The Impact on Body Signals of Smoking and Drinking: The Intersection

Yuesong Huang
University of Rochester
Rochester, NY
yhu116@u.rochester.edu

Junhua Huang
University of Rochester
Rochester, NY
jhuang77@u.rochester.edu

Yuyang Wang
University of Rochester
Rochester, NY
ywang383@u.rochester.edu

Boyang Wang
University of Rochester
Rochester, NY
bwang55@u.rochester.edu

***Abstract*—This study explores the complex relationship and impact of smoking, alcohol consumption, and health. We analyzed the *Smoking and Drinking Dataset with Body Signals* from Kaggle, a large dataset from South Korea [2]. We used correlation analysis, machine learning including logistic regression, SVM, XGBoost, and PCA to explore how these habits affect various health indicators. The results of the study show that smokers and drinkers of alcohol have significant physiological differences compared to non-smokers and non-drinkers, particularly in terms of blood pressure and cholesterol levels. These results provide health risks and important insights into the health consequences of these lifestyle choices.**

I. Introduction

Smoking and drinking are globally prevalent lifestyle habits with profound implications on individual health [9]. Smoking is a leading cause of several chronic diseases, while significant public wellness concerns are associated with alcohol consumption [10]. Studies from Tan et al. indicate that smoking not only severely increases the lay on the line of cardiovascular diseases (CVD) and all-cause death rate but also amplifies these risks when combined with high blood pressure [7], [8]. Particularly in populations like Chinese males, where smoke rates are high, the impact is notable with increased mortality rates from stroke and anemia heart undefined (IHD). This underscores the urgent need for effective smoke cessation programs, especially in high-prevalence regions, to mitigate the health charge posed by this international issue [8].

The interplay of behavioral mechanisms linking smoking and alcohol consumption is not neglectable [3]. It strongly influenced our team, combined with members with varied health backgrounds and subjective experiences. These factors contributed to a unique perspective in our research. Our team's direct experiences with the wellness consequences of smoking and alcohol significantly impelled our academic interest and dedication. In our research, we utilized health data from the Korean insurance sector [2]. Following thorough information cleaning and processing, we employed several sophisticated machine learning models to investigate potential correlations between subjective wellness data and an individual's smoking and drinking habits.

II. Literature Review

Although it's common sense tobacco and alcohol are harmful, in a study from China, it was discovered that current smokers are 65% less likely to seek health care compared to former smokers, and regular drinkers are also less inclined to seek medical assistance [1]. Moreover, another study from China shows that men who smoke heavily and have high blood pressure are at a significantly higher risk of dying from cardiovascular diseases, including heart disease and stroke, underscoring the urgent need for combined efforts in smoking cessation and blood pressure control [8]. A Romanian study found that tobacco and alcohol consumption among students can be predicted based on their physical activity levels and demographics, revealing a high prevalence of harmful tobacco use and alcohol consumption [6]. Similarly, the Research by Abirami et al. highlights how socio-economic and socio-cultural factors, like lifestyle and peer influence, significantly impact the early onset of smoking and drinking habits in adolescents [5]. The study at King Saud University found that 14.5% of undergraduate students smoke, with male students (32.7%) smoking more than female students (5.9%) [4].

These findings suggest a significant correlation between smoking and drinking and the propensity to seek health care. We also find significant work has been done to predict whether one smoke or drink which inspires us with the model.

III. Data

Our dataset includes 991,346 individual records, the research focuses on how these prevalent lifestyle choices influence various health indicators [2]. TABLE I provide the statistical information of the dataset.

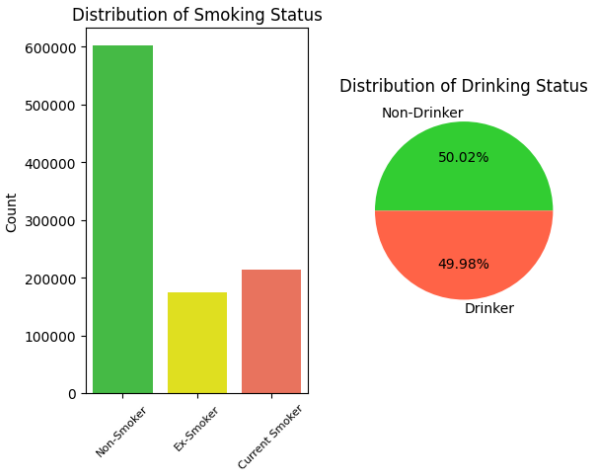
	age	height	weight	waistline
mean	47.61	162.24	63.28	81.23
std	14.18	9.28	12.51	11.85
50%	45.0	160.0	60.0	81.0
min	20.0	130.0	25.0	8.0
max	85.0	190.0	140.0	999.0

	sight_left	sight_right	hear_left	hear_right
mean	0.98	0.98	1.03	1.03
std	0.61	0.6	0.17	0.17
50%	1.0	1.0	1.0	1.0
min	0.1	0.1	1.0	1.0
max	9.9	9.9	2.0	2.0

