Analyzing the Impact of Smoking and Drinking on Health Metrics in Korea: A Data-driven Predictive Approach

Priyanka Logasubramanian, Md Amiruzzaman*, Ashik Ahmed Bhuiyan

Department of Computer Science, West Chester University, United States.

Email: {pl1032847, mamiruzzaman, abhuiyan}@wcupa.edu

Received: March 23, 2024; Revised: May 20, 2024; Accepted: June 26, 2024; Published: July 10, 2024

Abstract

This investigation thoroughly explores the effects of smoking and alcohol consumption on health measures within the population of South Korea, using a comprehensive dataset from the National Health Insurance Service of Korea. The study aims to build and validate predictive models by applying logistic regression methodologies, outlier detection, and cross-validation techniques. These models are designed to identify patterns in smoking and drinking behaviors, offering predictive accuracy. The analysis clarifies the relationship between lifestyle choices and key health indicators such as liver enzyme activity, blood fat levels, and heart and blood vessel measurements. The outcomes of this study are ready to significantly contribute to the development of targeted public health strategies, aiming to lessen the risks associated with lifestyle-induced health conditions.

Keywords: Behavioral Analysis, Drinking Behavior, Health Risk Factor, Lifestyle-Related Diseases, Smoking Patterns.

1 Introduction

Embarking on an empirical inquiry into the nexus between tobacco use, alcohol consumption, and their subsequent health outcomes, this research utilizes a comprehensive dataset provided by the National Health Insurance Service of South Korea. Centering on pivotal health metrics, including but not limited to arterial pressure, lipid concentrations, and hepatic enzyme levels, the study endeavors to shed light on the health consequences stemming from these lifestyle practices. The dataset undergoes meticulous processing and analytical scrutiny, emphasizing data sanitization, categorization of variables, and transformational processes to set a foundation for a sophisticated exploration of health indicators and their correlation with lifestyle behaviors.

Our study thoroughly examines the effects of smoking and drinking on health in South Korea. We aim to understand these habits and their impact on health better. This is important because it can help shape health policies that directly target the dangers of smoking and alcohol use. Using logistic regression analysis on detailed health data, we bring a new approach to exploring how these behaviors relate to health issues. Our work uses advanced statistical methods, such as ROC curves and AUC scores, to ensure our findings are reliable. The knowledge we gain will help build effective public health strategies specifically designed for South Korea.

Motivated by the challenges, the main contributions of this work are summarized as follows:

- Health Metrics Analysis: The study assessed the impact of smoking and drinking on key health indicators, such as liver health and cholesterol levels, in the South Korean population by analyzing national health insurance system data.

- Development of Prediction Models: Mathematical techniques, specifically logistic regression, were applied to create models that forecast whether a person smokes or drinks. These models underwent rigorous testing to verify their accuracy.

- Health Risks Assessment: The investigation examined how smoking and drinking relate to negative health indicators such as liver enzymes, cholesterol balance, and blood circulation metrics.

- Statistical Methodology Application: Statistical tests and calculations confirmed that the models provided accurate and trustworthy information.

- Public Health Strategy Contribution: Insights were provided that will help health officials design programs to reduce health problems caused by lifestyle habits.

- Health Policy Implications: Valuable information was supplied that could help shape health policies and strategies to reduce smoking and drinking.

- Novelty in Research: New understanding was brought to the complicated ways that lifestyle choices impact health, showing that a wide range of factors should be considered.

1.1 Problem Statement

1. Understanding Health Risks: The aim is to assess the severity of health risks posed by smoking and drinking within the Korean population, focusing on the correlation between these habits and adverse health metrics.

2. Predicting Unhealthy Habits: This aspect of the research seeks to determine if health data can be effectively used to predict individuals likely to engage in excessive smoking or drinking.

3. Improving Health Policies: The final objective is to leverage the findings to inform and enhance public health policies, specifically by developing strategies that can prevent the initiation of smoking and drinking habits.

2 Literature Review

This study investigates the transformative impact of statistical models on medical diagnostics and patient care, specifically exploring how various researchers' application of logistic regression and machine learning techniques contributes to the advancement of personalized treatment strategies and predictive healthcare analytics.

