Influential Factors For Tobacco Use Among The Youth

Kaixiang Liu, Bryan Alarcon, Jahn Tibayan, Kanika Sood

Abstract

Tobacco and e-cigarette use among youth is a critical public health concern, as the majority of adult smokers begin smoking during adolescence. This study aims to predict youth tobacco consumption based on survey responses from the National Youth Tobacco Survey provided by the U.S. Centers for Disease Control and Prevention. Below, we trained and evaluated several machine learning models to assist our findings. The potential of machine learning in research is vast. Through our findings, we identified key factors influencing youth tobacco use, providing a foundation for data-driven interventions and public health strategies aimed at preventing smoking initiation during adolescence.

Keywords

Classification Techniques
Decision Tree
Random Forest
Gradient Boost
Logistic Regression
Branching Logic In Survey Datasets
Smote Under/Over Sampling

Introduction

The growing use of tobacco and e-cigarette products has prompted significant concerns among the health industry. Understanding the patterns and behaviors associated with tobacco use is crucial for designing interventions and policies that can address this public challenge. We analyze a large dataset to uncover insights into smoking and vaping habits. Our goal is to predict whether an individual is a user of such products based on their survey responses. By applying machine learning techniques, we aim to identify the most influential factors contributing to tobacco use while ensuring that the models are robust and generalizable.

Background

Analyzing behavioral patterns related to tobacco and e-cigarette uses has been an area of growing interest in public health research. Previous studies have leveraged survey data to identify key demographic and behavioral factors associated with smoking and vaping. However, many existing analyses rely on traditional statistical methods, which may overlook complex interactions between variables. This work builds upon prior research by incorporating machine learning techniques to explore these relationships more comprehensively. By encoding categorical data and handling missing values systematically, our approach aims to provide deeper insights into the data while ensuring that the results are interpretable.

The National Tobacco Youth Survey (NTYS) [1] is an annual cross-sectional survey taken by U.S. middle school and high school students. The resulting dataset from this survey is used for this research. From this data, those interested in the surveillance, regulation, and prevention of tobacco use have access to valuable insights on tobacco use patterns and associated factors to help inform their strategies. Utilizing machine learning models helps bolster evidence-based control strategies. These strategies can be used along with proper regulation of tobacco products to lead to the ultimate goal of reducing all forms of tobacco product use among the U.S population particularly the younger generation.

There are various problem domains when it comes to Tobacco use, such as a focus on smoking prevention (such as our case), or a focus on smoking cessation (or quitting). These are important domains that help guide public health action. An example is a report by the Surgeon General in 2020 [2], details a comprehensive analysis of smoking cessation patterns in both the youth and adults, and highlights disparities in cession by age, ethnicity, level of education, and other demographics. The paper also utilized the NTYS dataset. Another similar report by Lai et al. developed machine learning models to predict smoking cessation outcomes utilizing algorithms combined with support vector machines and artificial neural networks to enhance cessation prediction accuracy [3]. The paper aimed to provide prediction models to assist evidence-based treatment and guidelines for quitting smoking by providing an individual smoking cessation success rate for patients of physicians.

This paper focuses on the problem domain of smoking prevention, especially in the youth, as research indicates that a majority of smokers initiate smoking during adolescence. In fact, the National Institute of Drug Abuse reports that nearly 90% of adult smokers started smoking before the age of 18, highlighting the criticality of the stage of adolescence in the onset of smoking patterns [4].

1 Data Set Data Processing

1.1 Data Set

The dataset used for this analysis is the survey result data from the National Tobacco Youth Survey [1]. It contains survey responses from individuals about their smoking and vaping habits. It includes questions covering demographics, behavior, and exposure to tobacco products. The dataset comprises over 20,000 rows and numerous categorical and numerical columns representing different survey questions. Similar datasets are available, such as PATH, the Population Assessment of Tobacco and Health (PATH) Study [12]. However, although it focuses on tobacco behaviors, attitudes, and demographic data, it doesn't focus on youth participants.

