Analyzing and Predicting Youth Substance Use Based on Family Engagement, Religious Background, Educational Environment, and Personal Experiences

Abstract:

This study leverages the National Survey on Drug Use and Health (NSDUH) to forecast youth substance consumption and the initial age of substance exposure using decision trees and ensemble models. Focusing on predictors like family engagement, educational influences, religious beliefs, and involvement in developmental group activities, this research underscores how early life experiences shape attitudes towards substance use. By integrating behavioral and environmental variables, our main aim through this study is to predict the type of substance that may be used or whether the substance would be used or not, and the age at which initial use occurs. The findings advocate for the critical role of familial and community support in mitigating risky behaviors, providing insights for targeted preventative strategies.

Introduction:

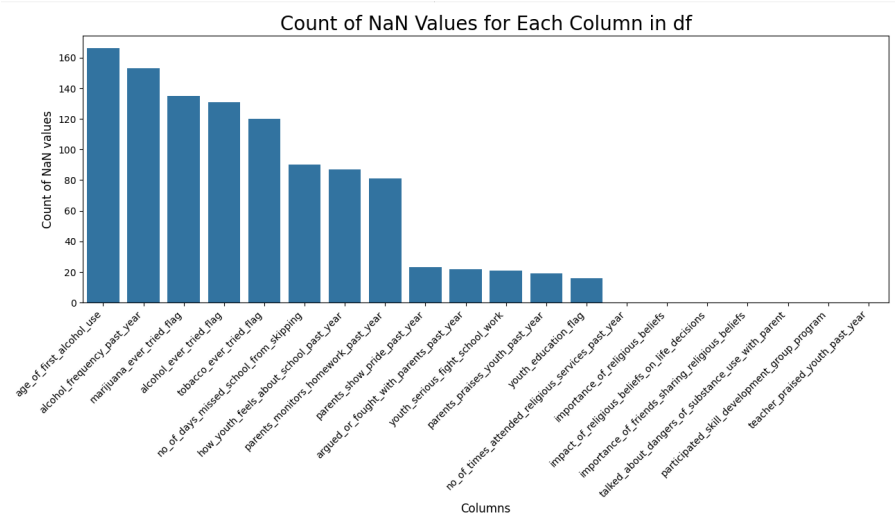
Alcohol consumption among youth is a significant public health issue that can lead to adverse effects on individual health and societal well-being. This study employs data from the National Survey on Drug Use and Health (NSDUH), an annual survey designed to gather information about substance use and related behaviors from U.S. residents aged 12 and older. Through this study, we aim to unearth potential risk and protective factors that are integral to understanding youth behavior towards alcohol. These insights are crucial for developing targeted prevention and intervention strategies that could significantly diminish alcohol consumption among young people.

