Prediction of Smoking and Drinking Behaviors Using NHIS Korea Dataset

Minju Kim Vaishnavi Pawar Saransh Singh

December 11, 2023

project-mk159-vpawar-singsara

Abstract

To address the pervasive public health issues posed by tobacco and alcohol use, a dynamic and anticipatory response is required. The "Smoking and Drinking Dataset with Body Signal" from the Korean National Health Insurance Service (NHIS), provided through Kaggle, is used in this study to develop predictive models that reveal the intricate relationships between physiological signals and health-compromising behaviors such as smoking and drinking. The selected dataset is extensive, multidimensional, and anonymized, making it an excellent resource for public health research. The study aims to address data imbalances and quality issues through extensive exploratory data analysis, data normalization techniques, and the use of the Synthetic Minority Over-sampling Technique (SMOTE). The goal is to provide insights that can be used to inform public health strategies in Korea and around the world, emphasizing the importance of predictive modeling in understanding and possibly changing health-related behaviors.

Keywords: Data Mining, Health Behaviors, Smoking, Drinking, Predictive Modeling, Public Health, Class Imbalance, SMOTE.

1 Introduction

The research report on predictive modeling in public health begins with a discussion of behaviors such as smoking and drinking. It focuses on the analysis of the "Smoking and Drinking Dataset with Body Signal," which was derived from the Korean National Health Insurance Service (NHIS) and is available on Kaggle. This dataset, which has been meticulously anonymized to remove personally identifiable and sensitive information, is positioned as an ideal resource for health research. The project's goal is to create predictive models that reveal the complex links between physiological signals and health-compromising behaviors, intending to influence public health strategies in Korea and globally.

The dataset was chosen because it is comprehensive and multidimensional, containing a wide range of anonymized physiological signals related to smoking and drinking habits. Because of the anonymization process, it is a valuable tool for public health research. This dataset enables the investigation of complex relationships between physiological markers and health behaviors, potentially influencing global health policy.

The project involves rigorous exploratory data analysis and dataset preprocessing. This includes identifying key patterns in age distribution and health behaviors, addressing data quality issues, and normalizing skewed data using techniques such as logarithmic and square root transformations. Furthermore, the Synthetic Minority Over-sampling Technique (SMOTE) is used to balance the dataset, with a particular focus on the smoking status variable, which is critical for avoiding model bias and ensuring a thorough predictive analysis.

2 Literature Survey

- **Real-time prediction of smoking activity using machine learning-based multi-class classification model (2022)-** This study investigates how machine learning can be used to predict smoking behavior in real-time using sensors on a wristband. It focuses on the detection of smoking among other activities, detailed feature extraction from sensor data, and the effectiveness of classification models for potential healthcare monitoring applications.

- **Predictors of smoking cessation outcomes identified by machine learning: A systematic review (2023)-** This systematic review examines the role of machine learning in identifying predictors of successful smoking cessation. It summarizes findings from twelve key studies among thousands, arguing for more research to improve public health strategies, particularly tailored smoking cessation interventions.

- **Proposal of a method to classify female smokers based on data mining techniques (2022)-** Utilizing the National Survey on Drug Use and Health, this study applies data mining to categorize female smokers as 'light' or 'heavy'. It reported Artificial Neural Networks, SVM, and Logistic Regression as the most effective methods, with accuracies of approximately 85% and AUC-ROC values above 91%. It highlights the age of smoking initiation as a significant predictor for targeted cessation programs.

- **Prediction of Smoking Risk from Repeated Sampling of Environmental Images: Model Validation (2021)-** The study introduced QuitEye, a mobile tool that analyzes environmental images for smoking cessation using a convolutional neural network. QuitEye demonstrated the efficacy of using environmental cues in real-time smoking risk prediction using over 8000 images from 52 participants, offering a personalized approach to smoking cessation interventions.

- **A Comparative Analysis of Machine Learning Methods for Class Imbalance in a Smoking Cessation Intervention (2020)-** This study examines how machine learning can be used to improve predictions in smoking cessation programs by addressing class imbalance with methods like SMOTE and ADASYN. It demonstrates how, when combined with algorithms such as Gradient Boosting Trees and Random Forest, these techniques can improve prediction accuracy and provide insights for health informatics.