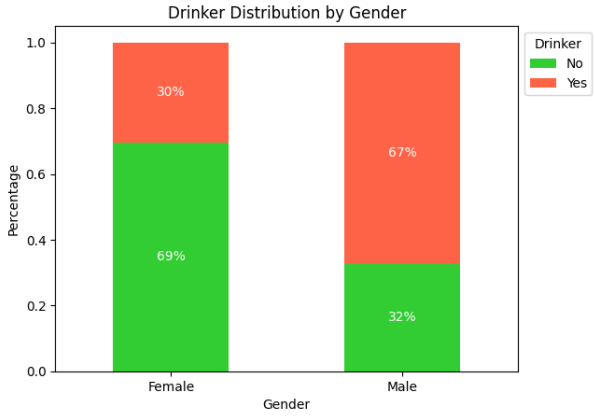
	SBP	DBP	BLDS	tot_chole
mean	122.43	76.05	100.42	195.56
std	14.54	9.89	24.18	38.66
50%	120.0	76.0	96.0	193.0
min	67.0	32.0	25.0	30.0
max	273.0	185.0	852.0	2344.0

	HDL_chole	LDL_chole	triglyceride	hemoglobin
mean	56.94	113.04	132.14	14.23
std	17.24	35.84	102.2	1.58
50%	55.0	111.0	106.0	14.3
min	1.0	1.0	1.0	1.0
max	8110.0	5119.0	9490.0	25.0

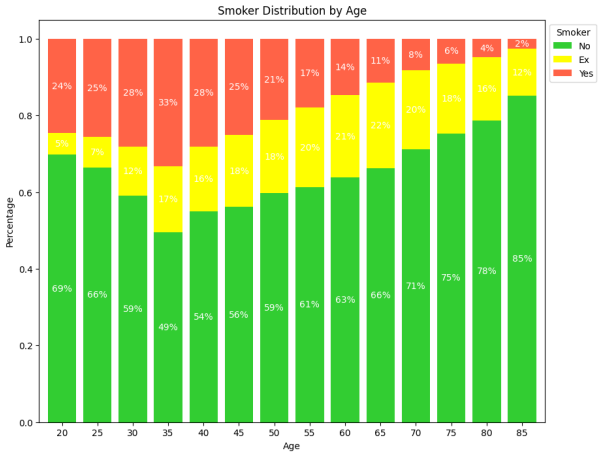
TABLE I: Feature Description



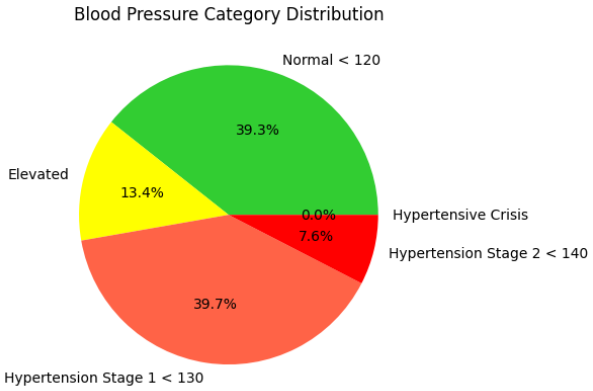
The two graphs above show the distribution of smoking and drinking. The bar graph on the left is broken down into three categories: non-smokers, former smokers, and current smokers. It shows that non-smokers are the most numerous. The pie chart on the right shows the percentage of drinkers and non-drinkers, which are close, but non-drinkers are slightly more numerous by a tiny margin of 0.02%. These two graphs show that people make different choices about smoking and drinking, which reflects their health habits and also indicates the fairness of the dataset.