1.2 Data Pre-Processing

To prepare the data for analysis, several steps are taken to ensure data quality and consistency:

1. Merged Columns: Many survey questions were spreaded across multiple columns (e.g., Q4a, Q4b, Q4c). There was a column for each available option for each question. All the columns that were correlated with the same question were merged into single columns to simplify analysis.

2. Index Encoding (Simple Questions): To handle varying responses within individual columns, each response was assigned a unique index. For simple questions, we used labels such as 'a,' 'b,' and 'c,' mapped to integer values starting with 1, 2, and 3, respectively.

For example, the response option 'A' for Question 4 corresponds to the numerical value 1, while option 'B' corresponds to the numerical value 2, continuing in sequential order.

3. Index Encoding (Complex Questions): For questions that allowed multiple selections (e.g., Q18a&b, Q18a&b&c), we encoded these combinations using additive values based on their corresponding numerical values. For example, given that the final single-letter designation 'Z' is assigned the numerical value 26, the multi-selection question Q18a&b was allocated the value 27, which represents the cumulative numerical value obtained by combining multiple response options.

4. Handling Missing Values: Special values such as 'N' (Not answered) were substituted with the median or mode based on the specific question context, 'S' (Skipped) was replaced by -1, and 'Z' (Not Displayed) entries were eliminated from the dataset to maintain data quality while preserving meaningful information.

5. Dimensionality Reduction: To address potential multicollinearity and complexity within the dataset, advanced dimensionality reduction techniques, specifically Principal Component Analysis (PCA), were systematically implemented. This process involved carefully transforming the high-dimensional feature space into a more compact, yet statistically representative lower-dimensional representation. By identifying and extracting the most significant orthogonal components that capture the maximum variance in the data, we significantly reduced redundant information while preserving the essential statistical characteristics. The PCA methodology allowed us to compress the original feature set, mitigating potential overfitting risks and computational inefficiencies, and creating a more streamlined dataset that maintained the core informational integrity required for robust predictive modeling.

2 Machine Learning Classification Technique

To analyze the dataset, we implemented several machine learning models such as Random Forest Classifier, Gradient Boost, Logistic Regression, and Decision Tree models. For the purpose of feature selection, we used a Random Machine as well as importance scores from Random Forest models in order to select the top 10 features. More specifically the 5 most important features selected were gathered from a Random Forest Model, and the other 5 columns were selected based on the columns with the most non-zero values to fight against class imbalance. Similar approaches were used in other adjacent papers, such as the 2023 study on smoking cessation predictions by Issabakhsh et al, or a 2021 paper by Choi et al, whom also employed a Random Forest model for feature selection [5][6]. Another large issue, known as class imbalance in ML literature, needs to be handled. Due to a much larger class size, or number of individuals that don't smoke, as compared to those who do, models may have a high chance of having high specificity, but lower sensitivity, and result in good prediction rates of the dominant class, but less so for the non-dominant class. In similar work, such as the study by Issabakhsh et al [5], ML techniques such as random over and undersampling have been used to deal with class imbalance. In our case, we have employed the use of random oversampling for the minority class and undersampling for the majority class to decrease the effect of the majority class on prediction results.

2.1 Decision Tree

The Decision Tree model is a supervised learning algorithm used for both regression and classification tasks. It works by splitting the dataset into smaller subsets based on various features, creating branches that lead to decision nodes and leaf nodes representing outcomes or classifications. Decision Trees are easy to implement and can capture complex non-linear relationships in the dataset. In our analysis, the Decision Tree is well-suited for our data due to its hierarchical structure, which allows relevant features to be captured at each split. They offer advantages over other models such as logistic regression due to being able to capture decision rules in a sequential form that could provide valuable insights [7]. Decision trees have been previously used in studies for these specific reasons, such as in Piper et al. [7].