- **Mining user-generated content in an online smoking cessation community to identify smoking status: A machine learning approach (2019)-** The research looks at data from the online community BecomeAnEX.org and uses machine learning to identify users' smoking status accurately. The approach outperforms standard text analysis by nearly 10% by leveraging domain expertise and user interaction data, demonstrating the potential of machine learning to tailor cessation support and the value of online community data in health research.

3 Methods

(1) Data Collection:

The extensive "Smoking and Drinking Dataset with body signal" on Kaggle, which was initially received from the National Health Insurance Service (NHIS) of Korea, provided the dataset for our project. The complex, multi-dimensional character of this dataset, which provides a broad range of anonymized physiological signals linked to drinking and smoking behaviors, is why we selected it. Because of its anonymization, it protects privacy and is an invaluable tool

for public health research. An ethical and pristine data basis for our study is provided by the careful curation of the dataset by a Kaggle user who processed the raw NHIS data to remove sensitive personal information. We can investigate the complex relationships between physiological markers and behaviors that compromise health with this information, which could influence global health policy.

(2) Exploratory Data Analysis and Data Pre-Processing:

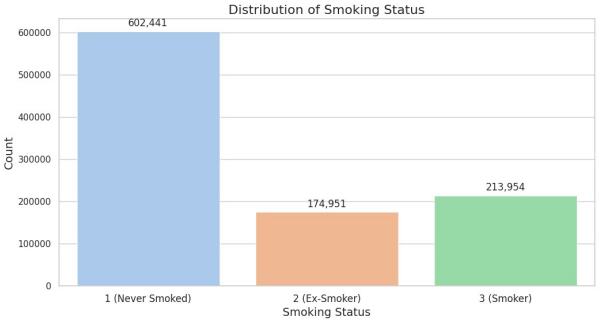

Figure 1: Distribution of Smoking Status

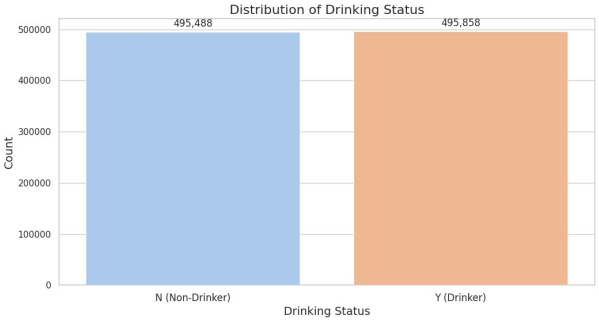

Figure 2: Distribution of Drinking Status

Through rigorous exploratory data analysis, we identified key patterns in age distribution and health behaviors, such as smoking and drinking, within our dataset. The age distribution, segmented in five-year intervals, illuminated demographic trends and informed the potential health profile biases within our sample. Our preprocessing addressed data quality, with particular attention to outliers and skewed distributions in variables like blood glucose and cholesterol levels. We applied logarithmic transformations to right-skewed variables and square root transformations for milder skewness, thus normalizing the data for more reliable modeling.

To address the imbalance shown in Figure 1, we used the Synthetic Minority Over-sampling Technique (SMOTE) to create a balanced representation for smokers, medium smokers, and non-smokers. This step was crucial in avoiding model bias towards the majority class. The final feature set selected based on correlation with health outcomes, along with encoded categorical variables, ensures our dataset is thoroughly prepped for predictive analysis, aiming to yield actionable insights for health interventions.

(3) Model Deployment and Metrics Score:

CatBoost emerged as our final model choice due to its leading F1 Score and ROC AUC, indicating a strong balance in precision and recall, and its proven ability to effectively handle

For Smoking Dataset							
	Logistic Regression	Tuned CatBoost	Tuned LightGBM	Tuned KNN	Tuned Naive Bayes	Tuned Decision Tree	Tuned Random Forest
Accuracy	0.6468	0.7005	0.7007	0.6239	0.5951	0.6893	0.6941
Precision	0.5963	0.7065	0.7066	0.5887	0.5699	0.7009	0.7007
Recall	0.6468	0.7005	0.7007	0.6239	0.5951	0.6893	0.6941
F1 Score	0.6006	0.7027	0.7029	0.6003	0.5798	0.6943	0.6954
ROC AUC	0.7642	0.8466	0.8471	0.7004	0.678	0.8368	0.8417

