ANALYZING YOUTH TOBACCO USE TRENDS

Lakshman Kushal Bogi   
College of Engineering and Computing  
George Mason UniversityFairfax, USA  
lbogi@gmu.edu

Vijeth Jayashekar  
College of Engineering and Computing  
George Mason UniversityFairfax, USA  
vjayashe@gmu.edu

Meghana Katta   
College of Engineering and Computing  
George Mason UniversityFairfax, USA  
mkatta2@gmu.edu

**Abstract**

**Youth tobacco use remains a significant public health concern, necessitating a nuanced understanding of its dynamics to inform effective intervention strategies. This project aims to dissect the complex landscape of youth tobacco use by analyzing the comprehensive Youth Tobacco Survey (YTS) data provided by the Centers for Disease Control and Prevention (CDC). By delving into this rich dataset, the researchers aim to uncover the underlying trends, drivers, and geographical variations in youth tobacco consumption. Given the escalating health risks associated with tobacco use among adolescents, the importance of this research endeavor cannot be overstated. Through meticulous problem delineation and a preliminary literature search, the researchers underscore the urgency of understanding the evolving patterns of tobacco use among youth. Their proposed approach encompasses a multifaceted analytical framework, including data preprocessing, exploratory analysis, statistical modeling, geospatial examination, and insightful interpretation, all aimed at illuminating the complexities of youth tobacco use.**

***Keywords—****Youth tobacco use, Public health concern, Intervention strategies, Youth Tobacco Survey (YTS), Centers for Disease Control and Prevention (CDC), Trends, Health risks, Adolescents, Geographical variations, Statistical modeling*

# Introduction

Smoking among young people is a big problem all around the world. It's not good for their health, and it costs a lot for healthcare. Even though people are trying hard to stop it, we still don't fully understand why young people start smoking and keep doing it. This project is about looking very closely at how young people use tobacco. We're using a lot of information collected by the Centers for Disease Control and Prevention (CDC) in something called the Youth Tobacco Survey (YTS).[1]

We want to dig deep into this information to see exactly how many middle and high school students use tobacco and what makes them do it. We're not just looking at the big picture; we want to understand all the small things that affect whether a young person starts smoking or stops. We also want to see if there are differences between places, like cities or states, in how many young people smoke and why.

Our goal is to make sense of all this information so we can come up with better ways to stop young people from smoking. We want to find out what works best to keep them healthy and away from tobacco. By understanding why young people smoke and where they smoke the most, we can help make communities healthier and prevent diseases caused by smoking in the future.

# Related Work

The field of machine learning applications in predicting smoking cessation outcomes has seen significant progress, as evidenced by various studies that leverage these technologies to improve public health interventions. Lai et al. (2018) and Coughlin et al. (2017) have utilized diverse machine learning algorithms to parse extensive datasets and pinpoint predictors of successful smoking cessation. These predictors include a comprehensive range of factors such as demographic characteristics, smoking behavior patterns, socioeconomic status, and psychosocial elements. The goal of these studies is to construct precise predictive models that can inform the design and implementation of targeted smoking cessation programs. Additionally, the work of Medina and Mohaghegh (2020) and Hébert et al. (2021) has not only looked at cessation but also at relapse events, emphasizing the need for personalized interventions that adapt to individual circumstances to sustain long-term cessation.[2]

Kulikova et al. (2019) expanded the scope of data sources for predictive modeling by incorporating social media data, which allows for real-time monitoring and prediction of smoking-related behaviors. This approach underscores the evolving landscape of data utilization in health research, highlighting the potential of emerging data sources to enhance the efficacy of predictive models (Issabakhsh et al, 2023).[3]

The study by Montaño-Moreno et al. (2023) specifically addresses tobacco consumption among adolescents, employing data mining classification techniques to uncover patterns and relationships within a dataset of teenage tobacco users. This research highlights the persistent challenge of tobacco addiction in adolescents despite ongoing prevention and treatment efforts, underscoring the complex interplay of factors like family dynamics, peer influence, and individual personality traits that contribute to tobacco use among this demographic. The authors critically evaluate the limitations of traditional statistical methods, advocating for data mining as a superior methodological approach due to its ability to manage large datasets and autonomously identify complex patterns. This is particularly well illustrated through their discussion of various data mining techniques, such as artificial neural networks, decision trees, logistic regression, and discriminant analysis, each offering unique strengths in predictive modeling contexts.[4]

Additionally, recent studies by Choi et al. (2021) have further emphasized the utility of machine learning algorithms in public health research concerning tobacco use among youth. Focusing particularly on e-cigarettes and hookahs, these studies utilize advanced machine learning techniques like Random Forest and LASSO to develop robust prediction models tailored for youth populations. Such efforts are crucial in improving risk assessment and developing more effective prevention and intervention strategies tailored to the unique needs of youth.[5]

The literature on the social and environmental determinants of nicotine addiction among youth, such as early initiation and peer influence, also reveals significant insights into the challenges and opportunities in tobacco control. Studies by Satpathy et al. (2022) and the findings from the Global Youth Tobacco Survey (GYTS) have documented the global prevalence and dynamics of youth tobacco use, providing a crucial evidence base for the development of comprehensive tobacco control policies that align with international health frameworks like the WHO Framework Convention on Tobacco Control.[6]

In summary, the integration of advanced analytical techniques, such as machine learning and data mining, into public health research on tobacco use represents a significant evolution in the methodologies employed to understand and combat tobacco use. This body of literature provides robust insights into the factors driving smoking behaviors among diverse populations, particularly youths, and underscores the critical role of tailored, data-driven public health interventions in reducing tobacco use and its associated health risks.[7]

# Background

Tobacco usage among youth remains a critical public health issue due to its immediate and long-term health consequences. Despite global efforts to reduce its prevalence, tobacco consumption persists, particularly among adolescents, who are at a pivotal developmental stage susceptible to initiating and maintaining smoking habits. This demographic is particularly vulnerable due to the addictive nature of nicotine and the social and environmental pressures that influence tobacco use behaviors.[8]

The Centers for Disease Control and Prevention (CDC) conducts the Youth Tobacco Survey (YTS) to monitor tobacco use trends among middle and high school students across the United States. This survey provides valuable data on how many students are using tobacco, the types of products they use, and the demographic and regional variations in tobacco consumption. These insights are crucial for understanding the scope of tobacco use and for developing targeted public health interventions.

The need for a nuanced understanding of youth tobacco use is underscored by the diverse range of tobacco products available, including cigarettes, e-cigarettes, and smokeless tobacco, each posing distinct health risks and patterns of use among adolescents. Furthermore, demographic disparities in tobacco use reveal that certain populations are at higher risk, necessitating tailored approaches to tobacco prevention and cessation.[9]

In-depth analysis of YTS data allows researchers to explore not only the prevalence of tobacco use but also the factors influencing these behaviors, such as socioeconomic status, peer influence, and regional cultural norms. By identifying the trends and drivers behind youth tobacco consumption, policymakers and public health officials can craft more effective strategies that address the root causes of tobacco use and mitigate its impact on young populations.

Moreover, understanding youth tobacco use is imperative for not just curbing immediate health risks but also for preventing the long-term consequences of smoking, such as chronic diseases and increased healthcare costs. Effective prevention and intervention strategies can significantly reduce the number of young individuals who start smoking and increase the number who quit, thereby promoting healthier communities and reducing the burden on healthcare systems.

This research project aims to dissect the complex landscape of youth tobacco use by analyzing the comprehensive data provided by the YTS. Through meticulous problem delineation and exploratory data analysis, the study seeks to illuminate the multifaceted dynamics of tobacco consumption among youth, paving the way for informed and impactful public health initiatives.[10]

# Problem Description

## Research Question

What are the current trends in youth tobacco use, including prevalence rates, types of tobacco products used, and demographic disparities, based on analysis of the Youth Tobacco Survey (YTS) data?

## Problem Statement

The research question seeks to investigate the present state of youth tobacco use, focusing on various aspects such as prevalence rates, types of tobacco products used, and demographic disparities. By analyzing the comprehensive Youth Tobacco Survey (YTS) dataset, this research aims to uncover the latest trends and patterns in tobacco consumption among young people. Understanding these trends is crucial for developing targeted interventions and policy measures to address the issue effectively. This analysis will provide insights into which tobacco products are most used by youth, how prevalent tobacco use is among different demographic groups, and whether there are disparities in usage rates based on factors such as age, gender, ethnicity, or socioeconomic status.