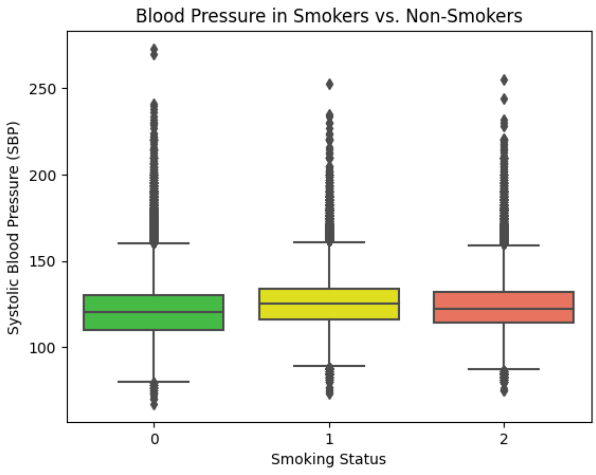


The bar graph above illustrates the proportion of individuals who drink alcohol inside each gender category. For females, a majority (69%) are indicated as non-drinkers, while 30% are drinkers. Conversely, the male person undefined shows a high portion of drinkers (67%) compared to non-drinkers (32%). This data suggests a marked difference in alcoholic beverage consumption between genders, with a substantially higher symmetry of males reporting as drinkers compared to females. The stark contrast shows the significance of gender as a factor in analyzing alcohol use patterns.

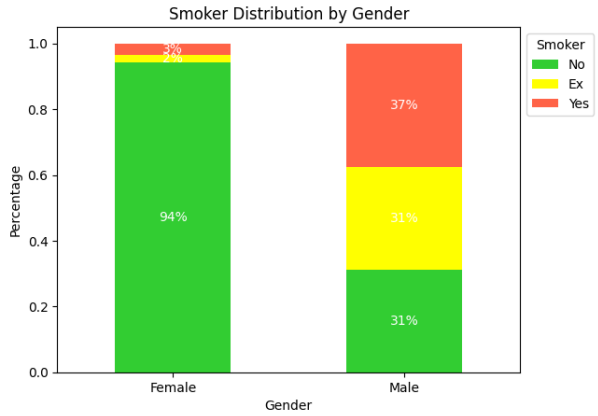


The bar graph above shows that the proportion of non-smokers has remained a major percentage with age. Also, the proportion of older people who quit smoking has increased, possibly because of health problems or knowledge of the long-term effects of smoking. The proportion of people who still smoke decreases with age from 35, possibly because older people quit more often or because smoking causes health problems that shorten life expectancy if there's not any selection bias.

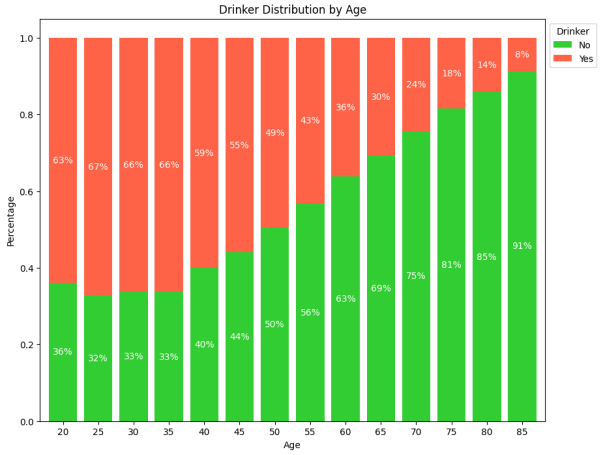




The graphs above show the relationship between smoking status and Systolic Blood Pressure(SBP). Despite the similar medians, the statistical distribution of SBP in current smokers (`SMK_stat_type_cd = 2`) is somewhat higher compared to the other groups. The pattern from the graph reveals the possible correlation between blood pressure and individuals' smoking status. Therefore, further statistical analysis would be required to test for meaning and control for potential confounding factors. The abundance of outliers also indicates a potentiality for extreme point SBP values within each category, which may justify extra investigation into individual wellness profiles.



This bar graph compares the smoke status among different genders. For females, the large majority (94%) are categorized as non-smokers, spell the remaining are evenly split between former smokers (Ex) and current smokers (Yes), both representing a very small percentage (3% and 2% respectively). In contrast, the distribution among males is more undefined spread: 31% are non-smokers, 31% are former smokers, and 37% are current smokers. This visual representation highlights a significant gender disparity in smoke habits, with a larger proportion of males being current or former smokers compared to females.