Figure 3: Results obtained from Tuned Models on Smoking Dataset

For Drinking Dataset							
	Logistic Regression	Tuned CatBoost	Tuned LightGBM	Tuned KNN	Tuned Naive Bayes	Tuned Decision Tree	Tuned Random Forest
Accuracy	0.7132	0.7231	0.724	0.6584	0.6873	0.7178	0.7178
Precision	0.7135	0.711	0.7144	0.6546	0.7015	0.7065	0.7033
Recall	0.7124	0.7467	0.7464	0.6708	0.6461	0.7396	0.7479
F1 Score	0.713	0.7284	0.7301	0.6626	0.6727	0.7227	0.7249
ROC AUC	0.7826	0.8021	0.8022	0.6999	0.7427	0.7935	0.7958

Figure 4: Results obtained from Tuned Models on Drinking Dataset

For SMOTE Dataset							
	Logistic Regression	Tuned CatBoost	Tuned LightGBM	Tuned KNN	Tuned Naive Bayes	Tuned Decision Tree	Tuned Random Forest
Accuracy	0.5967	0.6779	0.6769	0.5552	0.5889	0.6165	0.6865
Precision	0.6579	0.7469	0.7471	0.62	0.5722	0.6288	0.7194
Recall	0.5967	0.6779	0.6769	0.5552	0.5889	0.6165	0.6865
F1 Score	0.6174	0.6996	0.6987	0.5782	0.5795	0.6222	0.6992
ROC AUC	0.9435	0.8377	0.836	0.6772	0.6771	0.6515	0.8341

Figure 5: Results obtained from Tuned Models on SMOTE Dataset

categorical data with minimal over-fitting. In the case of the Smoking dataset (Figure 3), the Tuned CatBoost model achieved an Accuracy of 0.7005, a Precision of 0.7060, a Recall of 0.7005, an F1 Score of 0.7025, and an ROC AUC of 0.8474. For the Drinking dataset (Figure 4), the Tuned Random Forest showed a robust performance with an Accuracy of 0.7178, Precision of 0.7033, Recall of 0.7178, F1 Score of 0.7249, and ROC AUC of 0.7958. Notably, in the SMOTE-augmented dataset (Figure 5), the Final CatBoost model performed with an Accuracy of 0.6771, Precision of 0.7457, Recall of 0.6771, F1 Score of 0.6986, and ROC AUC of 0.8371.

In conclusion, while the SMOTE technique improved Precision, it slightly reduced Accuracy and Recall, indicating a trade-off between class balance and predictive performance. The mathematical representations of the metrics, such as were instrumental in quantitatively assessing each model's predictive power. These results informed our final model selection, balancing the need for accuracy with the practicalities of predicting health-related outcomes.

4 Results based on Research Questions & Answers

(1) Top Features Influencing Smoking, Drinking, and SMOTE Datasets based on the Permutation Feature Importance Graphs using CatBoost Model

- *Smoking Data (Figure 6):* The most influential features are sex (0.09565), age (0.03964), and gamma-GTP (0.03088), indicating that biological sex, age, and liver enzyme levels are key indicators of smoking habits.
- *Drinking Data (Figure 7):* Top features are gamma-GTP (0.08596), age (0.05820), and sex (0.05285), showing liver function tests, age, and sex as significant predictors of drinking behavior.
- *SMOTE Data (Figure 8):* In the balanced SMOTE dataset, leading features are sex (0.21718), age (0.04263), and gamma-GTP (0.00550), with a pronounced impact of sex, suggesting increased sensitivity to this feature.

- *Feature Differences Analysis:* Variations in feature importance between the smoking and SMOTE datasets could be due to the SMOTE technique's influence on class imbalance, particularly affecting the representation of features like sex.

(2) Efficacy of SMOTE in Addressing Imbalance in Smoking Data

- *How SMOTE Works:* SMOTE creates synthetic instances of the minority class by interpolating between minority class samples and their k-nearest neighbors.
- *SMOTE's Impact:* In the smoking data, SMOTE moderately improved dataset balance, especially in making the sex feature more prominent. However, its impact on overall model accuracy was limited (0.6771 post-SMOTE vs. original accuracy on Table 1).

