

# Project Scope Statement



**Body Signal Data**

## **SUBMITTED BY:**

**Yanyan Li  
Kaiyue Wei  
Haoqi Wang  
Xulu Wang**

## **DATE:**

**Oct. 19 2023**

# Table of Contents

<b>1. Project Summary.....</b>	<b>3</b>
1.1 Project #.....	3
1.2 Project Description.....	3
1.3 Date Submitted.....	3
1.4 Project Priority.....	3
1.5 Step 1 Project Deliverable.....	3
1.6 Step 2 List of Project Tasks.....	3
1.7 Step 3 Out of Scope.....	4
1.8 Step 4 Project Assumption.....	4
1.9 Step 5 Project Constraints.....	4
1.10 Step 6 Updated Estimate.....	4
1.11 Step 7 Approvals.....	5
<b>2. Introduction.....</b>	<b>5</b>
<b>3. Analysis of the dataset and Trained Model.....</b>	<b>5</b>
3.1 Exploratory Analysis and Visualization.....	5
3.2 Baseline Model.....	8
<b>4. Model Selection.....</b>	<b>9</b>
4.1 Model Performance Evaluation.....	10
4.2 Feature Importance.....	11

# 1. Project Summary

1.1 Project #	1.2 Project Description	1.3 Date Submitted	1.4 Project Priority
1	We're harnessing machine learning to analyze routine health data, aiming to classify individuals based on physiological signals related to smoking and drinking behaviors. Beyond traditional tobacco and alcohol biomarkers, we're exploring a wider range of signals. Our goal is to provide accurate identification of high-risk behaviors, enabling targeted interventions and using real-time feedback to promote positive change.	10/20/2023	Project II

