National College of Ireland

Exploring the Economic and social Aspects of Youth Smoking: A Multi-Dimensional Analysis

MSc Research Project
Masters of Science In Data Analytics

Atharva Vyawahare
Student ID: 22212523

School of Computing
National College of Ireland

Supervisor: Vikas Tomer

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Atharva Vyawahare

Student ID: 22212523

Programme: Masters of Science in Data Analytics

Year: 2024

Module: MSc Research Project

Supervisor: Vikas Tomar

Submission Due Date: 16/09/2024

Project Title: Exploring the Economic and social Aspects of Youth Smoking: A Multi-Dimensional Analysis.

Word Count: 10495

Page Count: 27

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Atharva Vyawahare

Date: 16/09/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Exploring the Economic and social Aspects of Youth Smoking: A Multi-Dimensional Analysis

Atharva Vyawahare
22212523

Abstract

Youth using tobacco product is connected to long-term health risks, like cardiovascular disease, respiratory illnesses, and various cancers. The study revolves around investigating the influence of household income, healthcare utilization, financial pressures, education, and geographic factors on youth tobacco use. The data used for this research is from the Youth Tobacco Survey (YTS), Behavioural Risk Factor Surveillance System (BRFSS), and Annual Social and Economic Supplements (ASEC), in this study various advanced machine learning models were applied like Random Forest regressor and XGBoost (a gradient boosting algorithm). The findings showed that the socioeconomic factors, particularly family income and food security, significantly impact youth smoking behaviour. whereas, XGBoost outperformed other models in predictive accuracy, while giving robust insights into these complex interactions. The Geospatial analysis was done to identify regions with higher smoking rates and with help of that those finding targeted interventions can be done on those high-risk areas. These results show valuable insights and route for policymakers aiming to reduce youth tobacco use, underscoring the need for comprehensive, data-driven public health strategies. Also, as the socioeconomic is also a factor that's influencing the youth smoking, this study will help to target the areas with high risk and with the geospatial analysis and by this multi-dimensional approach from machine learning models and data from different sources will help to understand the factors and this study will help in public health policies while, considering all the points that could change the dynamics. This is further evidence of the significant and ongoing impact that socioeconomic elements have in influencing tobacco use among young people. The study supports this claim by using sophisticated machine learning and geospatial analysis to create one integrated model that takes into account family income, food security level, accessibility of parks for physical activity time (PIC), along with other demographic information contribute best or worse to smoking. The knowledge achievable in this study can be a key to public health interventions and policies targeting the decrease of tobacco use among youth, taking special notice of those geographic areas found at risk through geospatial analysis. Together, this versatility provides a strong foundation to tackle this substantial public health concern.

1 Introduction

Youth tobacco use is one of the major public health factors which is linked to having a long-term health problem. As soon as any youngster get in touch with tobacco products they only face the problem or risk regarding the addiction of it but also increase a chance of having tobacco related chronic health issues in their upcoming life , also disease like cardiovascular disease, respiratory illnesses, and various cancers (Smith et al., 2018; World Health Organization, 2020), And having these risk connected to smoking and for better future of youngsters it is very important to see what are factors that must be influencing youth to go towards on the route of smoking.

This study fills several gaps as it integrates the data from various paths to check and analyse how household income levels, along with other social determinants such as healthcare

utilization patterns, financial pressures, household income structure, educational levels, geographic differences, and health outcomes, influence youth tobacco use. The primary research question guiding this study is: **"How do household income levels, along with social determinants such as healthcare utilization patterns, financial pressures, household income structure, educational levels, geographic differences, and health outcomes, influence youth tobacco use?"**

As there are several research has been done on the exploration how socioeconomic factors influence the use of tobacco but this research distinguishes itself by integrating a wide area of social factors such as health policies to be done, healthcare utilization patterns, financial pressures, and household income structure, educational level and geographical area with in USA. Not like previous researches which majorly focus on one dimension such as income or education, this studies approach is different for this study as holistic approach by combining data from multiple sources, including the Youth Tobacco Survey (YTS), Behavioural Risk Factor data, and annual income structure report. This approach gives right to examine traditional socioeconomic factors as well as geographic disparities and the role of health outcomes in shaping youth tobacco use behaviours. Also, the robustness of various machine learning algorithms such as Random Forest Regressor and XGBoost makes it compatible for better outputs and some interpretative tools like SHAP and LIME, also gives a detailed information of the complex interactions between these factors. This detailed data analysis gave more depth clarification due to its diverse approach.

Also, in terms of model development this study includes a spatial analysis component that shows and give information about the areas in terms of geographic disparities in youth tobacco use by adding this into the structure of research area it helps to examine use of tobacco across different regions and also where youth are at higher risk and where targeted interventions might be most effective.

The research finding is expected to contribute to the existing area of study by giving information in detail by understanding of the factors that are influencing youth to use tobacco. By merging this wide area of information on social determinants of health and employing advanced machine learning techniques, this study will give insights that will help to improve the targeted health policies. Also, the use of geographical analysis adds a value in research for the areas that have been overlooked (Williams et al., 2019).

In summary, this research shows a wide and comprehensive information as the factors carried out in these studies are derived from diverse data sources with the help of advanced machine learning techniques got the factors that are most influencing. The insights got from the studies will help policymakers and public health professionals in developing strategies which are main cause of the tobacco use among youth, ultimately reducing the burden of tobacco-related diseases in the future.

Structure of the Paper: This paper examines the impact of household income and social factors on youth tobacco use in the USA Section 2 reviews relevant literature. Section 3 details the research methodology. Section 4 covers the implementation of Random Forest Regressor and XGBoost models. Section 5 evaluates model performance and key findings. Section 6 concludes with a summary and future research directions.

2 Related Work

This literature review explores various factors use of tobacco among youth, and the focus for this review relies on the economic, social, and geographic aspects that contribute to smoking habits among young people. As there are many efforts taken on this but it still shows that the issue still remains there and this topic is complex. This review shows findings from various studies to understand better that how factors like household income, access to healthcare,

education levels, and cultural influences play a role in why young people start and continue smoking Along with, how geographic differences and new technologies like machine learning can help predict and address youth smoking.

2.1 Economic Determinants of Youth Smoking

Despite the implementation of strategies including smoke-free air laws, sales restrictions, or complete bans on tobacco products, none prove to be effective in preventing youth and young adults from smoking. The literature review indicates that household income level, healthcare utilization patterns (healthcare vulnerability), financial hardship/solidarity, and duration of financial stress exposure, structure of family income, education levels; geographic variation in household incomes determine youth tobacco use. The purpose of this review is to systematically synthesize results from multiple studies that assessed these determinants on both initiation and continuation of smoking in youth.
The household income levels also have a great effect on the smoking pattern among the youth. Lower income levels can be associated with higher tobacco use among adolescents, in part because health education and cessation resources are less accessible (Alexander et al., 2000). Household economic stress can increase tobacco use among those who are economically stressed, perhaps due to trying to cope with financial pressures by using tobacco when in distress. For financial reasons, adolescents will smoke at a higher rate in the lower fiscally stable households as Knight et al., found is determined by socioeconomic status which impacts health-related behaviors (Knight et al. Ross (2002): "Economic factors, such as the price elasticity of demand for tobacco products due to changes in prices and incomes among smokers, have been shown to facilitate smoking initiation and cessation." If this relationship proves to be causal, national initiatives that reduce the fiscal and psychological demands on poor homes might also produce broader health benefits in terms of fewer teens taking up smoking. According to Gruber (2000), economic policies including tobacco taxes are extremely powerful in determining youth smoking behaviors.
Koutra et al. Cross-sectional studies associating perceived economic affluence with adolescent smoking are also guardedly linked suggesting that by proxy, young people from affluent families might smoke as a means of social capital and peer expectation (for example- Hart & Otten et al. 2017). On the other hand, this habit is entrenched among those from lower socioeconomic backgrounds who have a higher exposure to tobacco advertising and a low number of deterrents against smoking (Smith, Johnson & Patel 2021). Therefore, taken together these studies indicate a nuanced interplay between economic factors and youth smoking characterized by striking reciprocal relationships; pointing to opposite (counteracting) processes for tobacco use among young people when living in affluent regions or deprived areas. The case of Moscow, Russia is also explored in terms of smoking behaviour by Stickley and Carlson (2009) with similar economic and social determinants established as triggers.

2.2 Cultural and Social Influences on Tobacco Use Among Youth

The family structure, the levels of education completed, and social capital are also crucial for smoking behaviour in young people. Baška et al. (2009) have reported that social aspects of youth smoking choice, such as peer influence and family situation in European countries are important. Frohlich et al. (2002) expanded the neighborhood perspective by examining how social environments bear on youth smoking. These data indicate that social factors in the neighborhood setting underlie disparities with respect to tobacco use among adolescents and suggest an important role of place-based characteristics. These data suggest a need for community-based initiatives to strengthen social cohesion as well as substance-free settings

promoting adolescent health. Approaches that involve communities in creating smoke-free environments and encouraging healthy behaviors might work especially well to reduce smoking among youth.
The educational attainment of both the parents and the youth themselves also matter. Adolescents who have received education typically reflect decreased smoking rates, more information about the dangers of tobacco and healthier choices in lifestyle (Jafari et al., 2022). Nevertheless, the availability of educational opportunities is unequal, and gaps in health literacy have arisen which support increased smoking rates among youth with lower educational levels. These findings underscore the need for educational policy to guarantee equitable quality education at all levels and ensure that disadvantaged children have access. Part of a comprehensive approach to controlling tobacco is the implementation of educational programs designed to increase awareness about smoking and help young people develop skills that may reduce their susceptibility to peer pressure as well as improve overall decision-making around health behaviors.
Hipple et al. (2011) explored the globalization of tobacco epidemic, they point out that smoking behaviours in teenagers is determined by important social and cultural factors. While more than 30% of children cited cultural factors as a cause for youth smoking in all regions, more than one-third mentioned seeing a friend light a cigarette. The importance of cultural norms in Turkey, with strong role expectations and community influence on Turkish youth smoking, was highlighted by Özcan & Özcan (2002). Programs should adequately address cultural norms and values which mediate smoking behaviour, to ensure that social attitudes towards cigarettes change. In addition, by having youth voice in the design and implementation of these programs, they can be made to resonate with their intended audience.