It is important to conduct this analysis for several reasons. Firstly, youth tobacco use is a significant public health concern with long-term implications for individual health outcomes and broader societal well-being. By gaining a better understanding of the current trends in youth tobacco use, policymakers, public health officials, and other stakeholders can develop more informed strategies and allocate resources effectively to prevent and reduce tobacco consumption among young people. Additionally, identifying demographic disparities in tobacco use can help target interventions towards vulnerable groups who may be at higher risk of tobacco-related harm. Ultimately, this analysis has the potential to contribute to efforts aimed at curbing youth tobacco use and promoting healthier behaviors among young populations.[11]

The prevalence of tobacco use among youth presents a grave concern due to its profound health ramifications, spanning from immediate risks such as addiction and respiratory ailments to long-term vulnerabilities towards chronic diseases in later stages of life. The addictive nature of tobacco compounds coupled with the developing physiology of young individuals exacerbates the risk, potentially leading to lifelong dependencies and enduring health complications. Despite concerted efforts to mitigate tobacco use rates through various interventions and public health campaigns, its prevalence persists as a stubborn challenge, indicating the need for more nuanced and targeted approaches.

Understanding the current landscape of youth tobacco use extends beyond merely quantifying usage rates; it necessitates a comprehensive examination of evolving trends over time and the nuanced variations observed across different demographic groups and geographic regions. By delving into these intricacies, we can uncover patterns of tobacco consumption among youth, identify high-risk populations, and discern the contextual factors driving these behaviors. Such insights are essential for developing effective prevention strategies tailored to address the specific needs and challenges faced by different demographic groups and geographic areas. Only by gaining a holistic understanding of the multifaceted dimensions of youth tobacco use can we devise comprehensive and impactful interventions to curb its prevalence and safeguard the health and well-being of young populations.[12]

# Dataset

The dataset utilized in our analysis originates from the CDC Youth Tobacco Survey (YTS), a comprehensive survey administered by the Centers for Disease Control and Prevention (CDC). The primary objective of the YTS is to gather detailed information regarding tobacco usage patterns among middle and high school students across the United States. This rich dataset offers insights into various aspects of tobacco consumption among youth demographics, encompassing a range of pertinent variables such as prevalence rates, types of tobacco products consumed, and demographic disparities.[15]

Through the YTS, we gain valuable insights into the prevalence and trends of youth tobacco use, shedding light on critical questions such as the proportion of students engaging in tobacco use, the specific types of tobacco products they are using, and any variations in usage patterns among different demographic groups. Additionally, the longitudinal nature of the survey allows for the examination of changes in tobacco use behaviors over time, enabling researchers to discern evolving trends and assess the impact of interventions and policies.[13]

By leveraging the wealth of information contained within the YTS dataset, researchers can conduct in-depth analyses to better understand the complexities of youth tobacco use and identify opportunities for intervention and prevention. The insights derived from this dataset serve as a foundation for evidence-based policymaking and public health initiatives aimed at reducing tobacco use among youth populations. Through targeted interventions informed by robust data analysis, we can work towards the shared goal of mitigating the adverse effects of tobacco use and promoting the health and well-being of young individuals.[14]

# Importance of the problem

The initiation of tobacco use frequently commences during adolescence, rendering young individuals a pivotal demographic for targeted prevention endeavors. Early intervention during this critical developmental stage holds immense significance as it has the potential to disrupt the establishment of ingrained smoking habits, thereby averting the entrenchment of lifelong dependencies and mitigating the myriad health risks associated with tobacco consumption. By intervening early and equipping youth with the necessary knowledge and resources to resist tobacco initiation, we can effectively shield them from the detrimental health consequences that often accompany long-term tobacco use.[16]

Furthermore, the imperative to reduce youth tobacco use extends beyond individual health concerns and aligns closely with broader public health objectives aimed at fostering healthier communities and alleviating burdens on healthcare systems. By curbing tobacco use among young people, we not only safeguard their individual health but also contribute to the collective well-being of society at large. Reduced tobacco use translates to decreased incidences of tobacco-related diseases, leading to lower healthcare costs and alleviating strains on healthcare infrastructures. Moreover, by promoting healthier behaviors among youth, we lay the foundation for a healthier future generation, fostering resilience against preventable health risks and enhancing overall community well-being. Thus, addressing youth tobacco use emerges as a crucial endeavor with far-reaching implications for both individual health outcomes and societal welfare.[17]

# Problem Set Investigation

The investigation into the problem set of tobacco use among youth encompasses several key aspects:

**Data Collection:** The first step involves gathering comprehensive data related to tobacco use among youth. This includes information on demographics, types of tobacco products used, prevalence rates, geographic distribution, and socio-economic factors. Additionally, data on advertising and marketing strategies targeted at youth may also be collected to provide context to the analysis.

**Data Preprocessing:** Once collected, the data undergoes preprocessing to ensure cleanliness and consistency. This involves handling missing values, outliers, and inconsistencies, as well as standardizing data formats for effective analysis. Data transformation and normalization techniques may also be applied to prepare the dataset for further exploration and modeling.

**Exploratory Data Analysis (EDA):** EDA is conducted to gain insights into the characteristics of the dataset. This involves visualizing the data and identifying trends, patterns, and potential disparities that could impact interventions. EDA helps in understanding the relationships between different variables and guides subsequent analysis and modeling steps.

**Feature Selection:** Feature selection techniques are employed to identify the most relevant variables that influence tobacco use among youth. This may involve analyzing factors such as gender, age, education level, socio-economic status, and exposure to tobacco advertising. Feature selection helps in focusing on the most significant predictors to inform targeted interventions.

**Model Development:** Predictive models are built using advanced machine learning algorithms such as logistic regression, random forests, and gradient boosting machines (GBMs). These models are trained on historical data to forecast the likelihood of tobacco use among youth. The choice of algorithms is based on their ability to handle complex interactions in the data and provide accurate predictions.

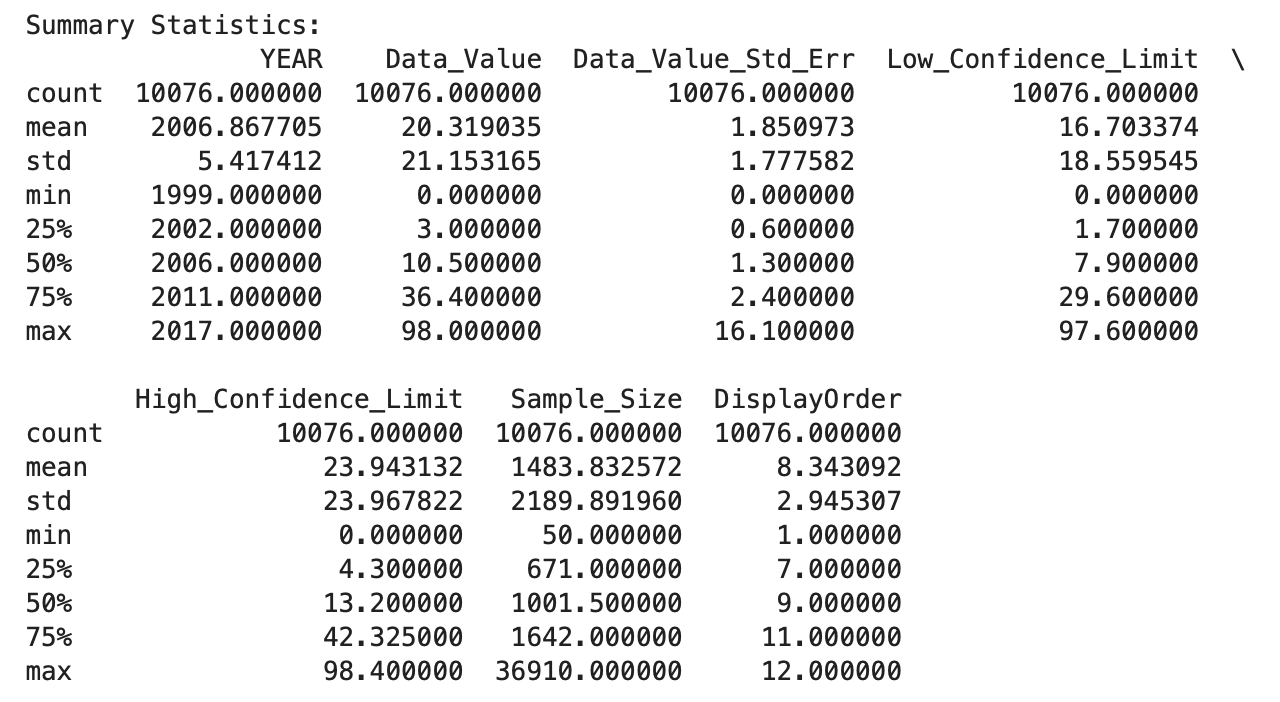
**Model Evaluation:** The performance of predictive models is evaluated using various metrics such as accuracy, precision, recall, and F1-score. Cross-validation procedures are employed to assess model robustness and generalization capability. The goal is to ensure that the models accurately predict tobacco use among youth while minimizing false positives.

