Prediction of Smoking and Drinking Behaviors Using NHIS Korea Dataset

Minju Kim, Vaishnavi Pawar, Saransh Singh

November 4, 2023

Abstract

Confronting the pervasive public health issues posed by tobacco and alcohol use requires a dynamic and anticipatory response. This project taps into a curated body signal dataset from the National Health Insurance Service of Korea, provided through Kaggle, to explore the potential of data mining in forecasting smoking and drinking behaviors. Recognizing the complexities of machine learning, particularly the challenge of class imbalance, we are considering synthetic oversampling techniques such as SMOTE and ADASYN, among other methods, to enhance the precision of our predictive models. The ambition of this research is to remain agile in our methodological approach, ultimately aiming to deliver a versatile analytic tool. Such a tool would not only discern patterns in health-compromising behaviors but could also serve as a cornerstone for developing targeted public health strategies. Our goal is to pave the way for data-driven insights that could inform and transform health intervention practices worldwide.

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

1 Introduction

Within the landscape of public health research, predictive modeling has emerged as a key instrument for preempting and understanding health-compromising behaviors such as smoking and drinking. This proposal centers around a sophisticated analysis of the "Smoking and Drinking Dataset with body signal," a dataset procured from Kaggle that originated from the National Health Insurance Service (NHIS) in Korea. The dataset was meticulously curated by a Kaggle user who processed the original NHIS data to remove any personal and sensitive information, thereby providing a clean, anonymized dataset suitable for public use and research applications.

We select this dataset for our data mining project with a clear purpose: to craft predictive models that encapsulate the intricate correlations between physiological signals and smoking and drinking behaviors. By doing so, we aim to generate insights with the potential to influence not only Korean public health strategies but also to extend implications and applications to a global context.

The project will employ synthetic oversampling methods and various machine learning classifiers as informed by relevant literature, which has identified the issue of class imbalance as a significant hurdle in health behavior prediction. We intend to navigate these challenges while enhancing the interpretability of our predictive models, aiming to contribute a meaningful tool for health policy makers and intervention programs worldwide.

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 accurately identify users' smoking status. 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 Methodology

Our study employs a robust methodology using data from the Korean National Health Insurance Service to predict smoking and drinking behaviors with high accuracy. The initial phase involves an in-depth exploration and visualization of the dataset to understand the distribution of these behaviors within the Korean population. We will identify and evaluate outliers to determine their relevance and conduct a correlation analysis to discern the relationships between various features and the target behaviors. This stage is crucial for isolating significant predictors and preparing the data for modeling.

In the preprocessing stage, we will address data imbalances with techniques such as SMOTE to ensure our predictive models are trained on a balanced dataset. This sets the stage for the application of various machine learning algorithms, including Random Forest, Gradient Boosting, and k-Nearest Neighbours (KNN). Each model will be fine-tuned and validated against performance metrics, with a particular focus on the F1 score, which balances precision and recall, ensuring a robust evaluation of the model's predictive capabilities.

The final stage involves a comparative analysis of the models to select the most effective one based on predictive accuracy. The chosen model will then be used to generate insights into the key factors influencing smoking and drinking behaviors. These insights are intended to inform targeted public health interventions and strategies, ultimately contributing to the broader goal of improving health outcomes within the Korean population.

Author Contributions

- Minju Kim: Identified the project topic, curated the dataset, conducted a comprehensive literature review, summarized key research findings, and formulated the Abstract, Keywords, Introduction, and Author Contribution sections of the proposal.
- Saransh Singh: Formulated the Methods section of the proposal.
- Vaishnavi Pawar: Responsible for the citation of the six papers - 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.