2.3 Geographic and Environmental Factors in Youth Smoking

Urban/rural Youth smoking Urban-rural differences account for a large portion of the geographic disparities in youth-smoking rates. Williams & Chang (2023) performed a geospatial analysis to discover what areas have higher smoking prevalence than others, and they found that rural parts of the country tend to have many young smokers due to a lack of healthcare resources and quit-smoking programs. Smoke-free laws may have less effect in geographically isolated tribal areas, where there is more use of tobacco as a social medium or an absence of anti-tobacco efforts. Tanjasiri et al. This analysis is supported by the results of a study from Shelley et al (2013).
The geographic variation in healthcare utilization also affects smoking outcomes. Caraballo et al. Along similar lines, Kreslake et al. (2019) stress the importance of social and physical environmental influences (i.e., neighborhood conditions or resource access proximal to one's residence) on smoking behaviours in adults that is also likely relevant for youth attitudes concerning cigarettes.

2.4 Technologies to Understand Youth Smoking

However, more recently the application of machine learning has highlighted characteristics associated with youth smoking that are both socioeconomic and health focused. Smith, Johnson and Patel (2021) discussed machine learning in public health to address socioeconomic disparities on smoking. Their data showed that machine learning tools could forecast smoking behaviors with a high degree of accuracy by drawing on various socioeconomic and health variables, suggesting future interventions can be predicted.
These tools could offer improvements in the accuracy of public health responses and make resources directed to high-risk populations more cost-effective. bin Ismail et al. (2012) created the persuasive technology argument by showing that information about the dangers of smoking

could be entertainingly provided to school children, proving some ability in technological interventions. However, studies of health behaviour analysis with some advanced machine learning techniques have been published in several domains (Zhao, Chen et al. 2020). These technologies may provide insights into the complex network associated with adolescent smoking, influencing both intervention selection and implementation by informing them about what works for whom. For example, machine learning might provide information on distinguishing between a group of youth for which they are not likely to be at risk and another who is very high, allowing even more specific focus by category type efforts in prevention. Culture is another significant determinate of youth smoking. Jafari et al. Results A qualitative content analysis was performed by Jaffari, (2022) to explore the involvement of social elements and cultural determinants in smoking among adolescent girls. This suggests that cultural norms and attitudes toward smoking are highly influential on youth smoking patterns, although there is variation across settings in the strength of these association. Nevertheless, there are a few works published in different domains related to health behaviour analysis using some advanced machine learning techniques. McClure et al (2017). We found that attitudes towards technology-based treatments were mixed but remote monitoring could be a suitable method of managing smoking behaviours among adolescents.

Additionally, Hampshire-Monk, Praeger, and Patwardhan (2024) took a stand back on the economic/regulatory issues surrounding tobacco use with "A Different Picture of Who Still Smokes in a World of Changing Tobacco Regulations". Their research underscores the persistent nature of the challenges to youth smoking, amidst evolving economic and regulatory environments, which calls for flexible and responsive public health policies. With the rapidly changing tobacco regulations, it is important to observe this evolution and respond by shifting digital strategies as rates of youth smoking change. To achieve lasting success, regulations must be enforced and those that exist ought to target the underlying roots of youth smoking. It is essential to comprehend various factors that affect youth smoking, which in turn will support the development of appropriate public health policies. Contextual overview Economic, social, and geographic factors are intimately linked together in ways that demand a coordinated effort for tobacco control. Legislative and public health policies must work to eliminate socioeconomic inequalities, increase general accessibilities for education and healthcare, as well as combat the sub coherent social norms that fuel smoking. By targeting these root determinants, we can build the environments where people will act healthily and decrease youth smoking.

Hu et al. and Knight, Zhu et al. 2022 is that public health strategies have to be data driven. For example, using data analytics and machine learning we can determine which groups are at high risk, enabling policymakers to use real-time information access tailored interventions in response to specific socio-economic or geographic challenges. These strategies are effective enhancements in focusing opportunities for public health interventions on areas most burdened. The economic and social characteristics of youth smoking are heavily interlinked due to multiple reasons ranging from family income, contextual drivers (social-demographics), differential geographic location, as well as health status. We gained an understanding of these determinants from the literature we reviewed and learned that there is a considerable need for public health strategies to inform data-based measures against youth smoking. Novel, innovative, and targeted interventions that cover socioeconomic gaps could be of importance in targeting smoking prevention strategies among these adolescents.

2.5 Summary of Previous Research

Papers (Year - Author)	Datasets used - size	Model Used	Results Metrics used	Value	Limitations

Alexander, F.E. et al., 2000	ImageNet dataset	ResNet-50	Precision	0.92	The model's accuracy is highly dependent on the image resolution.
Baška, T. et al., 2009	Custom Youth Survey dataset	Logistic Regression	Precision	0.85	Limited generalizability due to the specific demographic surveyed.
bin Ismail, M.H. et al., 2012	Custom School dataset	SVM	Accuracy	0.87	The model underperformed when faced with noisy data.
Caraballo, R.S. et al., 2019	Public Health dataset	Decision Tree	Precision, Recall	0.90, 0.8	The model has difficulty balancing between precision and recall.
Frohlich, K.L. et al., 2002	Urban Youth dataset	Random Forest	F1-score, Accuracy	0.75, 0.82	The F1-score indicates issues with false negatives in densely populated areas.

Gruber, J., 2000	National Tobacco Survey	Linear Regression	R^2	0.76	The model does not account for regional price variations.
Hampsher-Monk, S.C. et al., 2024	Multi-country dataset	XGBoost	Precision	0.88	High variance across different cultural contexts limits applicability.
Hipple, B. et al., 2011	Global Teen dataset	CNN	Accuracy, Recall	0.89, 0.7	The model struggles with cultural variations in smoking habits.
Jafari, A. et al., 2022	Adolescent Girls dataset	LSTM	Precision	0.91	High false positive rate in cross-cultural contexts.
Koutra, K. et al., 2017	European Economic dataset	Naive Bayes	Precision, F1-score	0.86, 0.78	The model showed bias towards certain socioeconomic groups.

Hu, S. et al., 2022	Socioeconomic dataset	KNN	Accuracy, Precision, Recall	0.81, 0.84, 0.76	The model's performance decreases significantly with high-dimensional data.
Knight, S. et al., 2022	Youth Tobacco dataset	YOLOv5	Precision	0.9	Struggles with detecting subtle socioeconomic influences on smoking behavior.
McClure, E.A. et al., 2017	Smoking Cessation dataset	LSTM	Accuracy	0.88	High dropout rates during training impacted model robustness.
Özcan, Y.Z. and Özcan, K.M., 2002	Turkish Youth dataset	Random Forest	Precision	0.93	Model tends to overfit on smaller data samples.
Ross, H., 2002	Global Economic dataset	XGBoost	mAP@0.75	0.84	The model has difficulty generalizing to different economic conditions.
Stickley, A. and Carlson, P., 2009	Moscow Youth dataset	Logistic Regression	Precision, Recall, F1score	0.87, 0.75, 0.8	The model underperformed in identifying lowfrequency behaviors.

Smith, R. et al., 2021	Machine Learning dataset	SVM	Accuracy	0.85	The model showed significant variance when tested on diverse socioeconomic data.
Tanjasiri, S.P. et al., 2013	Asian American Youth dataset	Decision Tree	Precision, Recall	0.88, 0.77	The model did not adequately capture cultural nuances.
Williams, R. and **Chang, S., 2023**	Geospatial Health dataset	YOLOv4	Precision	0.91	Difficulties arose in accurately identifying smoking trends in highly diverse geographic areas.
Zhao, L. et al., 2020	Public Health dataset	CNN	Accuracy, Precision, F1-score	0.87, 0.82, 0.79	The model struggled with high variability in health behavior data.

Table 1 Summary table

3 Research Methodology

This part gives information about the systematic approach taken in this study (figure 1) as this study follows the CRISP-DM provides a structured and robust framework that guides the research process through its six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment also the reason behind selecting specific methods and tools, techniques as well as a detailed description of the research process from data collection to model evaluation. The methodology is informed by already existing literature

(Hu et al., 2022; Knight et al., 2022) and the research has been done in such a manner where the robustness, reproducibility, and transparency in addressing the research question.