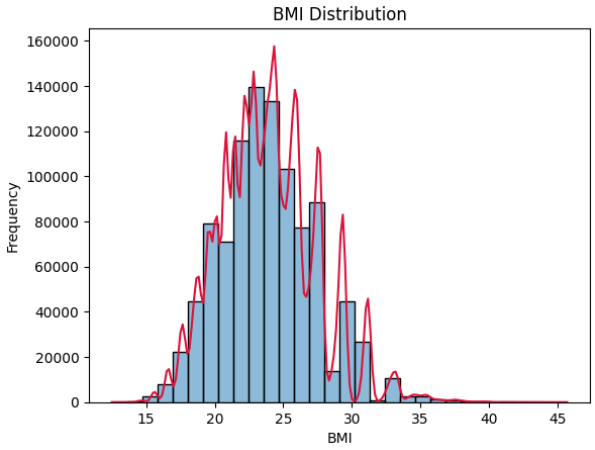


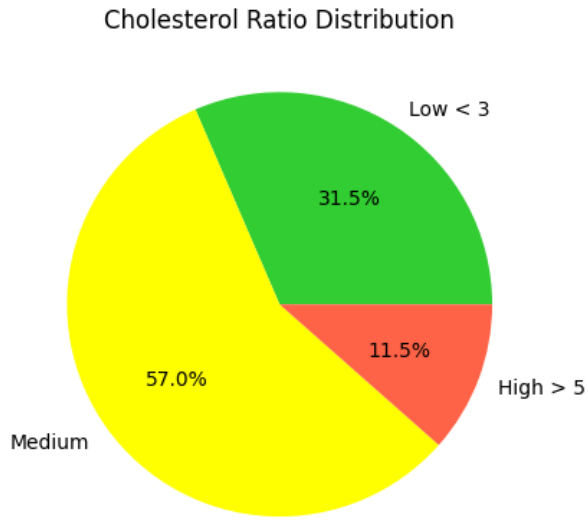
The graphs above show a certain proportion of individuals across all age groups continue to consume alcohol, but the younger age groups tend to have a higher rate of alcohol consumption. This could indicate a more widespread issue of young people consuming alcohol in social situations under the impact of cultural norms and peer pressure. The percentage of persons who consume alcohol declines with age, which could be an indication of people's propensity to alter their lives as they get older. In general, many people may decide to cut back on or give up drinking when health issues gain greater attention.

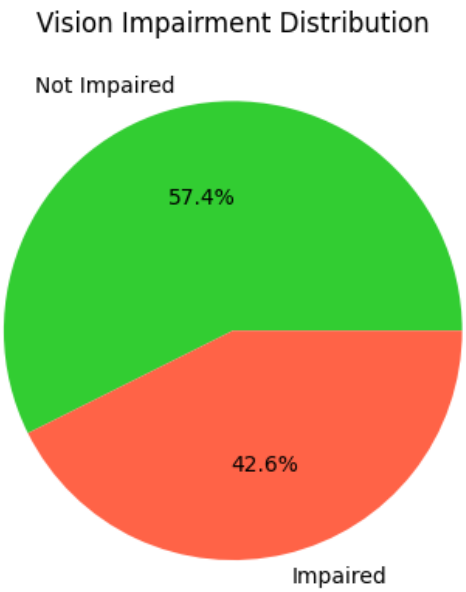
IV. Methods

In our analysis, we applied a combination of data mining and machine learning techniques to predict smoking and drinking status based on body signals. The methodology was chosen to effectively process, analyze, and interpret the complex dataset to achieve our goals.

A. Feature Engineering







The `SMK_stat_type_cd` and `DRK_YN` columns were converted into binary and ternary forms to facilitate the feature engineering process. In machine learning models, this transformation is essential for managing categorical data. Furthermore, we picked a variety of features, such as clinical measurements (blood pressure, cholesterol levels, liver function tests), health indicators (BMI, waistline), and demographic information (sex and age). To maintain consistency and improve model performance, these features were standardized using `StandardScaler` from `scikit-learn` library.

B. Dimensionality Reduction

Principal Component Analysis (PCA) was utilized as a dimensionality reduction technique. PCA helped in identifying the most significant features by transforming the original features into a set of linearly uncorrelated components. We plotted the cumulative sum of the explained variance ratio to determine the optimal number of components, ensuring that the reduced dataset retained the majority of the variance in the original dataset.

C. Classification Model

Our classification models included Support Vector Machine (SVM), XGBoost, and Logistic Regression. `GridSearchCV` was used to optimize the logistic regression model and determine the ideal set of hyperparameters. Because of their effectiveness in managing big datasets and high-dimensional spaces, XGBoost classifiers were deployed. Because SVM models with linear kernels perform well in binary classification tasks, they were employed. The performance of each model was assessed using confusion matrices, accuracy, and F1 score.