In medical research, the advancement of statistical models has sparked significant strides in diagnostics and personalized medicine. Boateng and Abaye (2019) juxtaposed with Sung et al. (2022), showcases a dedication to empirical precision. Boateng and Abaye (2019) harness logistic regression's flexibility and Sung et al. (2022) apply machine learning to gait analysis, underscoring methodological meticulousness. Similarly, Fazakis et al. (2021) on T2DM risk and Nopour et al. (2022) research using logistic regression for COVID-19 diagnosis emphasize the critical nature of accuracy in patient outcomes. Lamb et al.'s (2008), examination of the t-test, further solidify the universal call for statistical thoroughness. This collective effort reflects the overarching goal of enhancing predictive analytics in healthcare. Building on the detailed examination of statistical methodologies in the context of medical research, we can further elaborate on the distinct approaches and their collective impact:

In an elaborate study, Boateng and Abaye (2019) used logistic regression to delve into the complex interplay of health-related factors. Their analytical rigor brought to light distinct patterns and connections that enhance the comprehension of health issues. Such detailed analysis underpins the importance of clarity in

methodology, paving the way for further exploration in the field.

Similarly, in the detailed study by Nopour et al. (2022), logistic regression is employed alongside the Chi-square test to scrutinize data from 400 patients, aiming for a rapid and accurate COVID-19 diagnosis. This approach emphasizes swift virus identification, highlighting the model's precision in distinguishing positive cases critical during a health emergency.

When comparing both studies, it's evident that logistic regression is a versatile tool in the field of medical diagnostics. Nopour et al. (2022) study showcases the method's immediacy in crisis response, while Boateng and Abaye's research highlights its role in providing foundational knowledge that informs future scientific inquiries. Together, these studies exemplify the significant impact of logistic regression on advancing medical diagnostics and managing public health crises.

In the study by Sung et al. (2022), advanced machine learning methods are applied to evaluate how stroke severity affects walking patterns. Their technique, known as recursive feature elimination with cross-validation, helps pinpoint key walking features that signal the level of stroke impact. By focusing on how symmetrically a person walks, the researchers can accurately gauge stroke severity, potentially aiding in better patient diagnosis and treatment planning.

Sung et al. (2022) yielded impressive results in assessing stroke severity through gait analysis. Their Random Forest classifier achieved a remarkable classification accuracy of 96.01%. These results highlight the potential of machine learning to offer highly accurate and non-invasive diagnostic tools for stroke patients. The study's findings open new avenues for improving patient care and rehabilitation by providing precise insights into stroke severity.

Similarly, Fazakis et al. (2021), in their work titled "Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction," utilized a suite of machine learning algorithms, including data preprocessing and feature selection techniques, to tailor predictions of Type 2 Diabetes Mellitus risk to individual patient profiles. The precision of their predictive model was evaluated using metrics such as the Area Under the Receiver Operating Characteristic Curve (AUC), which gauges the model's accuracy. This methodological approach marks a significant step towards personalized medicine, highlighting the potential of machine learning to revolutionize preventative health strategies and treatment planning.

Incorporating Lamb et al.'s (2008) critical examination of the t-test, the synthesis emphasizes the need for statistical integrity across varied methodologies. While the t-test may differ from logistic regression and machine learning in its application, the underlying principle of methodological integrity is a common thread that ties these studies together. While not focused on specific medical outcomes, Lamb et al. (2008) examination of the t-test emphasized the importance of statistical rigor and clear reporting. Their results underscored the need for methodological integrity in all research endeavors. This finding serves as a reminder that the choice and application of statistical methods are fundamental to the credibility and reliability of research findings.

These research papers strongly align with the principles and analytical approaches pertinent to this project. Each study employs rigorous statistical methodologies to investigate various health-related phenomena, focusing on enhancing patient outcomes through refined diagnostic tools and treatment strategies. Concerning this project, these papers offer rich insights into the practical application of logistic regression and machine learning in healthcare. They collectively provide a framework for understanding how data-driven models can personalize care, a cornerstone of modern medical research. The studies by Boateng and Abaye (2019), Sung et al. (2022) underscore the necessity of methodological integrity and the potential of statistical analysis to inform and improve patient care practices, likely to be key considerations in this project's scope.

3 Case and Methodology

The research looks closely at how smoking and drinking affect health. It is divided into three main parts:

Statistical Exploration, Data Preparation, and Model Performance. Each part plays a key role in understanding how lifestyle choices like smoking and drinking can change health outcomes. Here's an overview of these sections:

3.1 Statistical Exploration

This section details the statistical examination to decode the data trends, particularly how health metrics behave and vary. Through statistical graphs and tests, it explores the effects of smoking and drinking on health, shedding light on patterns crucial for guiding public health decisions. This analytical phase is essential to ensure the data are interpreted correctly and to uncover any unexpected findings that could influence future health recommendations.

3.1.1 Normality Testing of Clinical Data

In the beginning stage of our normality assessment, we looked at histograms for various health measurements, including Gamma-GTP, SGOT_AST, SGOT_ALT, and HDL and LDL cholesterol levels, triglycerides, and blood pressure metrics. In contrast to the anticipated bell-shaped curve indicative of a normal distribution, the histograms demonstrated a significant skew towards lower values, with a notable rightward asymmetry, particularly in the variables associated with liver function and cholesterol.

