Substance Use Patterns in Young Adults: A Data-Driven Investigation Using Regularization Techniques and Tree-Based Methods.

Balaji Kolusu, M.S.

Abstract

Background

Substance use poses a substantial public health challenge, particularly among young adults—a demographic that frequently exhibits varied and evolving patterns of drug use. The complexity of factors contributing to substance use in this age group underscores the necessity for targeted research aimed at uncovering these dynamics. Such understanding is vital for developing interventions that can effectively address and mitigate the risk factors associated with increased substance use.

Purpose

This study leverages a cross-sectional survey of 261 adults delve into the predictors associated with lifetime drug use. Specifically, the research focuses on the variable TOTALDRUGS, which quantifies the number of different drugs used by individuals throughout their lifetime. By identifying key predictors, the study aims to illuminate the underlying factors that lead to substance use, providing a foundation for targeted preventive measures.

Methods: A combination of LASSO and Ridge regression, alongside tree-based methods, was applied to analyze survey data, exploring both continuous and categorical predictors of drug use.

Results: in the analysis using LASSO, Ridge Regression, and Random Forest, significant Predictors such as TOTAL ILLEGAL DRUGS1, ALCOHOL Yes, and TOBACCO Yes were identified as major influences on the total number of lifetime drugs used, with the Random Forest additionally emphasizing the role of psychological factors like ANXIETY alongside substance-specific influences.

Conclusions: The effectiveness of different statistical methods in identifying predictors suggests a multi-strategic analytical approach may be essential for fully understanding the determinants of substance use among young adults.

Keywords: Young Adults, Substance Use, LASSO Regression, Ridge Regression, Tree-Based Methods, Predictive Analytics, Public Health.

Introduction:

Substance use among young adults remains a pressing public health issue, with significant implications for individual health, societal well-being, and economic productivity. The aim of this study is to identify key predictors of lifetime drug use in young adults, utilizing advanced statistical models to provide a nuanced understanding of the factors influencing substance use behaviors. This research is important because understanding these predictors can help in developing targeted interventions aimed at preventing and reducing drug use in this vulnerable population.

Literature Review

Past research in the field of substance use has extensively explored demographic, psychological, and socio-economic factors as predictors of drug use among young adults. Studies such as those by Bachman et al. (1997) and Johnston et al. (2005) have primarily focused on the impact of factors like peer influence, parental supervision, and socioeconomic status on the likelihood of drug use. These studies have utilized traditional statistical methods, such as logistic regression and correlation analysis, to evaluate the relationship between these factors and drug use.

Moreover, research by Merline et al. (2004) explored the predictive power of early adolescent behavior on later drug use, highlighting the role of early exposure to substances as a significant predictor. However, these studies often did not incorporate advanced machine learning techniques that can handle large datasets with many predictors and capture complex nonlinear relationships between variables.

The present research seeks to address these limitations by utilizing contemporary statistical approaches like LASSO and Ridge regression, along with Random Forest analysis. These techniques were selected due to their proficiency in managing extensive predictor variables and effectively modeling intricate relationships and interactions. Employing these methods enables a thorough examination of the dataset contained in the **dataset**, which encompasses a diverse range of variables including demographic details, historical drug use, and other potential predictors of substance use.

Methods

Problem Statement

The primary objective of this study is to identify significant predictors of drug use among individuals, drawing from a dataset that includes various demographic, behavioral, and psychological characteristics. Given the complexity and potential multicollinearity among predictors, advanced statistical techniques, such as LASSO regression and tree-based methods, are employed to enhance the model's interpretability and predictive accuracy.

Dataset Description

The dataset comprises observations of 261 individuals, each described by 27 variables. These variables include both demographic factors (e.g., age, gender, race) and behavioral/psychological metrics (e.g., alcohol and tobacco use, depression, anxiety scores). Variables are categorized as follows:
Continuous Variables: Age, Depression, Anxiety, Loneliness.
Categorical Variables: Gender, Race, US Born, Employed, School, Therapy, Alcohol, Tobacco, and drug types like Ecstasy, Ketamine, Methamphetamine, etc.

Data Preprocessing

Data preprocessing involved several steps to prepare the dataset for analysis:

1. **Handling Missing Values:** Rows containing any missing values were omitted to maintain data integrity.
2. **Normalization of Numeric Data:** Missing numeric data were imputed with the mean of their respective columns to handle any remaining missing values effectively.
3. **Encoding of Categorical Variables:** Categorical variables were converted into dummy variables to facilitate their use in regression analysis. This step was critical for modeling as it transformed categorical text data into a numerical format that could be processed by the statistical models.