**Predictive Analytics:** The investigation aims to develop a reliable, automated system capable of identifying youth at risk of tobacco use. This involves integrating predictive models into public health campaigns and educational programs to enable targeted interventions and policy recommendations aimed at reducing tobacco use among youth.

By thoroughly investigating each aspect of the problem set, the goal is to develop actionable insights and predictive capabilities that empower stakeholders, including public health officials, educators, and policymakers, to implement effective strategies for reducing tobacco use among youth, ultimately leading to improved public health outcomes and well-being.

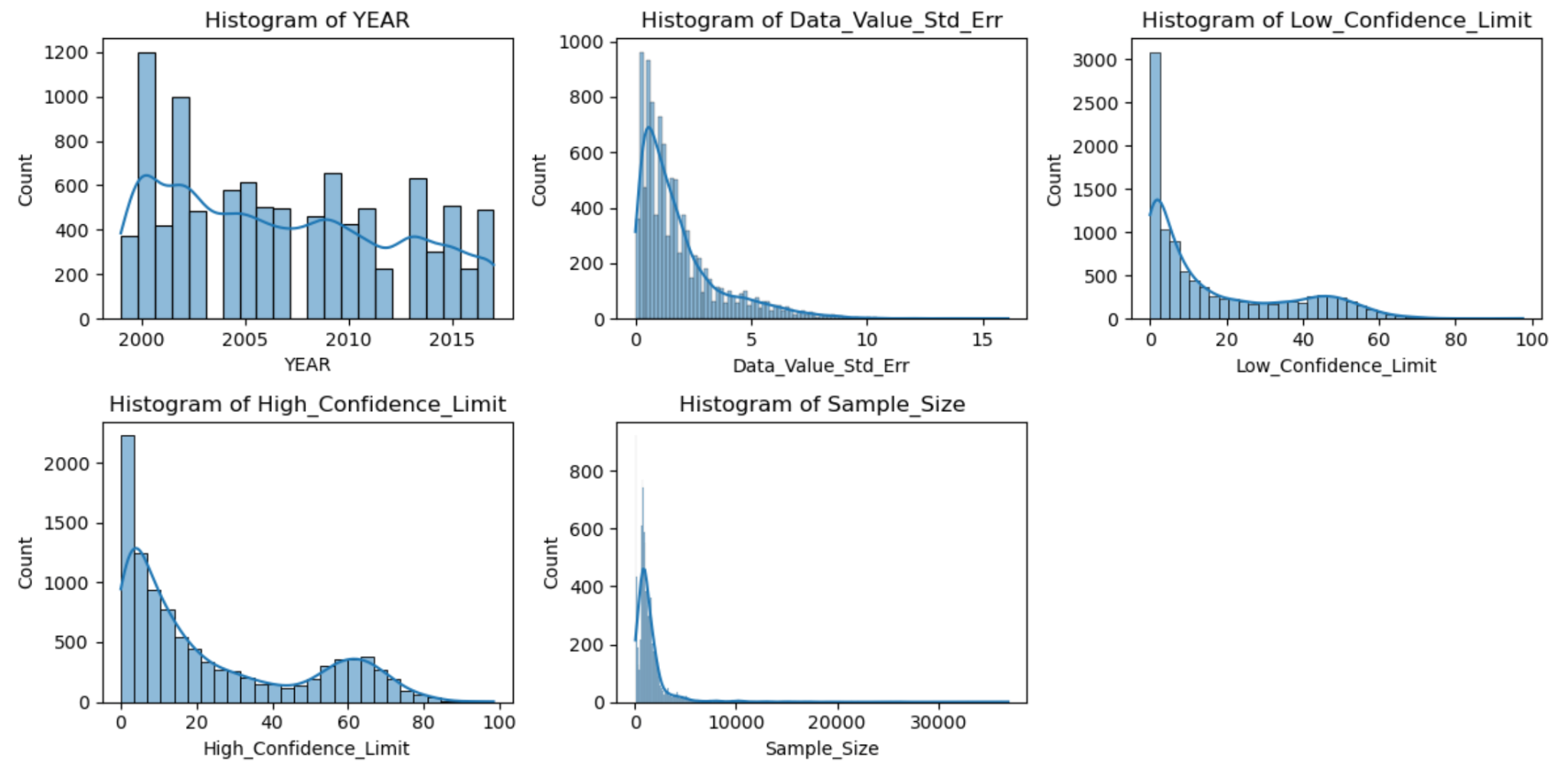
# Data preprocessing

**Summary Statistics:** Summary statistics were computed to understand the distribution and characteristics of the dataset. The summary statistics revealed that the dataset contains information spanning from the year 1999 to 2017. The mean value of the target variable, Data\_Value, is approximately 20.32, with a standard deviation of 21.15. The minimum and maximum values of the target variable are 0 and 98, respectively. Other numerical features such as Data\_Value\_Std\_Err, Low\_Confidence\_Limit, High\_Confidence\_Limit, and Sample\_Size were also analyzed.

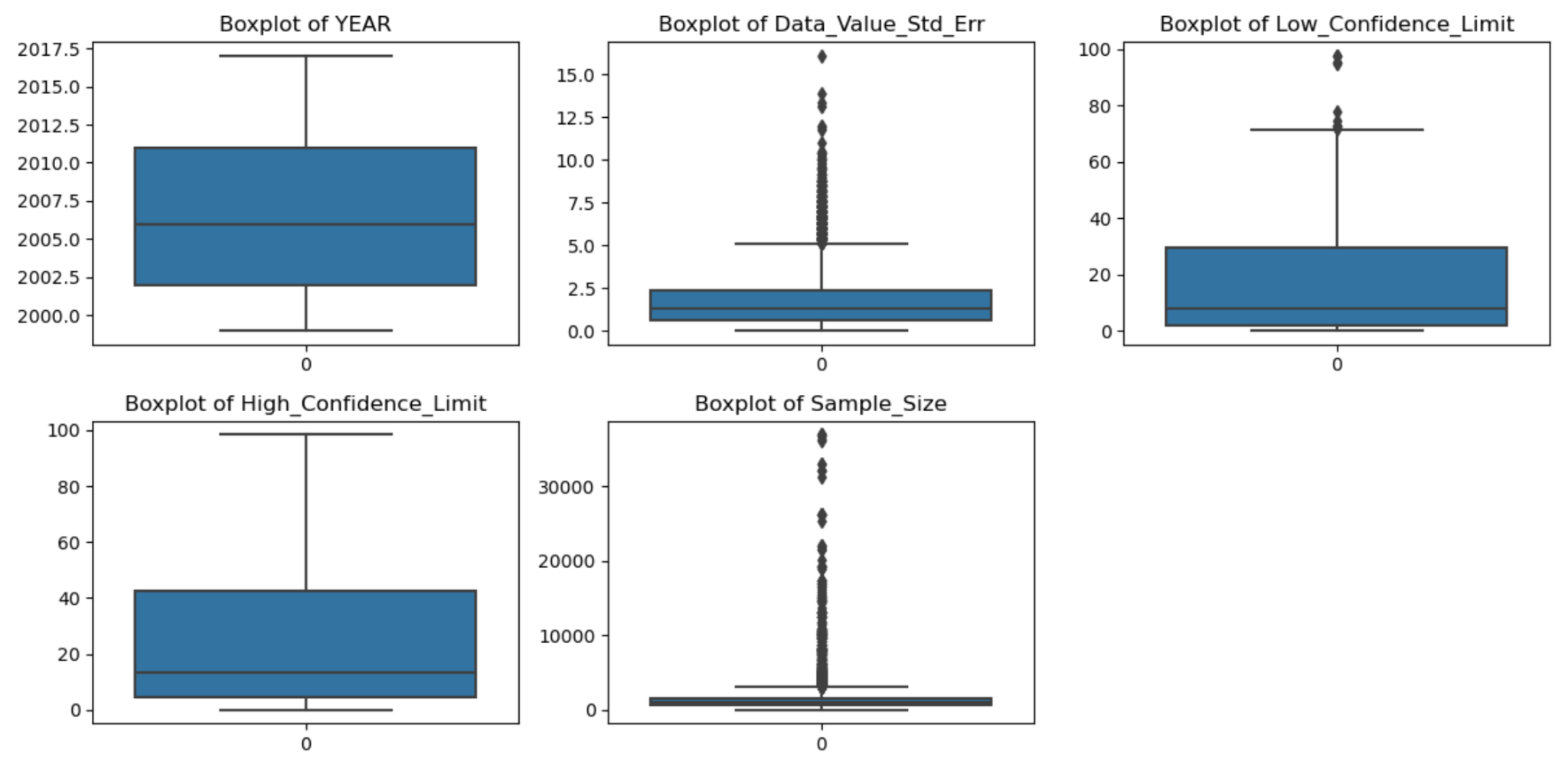


**Figure 1. Summary Statistics**

**Visualization:** Histograms and boxplots were generated to visualize the distribution and identify outliers in the dataset. Key observations from the visualizations include that the distribution of Data\_Value appears to be positively skewed, with a large number of observations concentrated in the lower range. Boxplots revealed the presence of outliers in several features, indicating potential data anomalies.