Subsequent Q-Q plots reinforced these findings, displaying clear departures from the diagonal line representing a normal distribution. The upper quantiles showed the greatest deviation, underscoring the non-normal nature of the data.

The overall assessment revealed that the clinical parameters under consideration do not follow a normal distribution. This necessitates of using non-parametric tests or data transformation techniques for any further statistical analysis to ensure an accurate interpretation of the clinical data.

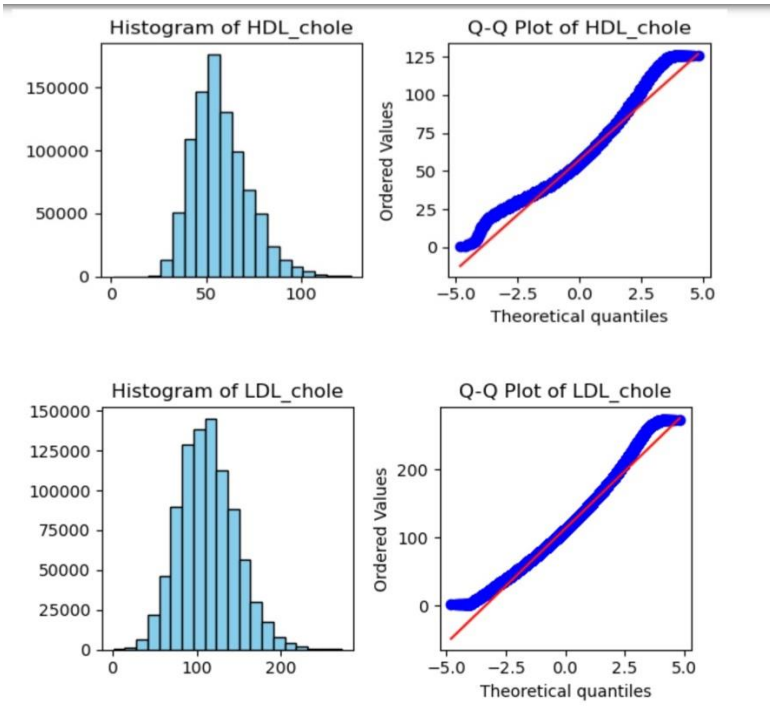

Figure 1. Normality Assessment of HDL & LDL Cholesterol Level

3.1.2 Correlation Matrix and Heatmap Insights:

- Liver Enzyme Correlation: Heatmap analysis indicated a strong positive correlation between 'SGOT_AST' and 'SGOT_ALT', suggesting linked liver health effects.

- Cardiovascular Indicators: Relationships between lipid levels and blood pressure were visualized, hinting at their collective cardiovascular implications.

3.1.3 Visualizing Lifestyle Distributions:

- Smoking Trends: Countplots revealed smoking status distributions across the population, which is crucial for public health planning.

- Drinking Habits: Similar visualizations for drinking status provided insights into behavioral patterns affecting liver and cardiovascular health.

3.1.4 Age-Related Smoking Analysis:

- Middle-Age Smoking Peak: Data indicated higher smoking rates among middle-aged individuals, pointing to a critical target for cessation programs.

- Shift in Smoking Habits: The trend of quitting smoking across age groups may reflect growing health awareness.

3.1.5 Boxplot Analysis of Health Metrics:

- Lifestyle and Health Metrics: Box plots illustrated the varying effects of smoking and drinking on HDL and LDL cholesterol, triglycerides, and blood pressure.

- Unexpected Cholesterol Levels: Non-smokers showed higher LDL cholesterol, challenging traditional views, and highlighting the need for broader health factor considerations.

3.1.6 Evaluating Health Metrics about Smoking Status: A T-Test and P-Value Analysis

Our statistical analysis using t-tests to compare health metrics between current smokers and non-smokers (including past smokers) yielded significant p-values across all variables. Notably, total cholesterol ($p < 0.0001$), HDL cholesterol ($p < 0.0001$), and LDL cholesterol ($p < 0.0001$) were lower in current smokers. Surprisingly, LDL cholesterol levels were higher in non-smokers, contradicting common health perceptions. Compensatory mechanisms or lifestyle factors could influence this not accounted for solely by smoking status. Elevated blood pressure (SBP and DBP with $p < 0.0001$) and liver enzymes (gamma_GTP, SGOT_AST, SGOT_ALT with $p < 0.0001$) in smokers indicate increased cardiovascular and hepatic risks, aligning with established health risks of smoking.