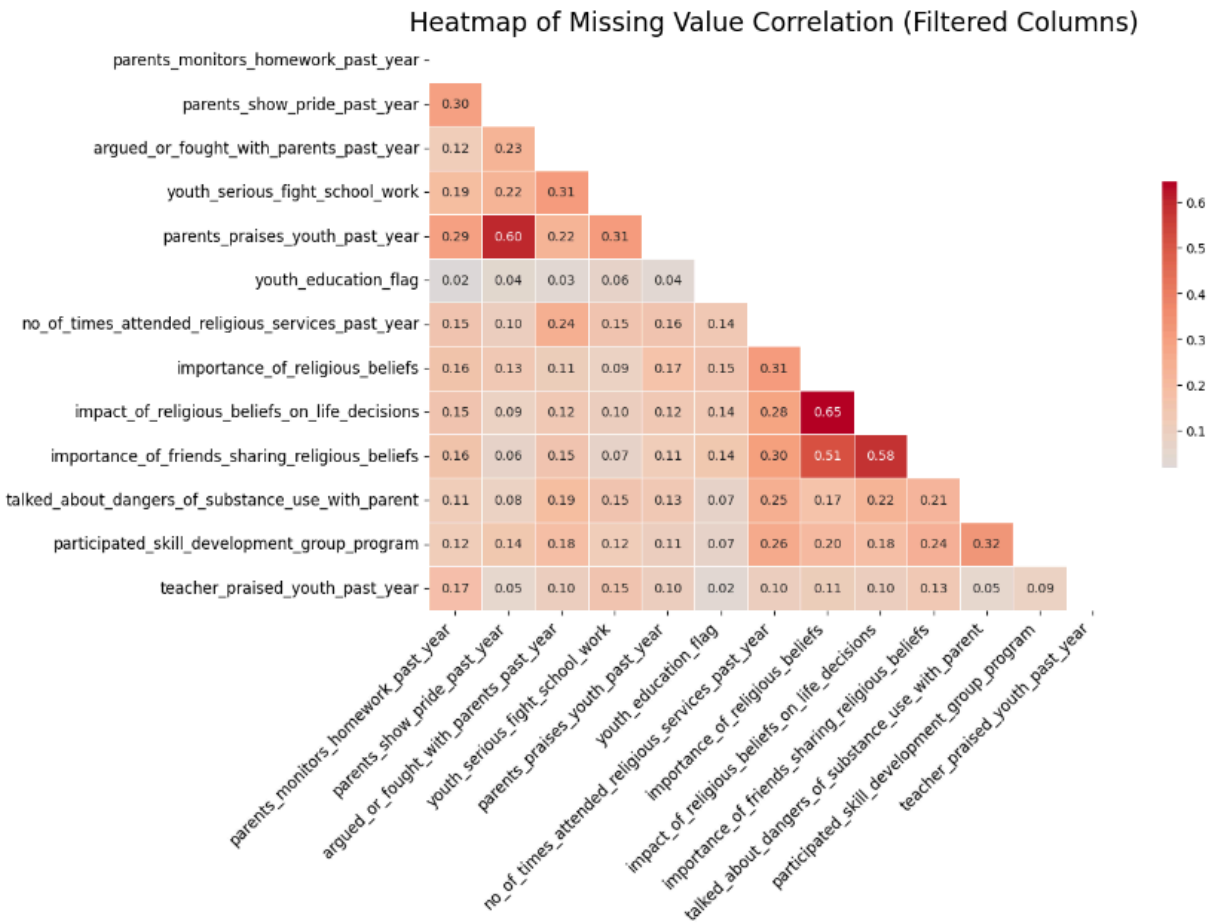
Our predictive models not only seek to identify the likelihood of alcohol consumption but also aim to categorize potential substance preferences among youth—ranging from alcohol to marijuana and tobacco. Additionally, we utilize regression techniques to estimate the age at which youths might first engage with alcohol. The outcomes of this research will provide valuable guidance for public health initiatives aimed at curbing substance use among adolescents.

Methodology:

The objective of this study was to predict both the likelihood of substance use among youth and the age at which they might first encounter substances such as alcohol, tobacco, and marijuana. The methodology was rigorously designed to analyze the National Survey on Drug Use and Health (NSDUH) dataset comprehensively, focusing particularly on cleaning the data and preparing it for accurate modeling.

Initial steps in the data preparation phase involved the removal of irrelevant variables that did not contribute to the prediction models. This was followed by a detailed examination of missing values within the dataset to ensure the integrity and completeness of the analysis. Variables that showed a strong correlation with alcohol consumption were specifically excluded to prevent multicollinearity, which could skew the predictive models.





Since removing these columns could introduce bias in the model, we imputed the missing values using the Simple Imputer method and replaced them with 0. The column EDUSCHGRD2 that describes the grade level of youth had many categories and for our analysis, knowing whether someone either went to school or not is enough. This variable was modified to a binary variable youth_education_flag with 98 denoting missing values or unknown response. We take care of this later in our analysis.

```
youth_education_flag
4     869
5     828
3     823
6     818
7     707
99    667
8     394
2     281
98     81
1      22
9       8
10      2
Name: count, dtype: int64
```

```
youth_education_flag
1     4752
0      667
98      81
Name: count, dtype: int64
```

The core of our methodology involved employing decision tree models to predict alcohol consumption based on a wide array of demographic and behavioral factors, including age, gender, race, educational attainment, and family income. These predictors were chosen due to their potential relevance to substance use patterns as indicated by prior research.

Classification Models:

Initially, we utilized decision trees for a simple binary classification task—to determine whether an individual has consumed alcohol. To enhance the robustness of our predictions, we applied cross-validation techniques, systematically partitioning the data into subsets, training the model on one subset, and validating it on another. This approach helps in minimizing bias and variance, ensuring that our model generalizes well to new data.

