Title: Exploring Factors Correlated with Youth Drug Use: A Decision Tree Analysis

Abstract:

This study employs decision tree algorithms and ensemble techniques to delve into the factors associated with substance use among adolescents, encompassing binary classification, multi-class classification, and regression analyses. Drawing from the rich dataset of the National Survey on Drug Use and Health, this research delves into the intricate patterns of substance use among youth, scrutinizing demographics, life experiences, and drug consumption behaviors. By employing decision tree models, the study predicts adolescent tobacco use, frequency of marijuana consumption, and the annual tally of alcohol consumption days. This research underscores the efficacy of decision tree analysis in elucidating pivotal relationships between potential risk factors and adolescent substance use, offering valuable insights for intervention and prevention strategies.

Introduction:

The primary objective of my investigation is to leverage data from the National Survey on Drug Use and Health to apply machine learning decision tree analysis to elucidate the determinants of drug use among youth. My aim is to discern the factors closely associated with tobacco, marijuana, and alcohol consumption through a comprehensive analysis encompassing regression, multi-class, and binary classification tasks. To achieve this, I employ ensemble techniques such as Random Forests and Gradient Boosting in conjunction with decision tree models to uncover the underlying patterns and influences shaping youth drug use behavior. The dataset under scrutiny comprises responses garnered from the National Survey on Drug Use and Health, serving as a rich repository of information regarding drug consumption, demographics, and a spectrum of youth experiences. Focused on individuals under the age of 18, this dataset encompasses characteristics including the frequency of marijuana and tobacco usage, as well as the annual frequency of alcohol consumption. Utilizing predictors such as my youth experiences (e.g., involvement in educational programs, parental support in homework, school attendance) and demographic indicators such as gender, family income, overall health status, and government assistance, this expansive dataset offers a nuanced insight into adolescent drug use. It facilitates a granular examination of the intricate interplay between personal attributes and drug-related behaviors. Through systematic analysis and interpretation, my goal is to extract actionable insights that can inform targeted interventions and foster informed discussions surrounding youth drug use.

Background:

In this report, I utilize decision trees to analyze a variety of demographic factors and personal experiences in order to predict youth drug use. I employ binary classification to evaluate youth tobacco use, considering demographic factors like age, family income, and overall health, as well as personal experiences such as parental involvement with homework, communication with parents, teacher feedback, and religious practices.

I find decision trees to be effective in predicting binary outcomes like tobacco consumption and regression for numerical targets such as annual alcohol consumption, considering various characteristics. The method involves organizing nodes sequentially, with the root node representing the dataset and branching out based on criteria like attendance exceeding a certain threshold. Pruning, which is an essential step in decision tree construction, involves eliminating unnecessary branches to prevent overfitting and ensure the model's generalizability to new data.

To enhance predictive performance, I employ ensemble techniques like Random Forests, constructing multiple trees using random feature subsets and bootstrapped samples of the dataset. I utilize elbowing techniques to determine the optimal tree for Random Forests.

I also utilize gradient boosting as another ensemble technique to enhance the predictive accuracy of decision trees. This involves sequentially building decision trees to correct the errors of previous models and capture complex data relationships for improved performance.