3.1.7 Evaluating Health Metrics in Relation to Drinking Status: A T-Test and P-Value Analysis

Similarly, drinkers versus non-drinkers showed significant differences in health metrics. HDL cholesterol was higher in drinkers ($p < 0.0001$), potentially reflecting the reported cardioprotective effects of moderate alcohol consumption. Conversely, LDL cholesterol was significantly lower in drinkers ($p < 0.0001$), challenging typical risk profiles. Blood pressure was higher in drinkers, with SBP showing a less pronounced increase ($p < 1e-20$) than DBP ($p < 0.0001$), suggesting alcohol's varied impact on vascular health. Liver enzyme levels (gamma_GTP, SGOT_AST, SGOT_ALT with $p < 0.0001$) and triglycerides ($p < 0.0001$) were higher in drinkers, indicating hepatic stress and dyslipidemia commonly associated with alcohol use.

The results reflect the complex interplay between lifestyle choices and health outcomes. While some findings align with conventional health knowledge, such as the association between smoking, increased blood pressure, and liver enzyme levels, others, like the unexpected LDL cholesterol pattern among non-

smokers, warrant further investigation. These contradictions could be due to a variety of confounders, including diet, exercise, genetic predispositions, and socio-economic status, which were not controlled for in the analysis. It's imperative to delve deeper into these relationships, possibly through multivariate analysis, to isolate the individual effects of smoking and drinking from other influencing factors. This nuanced understanding is critical for developing effective public health strategies and individualized patient guidance.

3.2 Data Preparation

Data preparation involves two key steps: splitting the dataset for training and testing, and standardizing the data. Splitting ensures model accuracy on separate sets, while standardizing ensures uniformity in feature scales, crucial for model performance.

3.2.1 Data Splitting

The dataset is divided into an 80-20 split for training and testing for both smoker and drinker models. This separation allows the models to be trained on a substantial portion of the data and then tested for accuracy on a separate set. The smoker model uses 'SMK_stat_type_cd' as the target variable, while the drinker model employs 'DRK_YN'.

3.2.2 Scaling Process

The data is standardized using StandardScaler to ensure uniformity in feature scales. This step is crucial for models like logistic regression, where diverse ranges in features like blood pressure and cholesterol can impact model performance. Scaling adjusts each feature to a common scale, enhancing the model's ability to learn effectively from the data.

3.3 Model Performance

Model performance plays a pivotal role in the study by offering a measure of how well logistic regression models can predict lifestyle behaviors from health metrics. The analysis provides valuable insights into the strength and direction of the relationships between health factors and smoking or drinking habits. With the help of AUC scores and cross-validation, the study confirms the models' accuracy and consistency, underscoring their importance in guiding public health decisions and crafting targeted health interventions.

3.3.1 Logistic Regression Analysis and Model Performance for Smoker Prediction

- Gamma-GTP & Smoking:

The coefficient for gamma GTP is 0.3148, indicating a relevant association with smoking status, suggestive of its impact on liver function.

- Liver Enzymes & Smoking:

SGOT_AST shows a negative association with a coefficient of -0.1835, suggesting that higher liver enzyme levels may be linked to non-smoking statuses.

- Blood Pressure & Smoking:

The model indicates a negative relationship between systolic blood pressure and smoking (Coefficient: -0.1449), while diastolic blood pressure presents a less significant coefficient (0.0786).

- Cholesterol & Smoking:

HDL cholesterol has a negative coefficient (-0.1263), while LDL cholesterol shows a smaller negative association (Coefficient: -0.0372), implying that smoking status may have varying impacts on cholesterol

levels.

Table 1. Coefficients For Each Feature in The Logistic Regression Smoker Model

Feature	Coefficient
HDL_cholestrol	-0.126362
LDL_cholestrol	-0.037282
triglyceride	0.060816
gamma_GTP	0.314801
SGOT_AST	0.060816
DBP	-0.078638

3.4 AUC Scores

In the one-vs-rest (OvR) approach, adopted for multiclass classification in our model, the AUC score for each class, 'current', 'never', and 'quit', is independently computed as though it were a binary classification. This involves applying the trapezoidal rule-based formula,

$$\text{AUC}= \sum_{i=1}^{n-1} \left(\frac{TPR_{i+1}+TPR_i}{2}\right) \times (FPR_{i+1} - FPR_I)$$

where TPR and FPR are the true positive and false positive rates at successive thresholds. The formula calculates the area under the ROC curve, plotted with FPR on the x-axis and TPR on the y-axis, summing the areas of trapezoids formed between the curve and the axis.