Subsequently, we expanded our analysis to a multi-class classification setting where the objective was to identify whether a respondent's substance use involved alcohol, marijuana, or tobacco. This was operationalized by integrating individual usage flags into a single 'substance use flag', thereby allowing for a comparative analysis across different types of substance use.

Regression Models for Age Prediction:

Parallel to our classification efforts, we developed a regression model aimed at predicting the age at which an individual first consumes alcohol. Recognizing outliers in the dataset—such as reports of alcohol consumption at improbably early ages—we corrected these by setting a reasonable minimum age threshold. This helped in maintaining the statistical integrity of our model.

```
age_of_first_alcohol_use
991.0    4138
15.0      329
14.0      291
13.0      225
16.0      166
12.0      112
11.0       61
17.0       57
10.0       36
9.0        21
8.0        19
7.0        17
6.0         9
5.0         6
3.0         5
4.0         5
2.0         2
1.0         1
Name: count, dtype: int64
```

Figure 1

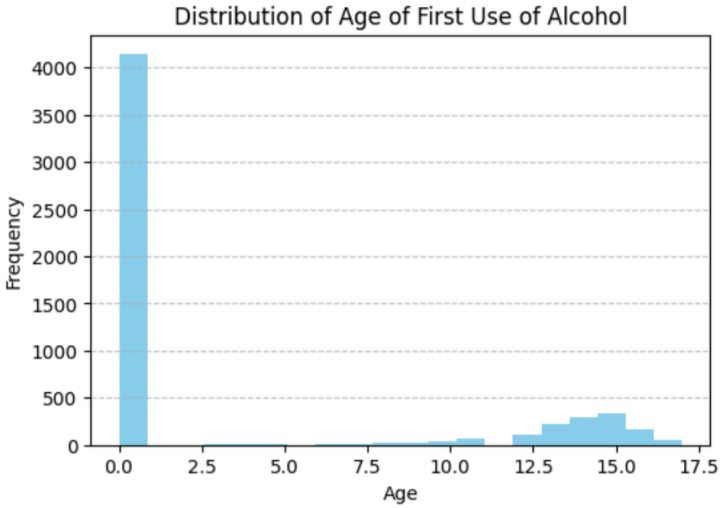
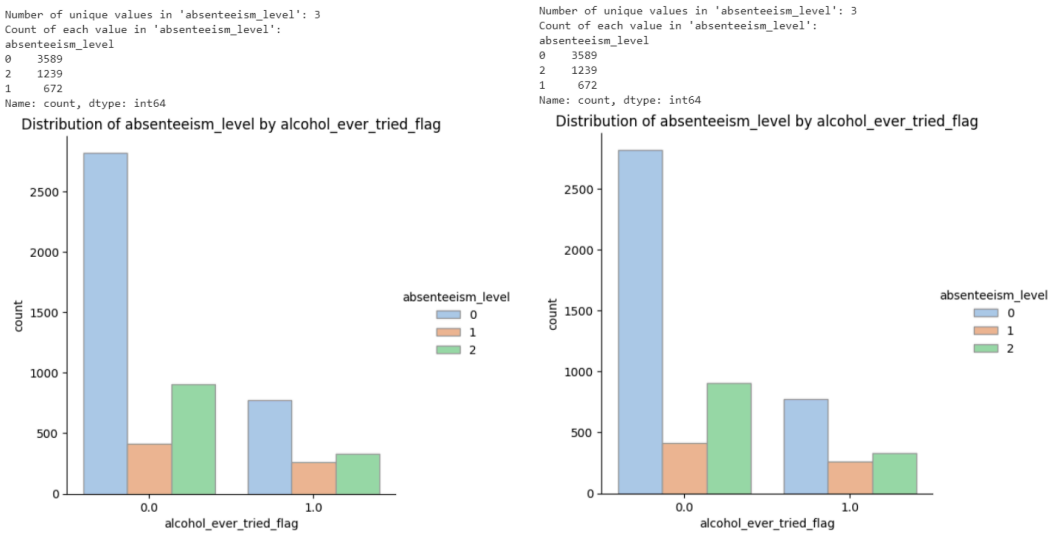


Figure 2

We thoroughly examined the distribution of ages at which alcohol consumption typically begins, as illustrated in Figure 2. Our observations indicated a common initiation age around 15 years,

with a noteworthy proportion of youths yet to initiate alcohol use. These insights were crucial for setting up our regression model, which aims to predict the age of first alcohol use accurately.