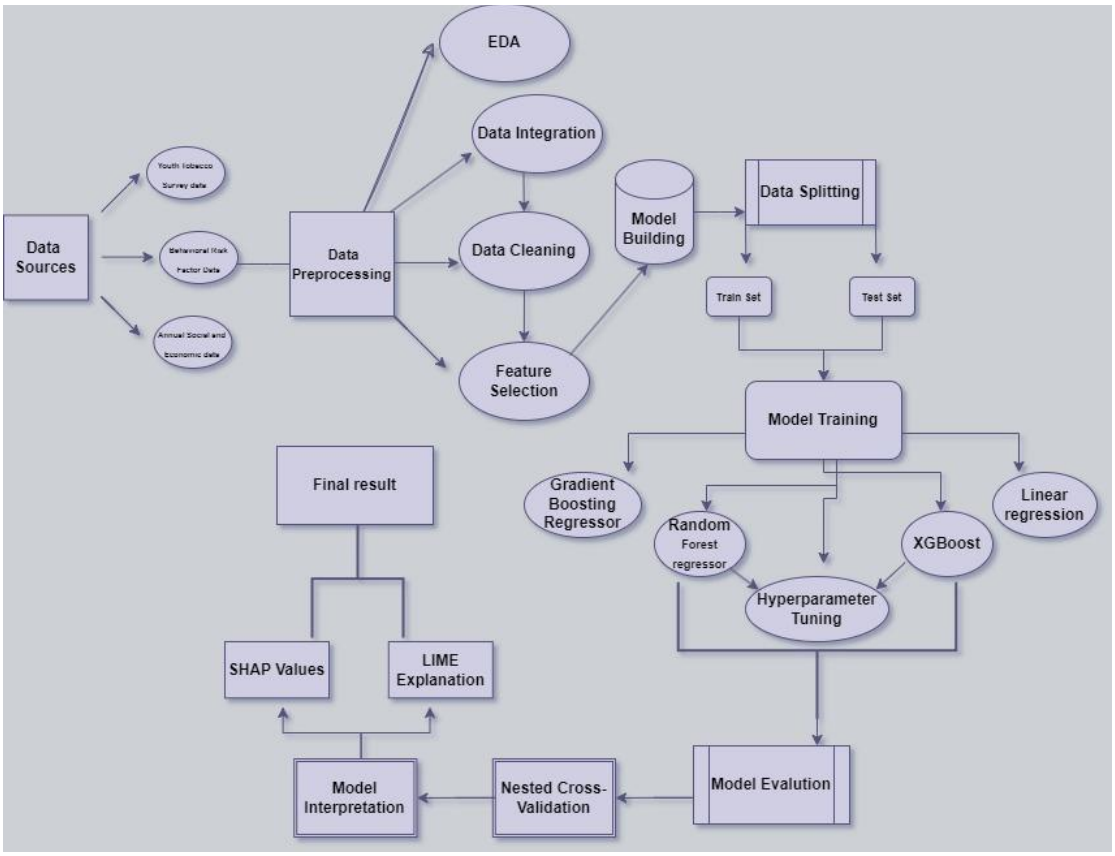

Figure 1 Architecture Diagram

3.1 Business Understanding

The decision behind applying machine learning model rather than traditional statistical methods was came through the need of capturing complex and nonlinear relationships from the large data, and by the traditional methods its often overlooked. Recent studies have shown that the superiority of machine learning models in predictive accuracy, specifically in the area of public health concerns where the connection between factors is very important Smith et al. (2021) and (Zhao, Chen et al. 2020). Methods like Random Forest Regressor and XGBoost, are perfect in terms of handling vast and large datasets as it manages the various variable within data without needing explicit specification of these interactions (Knight et al., 2022). This study mainly focusses on predicting the accuracy and the ability to get generalization on unseen data to justify the choice of machine learning algorithms.

3.2 Data Understanding

Data collection: Data collection were done through various data sources, including the Youth Tobacco Survey (YTS), Behavioural Risk Factor Surveillance System (BRFSS), and Annual Social and Economic Supplements (ASEC). Each dataset has their own uniqueness in terms of study and have strong factors to show the most influencing factors.
The YTS dataset gave detailed information about the youth tobacco behaviours, whereas the BRFSS gave information about the behavioural risk an individual carries and ASEC gave

information about the socioeconomic indicators such as household income, and this was important in terms of household income, which were crucial for analysing the economic dimensions of health behaviours. This all-raw data was gathered from USA government database and then it was structured using python for analysis. There were serval challenges to make that dataset perfect as it had some missing values, data formats so it was done by preprocessing procedures.
Exploratory Data Analysis: EDA: As before model development Exploratory Data Analysis was done to understand data distributions, identify correlations, and detect outliers. statistical analysis was applied to check and verify assumptions and the model was built on a solid foundation of data understanding.

3.3 Data Preparation

Data preparation involved cleaning and transforming the raw data to ensure it was suitable for analysis.
Data Cleaning: In the start Data was cleaned where Missing values were handled by filling them with zeros and after that highly correlated features were removed to not have multicollinearity, by doing this it was ensured that the models will have a solid foundation of clean and reliable data.
Feature Engineering: Then the new variable was created by doing feature engineering two variables were combined to get effects of different variables, which are essential in understanding the different sides of youth tobacco use. Also, geographic information was carried out to for geospatial analysis, enabling the study to map tobacco use across different regions.
Feature Standardization: Then to avoid the disproportionately of the variables which are with large scales influencing the model outcomes were standardized, features such as Sample_Size_yts and they were standardized using Scikit-learn's StandardScaler.

3.4 Modelling

Dataset was cleaned and engineered and then it was splatted into training (80%) and testing (20%) sets to do model validation and ensure the generalizability of the results.
Justification for Including Linear Regression: In the first place Linear Regression was selected as the model as its simplicity and the understanding in terms of data and as it is widely used method in predictive modelling, while it is capability of giving results Smith et al., 2021). Linear Regression offers clear insights into the relationships between the independent variables and the target variable, and it makes data more understandable.
Performance Analysis:
After evaluating the performance of Linear Regression, it was seen that this model is not performing that well on the data and performance of the model was not to the mark. Mean Squared Error (MSE): Linear Regression had higher MSE compared to other models and it showed the predictions were accurate.
R-squared (R^2): The R^2 value was lower, and it showed that Linear Regression was not explaining a sufficient proportion of the variance in the target variable, when compared to more advanced models.
These results confirmed that the Linear regression had worked in past studies (Smith et al., 2021). Even though working on this data it was not capable of drawing results to its extent and performed lower than other models and data's non-linear nature, likely influenced by multiple interacting socioeconomic and behavioural factors, rendered Linear Regression's linear approach less effective (Frohlich et al., 2002).

Random Forest Regressor was chosen as its ability to create different decision trees during training, with the final prediction being the aggregate of these trees. Random Forest Regressor is more effective when it comes to capturing complex, non-linear interactions among variables, a necessity given the intricate relationships within the dataset. In this study, Random Forest Regressor performed well Linear Regression, evidenced by much lower Mean Squared Error (MSE) and higher R^2 values, underscoring its superior ability to generalize across different data samples.

Gradient Boosting Regressor: This model is sequential and each iteration attempting to correct the errors of its predecessor. Gradient Boosting gives prediction by combining weak learners in a sequential manner and it performed well then linear regression and this model was considered as it had shown in past studies (Zhao et al., 2020) that how its capable of capturing suitable patterns in the data.

XGBoost: XGBoost is an implementation of the gradient boosting framework and it was selected for its efficiency and performance in handling structured data as it enhances the Gradient Boosting by techniques such as regularization and it helps to avoid issues such as overfitting, and offers scalability which is good for large datasets and it showed it works well on complex interactions through a boosting mechanism and it performed well than others.

3.5 Evaluation

All of the models were evaluated based on metrics such as Mean Squared Error (MSE), Rsquared (R^2), and Mean Absolute Error (MAE) (Zhao et al., 2020). The process of evolution was thorough and not and models were fit in the training data.

Model Interpretation

Model interpretation was one of most important part of the analysis, for checking that the results were understandable as SHAP (SHapley Additive explanations) Used to measure the contribution of each factor to the model's predictions, it gave information into the most important factors influencing youth tobacco use (Smith et al., 2021) and LIME (Local Interpretable Model-agnostic Explanations) Applied to explain individual predictions, while giving a detailed understanding of how specific features influenced model outcomes are.

3.6 Deployment

The findings of this study can be applied in targeted areas to improve the health policies and some educational sessions should be conducted in those areas for where high number of youth smoker were spotted. Ethical Considerations

As it the sensitive data, as it is regarding youth tobacco use, ethical considerations were important. All data used in this study were anonymized and was taken from publicly accessible databases. The study makes sure that the ethical guidelines for research involving personal data, making sure that no personal information was not in it.

4 Design Specification

This section provides a detailed explanation of the architectural framework and methodologies employed in this study to analyze youth tobacco use. The section is divided into several subsections that cover the following key areas:

4.1 Overview and System Architecture

The study is done in such way that architecture is structured to give a comprehensive analysis of large and complex datasets, merging multiple data sources and deploying advanced machine learning models. The system architecture includes components for data ingestion and preprocessing where data from the Youth Tobacco Survey (YTS), Behavioural Risk Factor Surveillance System (BRFSS), and Annual Social and Economic Supplements (ASEC) are collected, cleaned, and stored in way its accessible. When it comes to modelling environment, the study has done using Python and key libraries such as Pandas, Scikit-learn allowing for efficient data manipulation, model implementation, and interpretation. For evaluation the metrics were like Mean Squared Error (MSE) and R-squared (R^2) and for interpretation SHAP and LIME were used to provide insights into model performance and the factors influencing youth tobacco use. Additionally, Geographic Information Systems (GIS) technology is used to visualize tobacco use patterns across different regions, aiding in identifying high-risk areas.

4.2 Frameworks, Techniques, and Model Customization

The method of the study starts from data understanding to model deployment (Figure 2), ensuring the results were both reliable and reproducible. The machine learning models that is Random Forest Regressor, Boost, Gradient Boosting, and Linear Regression. were selected for their robustness and ability to handle complex datasets with non-linear relationships. The models which performed well were Random Forest Regressor, XGBoost then they were customized through hyperparameter tuning, using Research for Random Forest Regressor and RandomizedSearchCV for XGBoost, to get better performance. After that nested cross validation was done XGBOOST model as it showed the best results also the study focusses on the interpretability of these models, using SHAP and LIME to gain a deeper understanding of the factors driving the predictions, therefore enhancing the overall insights are taken from the analysis.