Our index encoding method, which transforms categorical data into unique numerical values, further enhances the Decision Tree's ability to process and learn from sparse datasets effectively. We also performed several fine-tuning steps, adjusting parameters to optimize the model's performance. To address data sparsity and imbalance, we applied a minority oversampling technique, ensuring that both majority and minority classes were adequately represented during model training. Synthetic Minority Oversampling, or SMOTE has been used for the same purpose of fighting class imbalance in studies such as Davagdorj et al. [8] and demonstrated a result of an increase in the precision of their models. In our case, the combination of techniques (index encoding and SMOTE) improved the model's overall accuracy and ability to generalize across different data subsets.

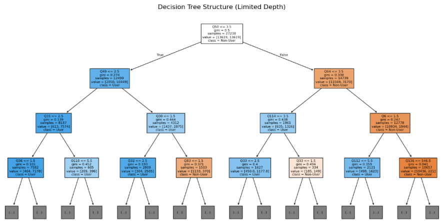
Decision Tree Structure (Limited Depth)
Q50 <= 3.5 gini = 0.5 samples = 27238 value = [13619, 13619] class = Non-User
True
False
Q49 <= 2.5 gini = 0.274 samples = 12499 value = [2050, 10449] class = User
Q64 <= 3.5 gini = 0.338 samples = 14739 value = [11569, 3170] class = Non-User
Q31 <= 2.5 gini = 0.139 samples = 8187 value = [613, 7574] class = User
Q30 <= 1.5 gini = 0.444 samples = 4312 value = [1437, 2875] class = User
Q114 <= 3.5 gini = 0.438 samples = 1961 value = [635, 1326] class = User
Q6 <= 1.5 gini = 0.247 samples = 12778 value = [10934, 1844] class = Non-User
Q36 <= 1.5 gini = 0.101 samples = 7582 value = [404, 7178] class = User
Q110 <= 5.5 gini = 0.452 samples = 605 value = [209, 396] class = User
Q32 <= 2.5 gini = 0.193 samples = 2809 value = [304, 2505] class = User
Q83 <= 1.5 gini = 0.371 samples = 1503 value = [1133, 370] class = Non-User
Q33 <= 2.5 gini = 0.4 samples = 1627 value = [450.0, 1177.0] class = User
Q33 <= 1.5 gini = 0.494 samples = 334 value = [185, 149] class = Non-User
Q112 <= 5.5 gini = 0.359 samples = 2121 value = [498, 1623] class = User
Q136 <= 546.5 gini = 0.041 samples = 10657 value = [10436, 221] class = Non-User
(...)

Figure 1. [A structural view of Decision Tree]

2.2 Random Forest

The Random Forest model is an ensemble learning method primarily used for classification and regression tasks. It builds multiple decision trees during training and merges their output via majority voting for classification or averaging for regression. By aggregating the results of each diverse decision tree, Random Forest models are robust to noise and variance, offering stability and reliable performance.

In our analysis, the chosen dataset is very sparse due to the nature of the majority of youth respondents reporting not engaging in smoking or vaping behaviors. This sparsity raises challenges for certain machine-learning models. However, Random Forest is suitable for this genre of dataset [9]. With the addition of our index encoding method, Random Forest can efficiently learn even when data contains many sparse values or encoded combinations.

We also performed extensive fine-tuning, ultimately settling on 500 random trees with a minimum of two leaf nodes per tree, which yielded the best results. Furthermore, to address the data sparsity, we applied both an undersampling technique, as well as SMOTE to balance the dataset, improving the model's ability to learn and generalize across the different classes. The use of undersampling the majority class in tandem with random forest models has been researched by Hasanin et al. [10] and shown to boost performance while still retaining a fair amount of original information in the resulting dataset. Finally, the combination of both undersampling and oversampling has also been utilized in previous studies, such as by Chawla et al. [11].