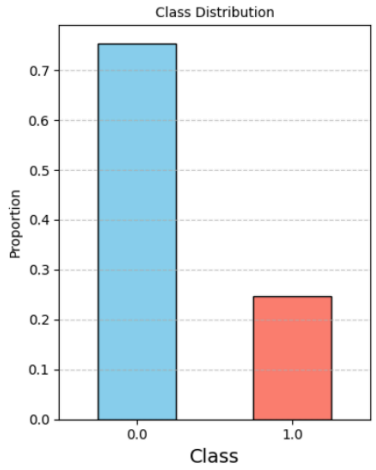
The variable *eduskpcom* denotes the number of days youth missed school from skipping. This variable was also distributed over a range of numeric values. We modified this variable and split it into three categories and renamed it to absenteeism_level: low, high, medium as shown in the plots below.



Most of the variables had classes 1 or 2. To make the data consistent throughout the dataset, we modified these columns to have binary values 0(where it was 1) or 1(where it was 2).

Computational Results:

Binary Classification: In our binary classification task, the objective was to predict whether or not a youth would engage in alcohol consumption. We encountered a significant class imbalance in the dataset: out of 5,550 data points, 4,138 indicated no alcohol consumption among adults.



To tackle this imbalance and enhance the reliability of our predictive models, we implemented stratified sampling techniques. This approach ensured that each class was adequately represented during the training of our decision tree and ensemble models, thus mitigating potential biases due to uneven class distributions. The figure below shows the important variables of decision tree classifiers when fitting the model on training data. It is evident that the relationship of youth with Parents and religious beliefs is a strong predictor.

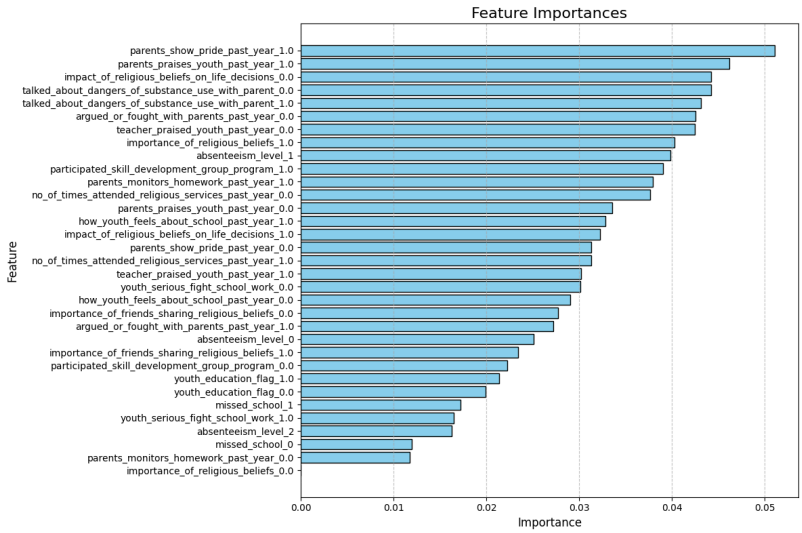
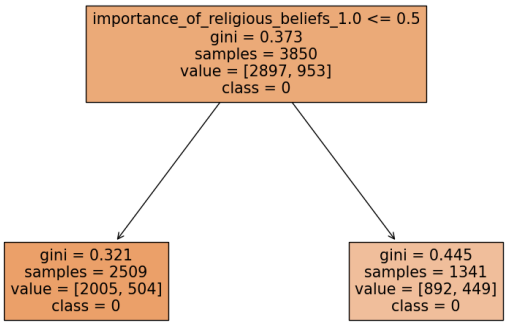
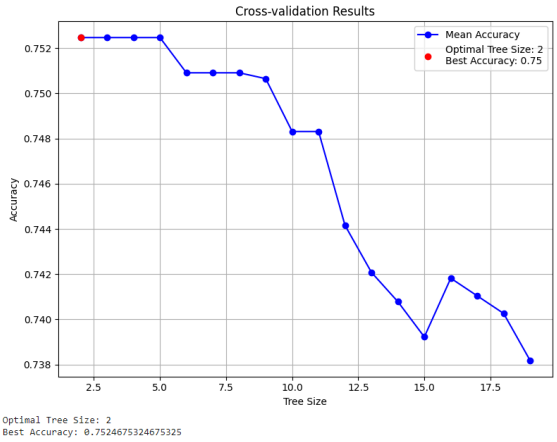