For each smoking status class, the AUC is computed against all other classes, and then these AUC scores are averaged, following the macro-averaging method, to yield an overall performance measure. The resulting AUC scores for the 'current' (0.72), 'never' (0.74), and 'quit' (0.66) classes, averaging to an overall AUC of 0.71, demonstrate the model's consistent and strong ability to differentiate between these smoking status categories. This macro-averaged AUC score not only reflects the model's effectiveness in classification tasks but also its potential applicability in epidemiological research, where accurate categorization of smoking status is crucial.

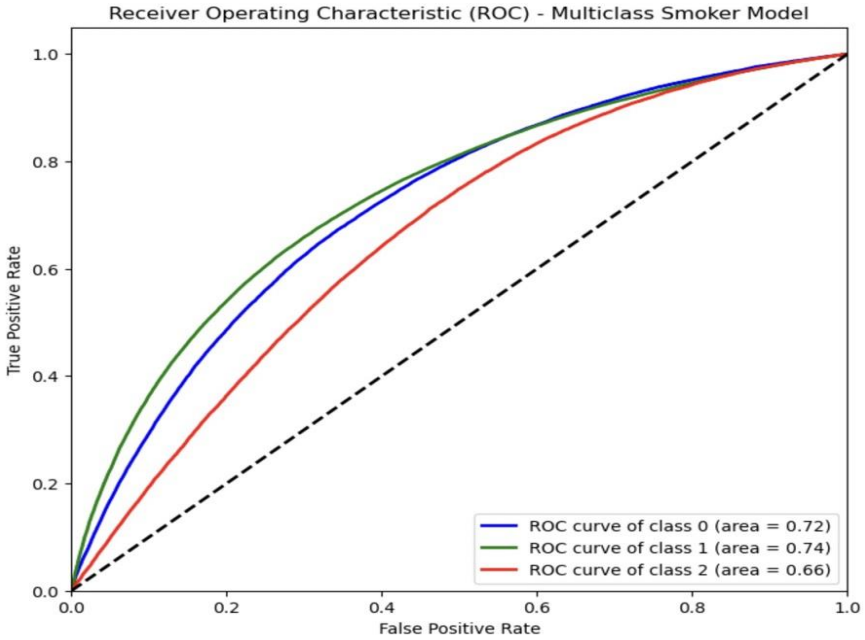

Figure 2. Receiver Operating Characteristic (ROC)-smoker model

In the displayed ROC curve, the solid lines for each class—showing the model's ability to differentiate between smoking statuses—lie above the diagonal dotted line, which represents random chance. This placement indicates that the model is making predictions with a level of accuracy that surpasses mere guessing. Areas under these curves (AUC) quantify this accuracy, with values closer to 1.0 reflecting higher predictive power. Conversely, any curve below the dotted line would suggest a performance worse than random, marking a model with consistently incorrect predictions.

3.5 Coefficients Analysis

The coefficients derived from the logistic regression highlight the differential impact of each health metric on the likelihood of being categorized within each smoking status.

3.6 Cross-Validation

The line graph for the ten-fold cross-validation shows AUC scores consistently around the mean of 0.7056, with expected minor fluctuations. This pattern suggests the model's reliable predictive performance across different data subsets, indicating robustness and a lack of overfitting. The graph underlines the model's dependability for real-world applications.

The logistic regression model for smoking status classification reveals key insights into the relationships between health metrics and smoking behavior. The AUC scores confirm the model's solid predictive strength, although there is room for improvement. Cross-validation results underscore the model's consistency, marking it as a potentially useful tool in clinical settings for risk assessment and in designing public health interventions targeting smoking-related health issues.