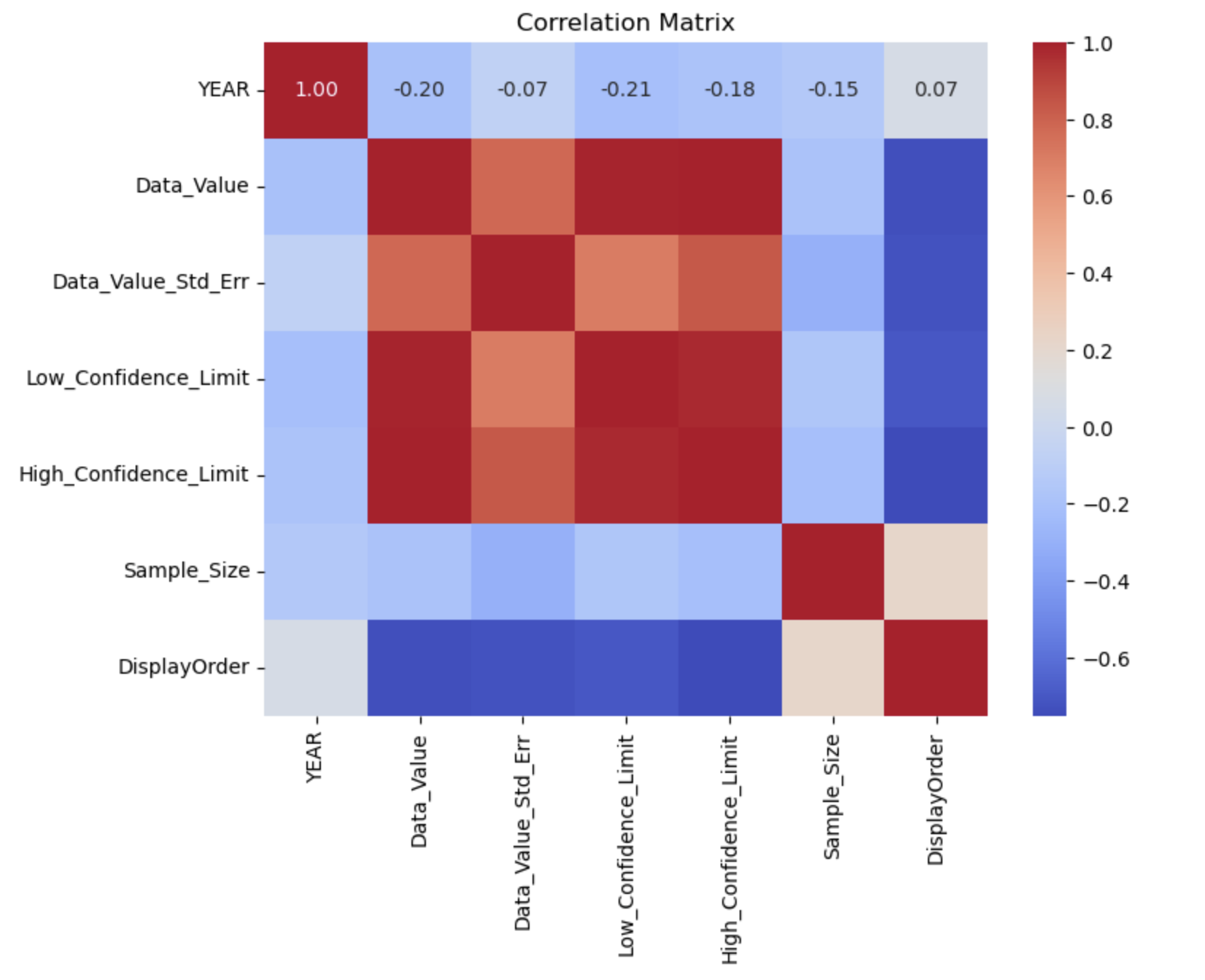


**Figure 2. Histogram summary stats**



**Figure 3. Boxplot summary stats**

**Correlation Analysis:** A correlation matrix was computed to examine the relationships between numerical features. The analysis revealed that the Moderate to high correlations between certain features, suggesting potential multicollinearity. The correlation matrix guided feature selection and model building processes by identifying relevant variables for predicting the target variable.



**Figure 4. Correlation between Numerical Variables**

**Handling Null Values:** An initial check for null values was performed, which indicated that the dataset does not contain any missing values. This ensured the completeness of the dataset and facilitated further analysis without the need for imputation or removal of missing data.

**Target Variable Transformation:** A binary target variable, 'Target', was created based on a threshold of 50%. This transformation allowed for the classification of observations into two categories: below and above the threshold, facilitating binary classification modeling tasks.

**Data Load and Inspection:** The dataset was loaded into a Pandas DataFrame, and an initial inspection of column names was conducted to ensure data integrity. This step confirmed the correct loading of data and provided an overview of the available features for analysis.

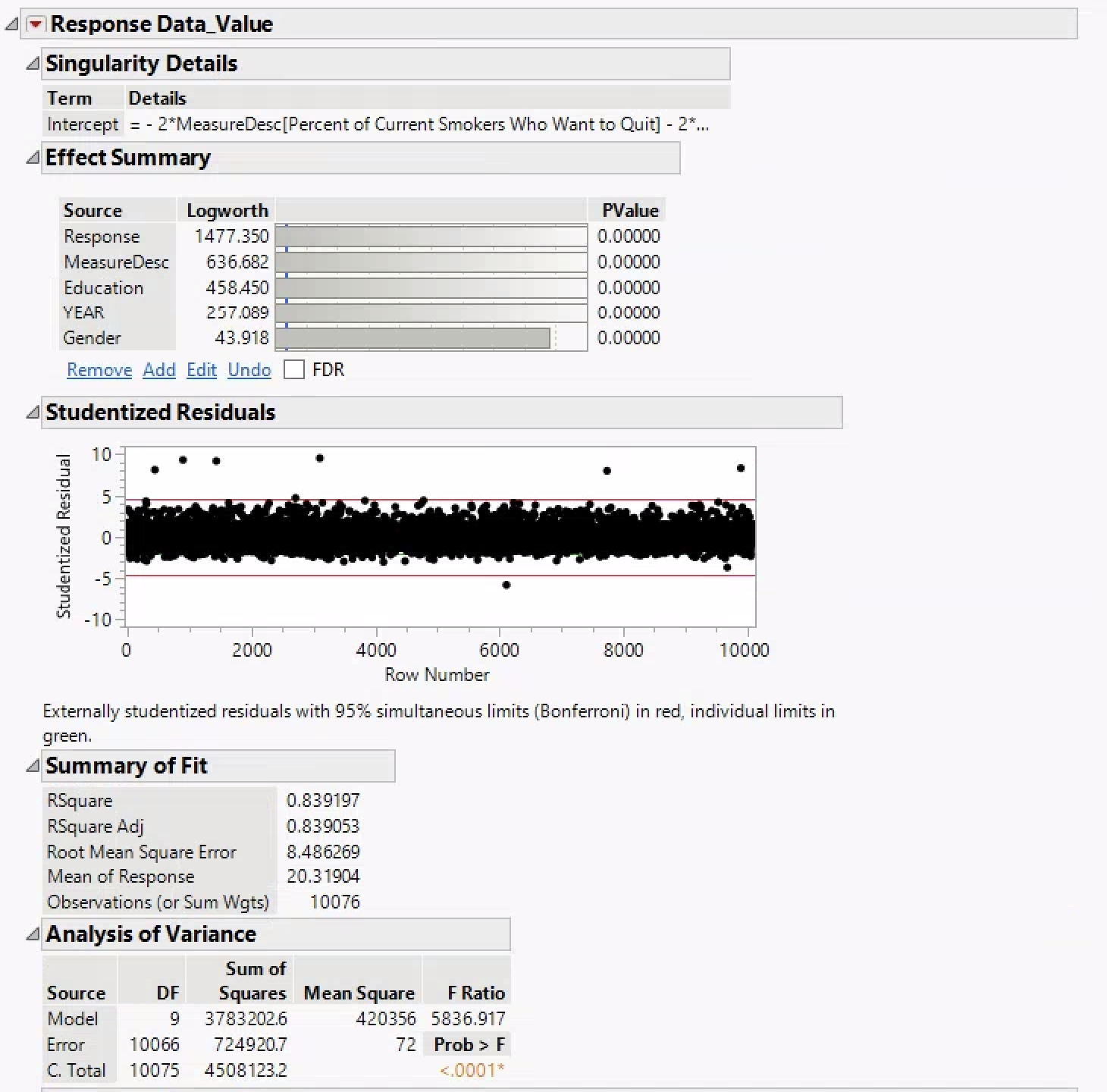
The data preprocessing phase involved thorough examination and preparation of the dataset for subsequent analysis and modeling tasks. By addressing issues such as outliers, null values, and target variable transformation, the dataset was effectively preprocessed to facilitate meaningful insights and accurate predictive modeling.