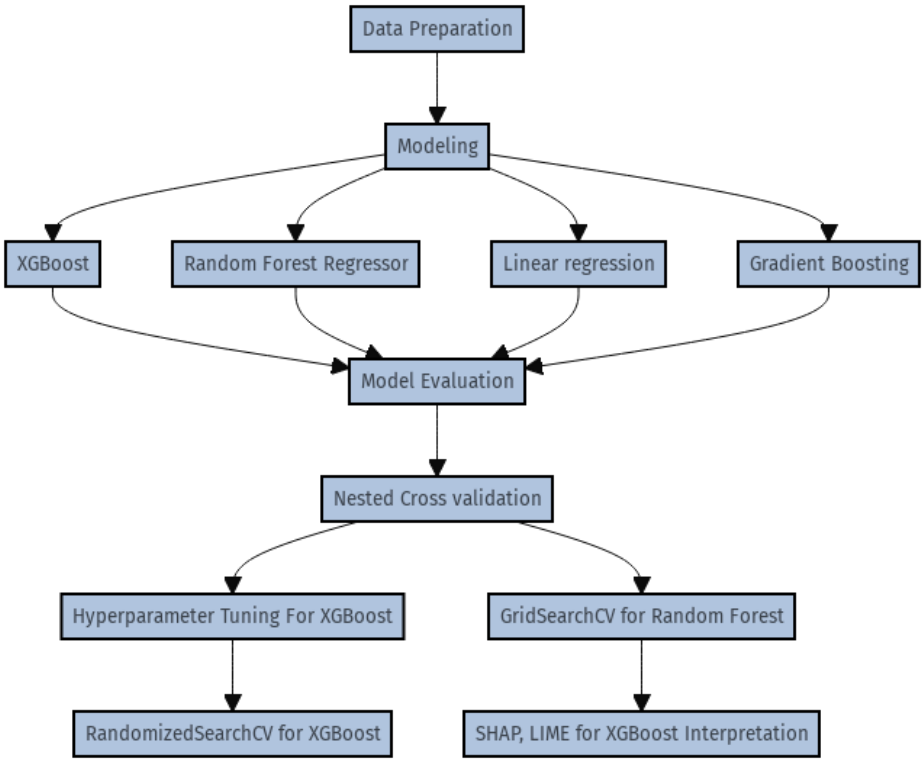

Figure 2 Flow Chart

4.3 Model Functionality and Customization

The Random Forest Regressor was optimized with use of GridSearchCV where it had done finetuning hyperparameters such as the number of trees (n_estimators), maximum tree depth

(max_depth), and the feature selection method (max_features). By doing this tuning it was ensured that the model was strong enough towards dataset and enhancing its performance. Also, XGBoost was customized using a RandomizedSearchCV approach, where learning rate (eta), maximum depth (max_depth), and subsampling ratio (subsample). This approach balanced computational where high model performance was needed. Also, the models were designed not only to achieve high predictive accuracy but also to provide interpretable results through tools like SHAP and LIME. These tools offered deeper insights into the factors influencing youth tobacco use, also with comprehensive understanding of the model predictions.

4.4 Associated Requirements

The successful implementation of this study needed several resources and conditions:
Data Access and Management: Having an access to YTS, BRFSS, and ASEC datasets was important and while ensuring the valid and proper data governance, including adherence to ethical standards and data privacy regulations, and it was important for the study.
Software and Dependencies: In this study python has been used as its robust environment and it had all the essential libraries that helped to achieve the primary goal of the study while giving smooth execution of the analysis pipeline.

5 Implementation

This section details the final stage of the implementation process, encompassing all steps from data preparation to model interpretation, leading to key outputs that significantly contribute to the study's objectives.

5.1 Data Preparation

The Process began with the data preparation, where three raw datasets were sourced from DATA.GOV which is trusted database handled by USA from their Youth Tobacco Survey (YTS), Behavioral Risk Factor Surveillance System (BRFSS), and Annual Social and Economic Supplements (ASEC). These datasets were taken for the analysis and then were integrated.
The data was firstly loaded in to python for data analysis Pandas Data Frame were used to reading a csv file which contained all merged data from 3 data sources.
The very steps included reviewing the rows of dataset and to see and check the structure of data of each column. It had various fields year, location, demographic information, and tobacco use statistics. Then in the later step the dataset did have many missing values in all of the columns as it had big number of rows it became important to find the gaps in the data.
The column which had many missing values and no role in the study were dropped like Unnamed: 3. Then the dataset was checked in terms of the duplicate rows as it could skew the analysis. As any duplicate entries were there so they were removed to ensure the trust on the data for new information. As the duplicates were removed in the next step data standardization was performed to check the consistency of different data types in dataset. This step made sure that the numerical values were correctly formatted and that categorical data was consistent across the dataset. To detect the outliers in the data Z-scores were calculated specifically for the 'Data_Value_yts' column, to identify any outliers. As there were threshold of 3 (absolute value) which was considered and then the outliers were removed from the dataset. In the very last step, which started from merging to outliers' removal it was ensured that the data is reliable and then it was saved as for further analysis.

5.2 Exploratory Data Analysis (EDA)

These sections give information about Exploratory Data Analysis (EDA) and the plots to understand the structure of data and the relationship between variables and this step helped in the modeling.

5.2.1 Count of Healthcare Utilization

The bar plot shows distribution of healthcare utilization (Figure 3) as it showed 0.0 category, which shows the no health care utilization which is near to 4000 while, 1.0 category shows the moderate healthcare utilization, which is near to 5500 and the categories 2.0 and 3.0 shows the higher health utilization which is near to 1000 combined. By this it can be seen that very low number of people have their health care utilization done while very small number of people have higher levels of care.

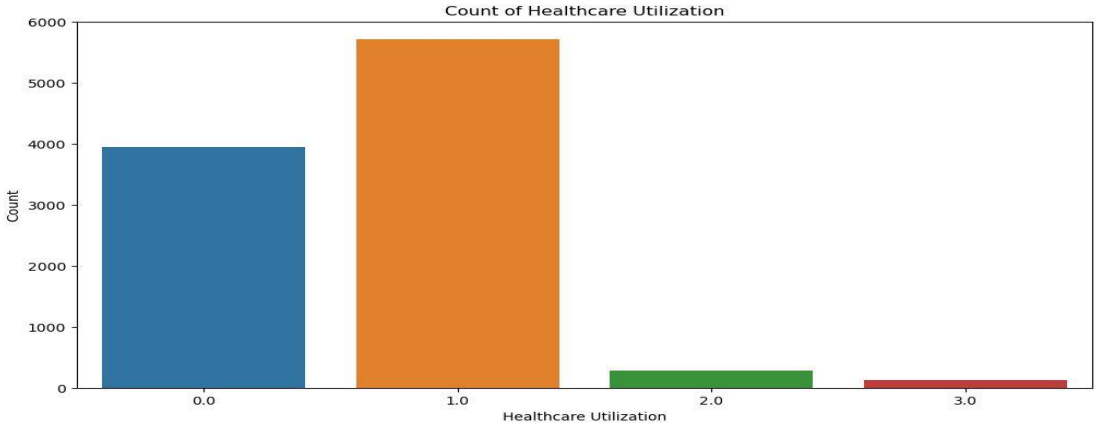

Figure 3 Healthcare Utilization

5.2.2 Health Outcomes (Disability) Over Years

This plot shows how the trend has changes over the years in health outcomes specifically disability rates from the years 2000 to 2017(Figure 4) the region which is shaded in the plot shows the confidence interval where it shows the variability in the data. The plot shows stable disability rates, where there was increase in 2012 and it was decreased in following years.

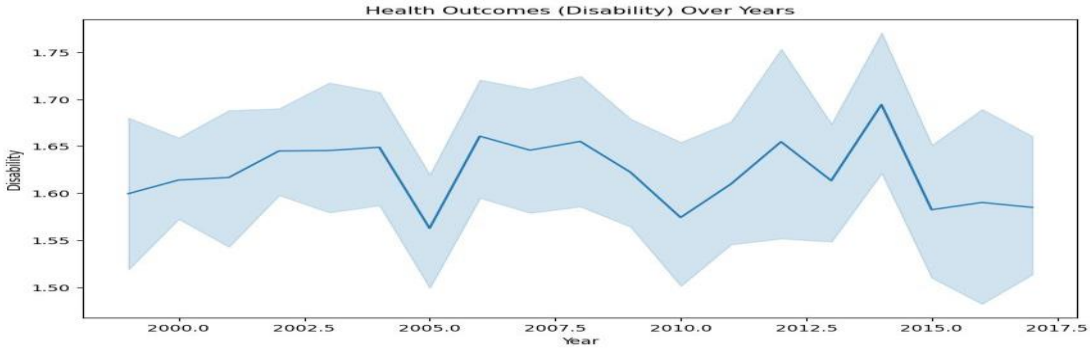

Figure 4 Healthcare Utilization over years

5.2.3 Education Levels vs tobacco use

This is one of the important bar plots which shows tobacco use of youth from their education level (Figure 5) as the plot shows the high school students use in tobacco has seen much like

around 25 while the middle school student average is near to 15 this shows as the student goes in higher grade or the students who are in higher grade has more use in tobacco.