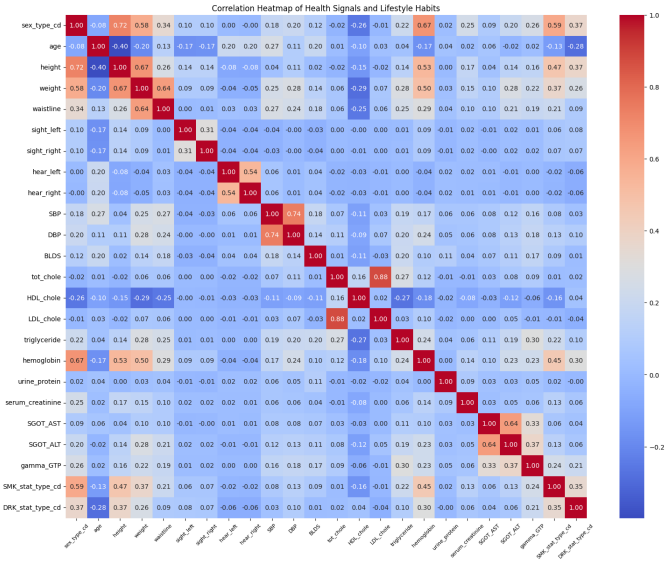
These models were selected due to their shown ability to perform binary classification tasks, particularly when predicting health outcomes. SVM gives a dependable method for classification with a defined margin, XGBoost uses gradient-boosting techniques for increased accuracy, and Logistic Regression delivers a strong baseline.

D. Evaluation

For each model, we created classification reports to provide insight into its performance. We generated accuracy and F1 scores to test the algorithms' ability to properly detect smoking and drinking status. Confusion Matrices also aid in demonstrating how well each model performs in terms of true positives, true negatives, false positives, and false negatives.

V. Results

A. Correlation Analysis



The Confusion Matrix shows a robust correlation (87.7%) between Low-Density Lipoprotein Cholesterol (LDL) and total cholesterol, highlighting their close association and the important role of LDL in cardiovascular risk. There were also moderate correlations between height and sex (72.3%), weight and sex (58.2%), and weight and height (66.9%), indicating that physical characteristics such as body stature and weight varied between genders and were proportional to height. Similarly, the correlation (53.7%) between the right and left ears indicated symmetry in hearing capacity. Blood pressure data, both systolic and diastolic, were found to be associated (74.1%), confirming many medical theories. Hemoglobin levels correlated moderately with height (53.2%) and sex (66.9%), indicating probable biological differences in blood health. The liver enzymes (SGOT ALT and SGOT AST) show a moderate correlation (64.2%), indicating their comorbidity in liver diseases. Smoking habits differed by gender, with a correlation of 59.1% indicating behavioral differences in smoking between genders.

B. Aprori Itemsets Analysis

Support	Itemsets
0.943339	urine_protein=1.0
0.607700	SMK_state=0.0
0.575224	SMK_state=0.0, urine_protein=1.0
0.531010	sex_type=1.0
0.500187	DRK_state=0.0
0.499813	DRK_state=1.0
0.498956	sex_type=1.0, urine_protein=1.0
0.472472	DRK_state=0.0, urine_protein=1.0
0.470867	DRK_state=1.0, urine_protein=1.0
0.468990	sex_type=0.0

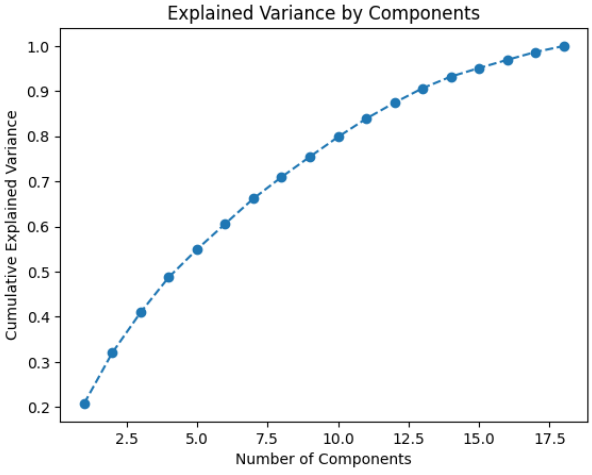
TABLE II: The 10 most frequent sets

Support	Itemsets
0.326586	DRK_state=0.0, sex_type=0.0
0.315914	SMK_state=0.0, DRK_state=0.0, sex_type=0.0
0.309762	DRK_state=0.0, sex_type=0.0, urine_protein=1.0
0.299769	SMK_state=0.0, DRK_state=0.0, sex_type=0.0, urine_protein=1.0
0.215822	SMK_state=2.0
0.215294	SMK_state=0.0, DRK_state=1.0
0.206279	sight_right=1.0
0.203424	SMK_state=0.0, DRK_state=1.0, urine_protein=1.0
0.203177	sight_left=1.0
0.202862	SMK_state=2.0, urine_protein=1.0