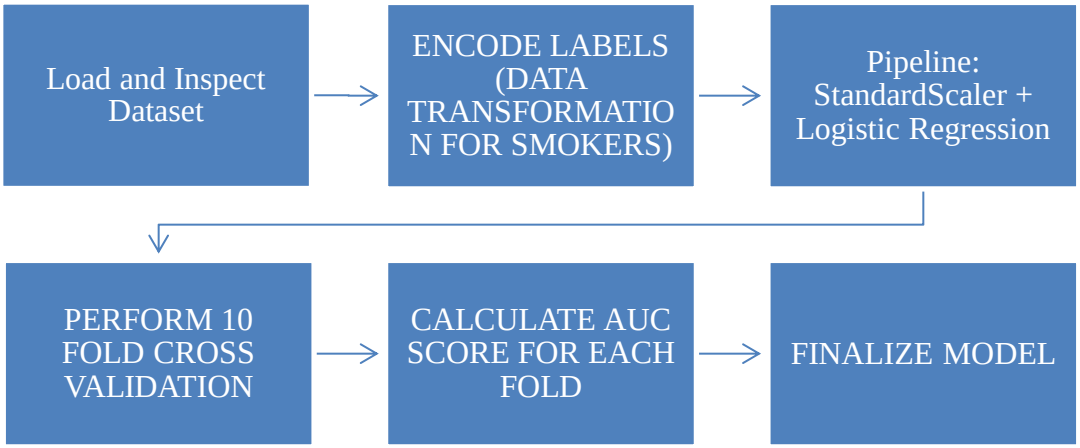

Figure 3. Cross-Validation Workflow for Smoking Classification Model

3.7 Logistic Regression Analysis and Model Performance for Drinkers Prediction

- Gamma-GTP & Drinking: Significant positive correlation (Coefficient: 0.7559), indicating a strong link between alcohol consumption and liver function.
- Blood Pressure & Drinking: Systolic blood pressure (SBP) showed a negative association (Coefficient: -0.2516), while diastolic blood pressure (DBP) had a positive relationship (Coefficient: 0.2552), suggesting complex effects of alcohol on blood pressure.
- Cholesterol & Drinking: Positive correlation with HDL cholesterol (Coefficient: 0.2071) and negative with LDL cholesterol (Coefficient: -0.1245), reflecting alcohol's varied impact on lipid profiles.

Table 2. Coefficients For Each Feature in The Logistic Regression Drinker Model

Feature	Coefficient
HDL_cholestrol	0.207087
LDL_cholestrol	0.124485
triglyceride	0.004123
SBP	0.251613
DBP	0.255192
gamma_GTP	0.755881
SGOT_AST	-0.117200
DBP	-0.108605

3.8 AUC Score

For the binary classification task of discerning drinkers in the dataset, our model calculates the AUC score using the trapezoidal rule, crucial for evaluating its discriminative performance. Achieving an AUC of 0.6872, the model proves moderately effective in distinguishing between drinker and non-drinker categories. This score results from the integration of the area beneath the ROC curve, graphed by plotting the True Positive Rate (TPR) versus the False Positive Rate (FPR) across various decision thresholds.

The formula applied is shown below:

$$\text{AUC}= \sum_{i=1}^{n-1} \left(\frac{TPR_{i+1}+TPR_i}{2}\right) \times (FPR_{i+1} - FPR_i)$$

Assessing the width of each trapezoid along the curve. While the AUC of 0.6872 is modest, it effectively signals the model's capability to separate drinkers from non-drinkers, underscoring its value in research areas where binary classification is essential.

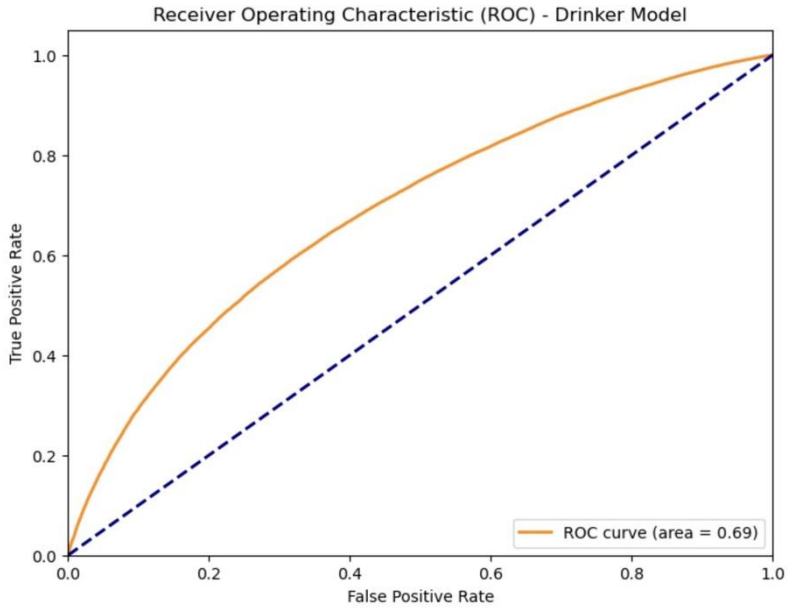

Figure 4. Receiver Operating Characteristic (ROC)-Drinker model

In the ROC curve for the drinker model, the solid orange line signifies the model's capability to identify drinkers, achieving an AUC of 0.69 which demonstrates a fair predictive accuracy beyond random guessing, represented by the dotted line. The area below this dotted line would imply a predictive performance worse than random, where the model's predictions are systematically incorrect.

3.9 Cross-Validation

The line graph illustrates the results of a ten-fold cross-validation for the drinker model, showing AUC scores demonstrating a reliable predictive accuracy level. The scores fluctuate slightly but remain close to a mean of 0.6865, highlighting the model's consistency and moderate predictive strength. Such performance indicates that the model is well-tuned to the variability within the dataset, ensuring its reliability when applied to different data segments in practical settings.

