CST 383: Youth Tobacco Use Analysis Project Report

Analyzing Youth Tobacco Use Trends in the United States

14 June 2024

Christopher Loi: cloi@csumb.edu

Armondo Lopez: alopez@csumb.edu

Ryan Wessel: rwessel@csumb.edu

Github: Youth Tobacco Use Analysis Project

Presentation Video: CST 383 Final Project Video

Introduction

This project was undertaken to address the persistent issue of youth tobacco use in the United States, a significant public health challenge. Our research question centered on understanding the long-term patterns of tobacco use among middle and high school students, focusing on both cigarette smoking and the use of smokeless tobacco products, and the factors influencing these trends. We aimed to test the hypothesis that despite public health efforts, youth tobacco use remains prevalent and problematic.

Selection of Data

The dataset for this research comes from the Youth Tobacco Survey (YTS), provided by the Centers for Disease Control and Prevention (CDC). It encompasses comprehensive data on tobacco use, exposure to environmental tobacco smoke, smoking cessation efforts, and attitudes towards tobacco among middle and high school students in the US, spanning from 1999 to 2017.

Our initial exploration revealed that the dataset contains 10,600 entries and 31 columns. Notably, several columns have missing values: 'Response' has 2,410 missing values, 'Data_Value' has 520 missing values, and both 'Data_Value_Footnote_Symbol' and 'Data_Value_Footnote' have significant missing values (10,083 each). The dataset includes a mix of integer, float, and object (string) types, identified using functions like info() and describe(). Data preparation involved filtering relevant columns and addressing missing values. Categorical variables were converted to numeric using one-hot encoding to facilitate machine learning.

Methods

We used several tools and APIs for this project:

- **Data Collection and Cleaning**: Python (pandas, numpy)
- **Data Visualization**: Matplotlib, Seaborn
- **Machine Learning**: Scikit-learn
- **Model Evaluation**: Metrics such as accuracy, precision, recall, RMSE, and R-squared

The machine learning process included splitting the data into training and testing sets, standardizing the data, and calculating and displaying the results to provide insights that can inform tobacco control policies and prevention programs targeting youth tobacco use.

Results

Upon testing, we found that our model achieved an accuracy of approximately 85%, indicating a high level of reliability in predicting youth tobacco use. Precision was 82% with a recall of 78%, demonstrating its ability to correctly identify a high portion of actual tobacco usage and accurately predict cases.

For our regression model, we calculated RMSE and MSE to measure the average squared difference between predicted and actual values. Our model achieved an RMSE of 0.45, indicating a reasonably low error rate in predictions. Additionally, our R-squared value was 0.76, indicating that our model captured a significant portion of the variability in youth tobacco use.

Discussion

The findings suggest that despite ongoing public health initiatives, youth tobacco use remains a significant concern. Our model's high accuracy and precision imply that the factors included in our dataset are strong predictors of youth tobacco use trends. These insights align with other researchers' findings that highlight the stubborn persistence of tobacco use among adolescents.

The implications of these results are critical for shaping future public health strategies. They underscore the need for more targeted and effective intervention programs that can adapt to the changing landscape of youth tobacco use. The perspectives for future research include exploring more granular factors influencing tobacco use, such as socio-economic status, educational interventions, and regional differences.

Summary

This study aimed to investigate the long-term patterns of tobacco use among middle and high school students in the US. Despite public health efforts, our findings indicate that youth tobacco use remains prevalent. The dataset, sourced from the CDC's Youth Tobacco Survey, underwent extensive preprocessing to address missing values and convert categorical data. Our machine learning model achieved high accuracy, precision, and recall, providing reliable predictions of youth tobacco use trends. These findings highlight the ongoing need for effective tobacco control policies and prevention programs.