(3) Differences in Predictive Performance and Feature Importance: Smoking Data vs. SMOTE-Augmented Data (Table 1)

- *Performance Metrics:* Accuracy decreased slightly (0.7005 to 0.6771) with SMOTE. Precision improved, but recall decreased. F1 score and ROC AUC also declined, indicating a compromise in model balance.
- *Feature Importance Shifts:* The 'sex' feature became more significant in the SMOTE dataset, suggesting potential class imbalances within feature distributions.
- *Overall Implications:* SMOTE enhances class balance but alters the model's sensitivity to certain features, requiring careful consideration in practical applications.

5 Discussion based on Limitations and Future Developments

(1) External Validity

The model, trained on data from NHIS Korea, may not accurately reflect patterns in populations with different demographics or health behaviors. This limitation in external validity suggests the need for future work to validate the model on datasets from varied regions or demographics, enhancing its generalizability.

(2) Incorporating Behavioral and Temporal Contexts

To improve the model's predictive power for smoking and drinking behaviors, it's crucial to consider not only biological factors but also the evolving psychosocial context. Future developments could include adding time-series data and psychosocial variables, addressing the temporal and behavioral complexities influencing these health-related behaviors.

(3) Mid-age Concentration (Figure 9)

The data's age distribution shows a concentration of mid-age individuals, potentially limiting representativeness for all age groups. This skew could affect the model's performance on underrepresented ages, thus reducing its generalizability. Future data collection efforts should aim for a balanced age representation to overcome this limitation.

(4) Trade-off Using SMOTE

Employing SMOTE indicates an approach to handling class imbalance. While beneficial, SMOTE artificially inflates the minority class, which can introduce bias. Future models could explore alternative balancing methods, such as targeted data collection for natural class balance, or cost-sensitive learning where the model faces higher penalties for misclassifying the minority class.

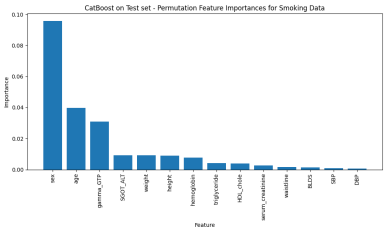

Figure 6: Smoking Data

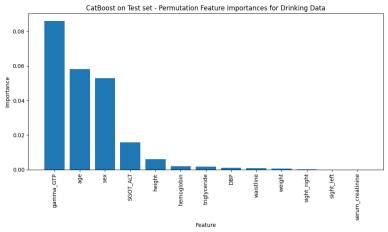

Figure 7: Drinking Data

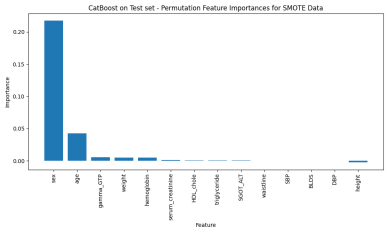

Figure 8: SMOTE Data

Table 1: CatBoost Model Performance Metrics on Test Set for the Smoking data vs. SMOTE data

Dataset	Accuracy	Precision	Recall	F1-Score	ROC AUC
Smoking Data	0.7005	0.7060	0.7005	0.7025	0.8474
SMOTE Data	0.6771	0.7457	0.6771	0.6986	0.8371