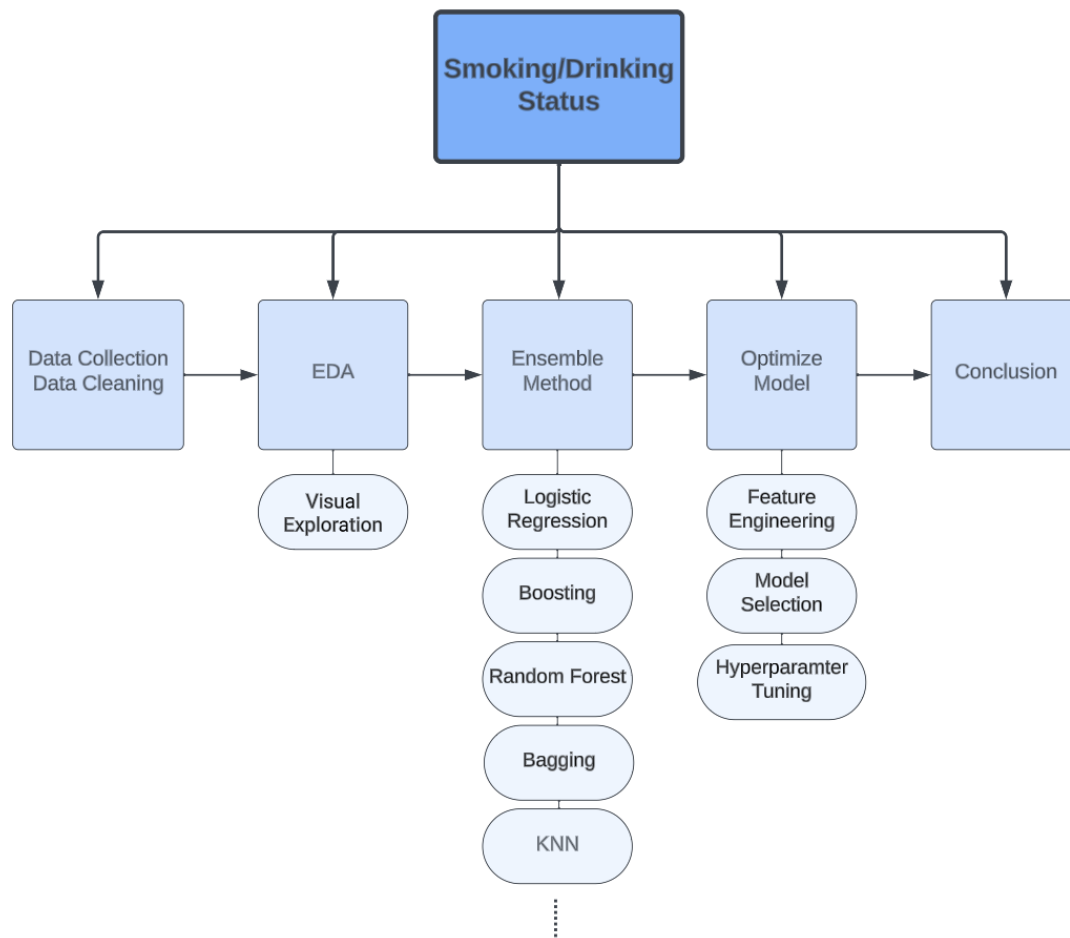
## 1.5 Step 1 Project Deliverable

Deliverable ID#	Description
1	Project Charter
2	Ensemble trained model with preliminary results of the training and testing.
3	Flask and Heroku application. (customer will choose whether to deploy it locally or on the cloud)
4	Final report of the project in addition to the cloud (heroku) deployed link

## 1.6 Step 2 List of Project Tasks

Task ID#	Task to be completed	Delivery Date	For Deliverable #
1	Submit Project Charter	09/16/2023	1
2	Generative Methods Based Analysis of the dataset (EDA visualization) <ul style="list-style-type: none"> <li>• Pair plots: Show pairwise relationships in a dataset. By using pair plots, we can immediately see the distributions of single variables and relationships between two variables.</li> <li>• Scatter plots: Useful for spotting structured relationships between variables.</li> <li>• Heatmaps: Useful for spotting correlations among multiple variables.</li> </ul>	11/12/2023	2
3	Machine learning model deployment	11/26/2023	3

	<ul style="list-style-type: none"> <li>• Perform the analysis by using traditional methods with generative and non-generative methods, and compare the results for different parameter sets.</li> <li>• Tune the parameters of the ML algorithm to achieve better results.</li> <li>• Report the evaluation metrics used and models' performance on the validation and test sets.</li> </ul>		
4	<p>Final report of the project</p> <ul style="list-style-type: none"> <li>• Summary the insights and give the audience an understandable explanation.</li> <li>• Answer the question in the introduction paragraph.</li> <li>• Based on the conclusion and give the advice for the people who smoke or drink.</li> </ul>	12/02/2023*	4



### 1.7 Step 3 Out of Scope

This project **will NOT accomplish or include** the following:

- Collection of new data or additional data sources beyond the provided dataset.
- Deployment of the final model to a production or operational environment.
- Building a full end-to-end pipeline for real-time predictions.
- Quantitative analysis of the downstream impacts of model predictions.
- Extensive error analysis or debugging of model limitations.

## 1.8 Step 4 Project Assumption

Project factors that will be considered to be true, real, or certain. Assumptions generally involve a certain degree of risk.

#	Assumption
1	The provided dataset is accurate and representative of the target population's health data related to smoking and drinking behaviors.
2	The machine learning models and algorithms chosen for the project are appropriate for the given dataset and can yield meaningful results.
3	Adequate resources, including computational resources and data processing capabilities, are available to complete the project within the specified timeline.

## 1.9 Step 5 Project Constraints

- Model accuracy should exceed 80% for both classifications.
- Model must be interpretable to understand predictive signals.
- Prediction speed should enable real-time use for APP.

## 1.10 Step 6 Updated Estimate

Estimate T&C hours required to complete project N/A	Enter total # of T&C hours N/A	If charge-back project, list total estimated T&C cost N/A	Enter N/A if not applicable.
--	-----------------------------------	--	------------------------------

## 1.11 Step 7 Approvals

Required For Project Class...	Role of Approver	Submitted for Approval on:	Approval Received on:
All classes	1. Client + Client Supervisor	Nakul R. Padalkar	09/16/2023
All classes	2. T&C Supervising Manager	Nakul R. Padalkar	11/12/2023
Class 3 + 4 only	4. VP for Technology & Communication	Nakul R. Padalkar	11/26/2023
Class 3 + 4 only	5. Project Review Board	Nakul R. Padalkar	12/02/2023