Figure 3: Importance of features in classifying whether a person will consume alcohol or not.






```
Optimal Tree Size: 2
Best Accuracy: 0.7524675324675325
```

Pruned tree and cross validation results

It's particularly noteworthy that the pruned decision tree model achieved an accuracy of 75.2% on the test set, which is identical to that of a simpler, single-node unpruned decision tree with 71.82%. This observation is intriguing as it suggests that both models—despite their differences

in complexity—encounter the same level of error when predicting alcohol consumption among youth.

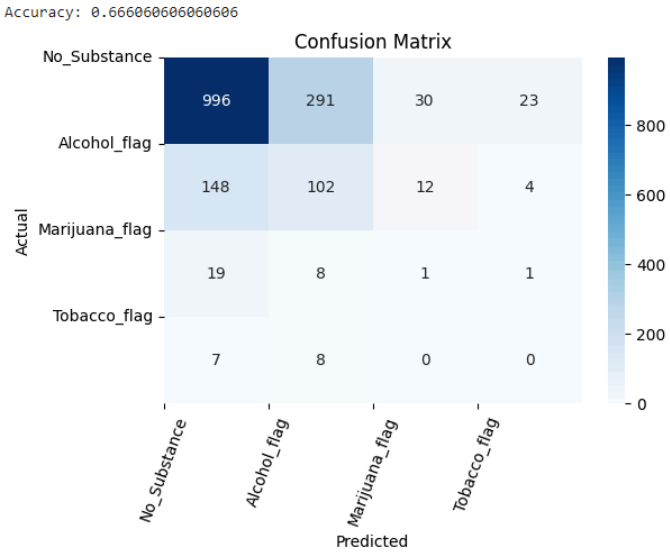
It indicates that simplifying the decision tree, through the process of pruning, did not compromise its predictive power. In fact, the pruned tree maintained high accuracy while potentially offering benefits in terms of model interpretability and generalizability. This similarity in performance also raises questions about the underlying distribution of the data and the factors most predictive of substance use, suggesting that a minimal set of features may be adequate for achieving significant predictive accuracy.

These findings underscore the importance of feature selection and model complexity in the context of predictive modeling. They highlight that more complex models, such as an extensive decision tree, do not necessarily provide better accuracy than their simpler counterparts. This insight is crucial for optimizing model performance in practical applications, where simplicity and efficiency are often as valued as accuracy.

Multiclass classification: In our multiclass classification analysis, the aim is to predict whether an individual might use any type of substance. For this purpose, we have organized the possible outcomes into four distinct categories as previously described.

```
substance_use_flag
0    3902
1    1362
2     142
3      94
Name: count, dtype: int64
```