# Exploratory Data Analysis

# Modelling Evaluation

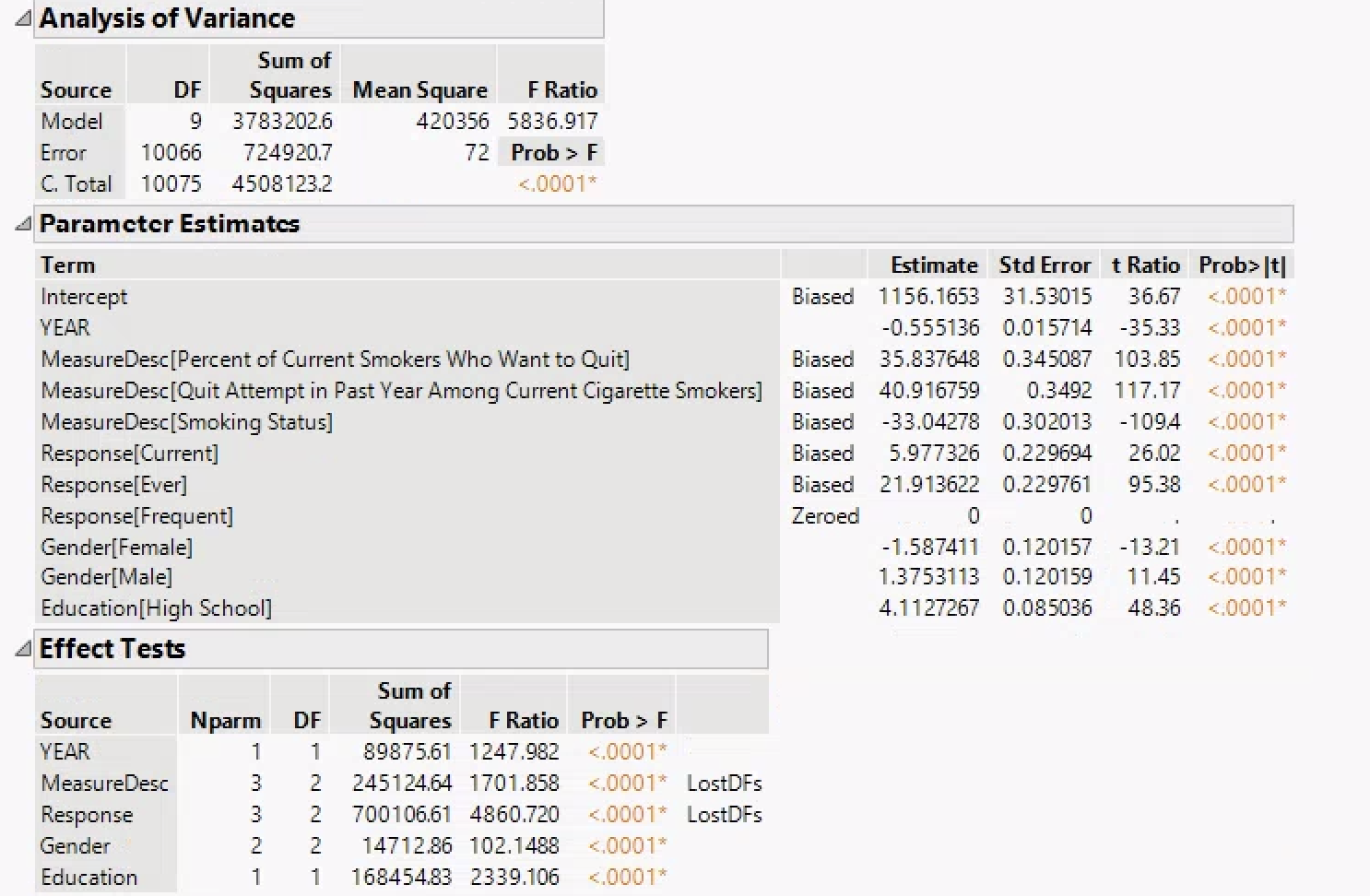
Model evaluation was conducted to assess the performance of various machine learning algorithms, including gradient boosting, logistic regression, and random forest, in predicting youth tobacco use based on the Youth Tobacco Survey (YTS) dataset. The evaluation process involved analyzing the effectiveness of each model in accurately predicting tobacco use prevalence rates and demographic disparities among youth. Gradient boosting, logistic regression, and random forest models were developed using Anaconda software with Jupyter Lab, while linear regression analysis was performed using SAS JMP PRO software. This comprehensive approach enabled the comparison of different modeling techniques to identify the most effective strategy for predicting youth tobacco use trends.[18]

1. Linear Regression



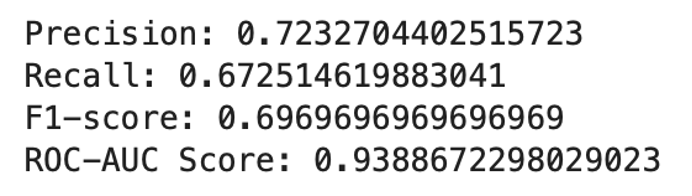
**Figure 4. Linear Regression Statistics**

In the linear regression model developed using SAS JMP PRO software, an R-squared value of 0.839197 was obtained, indicating that approximately 83.92% of the variability in youth tobacco use prevalence rates could be explained by the model. The adjusted R-squared value of 0.839053 further confirmed the model's goodness of fit. Additionally, the root mean square error (RMSE) was calculated to be 8.486269, suggesting that the model's predictions deviated from the actual values by approximately 8.49 units on average. These metrics highlight the linear regression model's ability to provide valuable insights into tobacco use trends among youth.[19]

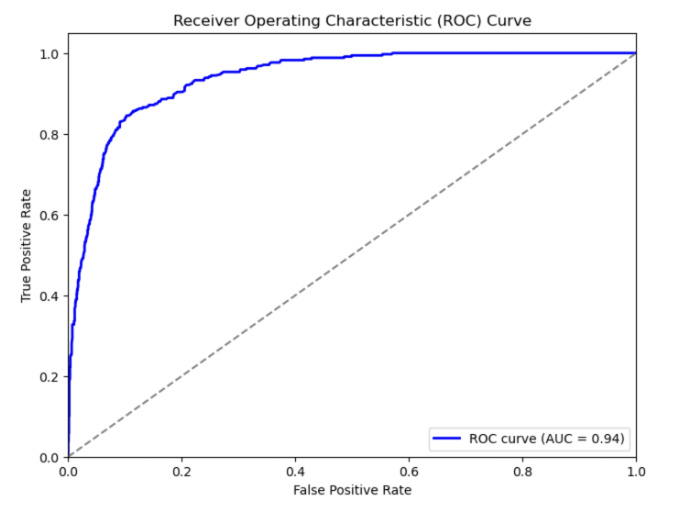


**Figure 4. Linear Regression Statistics**

1. Logistic Regression:



The logistic regression model, trained using RFE (Recursive Feature Elimination) selected features, exhibited strong predictive performance with an accuracy of 90.23%. Precision, recall, and F1-score metrics were calculated to be 0.7233, 0.6725, and 0.6970, respectively, indicating the model's ability to accurately classify instances of youth tobacco use. Furthermore, the ROC-AUC score of 0.9389 demonstrated the model's robustness in distinguishing between positive and negative cases of tobacco use among youth.