# 2. Introduction

Tobacco use and excessive alcohol consumption are major public health concerns globally, contributing to millions of preventable deaths each year. Being able to predict smoking and drinking behavior from

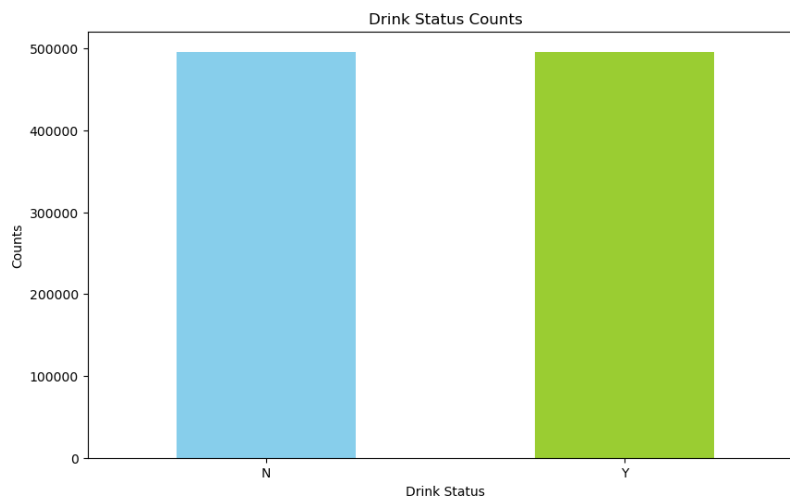
physiological data could help public health agencies target interventions more effectively and contribute to personalized medicine. For example, public health agencies can allocate resources for smoking cessation or alcohol abuse programs to communities with higher prevalence, and healthcare providers can tailor their recommendations and treatment plans based on individual risk factors. This project seeks to develop machine learning models that can classify individuals as smoking status and drinkers/non-drinkers based on body signal data. The body signal data provides a wide range of biomarkers that may be predictive of smoking and drinking habits. However, these relationships have not been extensively studied using machine learning approaches. This project could gain new insights into factors like blood pressure, cholesterol, liver enzymes, etc. correlate with and potentially predict smoking and drinking behavior.

Overall, this is a classification problem aiming to categorize individuals based on their physiological profile. This paper attempts to address the following questions: Can physical indicators in routine health screening data effectively predict smoking and drinking status of individuals? Which signals show the strongest predictive relationships?

### 3. Analysis of the dataset and Trained Model

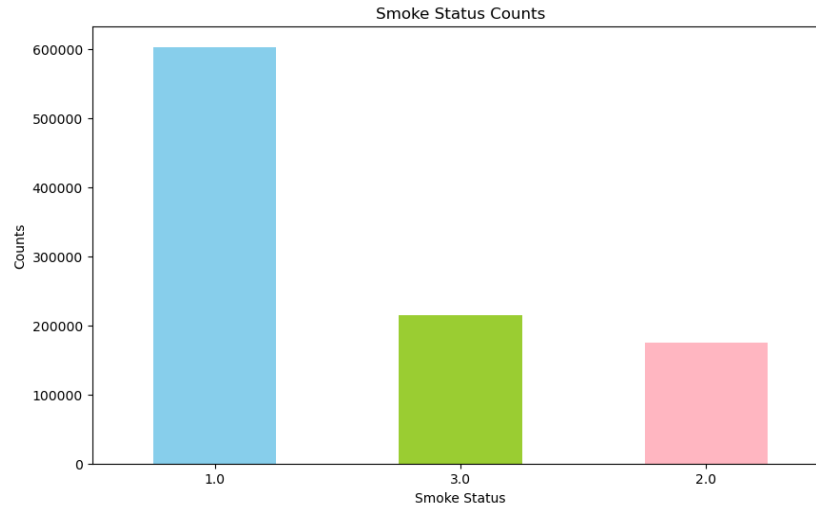
#### 3.1 Exploratory Analysis and Visualization

Before moving into model training, it's essential to gain insights into our dataset through exploratory analysis. Our dataset does not have any missing value. The following graphs show the distribution of selected target variables for analysis and modeling: Smoke Status and Drink Status.