Figure 4: Multiclass variable



Confusion matrix of multiclass classification

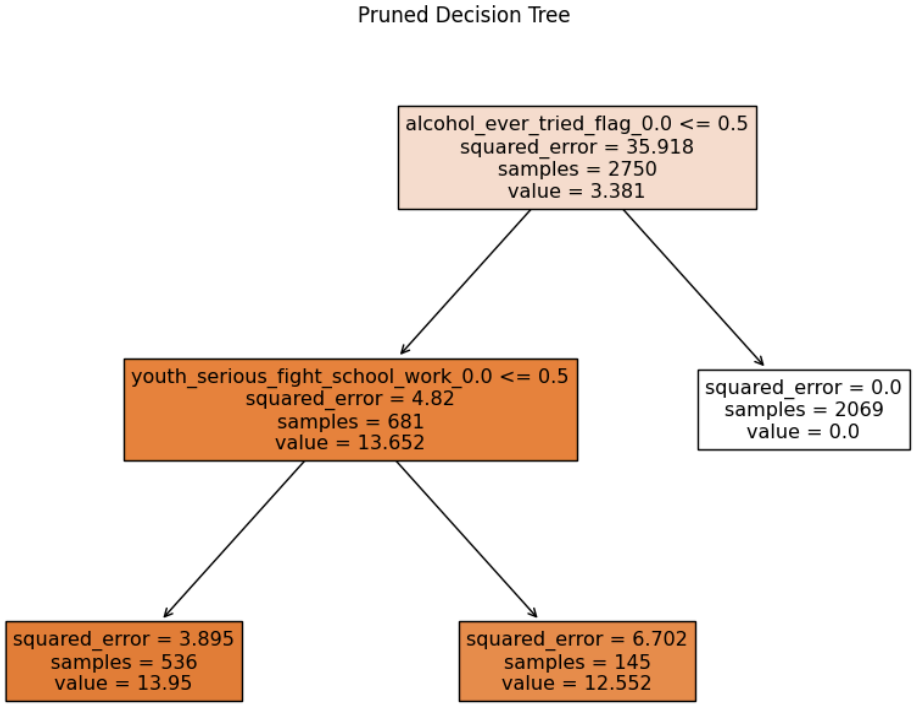
After exploring various modeling approaches, we found that the bagging model stood out as the most effective for analyzing our dataset. This ensemble method significantly improved prediction accuracy, particularly in scenarios involving non-use of substances, as evidenced by the detailed results shown in the confusion matrix.

The bagging model's strength lies in its robustness and ability to generalize across different cases, resulting in a commendable overall accuracy rate of 69.58%. It is important to highlight that while the model performs exceptionally well in predicting cases of non-substance use, there is still room for improvement in its predictive capabilities for other categories.

Further refinement of the model could potentially be achieved by incorporating more comprehensive data on youth behavior. By expanding our dataset to include additional behavioral indicators, we anticipate that the model's accuracy could be enhanced even further. This approach would allow for a more nuanced analysis of the factors influencing substance use among youth, thereby improving our ability to predict various outcomes more effectively.

Regression Model Analysis:

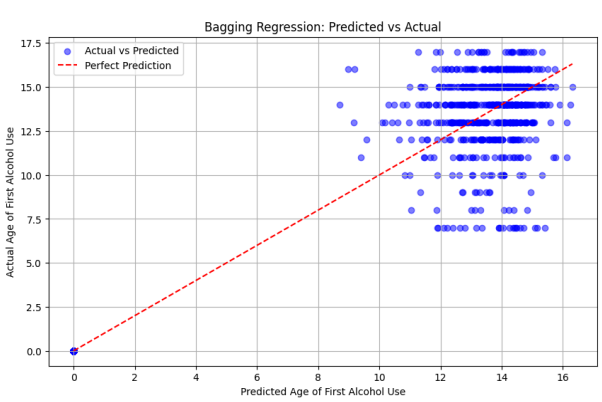
Our regression model aimed to predict the age at which young individuals are likely to have their first alcoholic drink. We employed a variety of modeling techniques to achieve this, with a focus on using a pruned decision tree for regression analysis. This method proved effective, allowing us to estimate the age of initial alcohol use with an impressive accuracy of approximately 96.84%, despite some inherent limitations in the data.

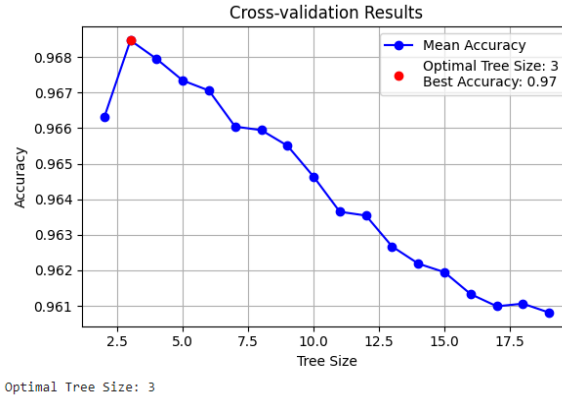


Throughout our analysis, it became evident that the presence of an "alcohol use" indicator within the pruned decision tree was a critical factor in predicting the onset age of drinking. Specifically, whether or not an individual had previously consumed alcohol was found to significantly influence the age at which they first experimented with alcohol.