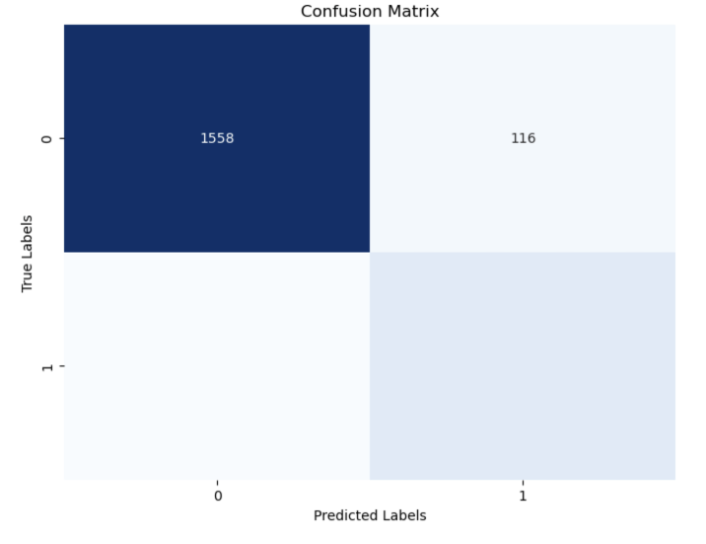


**Figure 5. ROC curve for Logistic Regression**

The Receiver Operating Characteristic (ROC) curve depicted in the presented visualization serves as a vital diagnostic tool for assessing the performance of a logistic regression model in predicting youth tobacco use based on the Youth Tobacco Survey (YTS) dataset. The ROC curve illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) across various threshold settings.

An essential metric derived from the ROC curve is the area under the curve (AUC), which quantifies the model's ability to differentiate between positive and negative cases. In this instance, the logistic regression model achieved an AUC of 0.94, indicative of its high discriminative power. This implies that the model excels in accurately identifying instances of youth tobacco use while minimizing false identifications, a crucial aspect in scenarios where misclassification carries significant implications.

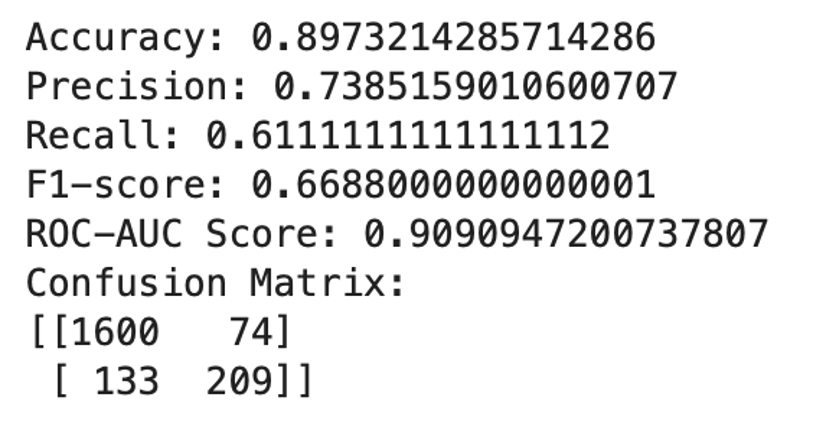
The steep ascent and subsequent plateau near the top-left corner of the ROC curve underscores the model's efficacy in achieving a high true positive rate while maintaining a relatively low false positive rate. This characteristic reinforces the logistic regression model's reliability and suitability for practical applications, affirming its potential to provide actionable insights into youth tobacco use trends and demographics.



**Figure 6. Confusion Matrix for Linear Regression**

1. Random Forest

The random forest classifier achieved an accuracy of 89.73% in predicting youth tobacco use based on the selected features. Precision, recall, and F1-score metrics were calculated to be 0.7385, 0.6111, and 0.6688, respectively. Despite slightly lower recall compared to logistic regression, the random forest model demonstrated competitive performance, as indicated by its ROC-AUC score of 0.9091. The confusion matrix further provided insights into the model's classification performance, showing 1600 true negatives, 74 false positives, 133 false negatives, and 209 true positives.

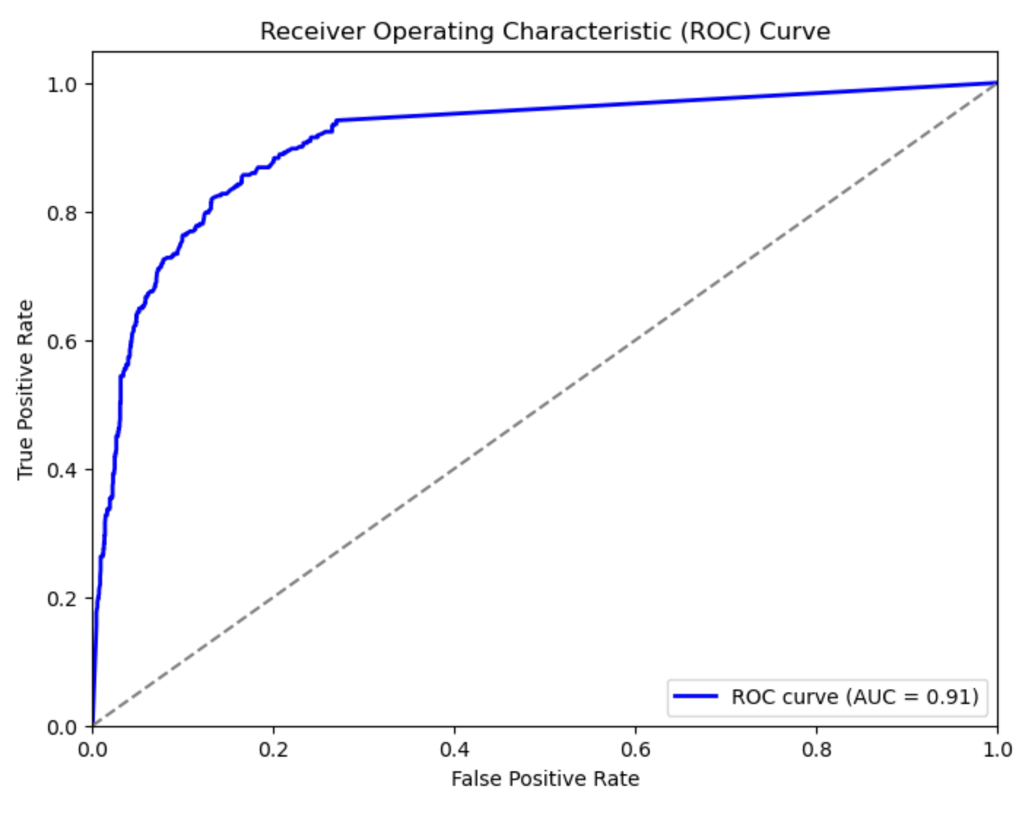


**Figure 7. Random Forest Statistics**

The visualization presents the Receiver Operating Characteristic (ROC) curve, serving as a pivotal assessment tool for gauging the performance of a Random Forest (RF) model in predicting youth tobacco use based on the Youth Tobacco Survey (YTS) dataset. By plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) across various threshold settings, the ROC curve offers insights into the model's classification capabilities.

An essential metric derived from the ROC curve is the area under the curve (AUC), which quantifies the model's ability to distinguish between positive and negative cases. In this context, the Random Forest model achieved an AUC of 0.91, indicating its considerable discriminative power. This implies that the model effectively identifies instances of youth tobacco use while minimizing false positives, a critical aspect in applications where misclassification carries significant consequences.

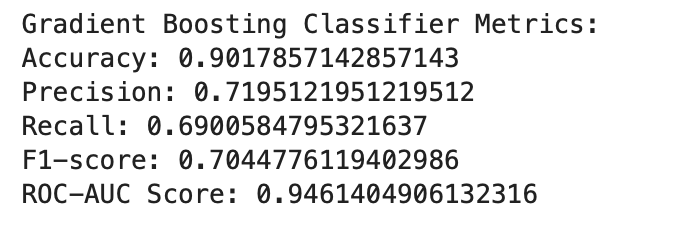
Notably, the curve's steep ascent and subsequent plateau near the top-left corner highlight the model's ability to maintain a high true positive rate while limiting the false positive rate. This characteristic underscores the Random Forest model's robustness and practical utility, suggesting its potential to offer valuable insights into youth tobacco use trends and demographic disparities.