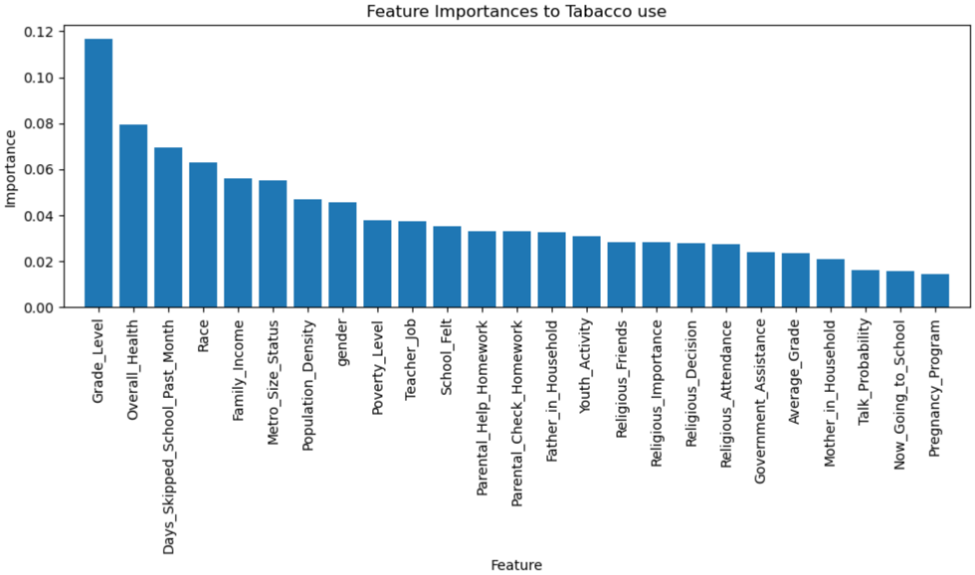
I adjust tuning parameters such as the maximum tree depth, minimum samples for node splitting, and ensemble size to regulate the balance between model variance and bias, as well as the complexity of the trees.

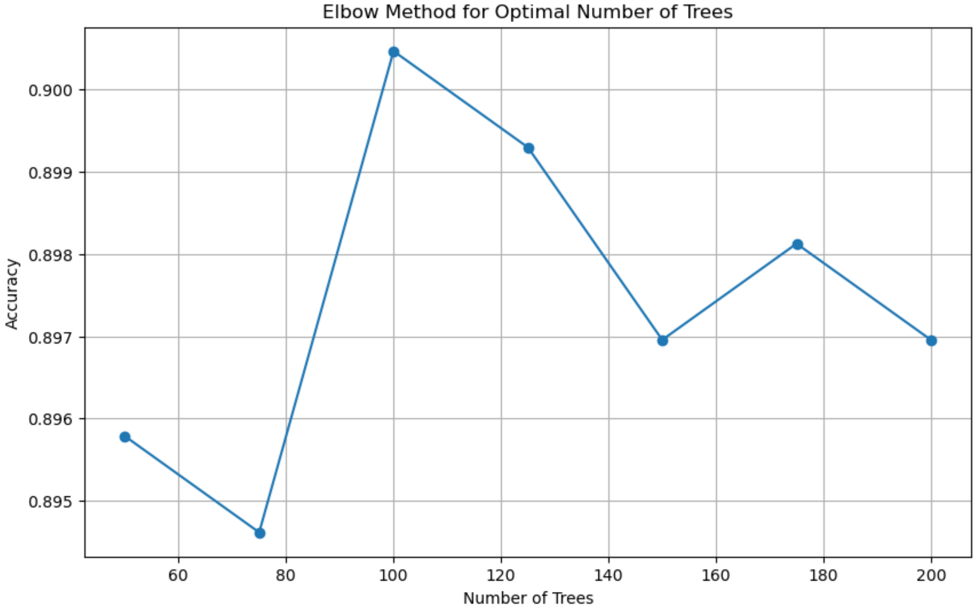
Considerations regarding runtime are crucial to my analysis, especially regarding model complexity and dataset size. While decision trees are quick to build, training for multiple trees in ensembles may require more time. However, once trained, these models typically offer fast prediction times, making them suitable for practical applications.