**Figure 1 Distribution of the dependent variable (Drink Status)**

Figure 1 illustrates the distribution of the dependent variable, Drink Status. 'N' represents 'No', and 'Y' represents 'Yes'. The distribution of Drink Status is intentionally balanced, with an equal representation of 'Yes' and 'No' to ensure a balanced dataset for analysis.



**Figure 2 Distribution of the dependent variable (Smoke Status):  
1 (never), 2(used to smoke but quit), 3(still smoke)**

Figure 2 shows the distribution of the dependent variable, Smoke Status. `1` represents 'never smoke', `2` represents 'used to smoke but quit', and `3` represents 'still smoke'. The distribution of Smoke Status is imbalanced, with more than half of the records never smoking.

**Table 1 & 2 Part of Data Summary**

	age	height	weight	waistline	sight_left	sight_right	hear_left	hear_right	SBP	DBP
count	991346	991346	991346	991346	991346	991346	991346	991346	991346	991346
mean	47.614	162.241	63.284	81.233	0.981	0.978	1.031	1.030	122.432	76.053
std	14.181	9.283	12.514	11.850	0.606	0.605	0.175	0.172	14.543	9.889
min	20	130	25	8	0.1	0.1	1	1	67	32
25%	35	155	55	74.1	0.7	0.7	1	1	112	70
50%	45	160	60	81	1	1	1	1	120	76
75%	60	170	70	87.8	1.2	1.2	1	1	131	82
max	85	190	140	999	9.9	9.9	2	2	273	185

	HDL_c hole	LDL_ch ole	triglyceri de	hemogl obin	urine_ protein	serum_creatinine	SGOT_AST	SGOT_ ALT	gamma_ GTP	SMK_stat _type_cd
count	991346	991346	991346	991346	991346	991346	991346	991346	991346	991346
mean	56.937	113.038	132.142	14.230	1.094	0.860	25.989	25.755	37.136	1.608
std	17.238	35.843	102.197	1.585	0.438	0.481	23.493	26.309	50.424	0.819
min	1	1	1	1	1	0.1	1	1	1	1

<b>25%</b>	46	89	73	13.2	1	0.7	19	15	16	1
<b>50%</b>	55	111	106	14.3	1	0.8	23	20	23	1
<b>75%</b>	66	135	159	15.4	1	1	28	29	39	2
<b>max</b>	8110	5119	9490	25	6	98	9999	7210	999	3

Table 1 & 2 above shows part of the summary of the variables including the two target variables, to provide a better idea of the data.

The below figure 3 is the heatmap for all columns including the categorical variables like Drink Status and Sex which have been converted to numerical variables. From the figure, we can observed In Figure 3 below, we present a heatmap encompassing all columns, including categorical variables like Drink Status and Sex, which have been converted into numerical variables. From the figure, we observe several noteworthy correlations:

- Height and weight display a significant correlation of 0.72. We will calculate the BMI (Body Mass Index) to alleviate the inner correlation between height and weight.
- The correlation between height and sex is also substantial, standing at 0.72. This is expected, given the natural height differences between men and women.
- Diastolic blood pressure (DBP) and Systolic blood pressure (SBP) exhibit a high correlation of 0.74. This correlation implies a strong relationship between these two blood pressure measurements.
- Total cholesterol (tot\_chole) and LDL cholesterol (LDL\_chole) are significantly correlated, with a coefficient of 0.887. To address the high correlation between these variables, we should consider excluding one of these variables during model training to avoid issues related to redundancy and multicollinearity. Finally we removed the feature 'tot\_chole', because this variable is able to be calculated from the other two variables based on a formula:  $HDL + LDL + 20\% \text{ triglycerides} = \text{Total cholesterol}$ ; and 'tot\_chole' has a higher variance.