TABLE III: The 10 least frequent sets

We also analyzed the frequent itemset using the Apriori method. As a result, we found that the vast majority of the population (94.33%) had normal urine protein levels, indicating generally good kidney health in the dataset. Non-smokers made up 60.77% of the overall population, with 57.52 % having normal urine protein levels, showing that there may be a relationship between not smoking and kidney health. The proportion of females in the dataset was slightly greater (53.10%), and a considerable number of females (49.90%) had normal urine protein levels. Drinking habits were spread uniformly, with non-drinkers and drinkers having equivalent healthy urine protein levels. The lower frequency itemset shows that males, who make up a smaller fraction of the population, are less likely to abstain from alcohol and smoking. There were also fewer people with good eyesight, but it was beyond the scope of our research.

C. Principal Component Analysis(PCA)



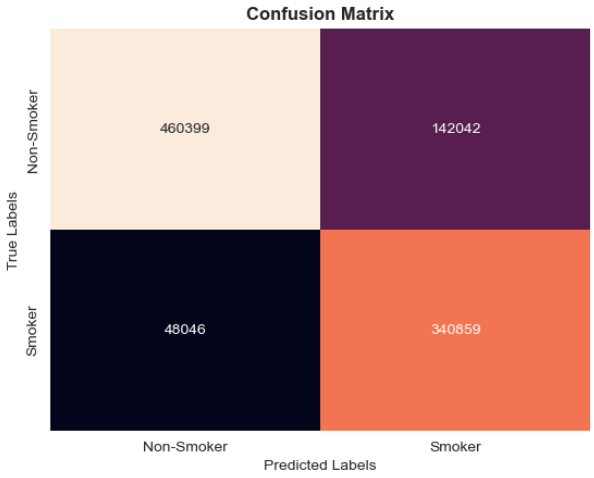
This graph indicates the cumulative explained variance, which measures how much information is kept regarding variance. This demonstrates how data dimensionality can be reduced while still retaining the majority of the information. The graph shows a steep increase in explained variance with the initial components, suggesting that they hold the most significant information. In contrast, the marginal increase in explained variance diminishes with additional components. These visualizations together underline the importance of identifying key variables that capture the essence of the dataset without the need for all original variables, hence optimizing the analytical efficiency.

D. Logistic Regression

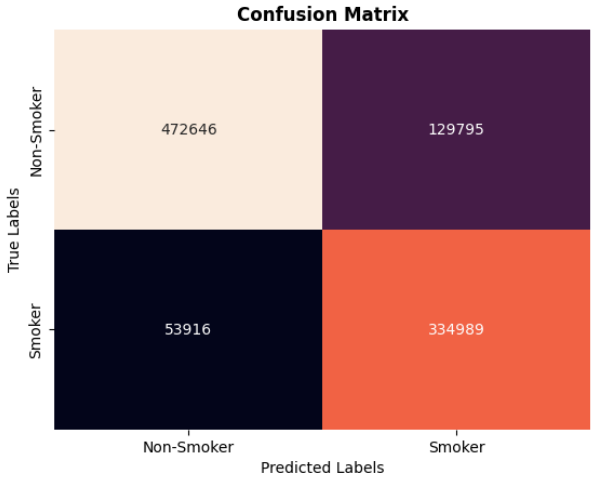
The Logistic Regression model yielded significant insights into the smoking and drinking habits of individuals based on the selected body signal indicators. The performance of the model was quantitatively assessed through a range of metrics, as detailed in the classification reports.

1) Smoking Status Prediction: The classification report for smokers and non-smokers showed the model's proficiency with a precision of 0.91 for non-smokers and 0.71 for smokers. The model was particularly effective in recall for smokers at 0.88, indicating a high sensitivity. The F1 score, a harmonic mean of precision and recall, was 0.83 for non-smokers and 0.78 for smokers, signaling a balanced performance. The accuracy for this category was calculated to be approximately 80.83%.

2) Drinking Status Prediction: For drinking status, the model achieved a balanced precision and recall of 0.72 for both non-drinkers and drinkers, with an F1 score of 0.72 for each group. The accuracy for predicting drinking status was slightly lower at 72.28% compared to smoking status.



3) Confusion Matrices: The Confusion Matrix for smoking status shows a true positive count of 340,859 for smokers, indicating a high number of correct predictions. However, the model also had a false negative count of 48,046, suggesting some smokers were incorrectly labeled as non-smokers.



The Confusion Matrix for drinking status presents a true positive count of 356,186 for drinkers, and a true negative count of 360,317 for non-drinkers, indicating a robust predictive ability for both classes.