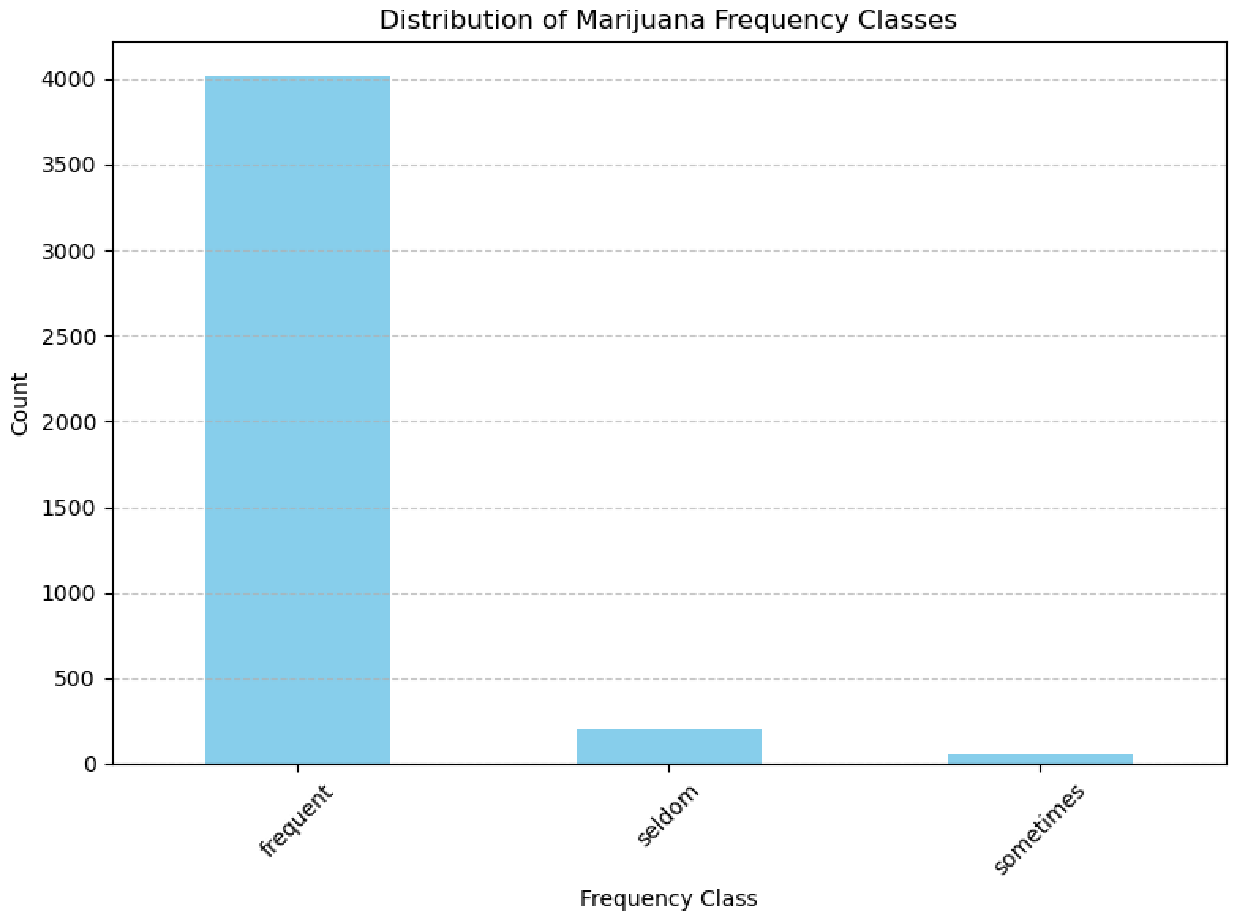
Methodology:

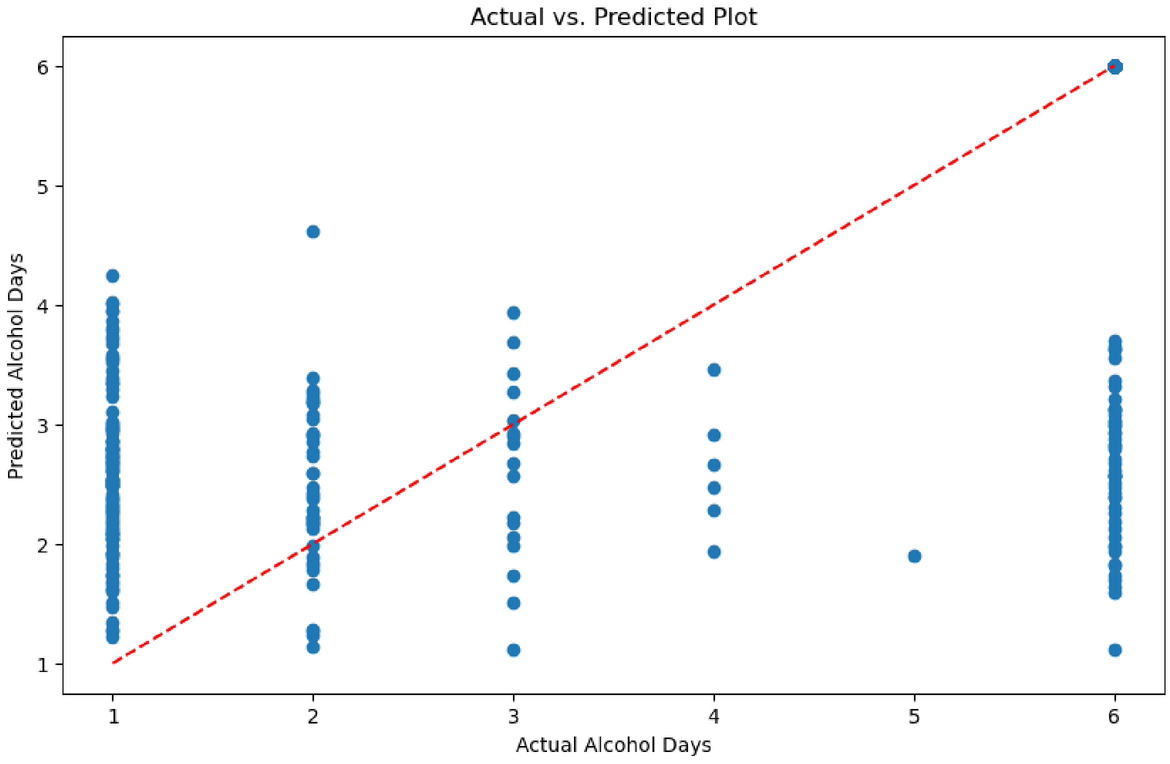
In this study, I implemented a comprehensive methodology focused on tobacco use to ensure the accuracy of my models and the quality of my data for predicting youth drug use. I took steps such as removing missing values, converting variables to categorical format, and renaming variables for clarity. By aggregating multiple demographic factors into a single column and dividing the dataset into training and testing sets, I ensured robust model training and validation. Utilizing ensemble methods like Random Forests and Gradient Boosting, I predicted various aspects of youth drug use, including tobacco use, marijuana frequency, and alcohol consumption. Pruning and cross-validation techniques, including the elbowing method for determining the optimal number of trees, were employed to refine the models. These evaluations provided insights into significant factors and underlying patterns of youth behavior, ultimately enabling the development of robust prediction models for identifying and predicting patterns in youth alcohol consumption.

Computational Results:


Feature Importances to Tabacco use
Importance
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Grade_Level
Overall_Health
Days_Skipped_School_Past_Month
Race
Family_Income
Metro_Size_Status
Population_Density
gender
Poverty_Level
Teacher_Job
School_Felt
Parental_Help_Homework
Parental_Check_Homework
Father_in_Household
Youth_Activity
Religious_Friends
Religious_Importance
Religious_Decision
Religious_Attendance
Government_Assistance
Average_Grade
Mother_in_Household
Talk_Probability
Now_Going_to_School
Pregnancy_Program
Feature


Elbow Method for Optimal Number of Trees
Accuracy
0.895
0.896
0.897
0.898
0.899
0.900
60
80
100
120
140
160
180
200
Number of Trees


Distribution of Marijuana Frequency Classes
Count
4000
3500
3000
2500
2000
1500
1000
500
0
frequent
seldom
sometimes
Frequency Class


Actual vs. Predicted Plot
Predicted Alcohol Days
6
5
4
3
2
1
1
2
3
4
5
6
Actual Alcohol Days

Discussion:

In my evaluation of various classification models for predicting youth tobacco use, the Random Forest Classifier with the optimal number of trees displayed the best overall performance, achieving an impressive accuracy of 90%. This optimized classifier also outperformed the standard Random Forest Classifier and the Gradient Boosting Classifier in terms of precision and recall for the minority class ("Used"), with a precision of 0.30 and recall of 0.04. This superior performance in correctly identifying instances of tobacco use makes the Random Forest Classifier with the optimal number of trees my preferred choice. I identified key predictors of tobacco use as grade level, overall health, family income, days skipped school, and population density. Analysis of the distribution of marijuana frequency classes revealed 4015 instances categorized as 'frequent', 201 instances as 'seldom', and 53 instances as 'sometimes'. Moreover, the classification model achieved an impressive accuracy of approximately 94.5%, indicating strong performance on the test data. Moving to regression analysis for predicting the number of days of alcohol consumption in the past year, my model yielded a respectable R-squared score of 0.6635, indicating a good fit of the model to the data. However, the Mean Squared Error (MSE) and Mean Absolute Error (MAE) were calculated at 1.1453 and 0.4620, respectively. While these metrics provide valuable insights into the model's predictive performance, further exploration and refinement may be necessary to improve accuracy and minimize errors. Overall, the findings suggest promising results in predicting youth drug use behaviors using machine learning techniques, but ongoing refinement and validation of models will be essential for enhancing predictive accuracy and reliability in real-world applications. Furthermore, The Recovery Village's article 'Drug Use in High School' highlights that during school, teenagers experience significant peer pressure, often leading to engagement in risky behaviors such as experimenting with drugs or alcohol in order to fit in socially. Concurrently, academic pressure mounts as students navigate challenging coursework and strive for success in preparation for college or career paths, fueled by expectations from both parents and teachers. The overwhelming workload and stressors of academic life may drive some teens to seek relief through performance-enhancing drugs to boost energy and concentration, or to aid sleep under stress. However, misuse of these substances, taken without prescription, can lead to addiction and pose serious health risks.

Conclusion:

Employing decision tree algorithms and ensemble techniques in this study provided significant insights into the factors influencing substance use among adolescents. The Random Forest Classifier with the optimal number of trees proved highly effective, displaying exceptional accuracy in predicting tobacco use and identifying key predictors such as grade level, overall health, family income, days skipped school, and population density. Additionally, the models successfully predicted marijuana frequency classes and the number of days of alcohol consumption with strong performance in regression analysis. These findings underscore the potential of machine learning techniques in understanding youth drug use behaviors and highlight the importance of addressing peer and academic pressures in efforts to mitigate substance experimentation and misuse among adolescents, ultimately promoting their overall well-being.

Reference:

The Recovery Village.Last updated: July 12, 2023. "Drug Use in High School."
https://www.therecoveryvillage.com/teen-addiction/drug/high-school-drug-use

Marsiglia FF, Kulis S, Nieri T, Parsai M. God forbid! Substance use among religious and non-religious youth. Am J Orthopsychiatry. 2005 Oct;75(4):585-98. doi: 10.1037/0002-9432.75.4.585. PMID: 16262516; PMCID: PMC3043382.

Grim BJ, Grim ME. Belief, Behavior, and Belonging: How Faith is Indispensable in Preventing and Recovering from Substance Abuse. J Relig Health. 2019 Oct;58(5):1713-1750. doi: 10.1007/s10943-019-00876-w. Erratum in: J Relig Health. 2019 Aug 21;: PMID: 31359242; PMCID: PMC6759672.