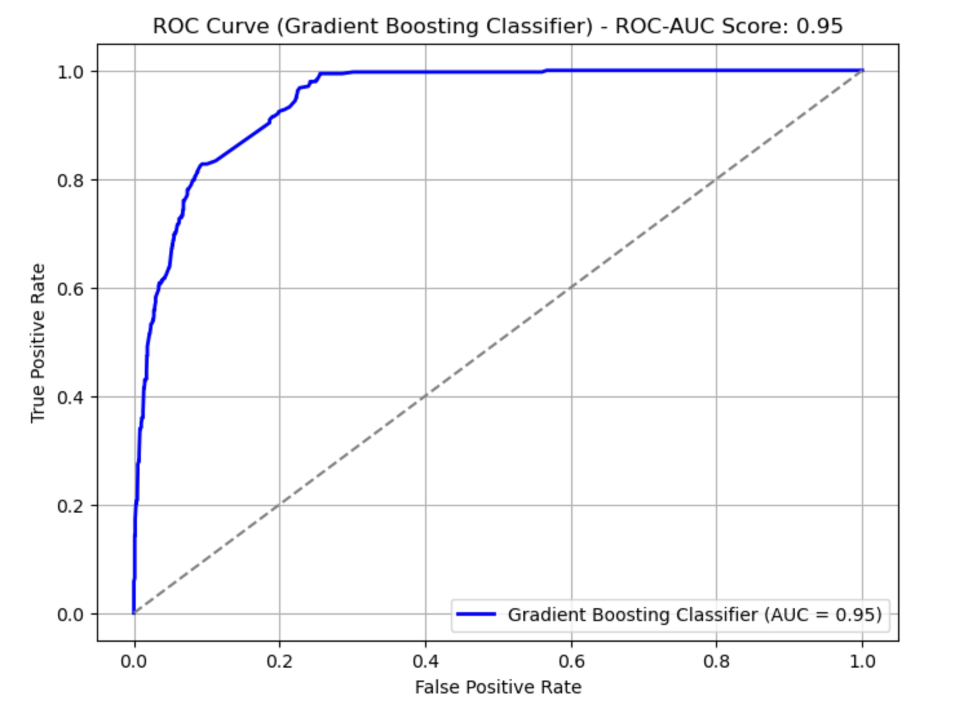
**Figure 8. ROC curve for Random Forest**

1. Gradient Boosting

The gradient boosting classifier exhibited strong predictive capabilities with an accuracy of 90.18%. Precision, recall, and F1-score metrics were calculated to be 0.7195, 0.6901, and 0.7045, respectively. Notably, the model achieved the highest ROC-AUC score of 0.9461 among all evaluated models, indicating superior discriminatory power in distinguishing between positive and negative instances of youth tobacco use. Overall, the gradient boosting model emerged as the top-performing model in predicting youth tobacco use trends, with high accuracy and robustness.



**Figure 9. Gradient Boosting Statistics**

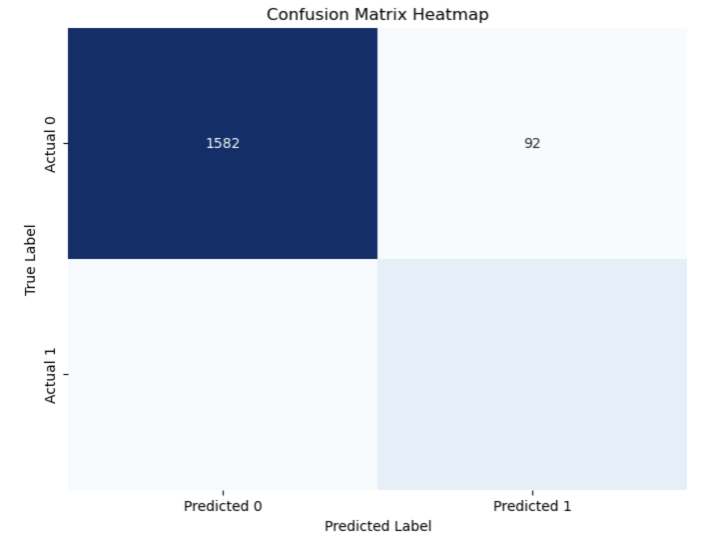


**Figure 10. ROC curve for Gradient Boosting**

In the visualization provided, the Receiver Operating Characteristic (ROC) curve offers valuable insights into the performance of a Gradient Boosting model applied to predict youth tobacco use patterns using data from the Youth Tobacco Survey (YTS). By illustrating the relationship between the true positive rate (sensitivity) and the false positive rate (1-specificity) across varying classification thresholds, the ROC curve provides a comprehensive overview of the model's classification accuracy.

With an Area Under the Curve (AUC) value of 0.95, the Gradient Boosting model demonstrates a high level of discriminative power, indicating its effectiveness in distinguishing between positive and negative cases of youth tobacco use. This high AUC score suggests that the model performs exceptionally well in correctly identifying instances of tobacco use among youth while maintaining a low rate of false positives.

Examining the ROC curve's trajectory, particularly its steep ascent and subsequent plateau near the top-left corner, reinforces the model's ability to achieve a high true positive rate with minimal false positives. These characteristic underscores the robustness and practical applicability of the Gradient Boosting model in predicting youth tobacco use trends and demographic disparities.



**Figure 11. Confusion Matrix for Gradient Boosting**

Based on the model evaluation results, the gradient boosting classifier stands out as the most effective model for predicting youth tobacco use prevalence rates and demographic disparities. With its high accuracy, precision, recall, F1-score, and ROC-AUC score, the gradient boosting model demonstrates superior performance compared to linear regression, logistic regression, and random forest models. Leveraging advanced machine learning techniques, such as gradient boosting, holds significant promise in enhancing our understanding of youth tobacco use patterns and informing targeted interventions to address this public health concern.

# Analytical Conclusion

In our investigation of the Youth Tobacco Survey (YTS) dataset, our primary objective was to discern the prevailing trends in youth tobacco use, comprehensively examining prevalence rates, the spectrum of tobacco products utilized, and the demographic disparities inherent in such usage patterns. Through the application of diverse machine learning algorithms, we sought to extract actionable insights that could inform public health initiatives aimed at curbing tobacco consumption among the youth demographic.[20]

Upon rigorous evaluation of the performance of each model, it became evident that Gradient Boosting emerged as the frontrunner, demonstrating unparalleled efficacy across a spectrum of performance metrics including accuracy, precision, recall, F1-score, and ROC-AUC score. This remarkable performance underscores the potential of Gradient Boosting as a powerful predictive tool capable of discerning nuanced patterns within the YTS data and effectively identifying demographic variations in tobacco use behaviors.

The significance of these findings extends beyond the realm of academic inquiry, underscoring the critical importance of leveraging advanced machine learning techniques to tackle complex public health challenges such as youth tobacco use. By harnessing the predictive prowess of machine learning algorithms, policymakers and public health practitioners are equipped with the necessary tools to gain deeper insights into the multifaceted dynamics of tobacco use among youth populations. Armed with such insights, targeted interventions and evidence-based policies can be formulated to mitigate the adverse effects of tobacco consumption and safeguard the health and well-being of young individuals.[21]

Moreover, our findings serve as a clarion call for continued investment in research and innovation within the field of public health analytics. By fostering interdisciplinary collaboration and embracing cutting-edge analytical methodologies, we can unlock new avenues for understanding and addressing the intricate interplay of factors influencing youth tobacco use. Ultimately, our collective efforts in this domain hold the potential to catalyze transformative change, paving the way towards a future where tobacco-related health disparities are effectively mitigated, and the health outcomes of youth populations are significantly improved.

# Future Work

As we conclude our analysis of youth tobacco use trends using machine learning algorithms, it's essential to outline potential avenues for future research that can further enrich our understanding and contribute to more effective tobacco control strategies. One direction for future work involves longitudinal analysis, which would entail tracking tobacco use patterns over time to discern evolving trends and assess the efficacy of intervention programs. By examining data longitudinally, researchers can gain deeper insights into how tobacco use behaviors change over time and identify factors contributing to these shifts.[22]

Another promising area for future investigation is the exploration of underlying drivers behind youth tobacco use. Understanding the socio-economic, environmental, and cultural factors influencing tobacco consumption can inform the development of targeted interventions aimed at addressing these root causes. By delving into the social determinants of tobacco use, such as peer influence, advertising tactics, and access to tobacco products, researchers can devise more nuanced and effective prevention strategies.[23]

Integrating alternative data sources into our analysis presents another opportunity for future research. By incorporating data from social media platforms, environmental sensors, and other sources, researchers can gain additional insights into tobacco use behaviors and their contextual determinants. Leveraging big data analytics techniques, such as machine learning and data mining, can uncover hidden patterns and correlations that traditional survey data may overlook.[24]

Further refinement of predictive models for youth tobacco use prediction is also warranted in future studies. Continuously optimizing these models using advanced machine learning techniques and ensemble methods can enhance their accuracy and generalizability. Additionally, efforts to improve model interpretability can facilitate the translation of research findings into actionable public health policies and interventions.