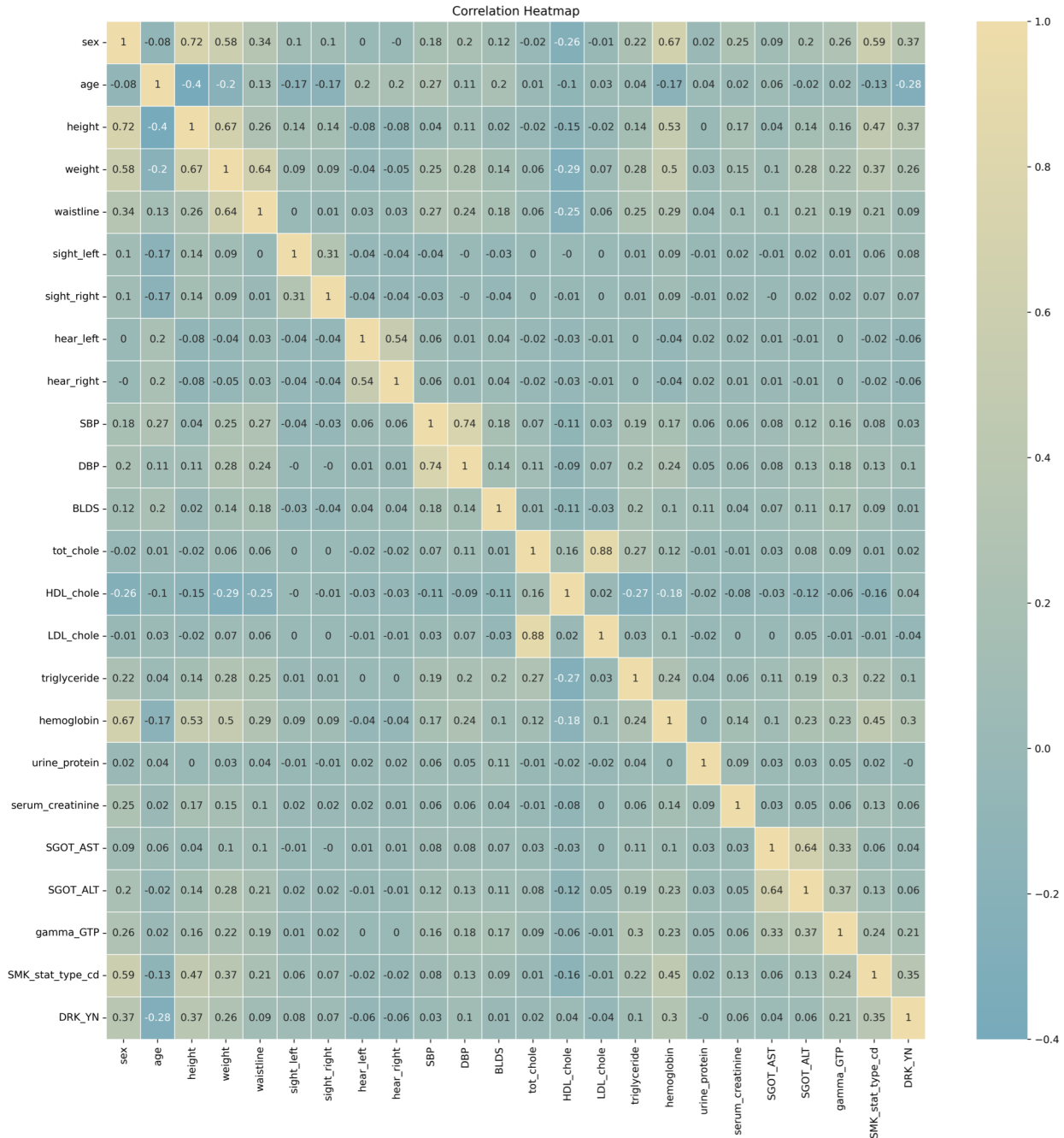


Figure 3 Correlation Matrix of all variables

### 3.2 Baseline Model

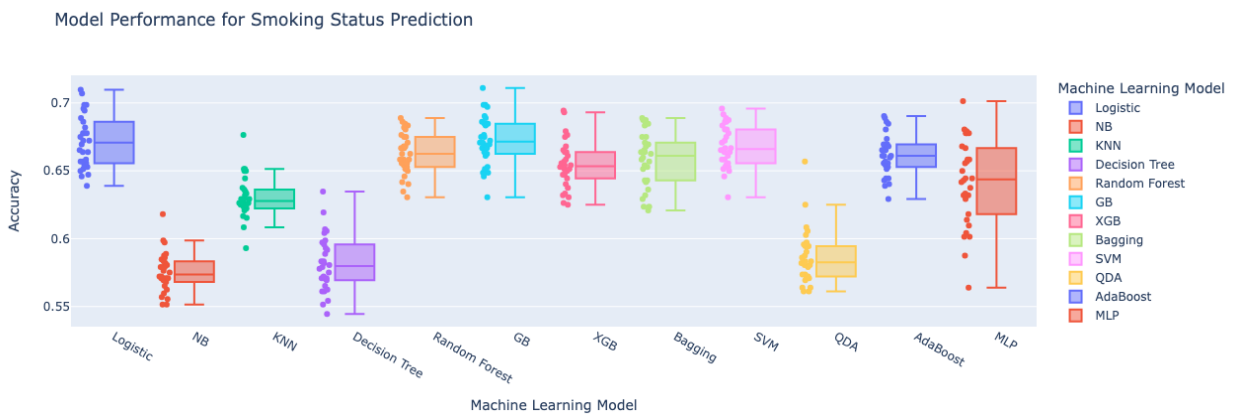
The dummy classifier is chosen as the baseline model. Dummy classifier makes predictions ignoring the input features, only based on the values of target variables, which makes it sensitive to the unbalanced dataset. If an advanced model fails to outperform the baseline, it indicates a failure to deal with unbalanced classes. The 'prior' strategy is used to generate predictions. It turns out that the classification accuracies of drinking status model and smoking status model are 49.5% and 32.5%, respectively. Considering that there are two classes for drinking status and three classes for smoking status. The