Analytical Methods

The analysis incorporated the following statistical methods:

LASSO Regression:

- **Purpose:** To perform variable selection and regularization to enhance the model's prediction accuracy. LASSO regression is particularly useful for reducing the complexity of the model by penalizing the absolute size of the regression coefficients. By doing so, it effectively reduces overfitting and selects significant variables by shrinking the less important variable's coefficient to zero.
- **Implementation:** LASSO was implemented using the **cv.glmnet** function from the **glmnet** package in R, which also facilitated the selection of the optimal lambda value through cross-validation.

Ridge Regression:

- **Purpose:** Used as a comparison to LASSO, Ridge regression addresses multicollinearity among predictors without eliminating them, by imposing a penalty on the size of coefficients.
- **Implementation:** Similar to LASSO, Ridge regression was applied using the **cv.glmnet** function with **alpha** set to 0, aiding in lambda selection through cross-validation.

Random Forest:

- **Purpose:** To assess variable importance and improve prediction accuracy through an ensemble method that mitigates the risk of overfitting associated with decision trees. Random Forest is beneficial for handling large datasets with higher dimensionality and provides insights into the importance of each variable.
- **Implementation:**
  The **randomForest** package in R was used to fit the model and compute variable importance, which helps in identifying the most significant predictors influencing drug use.

Variable Importance and Model Evaluation

Variable importance was evaluated using the **vip** package in R, which provides a clear graphical representation of the importance scores derived from the Random Forest model. This analysis is crucial for understanding which variables have the most influence on the predicted outcome and thus informs practical interventions or further research.

Conclusion

The combination of LASSO, Ridge, and Random Forest models offers a robust approach to understanding the factors that contribute to drug use. By employing both regularization and ensemble methods, the study aims to achieve a balance between predictive performance and model interpretability, ensuring that the findings are both statistically significant and practically relevant.

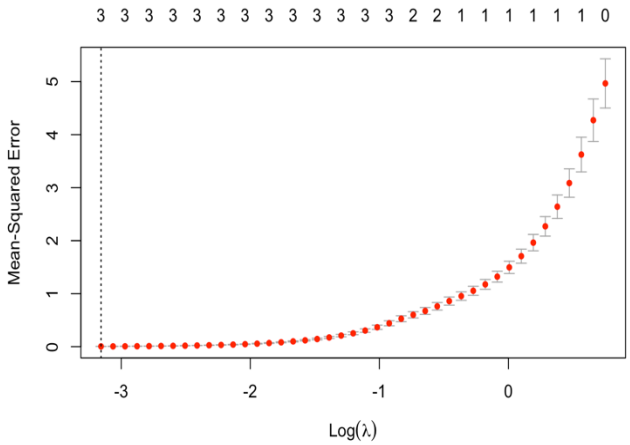
Results

Descriptive Statistics of Key Variables

Variable	Mean	Median	Std. Dev.	Min	25th Pctl.	75th Pctl.	Max
Age	21.44	21.00	2.22	18	20.00	23.00	25
Total Drugs	2.72	2.00	2.21	0	1.00	4.00	12
Illegal Total Illegal Drugs 1	1.75	1.00	2.02	0	0.00	3.00	11
Illegal Total Illegal Drugs 2	1.09	0.00	1.82	0	0.00	2.00	10
Depression	27.59	24.00	9.75	11	24.00	24.00	62
Anxiety	35.94	36.00	4.90	22	32.00	38.00	54
Loneliness (n=213)	6.51	6.00	4.04	3	3.00	6.00	15

The age of participants varied moderately and appeared slightly right-skewed, as indicated by the mean being higher than the median. Drug use also varied widely and was right-skewed, with most participants using fewer drugs but a few using many. Mental health assessments for depression, anxiety, and loneliness showed variability. Depression and anxiety scores were generally high, suggesting a right-skewed distribution. The loneliness scores, though fewer in number, also indicated a right skew with varying levels of loneliness among participants. These statistics provide a foundational understanding of the data, which can be further analyzed to explore the relationships and predictive abilities of these variables.