Appendix :

```
# Target variable
target_binary = 'Tobacco_Ever_Used'

# Prepare data
# Convert selected columns to categorical
X_binary[demographic_cols + youth_exp_cols] = X_binary[demographic_cols +
youth_exp_cols].astype('category')

y_binary = youth_data[target_binary]

# Train-test split
X_train_binary, X_test_binary, y_train_binary, y_test_binary = train_test_split(X_binary, y_binary,
test_size=0.2, random_state=42)

# Model training
binary_classifier = RandomForestClassifier()
binary_classifier.fit(X_train_binary, y_train_binary)

# Predict class labels
y_pred = binary_classifier.predict(X_test_binary)

# Generate the classification report
report = classification_report(y_test_binary, y_pred, target_names=["Not Used", "Used"])

# Print the classification report
print("Classification Report:")
print(report)

# Initialize GradientBoostingClassifier
gb_classifier = GradientBoostingClassifier()

# Train the GradientBoostingClassifier
gb_classifier.fit(X_train_binary, y_train_binary)

# Predict class labels
y_pred_binary_gb = gb_classifier.predict(X_test_binary)

# Evaluate the GradientBoostingClassifier
classification_rep_gb = classification_report(y_test_binary, y_pred_binary_gb)
```

```
print("\nGradient Boosting Classifier Classification Report:")
print(classification_rep_gb)

Define the range of number of trees to try
num_trees_range = range(50, 201, 25)

# Lists to store accuracy values
accuracy_scores = []


# Train-test split
X_train_binary, X_test_binary, y_train_binary, y_test_binary = train_test_split(X_binary, y_binary,
test_size=0.2, random_state=42)

# Iterate over each number of trees and train a Random Forest model
for num_trees in num_trees_range:
    # Initialize Random Forest classifier
    rf_classifier = RandomForestClassifier(n_estimators=num_trees, random_state=123)

    # Train the classifier
    rf_classifier.fit(X_train_binary, y_train_binary)

    # Predict class labels
    y_pred = rf_classifier.predict(X_test_binary)

    # Calculate accuracy and append to list
    accuracy = accuracy_score(y_test_binary, y_pred)
    accuracy_scores.append(accuracy)

# Plot the elbow method
plt.figure(figsize=(10, 6))
plt.plot(num_trees_range, accuracy_scores, marker='o', linestyle='-')
plt.title('Elbow Method for Optimal Number of Trees')
plt.xlabel('Number of Trees')
plt.ylabel('Accuracy')
plt.grid(True)
plt.show()

# Get feature importances
importances = binary_classifier.feature_importances_

# Get the names of the features
feature_names = X_binary.columns

# Sort feature importances in descending order
indices = np.argsort(importances)[::-1]

# Plot
```

```
plt.figure(figsize=(10, 6))
plt.title("Feature Importances to Tabacco use")
plt.bar(range(X_binary.shape[1]), importances[indices], align="center")
plt.xticks(range(X_binary.shape[1]), feature_names[indices], rotation=90)
plt.xlim([-1, X_binary.shape[1]])
plt.xlabel("Feature")
plt.ylabel("Importance")
plt.tight_layout()
plt.show()

Count the occurrences of each class
class_counts = youth_data['Marijuana_Frequency_Class'].value_counts()

# Plot the distribution
plt.figure(figsize=(8, 6))
class_counts.plot(kind='bar', color='skyblue')
plt.title('Distribution of Marijuana Frequency Classes')
plt.xlabel('Frequency Class')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()


# Target variable
target_regression = 'Alcohol_Days_Past_Year'

# Predictor columns
predictors = demographic_cols + youth_exp_cols

# Additional features
additional_features = ['Alcohol_Ever_Used', 'Tobacco_Ever_Used', 'Marijuana_Ever_Used']

# Concatenate predictors and additional features
X_reg = pd.concat([youth_data[predictors], youth_data[additional_features]], axis=1)

# Target variable
y_reg = youth_data[target_regression]

# Define column transformer for one-hot encoding
categorical_features = [column for column in X_reg.columns if X_reg[column].dtype == 'object']
column_transformer = ColumnTransformer([('encoder', OneHotEncoder(), categorical_features)],
remainder='passthrough')

# Apply column transformer to encode categorical predictors
X_reg_encoded = column_transformer.fit_transform(X_reg)
```

```
# Train-test split
X_train_reg, X_test_reg, y_train_reg, y_test_reg = train_test_split(X_reg_encoded, y_reg, test_size=0.2,
random_state=42)

# Model training
regressor = RandomForestRegressor()
regressor.fit(X_train_reg, y_train_reg)

# Model evaluation
y_pred_reg = regressor.predict(X_test_reg)
regression_score = r2_score(y_test_reg, y_pred_reg)
print("Regression R-squared Score (after encoding):", regression_score)
```