accuracy obtained by the dummy classifier is about the same as random guessing (50% and 33.3%, respectively), indicating that the dataset is not unbalanced.

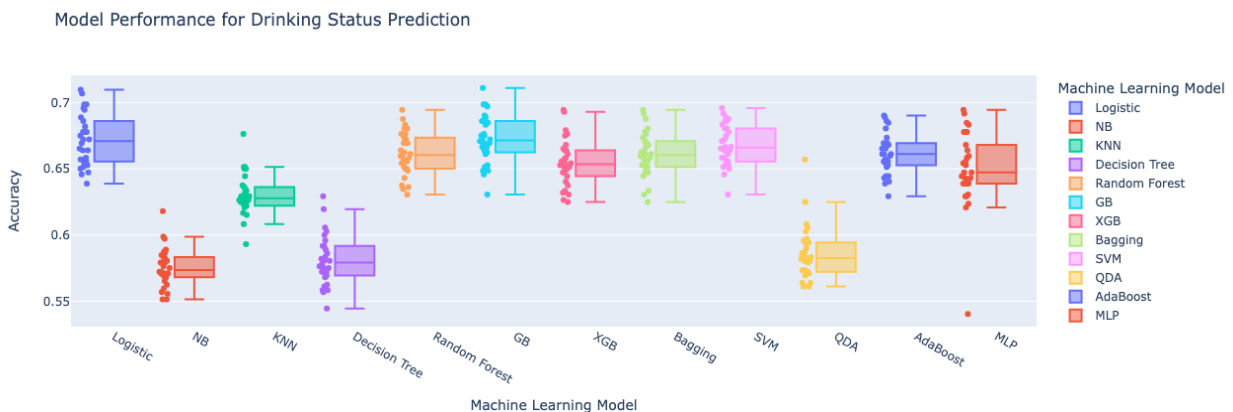
## 4. Model Selection

To guide our model selection process for both target variables, smoke and drink status, we began with a baseline analysis. Following this preliminary classification, we evaluated several candidate models, including LogisticRegression, GaussianNB, KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, GradientBoostingClassifier, XGBClassifier, BaggingClassifier, SVM, QuadraticDiscriminantAnalysis, AdaBoostClassifier, and MLPClassifier.