LASSO Regression Analysis



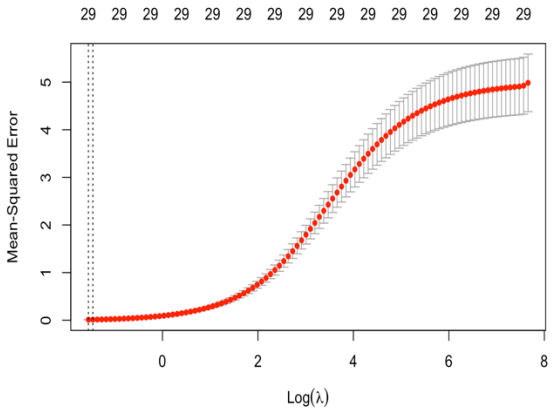
In the LASSO regression visualization, we analyze the mean squared error (MSE) plotted against the logarithm of the regularization parameter lambda (log(λ)). This plot helps us identify the optimal lambda value that minimizes the MSE.
From the plot:
- As log(λ) increases from highly negative towards zero, the MSE is initially low and stable, indicating good model performance even with substantial penalties applied.
- As log(λ) approaches zero, the MSE begins to rise sharply, suggesting that too high a penalty simplifies the model excessively, leading to underfitting and a loss of the ability to detect underlying data patterns.

The dots on the graph show the average MSE for each lambda value during cross-validation, while the bars display the variability of the MSE across different validation folds. The point where the MSE is lowest marks the optimal lambda (λ.min) that minimizes the error, providing the best balance between simplicity and predictive accuracy in the model.

Variable	Coefficient
(Intercept)	0.1055241
IllegalTotalillegalDrugs1	0.9821027
Alcohol Yes	0.9348273
Tobacco Yes	0.9061479

LASSO regression employs a penalty mechanism that reduces certain coefficients to zero, effectively simplifying the model by excluding less relevant predictors. In this analysis, only a select few predictors retained non-zero coefficients, underscoring their strong influence on the total drug use:
Illegal Total Illegal Drugs 1: Exhibited the highest coefficient at 0.982, reflecting a robust positive correlation with overall drug usage.
Alcohol Usage: Recorded a significant coefficient of 0.935, highlighting its substantial link to total drug consumption.
Tobacco Usage: Demonstrated a notable coefficient of 0.906, indicating a considerable positive influence.
Most other variables were minimized to zero coefficients, suggesting that within the data's framework and given the selected regularization strength, these factors do not significantly impact the model.

Ridge Regression Analysis



The Ridge regression graph depicts the Mean Squared Error (MSE) against various levels of the lambda parameter. Analyzing this graph helps us pick the best lambda to make our model accurate without being overly complex.
Beginning of the Graph:When lambda is small, the MSE is low, meaning the model predicts well without much penalty on its coefficients.
Middle of the Graph: As lambda gets larger, the MSE starts to go up. This suggests that the model's predictions are getting worse because the penalty is too strong and is overly reducing the coefficients.
End of the Graph:With a very large lambda, the MSE levels off but is higher than at the start. This tells us that the model has become too simple, possibly ignoring some of the subtler patterns in the data.
The red dots show the MSE for different lambdas during the cross-validation process. The spread of

the dots shows us how much the MSE changes across different test sets, indicating how consistent the model's performance is.

Variable	Coefficient	Impact on Drug Use
Alcohol Usage	0.890	Strong Positive
Tobacco Usage	0.873	Strong Positive
Illegal Total Illegal Drugs 1	0.178	Moderate Positive
Ecstasy Usage	0.585	Moderate Positive
Ketamine Usage	0.632	Moderate Positive
Methamphetamine Usage	0.600	Moderate Positive
Marijuana Usage	0.739	Moderate Positive
Heroin Usage	0.689	Moderate Positive
Gender (Male)	0.028	Slight Positive
US Born (Yes)	0.029	Slight Positive

Ridge regression differs from LASSO in that it reduces the coefficients of variables towards zero but never to zero. This approach is effective in managing issues with multicollinearity among predictors and gives a detailed view of the influence of all variables in the dataset. Our analysis using Ridge regression reveals varied contributions of different predictors to total drug use:

Key Findings on Predictors:

Alcohol and Tobacco Usage: Both have high coefficients of 0.890 and 0.873 respectively, indicating a strong positive relationship with drug use.

Illegal Total Illegal Drugs 1: With a coefficient of 0.178, this predictor has a noticeable but smaller impact compared to the coefficients in the LASSO model, yet it remains a significant factor.

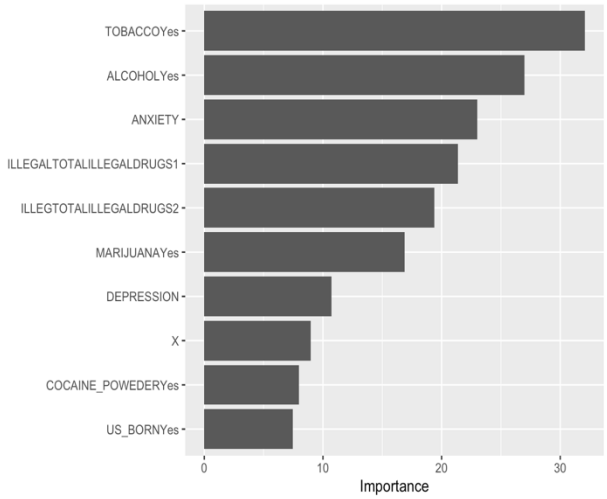
Other Substances: Drugs such as Ecstasy, Ketamine, and Heroin also demonstrate considerable influence with coefficients of 0.585, 0.632, and 0.689, indicating their substantial roles.