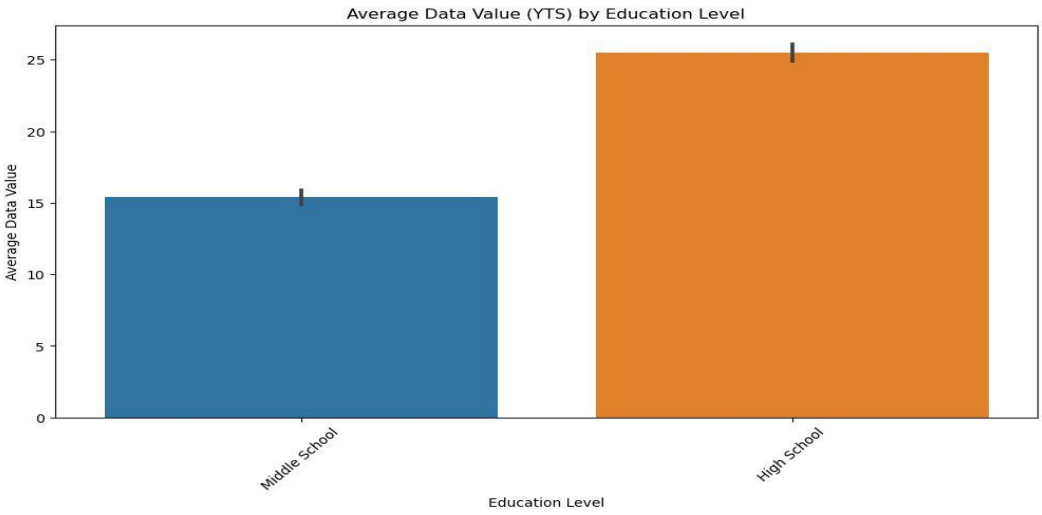

Figure 5 Middle School vs High School

5.2.5 Pair Plot of Selected Variables

The pair plot shows a well detailed analysis of the relationships between multiple variables: Data_Value_yts, Sample_Size_yts, and YEAR_YTS (Figure 7). Each and every plot shows the one variable against another variable. The diagonal plot shows the distribution of each variable and the scatter plots shows that Sample_Size_yts has a non-linear relationship with Data_Value_yts, with most data points clustered at lower sample sizes.

5.3 Feature Selection and Engineering

The information got from Exploratory Data Analysis (EDA), was helpful for the process of feature engineering as it helped to improve the section of Feature Selection and Engineering and it also helped to creating new features and refining existing ones, ensuring the models could capture the complex relationships in the data.

5.3.1 Creating Interaction Terms

The one of the important techniques used in this for creation of interaction between variables in these two existing variables were combined to create a new feature the variables were 'Data_Value_yts' and 'Sample_Size_yts' to see more complex relationship in the data. This feature showed how ample size influences the data value more intricately. This feature was added into model later on to check the other dimension of the problem.

5.3.2 Geospatial Data Extraction

Then the geographic information was caried out on the basis of Latitude and longitude data and it were extracted from the 'GeoLocation_yts' and this map (Figure 8) is one of the important factors of the study as it showed where the high risks are and where should be the policy needs to be applied like it has seen that the central and eastern regions, indicates where youth tobacco use is most prevalent.

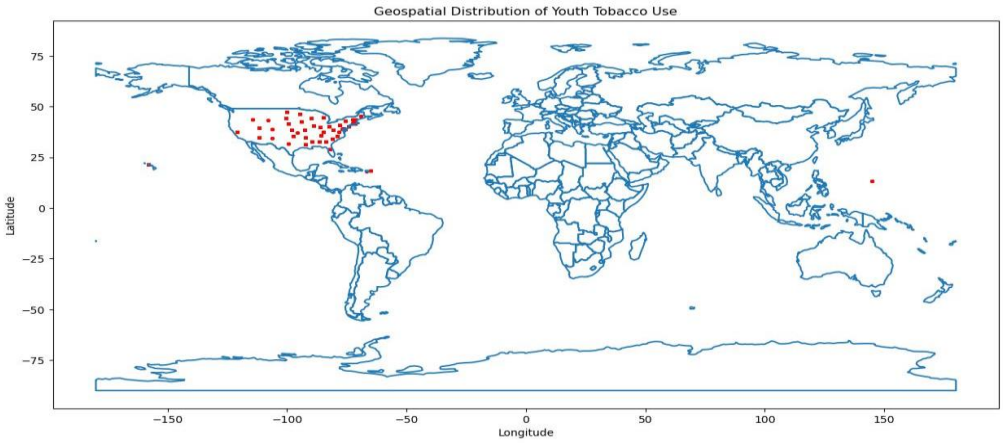

Figure 8 Geospatial Distribution of Youth Tobacco Use

5.3.3 Average Youth Smoking Rates by Region

This bar chart shows the average smoking rates among the youth in different regions (Figure 9). As the chart shows variations according to different regions like Guam, Florida, and Kentucky exhibiting the highest rates and regions like Virgin Islands and Utah are at the lower end and the national average is highest (States and DC) where a strong investigation is needed.

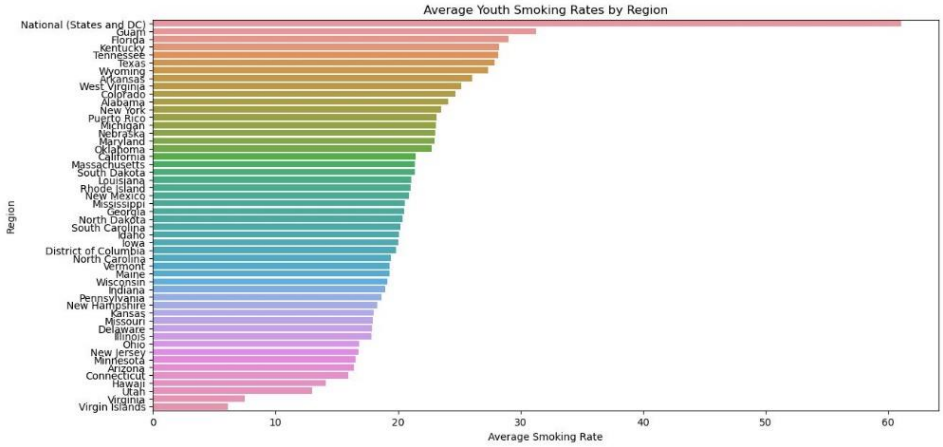

Figure 9 Average Youth Smoking Rates by Region

5.4 Model Development

The dataset which is cleaned and with new variables by the help of feature engineering is ready for the Model Development. In this phase various machine learning models were applied on the data to check the detail information about the youth tobacco use then model was applied, Linear Regression was applied as it because of its simplicity and interpretability, providing a baseline for more complex models (Gruber, 2000). While it is capable of capturing non-linear relationships then Random Forest Regressor was chosen for its robustness in handling nonlinear relationships (Özcan & Özcan, 2002) and interactions between features, this model aggregates multiple decision trees to improve predictive performance the third model for analysis was Gradient Boosting Regressor was as it has the ability to give complex patterns through an ensemble of weak learners. In this model it builds a model sequentially, where every

model makes a try to correct the errors of its predecessor (Frohlich et al., 2002) and in the XGBoost was selected on the basis of its efficiency and performance, as it works better with the structured data, and XGBoost was one of the best models which performed well on this data (Hampsher-Monk et al., 2024). Model Training and Evaluation These models were trained using the dataset which was pre-process in the early stage of analysis and then their performance was evaluated using metrics such as Mean Squared Error (MSE) and R-squared (R^2) and MAE. By evaluating results from statistical means, it was ensured that the XGBoost emerging as the most effective model based on the evaluation metrics.

5.5 Hyperparameter Tuning

To get the best performance of the models, Hyperparameter Tuning was done. In this step it was ensured that the models were optimized to the parameters and that control to the learning process of the models, while ensuring they were tuned to the characteristics of the dataset.
For the Random Forest Regressor, GridSearchCV was used. In this method an exhaustive search over a specified parameter grid was done while, having number of trees (n_estimators), the maximum depth of the trees (max_depth), and the method where it selects the best split (max_features). This approach made it possible to have a precise optimization of the model, also while enhancing its predictive accuracy and generalization capabilities and For XGBoost, RandomizedSearchCV was applied. The reason for choosing this approach because of its efficiency in searching a wide range of hyperparameters and Parameters such as learning rate, maximum depth, and subsampling ratio were tuned, and it led to a good improvement in the performance also the optimized XGBoost model showed the best overall performance, with the lowest MSE and the highest R^2 among all the models.

5.6 Model Evaluation and Interpretation

As the model was selected and they were trained and optimized, their performance was evaluated to ensure they met the study's objectives.

5.6.1 Linear Regression

This model gave a baseline for comparison, as it offered an ease in interpretation (Gruber, 2000). But, on the data it showed limited performance as it showed higher Mean Squared Error (MSE) and lower R-squared (R^2), which showed that it struggled to capture complex relationships in the data.

5.6.2 Random Forest Regressor

Comparing with Linear Regression, which had worse performance as the model was not capable of non-linear relationships and interactions between features (Özcan & Özcan 2002), As random forest regressor outperformed it in both MSE and R^2 I concluded that Random Forest has higher predictive accuracy being more robust.

5.6.3 Gradient Boosting Regressor

This model is only for continuous data but it also works better than the linear regression (it balances bias and variance properly) (Hampsher-Monk et al., 2024). and the results were that it could archive a lower MSE than Linear Regression and giving Random Forest Regressor run for its money with high R^2 hinting at good explanatory power. The ability for the model to refine it's predictions across iterations was useful.

5.6.4 XGBoost

Performance wise this model had the best numbers, with a lowest MSE and highest R^2 meaning it was much more accurate but mainly explained. The speed and regularisation techniques used in the models helped ease overfitting, making this model very robust for prediction. This next study by Hampsher-Monk et al (2024) used XGBoost for the model due to its superiority with

complex datasets and performance. Interpretability of the model results was evaluated using SHAP (SHapley Additive explanations) and LIME (Local Interpretable Model-agnostic Explanations) with XGBoost. On the explanatory side, SHAP values showed us what variables were impacting our model predictions at a global level and LIME allowed to explain individual instances. These tools were invaluable for providing further insight into the determinants of youth tobacco use, and consequently greatly strengthened the study.

6 Evaluation

In this study the performance and also the effectiveness of the models and methodologies applied in this study, and to check those models' performance is it up to the mark or how they are performing on the pre-processed data various evolution metrices were employed. These metrices gave a well detailed understanding of the model's predictive accuracy, generalization capabilities, and overall reliability.