2.3 Logistic Regression

Logistic Regression emerged as a fundamental predictive modeling technique in our analysis, serving as a critical baseline model for binary classification challenges. This probabilistic algorithm operates by establishing a linear decision boundary that maps input features to the logarithmic odds of the target variable's occurrence, utilizing the sigmoid function to transform linear combinations of features into probabilities between 0 and 1. In our specific implementation, we leveraged logistic regression's inherent strengths in handling sparse and encoded feature spaces. The preprocessing steps were crucial: we first applied standardization techniques to normalize the feature distributions, ensuring that each predictor contributed proportionally to the model's decision-making process without being unduly influenced by scale differences. To mitigate potential model limitations, we approached this concern by exploring ensemble resampling techniques that ultimately proved ineffective and computationally-expensive in addressing the dataset's complexity. Eventually we implemented weighted sampling and minority oversampling (SMOTE) to address the inherent class imbalance in our dataset. This approach prevented the model from becoming biased towards the majority class and improved its predictive performance across all classification thresholds. The model's interpretability was a significant advantage, allowing us to directly examine the coefficients and understand the directional and magnitude of each feature's impact on the predicted outcome. This transparency provided critical insights into the underlying data relationships, serving not just as a predictive tool but also as an exploratory mechanism for feature importance. By carefully tuning hyperparameters through cross-validation and employing robust evaluation metrics like precision, recall, F1-score, and area under the ROC curve, we established logistic regression as a reliable baseline against which more complex machine learning algorithms could be compared. This methodical approach ensured a comprehensive and statistically sound initial modeling strategy.

2.4 Gradient Boost

Gradient Boosting is an ensemble technique that builds a series of decision trees sequentially, where each new tree attempts to correct the errors of the previous trees. Unlike Random Forests, which build trees independently and aggregate their results, Gradient Boosting incrementally improves the model's predictive accuracy by optimizing a loss function through gradient descent. This approach makes it highly effective for capturing complex, non-linear relationships in data. In order to handle our imbalanced dataset, we utilized Gradient Boosting to identify meaningful patterns among sparse data points. Gradient Boosting addresses our challenge by focusing on minimizing errors iteratively, prioritizing the minority class through targeted improvements in each boosting iteration. Our use of index encoding further complements the Gradient Boosting approach by transforming categorical data into numerical values. The nature of Gradient Boosting offers great ability to handle encoded features, particularly those with sparse representations.

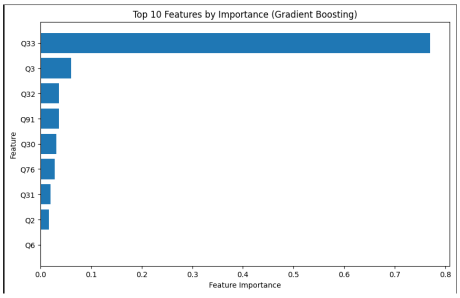

Figure 2. [Top 10 most influential features in Gradient-Boosting]

3 Results

The performance of the four machine learning models was evaluated using accuracy, recall, F1-scores, and visualizations. These metrics help assess the precision of each of the models.

The **Decision Tree** model achieved the highest accuracy at **92%** and a macro-average F1-score of **0.74**. It performed well on the majority class(non-users), with an F1-score of **0.96**, but struggled with the minority class(users), where it achieved a recall of **62%** and a precision of **44%** The imbalance in this performance demonstrates the Decision Tree's tendency to overfit, especially with data sets that are sparse or imbalanced, which occurred in this case. While it accurately classified the majority class, it still was limited in differentiating between the two classes, causing a high bias. The lack of accuracy in the minority class makes it less effective in capturing trends within the target group.

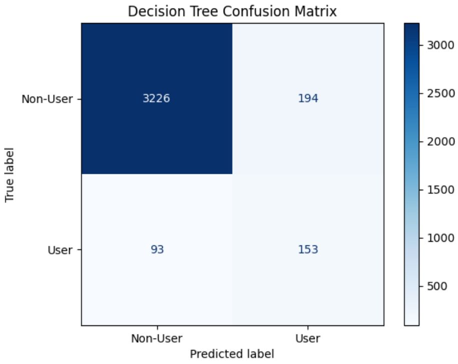

Figure 3. [Confusion Matrix for Decision Tree]

The **Random Forest** model has an accuracy of **91%** and a macro-average F1-score of **0.75**. It's performance on the majority class(non-users) was strong, with a precision of **99%** and an F1-score of **0.95**. However, the model struggled with the minority class(users) as seen in the **42%** precision and F1-score of **0.56**. These results suggest that the model frequently misclassified non-users as users, which is likely due to the class imbalance. While the model worked well for non-users, it's effectiveness was reduced due to the inability to generalize well for users.