Moreover, our study uncovered that experiences involving serious fights at school or work were the second most influential predictor in our model. This finding highlights the potential psychological impact of such violent incidents on young people's behavior, suggesting a possible link to earlier substance experimentation and subsequent long-term mental health issues, such as depression. This insight underscores the importance of considering social and behavioral contexts when assessing risk factors for early alcohol use.

Bagging proved to be highly accurate but the pruned decision tree has outperformed it with an accuracy of 96.84% as shown below.






```
Optimal Tree Size: 3
```

Discussion:

The NSDUH dataset presents a comprehensive overview of behavioral and demographic variables that offer valuable insights into substance use patterns. While our study primarily focused on the impact of religious beliefs and educational experiences on youth alcohol consumption, these are just a fraction of the potential analyses that this rich dataset supports. Our findings highlight significant risk and protective factors that could be instrumental in shaping effective prevention and intervention strategies targeted at reducing substance use among youth.

Influential Factors in Substance Use: Our analysis revealed that parental communication about substance use, the child's academic achievements, and their attitudes towards school play pivotal roles in influencing their likelihood of consuming alcohol. Specifically, positive school experiences and open conversations about substance risks are linked to a lower likelihood of substance use, underscoring the importance of supportive educational and familial environments.

The decision tree model, which was pruned to optimize performance without compromising accuracy, demonstrated a commendable accuracy rate of 75.2% in predicting whether an

individual would consume alcohol. This indicates the model's effectiveness in capturing the complex interplay of factors that contribute to alcohol consumption among youth.

Multiclass Classification Findings: In terms of multiclass classification, the goal was to ascertain broader substance use patterns, including alcohol, tobacco, and marijuana use. The results indicated that besides the factors mentioned above, participation in self-esteem and problem-solving groups also emerged as crucial predictors. This suggests that interventions aimed at building these skills could significantly mitigate substance use incidences.

However, it is important to acknowledge certain limitations in our study. The cross-sectional design of the NSDUH dataset restricts our ability to draw causal inferences about the relationships between behavioral factors and substance use. Additionally, the reliance on self-reported data might introduce biases, such as underreporting due to social desirability, which can affect the accuracy of our predictions. The relatively small sample size could also limit the generalizability of our findings to a broader population.

Despite these limitations, our study provides critical insights into the factors that can reduce the likelihood of alcohol and other substance use among youth. These insights reinforce the value of targeted educational and family-oriented interventions. Looking forward, extending this research to include longitudinal data could help clarify causal relationships and enhance the predictive power of our models. Furthermore, incorporating a larger and more diverse dataset could improve the generalizability of our results, providing a stronger basis for developing nationwide substance abuse prevention programs.

Conclusion:

Our research utilized decision tree models and tree-based ensemble methods, including Random Forest, Bagging, and Boosting, to predict alcohol consumption among youth. Employing data from the NSDUH survey, our models achieved commendable accuracies, highlighting their effectiveness in substance use prediction.

In binary classification tasks, the pruned decision tree model notably achieved an accuracy of 75.24%, while in a more focused analysis, it reached up to 96% accuracy using less than half of the initial predictors. These results not only demonstrate the efficiency of pruning in reducing model complexity but also underscore the potential of these models to identify early alcohol use accurately. Our findings are significant as they affirm that both decision tree models and ensemble techniques are robust tools for predicting whether an individual might engage in alcohol consumption. More importantly, these methods can accurately estimate the age of first alcohol use, providing crucial data for preventive health strategies.

Additional studies could expand on this work by incorporating more diverse datasets or by applying these techniques to predict other types of substance use and behavioral outcomes.

Overall, the effectiveness of our predictive models provides valuable insights for public health officials and educators seeking to implement evidence-based strategies to reduce alcohol misuse among youth. As we continue to refine these models and enhance their accuracy, they hold the potential to significantly impact public health initiatives by enabling more precise and timely interventions.

References:

Substance Abuse and Mental Health Services Administration. (2020). National Survey on Drug Use and Health (NSDUH) 2020. [Codebook]. Retrieved from https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/studies/NSDUH2020/NSDUH-2020-datasets/NSDUH-2020-DS0001/NSDUH-2020-DS0001-info/NSDUH2020-DS0001-info-codebook.pdf.