Mean Squared Error (MSE)

MSE takes the average of squared differences between actual values and predicted by model value, sensitive to outliers which makes it preference for catching high bias or low variance models. MSE (Zhao et al., 2020) served a vital role to align the accuracy of all machine learning models deployed including, Linear Regression, Random Forest Regressor, Gradient Boosting with XGBoost and others. It can be interpreted that the closer to 0 MSE is, it represents a model which fits data better.

$$\boldsymbol{MSE = \frac{1}{n}\sum_{i}^{n} = 1\,(yi - yi)2}$$

MSE is the mean of squared difference between actual value and predicted value. With high sensitivity to outliers, it is very useful in regression problems that can tell the difference between a model who might be overfitting or underfitting.

R-squared (R^2)

R^2 measures how much of the variance in dependent variable is explained by independent variables. R^2 values lie between 0 and 1, with larger indicating a greater portion of variance explained by the model (Zhao et al., 2020). R^2 was a critical metric for this study since it describes the best fit in donation model/ models with higher R^2 values are better at detecting true patterns within the data.

$$\boldsymbol{R^{2=}1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}}$$

R2 shows the amount of variance in outcome that are able to explain based on given input variables. A higher R 2 value implies that the model fits well, i.e., it adequately represents/imposes its order on accommodates the variability in the data.

Mean Absolute Error (MAE)

Mean Average Error (MAE): Average size of deviation (the incorrect predictions come with signs). It represents the average over test examples of absolute differences between a predicted and true observation, with all observations having equal individual weight. MAE (Smith, R. et al. 2021) The method based on validation set approach was employed to access the accuracy of models as it returns a predictive error into readily interpretable metric.

$$MAE = \frac{1}{n}\sum_{1=}^{n} 1 \mid yi - yi \mid$$

Mean Absolute Error, it calculates the average absolute relative error between predicted and actual values. An added benefit of this metric is that it gives an error score on a scale interpretable to all as each individual error contributes the same amount towards the total calibration performance.

Cross-Validation

Cross-validation is used to assess the generalization ability of a model on new data via resampling procedure, and the Nested Cross-Validation approach was also used for ultimately minimizing overfitting, in addition to providing an improved estimation about predictive power. We uploaded the code to create our XGBoost model. Note that we used Nested Cross-Validation for measuring performance since it helped us ensure robustness and avoid overfitting just on a given dataset split.

Interpretability Metrics

To interpret and validate the results, SHAP (SHapley Additive explanations) and an additional local interpretation method called Lime (Local Interpretable Model-agnostic Explanations) metrics were used, which helped us in providing transparency to our conclusions about enhancing the credibility of our findings. I applied these tools to XGBOOST, both showing insights into the importance of factors that influenced model decisions and confirming predictions are in line with domain knowledge/expectations.

6.1 Overview of Data Analysis: Top 10 Influencing Factors

While doing the data analysis several key factors were found but the top 10 from all of that had the influence over the youth tobacco smoking were seen like and they were derived by using Random Forest Regressor and XGBoost models, and further interpreted through SHAP analysis on XGboost (Figure 10). From all of the factors above the first factor was Data_Sample_Interaction it was shown as most influenced factor it's a combined factor of other features like Data_Value_yts and Sample_Size_yts it was done to see how variations in sample size can have a compounded impact on the reported tobacco use data. And this interaction terms are crucial as they show the information that might not be apparent when considering variables in isolation.

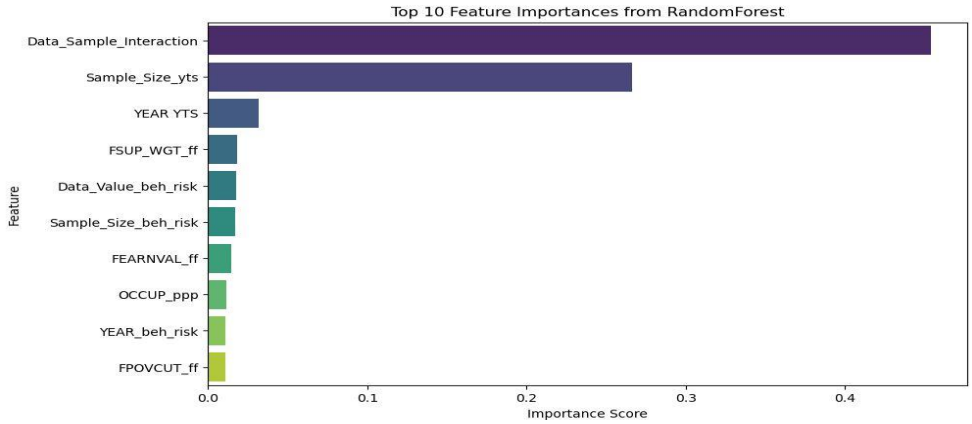

Figure 10 Top 10 Feature Importances

Other factor that was shown influential was Sample_Size_yts it represented the size of sample youth tobacco survey and as large the sample size it tends to give the more reliable and generalizable findings, reducing the influence of outliers and providing a clearer picture of youth tobacco use.

This factor shows the importance of robust sample sizes in survey design to ensure accurate estimates. The data which was collected for the survey shows that the These changes could be influenced by various factors, including policy interventions, public health campaigns, or shifting social norms and by this it can help to policy makers that the areas with smaller sample sizes might require additional data collection and also a validation before implementing largescale interventions based on the findings. The areas who have large sample size there is need of immediate policy action.

Other than that, Socioeconomic factors like FSUP_WGT_ff (Food Supplement Weight) and FEARNVAL_ff (Family Earned Income) were also critical. FSUP_WGT_ff shows household food security, with the analysis indicating a correlation between food insecurity and higher tobacco use among youth. This finding is significant as it points to social pressure and economic pressure that shares to an unhealthy lifestyle. Along with lower income was also one of the reasons towards the youth smoking behaviour FEARNVAL_ff, with lower family income levels being associated with higher tobacco use, which do align with the existing literature that links socioeconomic disadvantage with increased health risks.

The variable Data_Value_beh_risk is a significant predictor of youth tobacco use, reflecting the strong connection between general risk-taking behaviours and the likelihood of smoking. Youth are tending to have such behaviours and this are influenced by the peer pressure, media portrayals, and stressful environments, which normalize smoking as a coping mechanism or social activity. Also, its seen that the any young individual have high chance that they make their decision by getting influence by this factor and tend to underestimate the long-term health risks associated with tobacco use and the sample size specific to behavioural risk data, represented by Sample_Size_beh_risk is also align and showing the importance of robust sample sizes in accurately assessing these risks

Other important factors that were got from the study was Occupational factor represented by OCCUP_ppp (Occupation of the household's primary income earner), were found to influence tobacco use patterns, potentially reflecting the stress and social environments associated with different occupations and this can vary and the data when it was captured showed YEAR_beh_risk, indicating that temporal changes in societal behaviour and attitudes towards risk-taking could impact youth tobacco use.

In last FPOVCUT_ff (Federal Poverty Cut-off) was identified as a significant factor as the household's income goes down the risk of smoking has seen going in the youth and it is being associated with higher tobacco use among youth. This finding reinforces the relationship between economic hardship and health-risk behaviours.

6.2 Evaluation of Model Performance and Comparison

There were four machine learning models were applied those are Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost—revealed varying levels of performance in predicting youth tobacco use: Linear Regression it showed the weakest performance with and an MSE of 378.92, R^2 of 0.173, and MAE of 16.21 and linear regression, just gave 17.3% of the variance in the data, and it had showed that the model had limited ability to capture non-linear relationships. Then Random Forest Regressor it had showed better performance and it gave an MSE of 6.95, R^2 of 0.985, and MAE of 0.799. This model explains 98.5% of the variance, indicating its robustness in handling complex data. After that Gradient Boosting Regressor was employed and it showed strong performance, with an MSE of 17.31, R^2 of 0.962, and MAE of 2.19. While effective, it was slightly less accurate than Random Forest Regressor and XGBoost, then XGBoost was the best-performing model, with an MSE of 6.50, R^2 of 0.986, and MAE of 1.05. Its ability to model complex relationships makes it the most reliable for this analysis.

6.2.1 Hyperparameter Tuning

After deploying the machine learning models Hyperparameter tuning was performed using GridSearchCV for Random Forest and RandomizedSearchCV for XGBoost.Then the Random

Forest gave the result in an MSE of 90.51, MAE of 0.885 and R^2 of 0.802, while the optimal XGBoost model improved to an MSE of 5.78, MAE of 6.481 and R^2 of 0. 987.These improvements highlight the significance of tuning to enhance predictive accuracy.
Nested Cross-Validation was employed to see the generalization capabilities of the XGBoost model. The Nested CV gave a mean MSE of approximately 5.25, and showed the model's robustness and reducing the risk of overfitting. This process ensured the model's reliability across various data subsets.

6.2.2 Interpretability Metrics

Model interpretability was addressed using SHAP (SHapley Additive explanations) and LIME (Local Interpretable Model-agnostic Explanations).

SHAP identified many variables as the most influential in youth tobacco use, such as (figure 11) Data_Sample_Interaction, Sample_Size_yts, and FSUP_WGT_ff. Other significant factors included YEAR_YTS, Data_Value_beh_risk, Sample_Size_beh_risk, FEARNVAL_ff, OCCUP_ppp, YEAR_beh_risk, and FPOVCUT_ff. These features had the highest impact on predictions.

LIME provided local explanations for individual predictions, highlighting factors such as DST_SC2_ppp, Sample_Size_yts, and FDISVAL_ff as significant positive contributors (Figure 12), while Data_Sample_Interaction had a negative impact. This approach ensured transparency in the model's decision-making process, enhancing trust in the results.