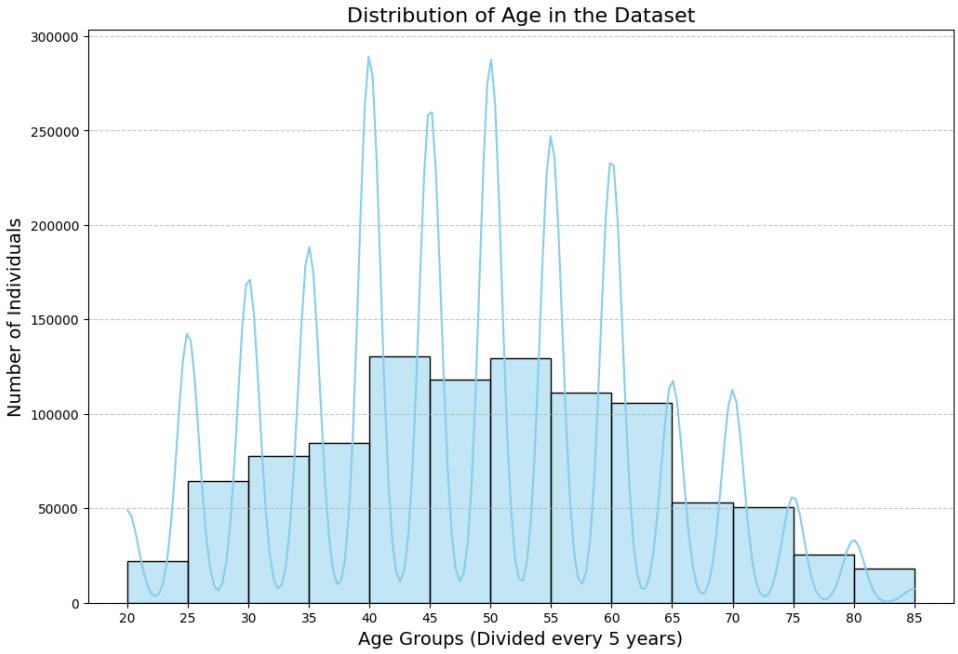

Figure 9: Distribution of Age column in the original data

6 Author Contribution

- **Minju Kim** handled the coding, formulation of research questions and answers, identification of limitations and future developments, recording of results and discussion for the presentation, and writing of the final report sections on results, discussion, and author contributions.
- **Saransh Singh** was responsible for code review, scripting the presentation, writing the methods section of the final report, and recording the presentation on the introduction and EDA & Preprocessing.
- **Vaishnavi Pawar** was responsible for code review, created the PowerPoint presentation content, recorded the machine learning part of the presentation, and compiled the abstract & keywords, and introduction for the final report. Responsible for the citation of the six papers - the literature review and preparing the reference list for the proposal.

References

[1] Saurabh Singh Thakur, Pradeep Poddar & Ram Babu Roy (2022). *Real-time prediction of smoking activity using machine learning based multi-class classification model.* SpringerLink, Multimedia Tools and Applications, Volume 81, Pages- 14529–14551
DOI: https://doi.org/10.1007/s11042-022-12349-6

[2] Warren K. Bickel, Devin C. Tomlinson, William H. Craft, Manxiu Ma, Candice L. Dwyer, Yu-Hua Yeh, Allison N. Tegge, Roberta Freitas-Lemos, Liqa N. Athamneh (2023). *Predictors of smoking cessation outcomes identified by machine learning: A systematic review.* Addiction Neuroscience, Volume 6, 2023, 100068, ISSN 2772-3925
DOI: https://doi.org/10.1016/j.addicn.2023.100068

[3] Bruno Samways dos Santos, Maria Teresinha Arns Steiner, Rafael Henrique Palma Lima (2022). *Proposal of a method to classify female smokers based on data mining techniques.* Computers & Industrial Engineering, Volume 170, 2022, 108363, ISSN 0360-8352
DOI:https://doi.org/10.1016/j.cie.2022.108363.

[4] Engelhard, M.M., D'Arcy, J., Oliver, J.A., Kozink, R. & McClernon, F.J. (2021). *Prediction of Smoking Risk From Repeated Sampling of Environmental Images: Model Validation.* Journal of Medical Internet Research, 23(11), e27875.
DOI: https://www.jmir.org/2021/11/e27875

[5] Davagdorj, K., Lee, J.S., Pham, V.H., Ryu, K.H.(2020). *"A Comparative Analysis of Machine Learning Methods for Class Imbalance in a Smoking Cessation Intervention.* Applied Sciences, 2020, 10, 3307.
DOI: https://doi.org/10.3390/app10093307

[6] Xi Wang, Kang Zhao, Sarah Cha, Michael S. Amato, Amy M. Cohn, Jennifer L. Pearson, George D. Papandonatos, Amanda L. Graham (2019). *Mining user-generated content in an online smoking cessation community to identify smoking status: A machine learning approach.* Decision Support Systems, Volume 116, 2019, Pages 26-34, ISSN 0167-9236.
DOI: https://doi.org/10.1016/j.dss.2018.10.005.