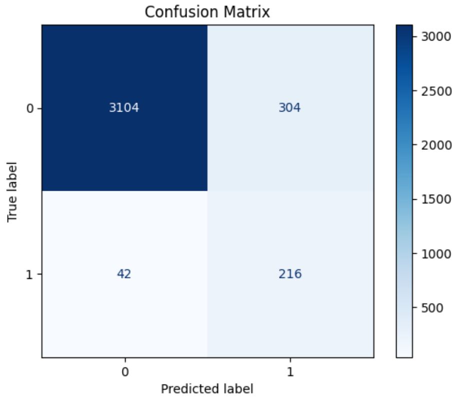

Figure 4. [Confusion Matrix for Random Forest]

Logistics Regression showed strong potential for our dataset, with an accuracy of **90%** and a macro average F1-score of **0.76**. It exceeded in recall for the minority class(users), by achieving a recall score of **93%**, which means it correctly identified most users. However, the precision was a low **41%**, indicating a high rate of false positives where non-users were classified as users. Logistic Regression's balance between identifying users and generalizing to unseen data made it the most reliable model for addressing the goal of this study.

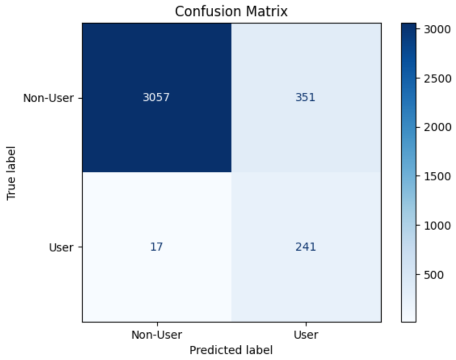

Figure 5. [Confusion Matrix for Logistic Regression]

The **Gradient Boosting** model has an accuracy of **84%** and a macro-average F1-score of **0.66**. It performed well in recalling the minority class(users0, with a recall of **84%**, ensuring most users were identified. Unfortunately, the precision for this class was only **28%**, which create frequent misclassification of non-users as users. The model's boosting process likely overfit to noise in the dataset, especially since there is already an imbalance in the data. Gradient Boosting did capture some relationships, but its high false positive rate makes it lass suitable in identifying between users and non-users.

4 Conclusions

Among all the models that were tested, **Logistics Regression** provided the most balanced performance for our study. With a recall of **93%** for users, it was the most effective in identifying individuals that are at the risk of tobacco usage. While it did have a lower precision for users, the model's ability to generalize the dataset and low risk of missing actual users made it the most reliable option. Other models like **Random Forest** and **Decision Tree** have higher overall accuracy, but they were limited with the minority group due to bias towards the majority group. Logistic Regression's combination of high recall, generalization, and identifying users make it the most effective with this study's objective in analyzing trends in youth tobacco usage.

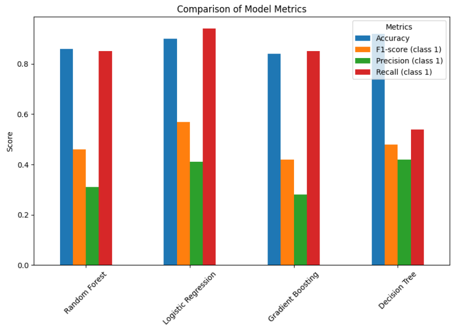

Figure 6. [Bar Graph Comparison of Models]

5 Future Work

In the Future, we will enhance our models and analysis by addressing some limitations in the current study. First, we plan to improve our preprocessing steps by using better techniques to handle class imbalance. Additionally, incorporating more comprehensive datasets with better representation of users will reduce the current imbalance and help models learn more refined patterns. We also intend to explore additional machine learning algorithms, such as support vector machines, neural networks, and XGBoost, to evaluate their effectiveness on the problem. Finally, we plan to include external data sources, such as regional or demographic-specific datasets, to help our findings and generalize the models with a broader population. The enhancements are aimed to create a more refined and reliable system for identifying trends in youth tobacco usage.