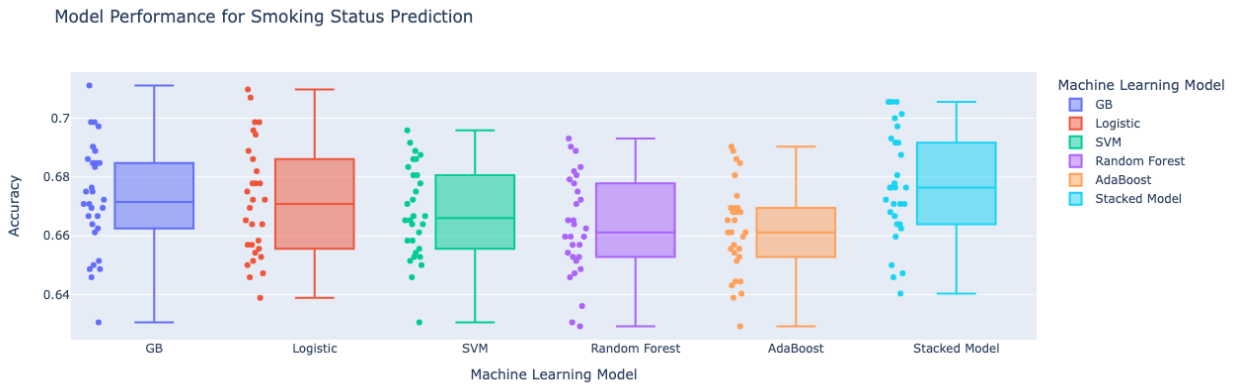
Figures 4 and 5 below display the model results for smoke status and drink status, indicating that GradientBoostingClassifier, LogisticRegression, SVM, AdaBoostClassifier, and RandomForestClassifier exhibit the best performance with relatively high prediction accuracy on test data. Thus, we have selected these models as the level 0 models in the stacking process. Logistic regression is used as the level 1 combiner or metamodel to aggregate the results of the level 0 models.



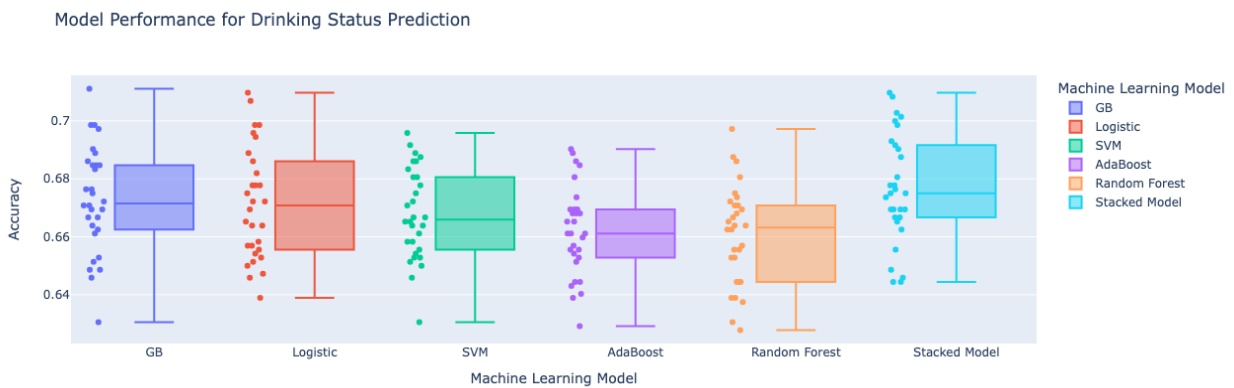
**Figure 4 Machine Learning Model Result for Smoke Status**