4) Overall Model Performance: The Logistic Regression model took approximately 3 hours to complete the training and prediction processes of 320 fits on our dataset. When aggregating the results, the model presented an overall accuracy of approximately 76.55% and an F1 score of 75.18%. These results underscore the Logistic Regression model's utility as a reliable classifier in the domain of public health research, with substantial implications for preventive health measures.

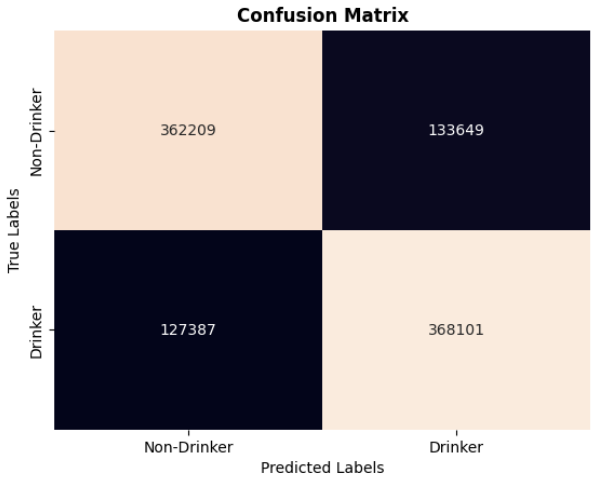
E. XGBoost

The XGBoost model, an advanced gradient boosting framework, was implemented to predict the smoking and drinking status from the health-related features dataset. The model's predictions were analyzed and quantified through classification reports and confusion matrices.

1) Smoking Status Prediction: The XGBoost model achieved a precision of 0.90 for non-smokers and 0.72 for smokers, suggesting a higher reliability in predicting non-smokers. The recall was higher for smokers at 0.86, indicating that the model is quite sensitive to identifying smokers. The F1 score for non-smokers stood at 0.84, while smokers had an F1 score of 0.78, demonstrating a balanced performance between precision and recall for both classes. The overall accuracy of the model for smoking prediction was approximately 81.47%.

2) Drinking Status Prediction: For the prediction of drinking status, the model's precision and recall were relatively balanced with scores of 0.74 for non-drinkers and 0.73 for drinkers. The F1 score mirrored this balance, standing at 0.74 for both classes. The model's accuracy for drinking status prediction was 73.67%.

3) Confusion Matrices: The Confusion Matrix for smoking status (first image) indicates a true positive rate of 334,989 for smokers and a true negative rate of 472,646 for non-smokers.



The Confusion Matrix for drinking status (second image) shows a true positive rate of 368,101 for drinkers and a true negative rate of 362,209 for non-drinkers.

4) Overall Model Performance: The Logistic Regression model only took approximately 30 minutes to complete the training and prediction processes of 320 grid fits on our dataset. The aggregate results from the model give us an average accuracy of approximately 77.57% and an average F1 score of 76.15% across the predictions for both smoking and drinking status. These figures highlight the effectiveness of the XGBoost model in classifying individuals based on the provided health indicators.

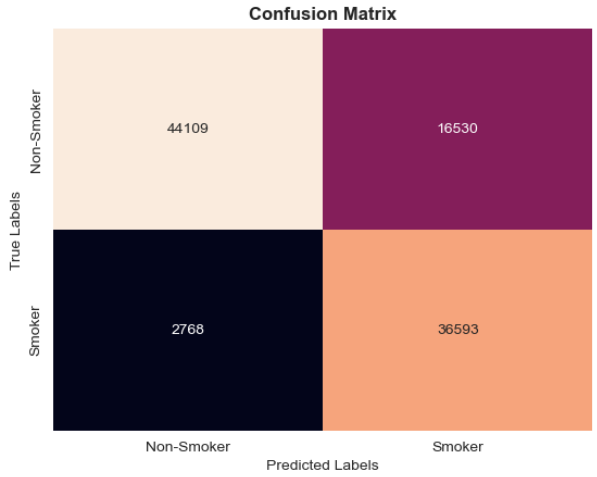
The findings show the capability of ensemble learning methods like XGBoost in handling complex predictive tasks and emphasize their potential utility in public health domains for identifying individuals at risk due to their smoking and drinking behaviors.

F. Support Vector Machine (SVM)

The SVM model was also tried to understand its efficacy in predicting smoking and drinking behaviors based on health indicators.

1) Drinking Status Prediction: The SVM produced balanced results with a precision of 0.74 for both non-drinkers and drinkers. The recall was similarly equitable, with non-drinkers at 0.73 and drinkers at 0.74, leading to an F1 score of 0.74 for both categories. The accuracy of the drinking status predictions was 73.67%.