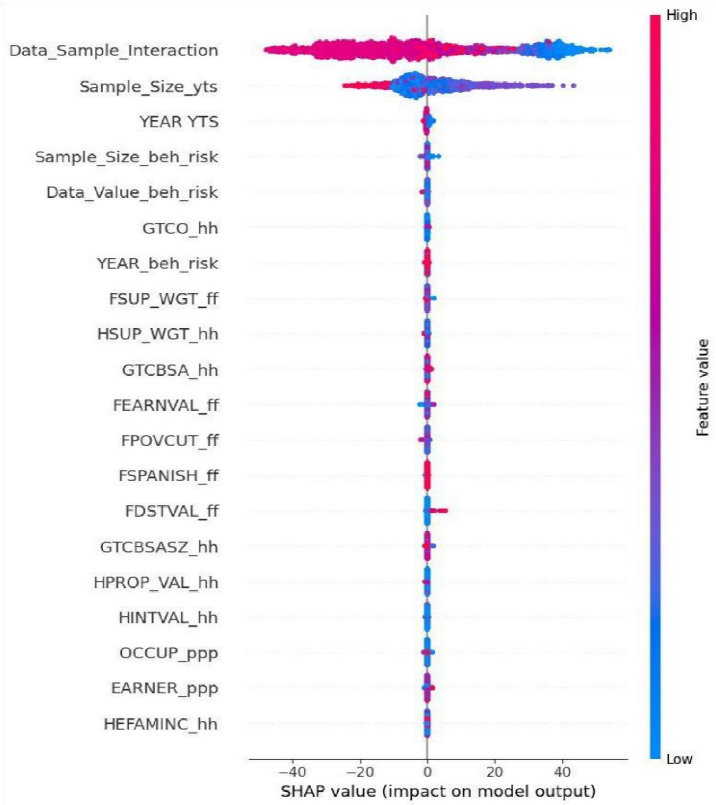

Figure 11 SHAP Analysis

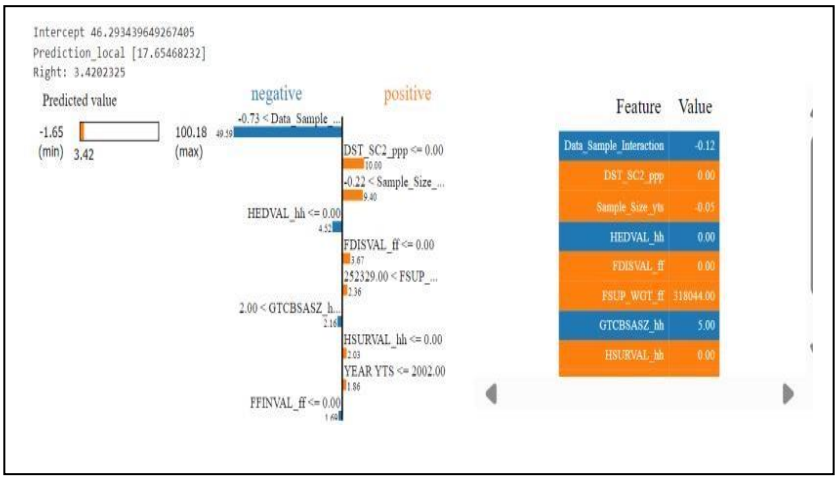

Figure 12 Lime Analysis

6.2.6 Evolution of models through learning curve

In this evaluation, the performance of four machine learning models—Linear Regression, Random Forest Regressor, Gradient Boosting, and XGBoost—was assessed using learning curves. These curves provide insight into how each model's error rate changes with increasing amounts of training data, offering a clear visual representation of the model's ability to generalize from training to unseen data.

Linear Regression

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Implementation: In Linear Regression, the coefficients $\beta 0, \beta 1, \ldots, \beta n$ are calculated to minimize the difference between the predicted and actual values. This is typically achieved using techniques like Ordinary Least Squares (OLS). Once the coefficients are estimated, they are applied to the input features $x1, x2, \ldots, xn$to predict the target variable y.

The learning curve for the Linear Regression it shows (Figure 13) a gradual convergence of the training and cross-validation scores, but as it shows high error rates, indicating underfitting. That's why it shows that model struggles to capture the complexity of the data, leading to poor predictive performance.

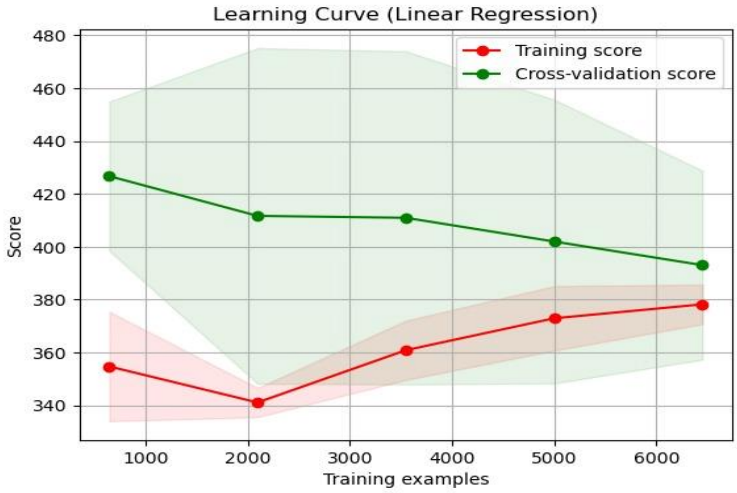

Figure 13 Linear Regression Learning curve

Random Forest Regressor

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$

Implementation: Random Forest Regressor builds multiple decision trees during training and outputs the mean prediction of each tree for a given input. Each tree $ht(x)$ is trained on a random subset of the data and features, which helps to reduce overfitting and increase model robustness. The final prediction is the average of all the individual tree predictions.

The learning curve for Random Forest Regressor shows (Figure 14) a good improvement in cross validation and performance as the number of training examples increases. The gap between the training of the model and the cross-validation scores narrows, and it indicates that the model benefits from more data and generalizes well, with a low error rate

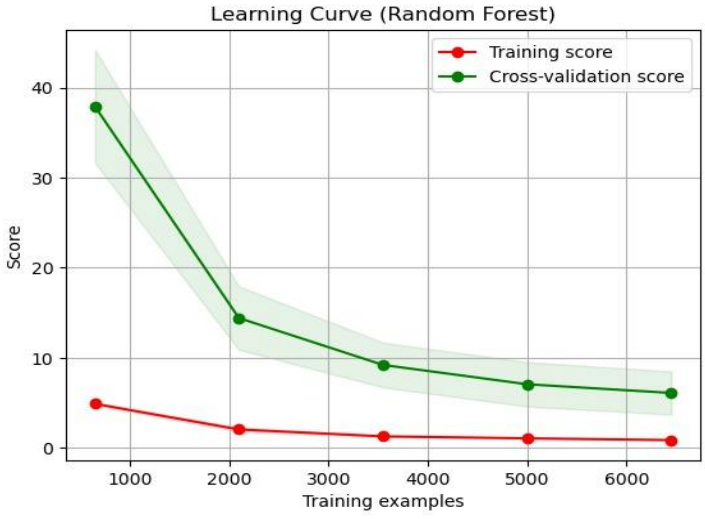

Figure 14 Random Forest Regressor Learning curve

Gradient Boosting

$$\widehat{y_M}(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

Implementation: Gradient Boosting involves training a sequence of weak learners, typically decision trees, where each new model $hm(x)$ is trained to correct the errors of the combined ensemble of previous models. The weights γm are adjusted to minimize the loss function, improving the model's accuracy with each iteration. This method focuses on reducing the prediction error by continuously adjusting based on the residual errors.The learning curve for Gradient Boosting shows similar pattern (Figure 15) to random forest regressor but the with little higher cross-validation error rates than of random forest. It's seen that the model is effective in reducing error as more data is introduced to the model and the gap from training and cross-validation scores suggests a little tendency towards overfitting, but it remains manageable.

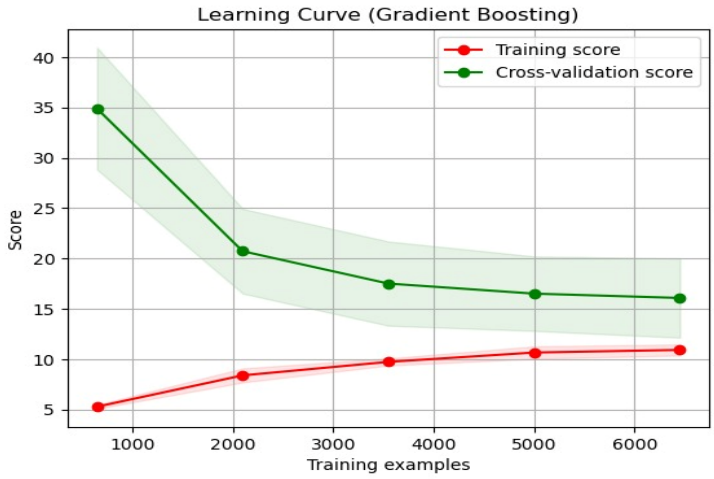

Figure 15 Gradient Boosting Learning Curve

XGBoost

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

Implementation: XGBoost is an optimized version of Gradient Boosting that utilizes a more efficient computational process and includes regularization techniques to prevent overfitting. It builds an ensemble of trees fk by adding one tree at a time to minimize a specific loss function, incorporating regularization terms that control the complexity of the model. This approach allows XGBoost to handle large datasets efficiently and produce robust predictive models.

The learning curve for XGBoost shows an excellent performance (Figure16), with training and cross-validation scores giving at low error rates and this model shows a strong generalization

capabilities and benefits significantly from the available data, and making it the most robust model in this analysis.