6 References

1. A. S. Gentzke, T. W. Wang, M. Cornelius, E. Park-Lee, C. Ren, M. D. Sawdey, K. A. Cullen, C. Loretan, A. Jamal, and D. M. Homa, "Tobacco product use and associated factors among middle and high school students—National Youth Tobacco Survey, United States, 2021," *MMWR Surveill. Summ.*, vol. 71, no. 5, pp. 1–29, Mar. 2022, doi: 10.15585/mmwr.ss7105a1.

2. U.S. Public Health Service Office of the Surgeon General and National Center for Chronic Disease Prevention and Health Promotion, *Smoking Cessation: A Report of the Surgeon General*. U.S. Department of Health and Human Services, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK555598/.

3. C.-C. Lai, W.-H. Huang, B. C.-C. Chang, and L.-C. Hwang, "Development of Machine Learning Models for Prediction of Smoking Cessation Outcome," *Int. J. Environ. Res. Public Health*, vol. 18, no. 5, p. 2584, Mar. 2021, doi: 10.3390/ijerph18052584.

4. National Institute on Drug Abuse, *Preventing Tobacco Use Among Youth and Young Adults: A Report of the Surgeon General*. National Institutes of Health, Bethesda, MD, 2021. [Online]. Available: https://www.drugabuse.gov.

5. M. Issabakhsh, L. M. Sánchez-Romero, T. T. T. Le, A. C. Liber, J. Tan, Y. Li, R. Meza, D. Mendez, and D. T. Levy, "Machine learning application for predicting smoking cessation among US adults: An analysis of waves 1-3 of the PATH study," *PLoS One*, vol. 18, no. 6, p. e0286883, Jun. 2023, doi: 10.1371/journal.pone.0286883.

6. J. Choi, H.-T. Jung, A. Ferrell, S. Woo, and L. Haddad, "Machine Learning-Based Nicotine Addiction Prediction Models for Youth E-Cigarette and Waterpipe (Hookah) Users," *J. Clin. Med.*, vol. 10, no. 5, p. 972, Mar. 2021, doi: 10.3390/jcm10050972.

7. M. E. Piper, W. Y. Loh, S. S. Smith, S. J. Japuntich, and T. B. Baker, "Using decision tree analysis to identify risk factors for relapse to smoking," Substance Use Misuse, vol. 46, no. 4, pp. 492–510, 2011, doi: 10.3109/10826081003682222.

8. K. Davagdorj, J. S. Lee, K. H. Park and K. H. Ryu, "A machine-learning approach for predicting success in smoking cessation intervention," 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), Morioka, Japan, 2019, pp. 1-6, doi: 10.1109/ICAwST.2019.8923252. keywords: Feature extraction;Logistics;Machine learning;Education;Predictive models;Public healthcare;Data models;Smoking cessation;Class imbalance;Factor analysis;SMOTE;machine learning classifiers,

9. C. Chen, A. Liaw, and L. Breiman, "Using Random Forest to Learn Imbalanced Data," Department of Statistics, UC Berkeley, and Biometrics Research, Merck Research Labs, Tech. Rep. 666, 2004. [Online]. Available: https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf

10. T. Hasanin and T. Khoshgoftaar, "The Effects of Random Undersampling with Simulated Class Imbalance for Big Data," 2018 IEEE International Con-

ference on Information Reuse and Integration (IRI), Salt Lake City, UT, USA, 2018, pp. 70-79, doi: 10.1109/IRI.2018.00018. keywords: Big Data;Machine learning algorithms;Forestry;Machine learning;Radio frequency;Data models;Libraries;Big Data, Machine Learning, Imbalanced Data, Random Undersampling, Random Forest, Apache Spark, H2O,

11. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

12. National Institutes of Health, "About the PATH Study," [Online]. Available: https://pathstudyinfo.nih.gov/about. [Accessed: Dec. 12, 2024].