The logistic regression model for drinker classification suggests the significant influence of liver enzymes and blood pressure on drinking behavior. While it demonstrates a reasonable ability to predict alcohol consumption habits, the moderate AUC scores point to the potential for further refinement. Enhancements could include integrating additional health metrics or applying more advanced analytical methods to better capture the nuanced effects of alcohol on health.

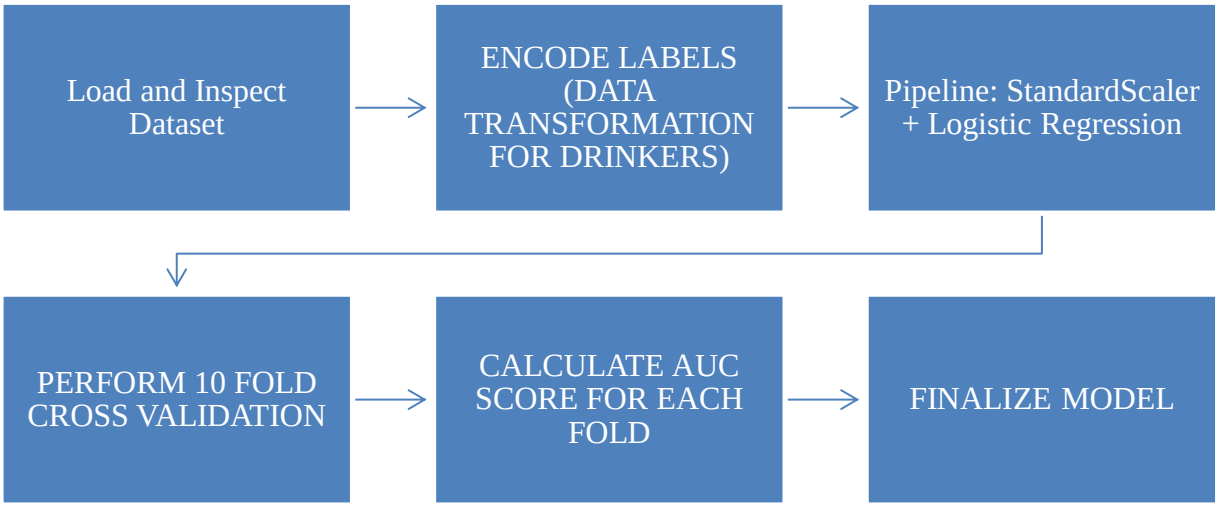

Figure 5. Cross-Validation Workflow for Drinking Classification Model

4 Results

The logistic regression analysis revealed significant correlations between lifestyle habits and various health metrics, shedding light on the potential risks associated with smoking and alcohol consumption. Notably, smoking was strongly correlated with increased Gamma-GTP levels and decreased HDL cholesterol, signifying potential liver and cardiovascular health risks. In contrast, alcohol consumption positively correlated with Gamma-GTP and variable blood pressure readings, underlining its impact on physiological functions. One of the more unexpected findings was the higher LDL cholesterol levels observed in non-smokers, which challenges traditional health perceptions. Regarding model performance, the smoker model displayed good predictive capabilities with AUC scores of 0.72 for current smokers, 0.74 for never smokers, and 0.66 for former smokers. The drinker model achieved a moderate AUC score of 0.6872, demonstrating reasonable effectiveness in distinguishing between drinkers and non-drinkers.

5 Discussion

Our study's findings are crucial for understanding health dynamics in South Korea and have wider relevance for other populations. The correlation between smoking and increased liver and heart health risks raises a global health concern, considering the widespread prevalence of smoking. This aspect of our research could influence international health guidelines and inform smoking cessation programs worldwide.

Moreover, our observation of the complex impact of alcohol consumption on health metrics like HDL cholesterol and liver enzymes has significant implications. These results can aid health professionals in advising patients on the nuanced effects of moderate alcohol consumption, challenging simplistic views on drinking habits.

This research also contributes to a growing body of evidence supporting the importance of personalized healthcare. The differences observed in health impacts among smokers and non-smokers point towards the need for individualized health strategies, considering personal habits, dietary patterns, and genetic predispositions.

In addition, our study provides a framework for public health officials and policymakers. The insights gained can help design more effective, culturally sensitive health promotion campaigns, particularly in countries with similar cultural and lifestyle patterns as South Korea.