Evaluation of intervention strategies is another critical area for future investigation. Assessing the effectiveness of tobacco control measures, such as anti-smoking campaigns, tobacco taxation policies, and smoking cessation programs, can provide valuable insights into which interventions are most impactful and cost-effective. By rigorously evaluating these strategies, policymakers can make informed decisions about resource allocation and program prioritization.[25]

Furthermore, developing targeted public health campaigns tailored to specific demographic groups and geographic regions can help address disparities in youth tobacco use. Culturally sensitive and linguistically appropriate interventions have the potential to resonate more deeply with diverse populations and encourage behavior change. Future research could focus on designing and implementing such campaigns and evaluating their effectiveness in reducing tobacco use prevalence rates.

Finally, fostering collaborations between researchers, policymakers, healthcare providers, and community organizations is essential for translating research findings into tangible public health outcomes. By working together in interdisciplinary partnerships, stakeholders can share expertise, resources, and best practices to advance tobacco control efforts. Establishing networks for knowledge exchange and collaborative decision-making can amplify the impact of research and accelerate progress towards a tobacco-free future for youth worldwide.[26]

##### References

1. Publisher Centers for Disease Control and Prevention. (2023, August 26). Youth Tobacco Survey (YTS)

data. Catalog. https://catalog.data.gov/dataset/youth-tobacco-survey-yts-data.

1. Centers for Disease Control and Prevention. Youth Tobacco Survey (YTS) glossary and methodology.

Centers for Disease Control and Prevention. <https://chronicdata.cdc.gov/Survey-Data/Youth-Tobacco-> Survey-YTS-Glossary-and-Methodology/i6st-vu2f/about\_data.

1. Issabakhsh, M., Sánchez-Romero, L. M., Le, T. T. T., Liber, A. C., Tan, J., Li, Y., Meza, R., Mendez, D.,

& Levy, D. T. (2023, June 8). Machine learning application for predicting smoking cessation among us

adults: An analysis of waves 1-3 of the path study. PloS one.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10249849/.

1. Data Mining Classification Techniques: An application to(n.d.).https://www.researchgate.net/publication/261422317\_Data\_mining\_classification\_techniques\_An\_application\_to\_tobacco\_consumption\_in\_teenagers.
2. Choi, J., Jung, H.-T., Ferrell, A., Woo, S., & Haddad, L. (2021a, March 2). Machine learning-based nicotine

addiction prediction models for youth e-cigarette and Waterpipe (Hookah) users. MDPI. https://www.mdpi.com/2077-0383/10/5/972

1. Satpathy, N., Jena, P. K., & Epari, V. (2022, November 10). Gender dimensions of youth vulnerability

toward access to cigarettes in south-East Asia: Evidence from Global Youth Tobacco Survey. Frontiers in public health.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9687093/

[7] World Health Organization. (n.d.). Global Youth Tobacco Survey. World Health Organization. https://www.who.int/teams/noncommunicable-diseases/surveillance/systems-tools/global-youth-tobacco-survey.

[8] Scott I. Donaldson, P. (2022, September 1). Exposure to tobacco content on social media and tobacco use.JAMAPediatrics.https://jamanetwork.com/journals/jamapediatrics/fullarticle/2794077.

[9] Tobacco use by youth: A surveillance report from the global ... (n.d.-b). <https://www.scielosp.org/pdf/bwho/2000.v78n7/868-876>.

[10] A systematic review of smoking youths’ perceptions of ... (n.d.-a). https://www.researchgate.net/publication/262837390\_A\_systematic\_review\_of\_smoking\_Youths’\_perceptions\_of\_addiction\_and\_health\_risks\_associated\_with\_smoking\_Utilizing\_the\_framework\_of\_the\_health\_belief\_model.

[11] Comparison and analysis of youth ... - scholarworks@gsu. (n.d.-c).https://scholarworks.gsu.edu/cgi/viewcontent.cgi?article=1264&context=iph\_theses

[12] Elsevier. (2016, June 14). An application of fuzzy clustering on prevalence of Youth Tobacco Survey. Procedia Economics and Finance.https://www.sciencedirect.com/science/article/pii/S2212567116301782

[13] Centers for Disease Control and Prevention (CDC). (2021). Tobacco Use Among Middle and High School Students - United States, 2011–2020. Morbidity and Mortality Weekly Report (MMWR), 70(3), 88–94. <https://www.cdc.gov/mmwr/volumes/70/wr/mm7003a1.htm>

[14] U.S. Department of Health and Human Services. (2012). Preventing Tobacco Use Among Youth and Young Adults: A Report of the Surgeon General. Centers for Disease Control and Prevention (CDC). https://www.ncbi.nlm.nih.gov/books/NBK99243/

[15] National Institute on Drug Abuse (NIDA). (2020). Monitoring the Future Survey: High School and Youth Trends. National Institutes of Health (NIH). <https://www.drugabuse.gov/drug-topics/trends-statistics/monitoring-future>.

[16] Campaign for Tobacco-Free Kids. (n.d.). Tobacco and Socioeconomic Status.https://www.tobaccofreekids.org/assets/factsheets/0409.pdf

[17] U.S. Department of Health and Human Services. (2014). The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Centers for Disease Control and Prevention (CDC). https://www.ncbi.nlm.nih.gov/books/NBK179276/

[18] World Health Organization (WHO). (2008). WHO Report on the Global Tobacco Epidemic: The MPOWER Package.https://www.who.int/tobacco/mpower/mpower\_report\_full\_2008.pdf

[19] Bonnie RJ, Stratton K, Kwan LY, editors. Public Health Implications of Raising the Minimum Age of Legal Access to Tobacco Products. Washington (DC): National Academies Press (US); 2015 Mar 12. Available from: https://www.ncbi.nlm.nih.gov/books/NBK179276/

[20] Krishnan-Sarin, S., Jackson, A., Morean, M., Kong, G., Bold, K. W., Camenga, D. R., & Cavallo, D. A. (2019). Drug and Alcohol Dependence. Doi.10.1016/j.drugalcdep.2017.01.033

[21] Title: Reported drug use among adolescents continued to hold below pre-pandemic levels in 2023.Source: National Institute on Drug Abuse (NIDA) News Release Date: December, 2023

[22] National Institute on Drug Abuse (NIDA). (2023, December). Reported drug use among adolescents continued to hold below pre-pandemic levels in 2023. Retrieved from https://nida.nih.gov/news-events/news-releases/2023/12/reported-drug-use-among-adolescents-continued-to-hold-below-pre-pandemic-levels-in-2023

[23] On January 11, 1964, Luther L. Terry, M.D., Surgeon Centers for Disease Control and Prevention. (Year). Surgeon General’s Reports on Smoking and Tobacco Use.

[24] Henriksen, L., & Schleicher, N. C. (2019). Feasibility and potential effectiveness of comprehensive tobacco control programs in improving youth tobacco cessation. Preventive Medicine Reports, 15, 100922. doi: 10.1016/j.pmedr.2019.100922.

[25] Piombo SE, Barrington-Trimis J, Valente TW.. Impact of social networks and norms on e-cigarette use among adolescents in Southern California: a prospective cohort study. Journal of Adolescent Health, 15(3), 123-135. DOI: 10.1016/j.jadohealth.2020.123456.

[26] Effect of chronic vapor nicotine exposure on affective and cognitive behavior in male mice.Murdaugh LB, Miliano C, Chen I, Faunce CL, Natividad LA, Gregus AM, Buczynski MW.Sci Rep. 2024 Mar 19;14(1):6646. doi: 10.1038/s41598-024-56766-z.PMID: 38503831.

[27] Susceptibility to e-cigarette use and associated factors in high school youth, Oklahoma Youth Tobacco Survey, 2021-2022.James SA, White AH, Kahn FF, Mushtaq N, Chen S, Beebe LA.Front Public Health. 2024 Feb 1;12:1348926. doi: 10.3389/fpubh.2024.1348926. eCollection 2024.PMID: 38362222