2) Smoking Status Prediction: For smoking status, the SVM showcased a high precision of 0.94 for non-smokers and a notable recall of 0.93 for smokers, leading to an F1 score of 0.82 for non-smokers and 0.79 for smokers. The accuracy for this subset was 80.58%.



3) Confusion Matrices: The confusion matrix for drinking status displays a true positive count of 368,101 for drinkers, affirming the SVM's predictive strength, while The confusion matrix for smoking status reveals a strong ability to correctly identify smokers, with a true positive count of 36,295.

4) Overall Model Performance: The SVM model's computational time was over one hour for a subset of 100,000 samples with 10 fits, a significant increase in time per sample compared to the full dataset analyses done by Logistic Regression and XGBoost. Specifically, Logistic Regression processed the entire dataset of 1 million samples in three hours with 320 grid fits, and XGBoost completed its processing in 30 minutes for the same dataset size and number of grid fits. This indicates a higher computational cost for the SVM, which may impact its scalability and practical application for larger datasets.

G. Discussion

The SVM's results, while robust, come with a computational cost that suggests a trade-off between predictive performance and efficiency. This consideration is vital for real-world applications where time constraints are a factor.

The comparison of accuracy and F1 scores across different models emphasizes the SVM's capabilities in classification tasks. However, the extended computational time highlights the need for more efficient implementations or the use of a reduced feature set for SVM applications. For larger datasets, alternative methods such as Logistic Regression or XGBoost, which offer faster processing times, might be more practical.

VI. Conclusion

This journey of discovery has revolutionized the prespect we think about smoking and drinking. What started as curiosity evolved into a personal awakening to the effects of these habits on our bodies. Driven by machine learning capabilities, our research not only reveals changes in body signals associated with prevalent lifestyle choices, but also motivates us to pursue healthier living.

Our findings are not only exciting but also highly valuable, providing tangible insights into the relationship between lifestyle choices and health indicators. Logistic regression and XGBoost models showed high predictive performance and efficiency, revealing individual physiological differences based on smoking and drinking status. These results may help to raise awareness about public health.

Looking forward, we decided to make many prevalent lifestyle changes positively in the future and continue our research on the impact on body signals of smoking and drinking.

REFERENCES

[1] C. Li and J. Sun, "The impact of current smoking, regular drinking, and physical inactivity on healthcare-seeking behavior in China," *BMC Health Services Research*, vol. 22, no. 1, 2022. doi:10.1186/s12913-022-07462-z.

[2] Soo.Y, "Smoking and drinking dataset with body signal," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset. [Accessed Dec. 8, 2023].

[3] H. J. Little, "Behavioral mechanisms underlying the link between smoking and drinking," *Alcohol research & health: the journal of the National Institute on Alcohol Abuse and Alcoholism*. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6709747/. [Accessed Dec. 8, 2023].

[4] A. Mandil et al., "Smoking among university students: A gender analysis," *Journal of Infection and Public Health*, vol. 3, no. 4, pp. 179–187, 2010. doi:10.1016/j.jiph.2010.10.003.

[5] M. S. Abirami, B. Vennila, E. L. Chilukalapalli, and R. Kuriyedath, "Retracted article: A classification model to predict onset of smoking and drinking habits based on socio-economic and sociocultural factors," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 3, pp. 4171–4179, 2020. doi:10.1007/s12652-020-01796-4.

[6] G. Badicu, S. H. Zamani Sani, and Z. Fathirezaie, "Predicting tobacco and alcohol consumption based on physical activity level and demographic characteristics in Romanian students," *Children*, vol. 7, no. 7, p. 71, 2020. doi:10.3390/children7070071.

[7] X. Dai et al., "Health effects associated with smoking: A burden of proof study," *Nature Medicine*, vol. 28, no. 10, pp. 2045–2055, 2022. doi:10.1038/s41591-022-01978-x.

[8] J. Tan et al., "Smoking, blood pressure, and cardiovascular disease mortality in a large cohort of Chinese men with 15 years follow-up," *International Journal of Environmental Research and Public Health*, vol. 15, no. 5, p. 1026, 2018. doi:10.3390/ijerph15051026.

[9] C. Stanley, "How smoking and drinking affect the body," MEH, [Online]. Available: https://www.mountelizabeth.com.sg/health-plus/article/how-smoking-and-drinking-affects-the-body. [Accessed Dec. 8, 2023].

[10] K. J. Mukamal, "The effects of smoking and drinking on cardiovascular disease and risk factors," *Alcohol research & health: the journal of the National Institute on Alcohol Abuse and Alcoholism*. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6527044/. [Accessed Dec. 8, 2023].