataset due to its compatibility with binary outcomes. There are 3 types of patients in the target class for smokers (non – smokers : 1, stopped smoking: 2, smoker : 3). However, Logistic Regression is generally used for binary classification. But, in the scikit learn library, the model has been extended to include multi – class classification.

The model has been used in its default mode ('auto' for multi – class). Here is an exerpt from the documentation: "'*auto' selects 'ovr' if the data is binary, or if solver='liblinear', and otherwise selects 'multinomial'.*"

4.2. Gaussian NB

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. More can be found here. Its capacity to handle both continuous and categorical data makes it suitable for variables like age, health metrics, and lifestyle indicators. Despite the assumption of feature independence, Gaussian NB performs well, particularly with limited dependencies. Its efficiency with smaller datasets, simplicity, and interpretability make it a practical choice.

4.3. Artificial Neural Network

The choice of a deep neural network with multiple hidden layers and specific activation functions (ReLU for intermediate layers and Sigmoid for the output layer) was taken for the following reasons:

- **Complex Relationships:** The dataset includes various health and lifestyle features, suggesting potential complex relationships. The deep architecture allows the model to learn hierarchical and intricate patterns, capturing nuanced dependencies among predictors.
- **Non – linearity:** ReLU (Rectified Linear Unit) is commonly used for hidden layers, introducing non-linearity essential for learning complex mappings. It enables the model to represent more intricate relationships between input features.
- **Layer Size Variation:** Gradually increasing the number of units in hidden layers (e.g., from 16 to 256) allows the model to capture increasingly complex representations of the data, promoting feature learning at different abstraction levels.
- **Sigmoid Activation for Binary Classification:** Using Sigmoid activation in the output layer is appropriate for binary classification tasks like predicting "is_Smoking" and "isDrinker." It squashes the output to a range between 0 and 1, representing probabilities.
- **Categorical Crossentropy Loss:** Categorical Crossentropy loss is appropriate when dealing with categorical classification tasks. It penalizes the model based on the difference between predicted and true class probabilities.
- **Training and Evaluation:** The model is trained for 30 epochs with a batch size of 128, and performance is evaluated on a separate test set. This ensures a balance between model training and evaluation, helping to identify potential overfitting.

5. RESULTS

Smoking

```
Logistic Regression:
Accuracy: 0.6557
Classification Report:
              precision    recall  f1-score   support

         1.0       0.88      0.73      0.79      7906
         2.0       0.53      0.50      0.51      5453
         3.0       0.56      0.70      0.62      6641

    accuracy                           0.66     20000
   macro avg       0.65      0.64      0.64     20000
weighted avg       0.68      0.66      0.66     20000

confusion_matrix:
 [[5754  919 1233]
 [ 355 2719 2379]
 [ 467 1533 4641]]
```

```
Naive Bayes:
Accuracy: 0.6216
Classification Report:
              precision    recall  f1-score   support

         1.0       0.77      0.75      0.76      7906
         2.0       0.48      0.54      0.51      5453
         3.0       0.57      0.53      0.55      6641

    accuracy                           0.62     20000
   macro avg       0.61      0.61      0.61     20000
weighted avg       0.63      0.62      0.62     20000

confusion_matrix:
 [[5962  978  966]
 [ 815 2949 1689]
 [ 920 2200 3521]]
```

```
ANN
Accuracy: 0.6496999859809875

              precision    recall  f1-score   support

           0       0.87      0.73      0.79      7906
           1       0.49      0.66      0.57      5453
           2       0.60      0.54      0.57      6641

    accuracy                           0.65     20000
   macro avg       0.65      0.65      0.64     20000
weighted avg       0.68      0.65      0.66     20000


Confusion matrix tf.Tensor(
[[5779 1203  924]
 [ 368 3626 1459]
 [ 500 2552 3589]], shape=(3, 3), dtype=int32)
```

Drinking

```
Logistic Regression:
Accuracy: 0.71945
Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.72      0.72     10030
           1       0.72      0.72      0.72      9970

    accuracy                           0.72     20000
   macro avg       0.72      0.72      0.72     20000
weighted avg       0.72      0.72      0.72     20000

confusion_matrix:
 [[7220 2810]
 [2801 7169]]
```

```
Naive Bayes:
Accuracy: 0.68815
Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.73      0.70     10030
           1       0.70      0.65      0.67      9970

    accuracy                           0.69     20000
   macro avg       0.69      0.69      0.69     20000
weighted avg       0.69      0.69      0.69     20000

confusion_matrix:
 [[7303 2727]
 [3510 6460]]
```

```
ANN
Accuracy: 0.7228999733924866

              precision    recall  f1-score   support

           0       0.73      0.72      0.72     10030
           1       0.72      0.73      0.72      9970

    accuracy                           0.72     20000
   macro avg       0.72      0.72      0.72     20000
weighted avg       0.72      0.72      0.72     20000


Confusion matrix tf.Tensor(
[[7180 2850]
 [2692 7278]], shape=(2, 2), dtype=int32)
```

6. REFERENCES

[1] Smoking and Drinking Dataset with body signal

[2] Gaussian NB

[3] Github Repository