**Figure 5 Machine Learning Model Result for Drink Status**



**Figure 6 Machine Learning Best Models and Stacked Model Result for Smoke Status**

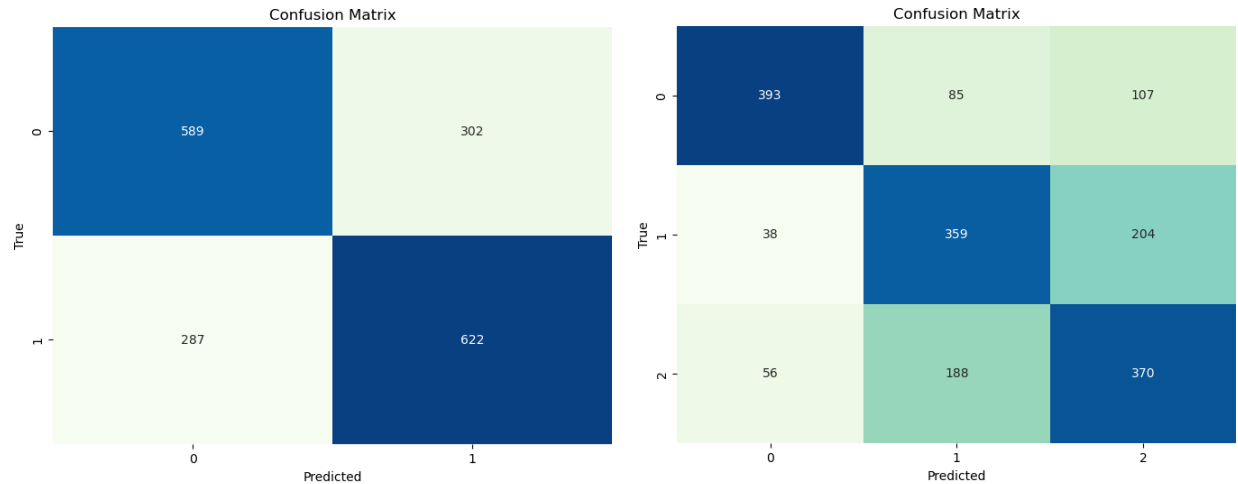


**Figure 7 Machine Learning Best Models and Stacked Model Result for Drink Status**

In addition to model selection, we have also chosen to implement a 10-fold cross-validation, repeated five times for each model. Figure 6 and Figure 7 above show the performance of the standalone and stacked models. Notably, the stacked model outperforms the individual candidate models, prompting us to focus on predicting smoke and drink status based on body signals using the stacked model. Additionally, to further improve the stacked models performance, we can fine-tune the hyperparameters to achieve even better results.

#### 4.1 Model Performance Evaluation

Based on the stacked models, the prediction accuracy of the test dataset is 67.28% and 62.33% for drinking and smoking status models, respectively, which both outperform the baseline model.



**Figure 8 Confusion Matrix of Test Predictions for Drinking Status (left) and Smoking Status (right)**

## 4.2 Feature Importance

For each of the selected best models, we analyze their feature importance. Among the top 5 important features of the drinking status prediction model, LDL cholesterol (HDL\_chole) and age, followed by 'hemoglobin,' 'SGOT\_ALT,' and 'gamma\_GTP,' are the most significant contributors. These features prominently influenced four out of the five best base models.

In the case of the smoking status prediction model, 'gamma\_GTP' is considered important by four out of five models, and age is deemed significant by three out of five models.

In summary, LDL cholesterol, gamma glutamyl transpeptidase (gamma\_GTP), and age emerge as the primary features that play a pivotal role in predicting both drinking and smoking status.

**Table 3 Top 5 Important Features for Drinking Status Prediction**

Models	1	2	3	4	5
<b>Logistic Regression</b>	sex	hemoglobin	HDL_chole	SGOT_ALT	sight_left
<b>GB</b>	gamma_GTP	age	HDL_chole	SGOT_ALT	hemoglobin
<b>SVM</b>	gamma_GTP	age	SGOT_ALT	HDL_chole	triglyceride
<b>Random Forest</b>	gamma_GTP	HDL_chole	age	triglyceride	hemoglobin
<b>AdaBoost</b>	SGOT_ALT	HDL_chole	age	SGOT_AST	serum_creatinine

**Table 4 Top 5 Important Features for Smoking Status Prediction**

Models	1	2	3	4	5
<b>Logistic Regression</b>	hear_left	sight_left	SBP	bmi	HDL_chole
<b>GB</b>	sex	age	hemoglobin	gamma_GTP	bmi

---

<b>SVM</b>	age	gamma_GTP	waistline	HDL_chole	triglyceride
<b>Random Forest</b>	sex	hemoglobin	gamma_GTP	triglyceride	waistline
<b>AdaBoost</b>	age	SBP	BLDS	SGOT_ALT	gamma_GTP