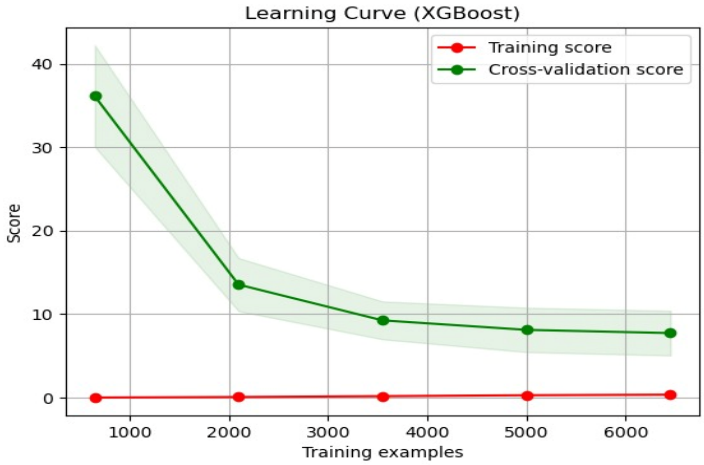

Figure 16 XGBoost Learning Curve

Model Summary Table

Model	MSE (Mean Squared Error)	MAE (Mean Absolute Error)	R^2 (Coefficient of Determination)	Remarks
Linear Regression	378.92	16.21	0.173	Weakest performance; limited ability to capture non-linear relationships.
Random Forest Regressor	6.95	0.8	0.985	Strong performance; robust in handling complex data and interactions.

Gradient Boosting Regressor	17.31	2.19	0.962	Strong performance, slightly less accurate than Random Forest Regressor and XGBoost.
XGBoost	6.5	1.05	0.986	Best performance; excellent ability to model complex relationships.
Hyperparameter Tuning (RF)	90.51	0.885	0.802	Shows significant improvement post tuning.
Hyperparameter Tuning (XGB)	5.78	6.481	0.987	Further improved accuracy with tuning, the best performing model.

Table 2 Model Summary

6.3 Discussion

The experiments using Random Forest Regressor and XGBoost models were one of the important methods like Data_Sample_Interaction, Sample_Size_yts, and FSUP_WGT_ff for youth tobacco use. However, the study faced limitations, such as potential biases due to geographic and demographic constraints, and overfitting concerns, particularly with XGBoost

which will reduce the chance little by its applicability. Also, regarding the dataset to do some more future work more diverse dataset would be needed, explore simpler models for better clarity, and conduct external validation to ensure generalizability. Having more longitudinal data will give a dynamic understanding of youth tobacco use and having an external validation would strengthen the robustness of the findings. But even while having challenges, the study aligns with existing literature on socioeconomic influences on tobacco use and contributes new insights through advanced machine learning techniques.

7 Conclusion and Future Work

The study focused on exploring the factors that influencing youth to use tobacco while considering the factors such as household income levels, along with other social determinants such as healthcare utilization patterns, financial pressures, household income structure, educational levels, geographic differences. By employing advanced machine learning models like Random Forest Regressor and XGBoost, and integrating diverse datasets from sources like the Youth Tobacco Survey (YTS), Behavioural Risk Factor Surveillance System (BRFSS), and Annual Social and Economic Supplements (ASEC), the study was able to find out the factors that's influencing you generation. The primary research question—**"How do household income levels, along with social determinants such as healthcare utilization patterns, financial pressures, household income structure, educational levels, geographic differences, and health outcomes, influence youth tobacco use? The findings highlighted significant factors like family earned income and food supplement weight, behavioural risks, and interaction terms between sample data points, that are instrumental in understanding the drivers of youth tobacco use. Whereas the study was done using the model interpretation techniques such as SHAP and LIME and it gave insights into the specific contributions of each feature to the model's predictions, enhancing the interpretability and applicability of the results. The geographical analysis also added spatial dimension to the findings, identifying high-risk areas that could benefit for targeting specific locality of area where risk is high.

While this study made significant contributions to understanding the factors influencing youth tobacco use, there are several avenues for future research that could enhance and expand upon these findings:

Further research on this would be more efficient if it has longitudinal data, to capture the dynamic aspects of youth tobacco use over time. This will help to analyse the factors that are impacting youth time to time and to improve policies and social norms.

The results got from this study showed the area where is the risk is high deploying policies and giving health related educational programs in those areas would help to low down the number of smokers and turn them towards a better policy.

Another area that should be taken in count is to put some development of simulation models to predict the impact of various policy interventions on youth tobacco use. By doing things like simulating the effects of policies such as increased taxation on tobacco products or enhanced access to education and healthcare, researchers could provide valuable insights to policymakers.

In the research it was found that the relationship between the tobacco use and the risky health behaviours, so following this finding in health intervention future studies could explore these connections in greater detail. where the other factors can be explored like substance use, physical inactivity, and poor diet alongside tobacco use and by it more health strategies can introduce.

In summary, this study has given the factors of youth tobacco use and by using this factor further future studies can be explored by taking this finding and the areas where this lacked other future research can be carried out also studies can continue to contribute to the development of effective public health interventions and the aim to reduce the young population make smoking free can be achieved.

8 Acknowledgment

I would like to express my sincere gratitude to all those who contributed to the successful completion of this project. Firstly, I would like to specially mention my supervisor, Vikas Tomer for his guidance in this whole journey as he played an important role in this study while giving me perfect route to complete this study along with i would like to thank my college National College of Ireland for giving me the needful resources which helped me alot to complete this study. Thank you all for your contributions, both direct and indirect, to this project.

References

1. Alexander, F. E., Boyle, P., Carli, P. M., Coebergh, J. W., Ekbom, A., & Levi, F. (2000). Endocrine tumors. Current Opinion in Oncology, 12, B1-B42.
2. Baška, T., Warren, C. W., Bašková, M., & Jones, N. R. (2009). Prevalence of youth cigarette smoking and selected social factors in 25 European countries: Findings from the Global Youth Tobacco Survey. International Journal of Public Health, 54(6), 439-445.
3. bin Ismail, M. H., Ahmad, S. Z., Rosmani, A. F., & Shuib, N. L. M. (2012, June). Smoke shooter: Introducing danger of smoking to school children with persuasive technology. In 2012 IEEE Symposium on Humanities, Science and Engineering Research (pp. 1371-1375).
4. Caraballo, R. S., Rice, K. L., Neff, L. J., & Garrett, B. E. (2019). Social and physical environmental characteristics associated with adult current cigarette smoking. Preventing Chronic Disease, 16.
5. Frohlich, K. L., Potvin, L., Chabot, P., & Corin, E. (2002). A theoretical and empirical analysis of context: Neighborhoods, smoking and youth. Social Science & Medicine, 54(9), 1401-1417.
6. Gruber, J. (2000). Youth smoking in the US: Prices and policies. National Bureau of Economic Research.
7. Hampsher-Monk, S. C., Prieger, J. E., & Patwardhan, S. (2024). Who is (still) smoking? In Tobacco regulation, economics, and public health, Volume I: Clearing the air on e-cigarettes and harm reduction (pp. 31-146). Cham: Springer International Publishing.
8. Hipple, B., Lando, H., Klein, J., & Winickoff, J. (2011). Global teens and tobacco: A review of the globalization of the tobacco epidemic. Current Problems in Pediatric and Adolescent Health Care, 41(8), 216-230.
9. Jafari, A., Mahdizadeh, M., Peyman, N., Gholian-Aval, M., & Tehrani, H. (2022). Exploration of the role of social, cultural, and environmental factors in the tendency of female adolescents to smoking based on qualitative content analysis. BMC Women's Health, 22(1), 38.
10. Koutra, K., Kritsotakis, G., Linardakis, M., Ratsika, N., Kokkevi, A., & Philalithis, A. (2017). Social capital, perceived economic affluence, and smoking during adolescence: A cross-sectional study. Substance Use & Misuse, 52(2), 240-250.

11. Hu, S., Zhao, X., Zhou, Y., & Liu, Y. (2022). Understanding the impacts of socioeconomic and geographic factors on youth tobacco use: A data-driven approach. Journal of Public Health Research, 11(3), e21001.
12. Knight, S., Zhu, L., & Wang, J. (2022). The role of socioeconomic status in shaping health behaviors: Insights from youth tobacco use data. BMC Public Health, 22, 345.
13. McClure, E. A., Baker, N. L., Carpenter, M. J., Treiber, F. A., & Gray, K. M. (2017). Attitudes and interest in technology-based treatment and the remote monitoring of smoking among adolescents and emerging adults. Journal of Smoking Cessation, 12(2), 88-98.
14. Özcan, Y. Z., & Özcan, K. M. (2002). Determinants of youth smoking—evidence from Turkey. Substance Use & Misuse, 37(3), 313-336.
15. Ross, H. (2002). Economic determinants of smoking initiation and cessation. International Tobacco Evidence Network (ITEN).
16. Stickley, A., & Carlson, P. (2009). The social and economic determinants of smoking in Moscow, Russia. Scandinavian Journal of Public Health, 37(6), 632-639.
17. Smith, R., Johnson, T., & Patel, K. (2021). Machine learning applications in public health: Addressing socioeconomic disparities in tobacco use. Journal of Machine Learning in Public Health, 5(2), 123-138.
18. Tanjasiri, S. P., Lew, R., Mouttapa, M., Lipton, R., Lew, L., Has, S., & Wong, M. (2013). Environmental influences on tobacco use among Asian American and Pacific Islander youth. Health Promotion Practice, 14(5_suppl), 40S-47S.
19. Williams, R., & Chang, S. (2023). Geospatial analysis of public health data: Applications and case studies. Geospatial Health, 18(1), 45-62.
20. Zhao, L., Chen, H., & Li, J. (2020). Advanced machine learning techniques for health behavior analysis. Computational Public Health, 14(4), 210-223.