Finally, this research holds significant potential for the public, offering crucial insights that could shape the future of health promotion and disease prevention strategies. By understanding the specific health risks associated with smoking and drinking, individuals can make more informed choices about their lifestyle habits. Additionally, public health officials might leverage this data to craft targeted campaigns, which could decrease the prevalence of associated health issues. Healthcare providers could also utilize these findings to offer personalized advice to patients, potentially improving health outcomes on a broad scale. Therefore, this study contributes to the academic discourse and serves as a beacon for public health enhancement.

6 Conclusion

Our research has shown a light on the complex relationship between lifestyle choices and health outcomes, highlighting how smoking and drinking contribute to various health risks in South Korea. These findings underscore the necessity of public health strategies that are both personalized and culturally attuned.

The study's limitations are rooted in its demographic specificity and cross-sectional data analysis. Future studies would benefit from a longitudinal approach to better ascertain the long-term effects of these habits. Moreover, including a more diverse participant pool would enhance the generalizability of the results.

The implications of this study are far-reaching. It provides a foundation for policymakers and health professionals to devise interventions that are sensitive to individual and cultural nuances. Additionally, the research underlines the importance of considering genetic predispositions alongside lifestyle choices in healthcare strategies.

For future research, we advocate a multifaceted approach that includes a broader array of variables and accounts for potential confounders. This would provide a more comprehensive understanding of the health impacts of lifestyle choices.

In conclusion, while acknowledging the constraints of our methodological approach, this study contributes valuable insights to the discourse on public health. It lays down a critical groundwork for future research that promises to refine our understanding and enhance the well-being of individuals both within and beyond the borders of South Korea.

References

[1] Boateng, E. Y., & Abaye, D. A. (2019). A review of the logistic regression model with emphasis on medical research. Journal of data analysis and information processing, 7(4), 190-207. https://www.scirp.org/journal/paperinformation.aspx?paperid=95655

[2] Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N., & Moustakas, K. (2021). Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access, 9, 103737-103757.

https://ieeexplore.ieee.org/document/9491154 /
[3] Lamb, T. J., Graham, A. L., & Petrie, A. (2008). T testing the immune system. Immunity, 28(3), 288-292. https://www.sciencedirect.com/science/article/pii/S1074761308000745
[4] Nopour, R., Shanbehzadeh, M., & Kazemi-Arpanahi, H. (2022). Using logistic regression to develop a diagnostic model for COVID-19: A single-center study. Journal of education and health promotion, 11. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9277749/
[5] Sung, J., Han, S., Park, H., Hwang, S., Lee, S. J., Park, J. W., & Youn, I. (2022). Classification of Stroke Severity Using Clinically Relevant Symmetric Gait Features Based on Recursive Feature Elimination With Cross-Validation. IEEE Access, 10, 119437-119447. https://ieeexplore.ieee.org/document/9932572

Author's Biography

Priyanka Logasubramanian is currently pursuing a master's degree in the Department of Computer Science at West Chester University, Pennsylvania, USA. She completed her Bachelor of Computer Applications (BCA) from Madras University, India, in 2017. Her research interests include machine learning, deep learning, data science, and artificial intelligence.

Md Amiruzzaman is an Assistant Professor in the Department of Computer Science at West Chester University. Before joining WCU, he worked as a software developer for almost 10 years for several companies. He has also held the position of Assistant Professor at Kent State University. He has completed a Bachelor's Degree in Computer Science from National University. Along with that, he has completed four Master's degrees with major in Computer Engineering in 2008 from Sejong University, Computer Science in 2011 from Kent State University (also, partly at Korea University), and Technology in 2015, also from Kent State University, and a Master's in Cybersecurity in 2023 from Georgia Institute of Technology. He received his Ph.D. degrees from Kent State University in 2016 (Mathematics Edu), 2019 (Evaluation and Measurement) and 2021 (Computer Science). In the past, he has worked as a Research Assistant at Sejong University and Korea University. Along with that, he gained the opportunity to teach at both National University and Korea University. His research interests include Visual Analytics of urban data, Data Mining, Machine Learning, Deep Learning, and Data Hiding.

Ashikahmed Bhuiyan, an assistant professor in the Computer Science Department at West Chester University, is a recognized authority in the field. He earned his PhD degree from the University of Central Florida (UCF) under the guidance of Zhishan Guo and Abusayeed Saifullah (from Wayne State University). His Bachelor of Science degree in Computer Science and Engineering from Bangladesh University of Engineering and Technology (BUET), Bangladesh, in 2013 laid the foundation for his research. His work primarily focuses on improving energy efficiency in real-time embedded systems, parallel computing, and mixed-criticality scheduling. His dedication and expertise have been recognized with the Best Student Paper Award at the 40th IEEE Real-Time Systems Symposium (RTSS 2019).