Demographic Factors: Variables like gender and being born in the US show smaller effects with coefficients around 0.028 and 0.029, suggesting a modest impact on drug use.

This simplified analysis underscores that alcohol and tobacco are the primary predictors of drug use, though other drugs and certain demographic characteristics also contribute significantly. Ridge regression's Comprehensive evaluation helps us understand the extensive range of factors affecting drug use.

Random Forest Analysis on Drug Use

The visualization from the Random Forest output is a bar chart representing the importance of each variable in predicting total drug use. The length of the bars corresponds to the importance score for each variable, with longer bars indicating a greater importance.



Interpreting the Random Forest Visualization:

- **Tobacco Usage (TOBACCO Yes)**: With the longest bar, it's the most significant predictor, showing the highest importance score in predicting drug use.
- **Alcohol Usage (ALCOHOL Yes)**: The second longest bar suggests that alcohol is also a highly significant predictor.Illegal

Drug Scores(ILLEGALTOTALILLEGALDRUGS1 and ILLEGTOTALILLEGALDRUGS2): These have substantial bars, indicating they are important predictors as well.

- **Psychological Factors (ANXIETY)**: The considerable length of the bar for anxiety indicates it’s an important predictor, likely reflecting the psychological component's impact on drug use.
- **Demographics (US_BORNYes)**: A smaller bar for being born in the U.S. suggests a modest impact compared to substance-

related predictors.The Random Forest model demonstrated high predictive accuracy, accounting for 97.5% of the variance in total drug use. It offered insightful details on variable importance, measured by increases in mean squared error (%IncMSE) and total node purity (IncNodePurity). Here are the most significant predictors identified:
This integration emphasizes the model's effectiveness in determining the impact of various predictors on drug use, highlighting those with the most substantial effects.

Variable	%IncMSE	IncNodePurity	Influence on Drug Use
Alcohol Usage	26.01	35.09	Strong Positive
Tobacco Usage	34.21	15.96	Strong Positive
Illegal Total Illegal Drugs 1	20.92	236.71	Strong Positive
Illegal Total Illegal Drugs 2	19.53	264.64	Strong Positive
Marijuana Usage	16.25	29.88	Strong Positive
Anxiety	24.40	154.00	Strong Positive
Cocaine Powder Usage	7.73	53.25	Moderate Positive
Ecstasy Usage	6.10	32.27	Moderate Positive
Depression	10.18	95.16	Moderate Positive
Loneliness	5.83	8.03	Moderate Positive

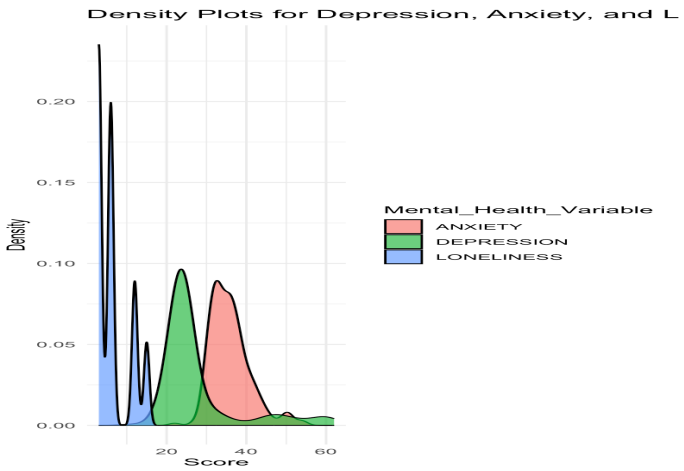
The analysis revealed that substance use predictors such as Alcohol, Tobacco, and various illegal drugs (Illegal Total Illegal Drugs 1 and 2, Marijuana) are the most influential on total drug use. Additionally, mental health indicators like Anxiety and Depression also show significant contributions, highlighting the complex interplay between substance use and mental health.

Recommendations

- **Policy Interventions**: Focus on addressing not only illegal substance use. but also, legal substances like alcohol and tobacco which are significant predictors of broader drug use.
- **Mental Health Services**: Enhance support for mental health as its strong association with drug use suggests a need for integrated treatment approaches.

Graphical Analysis and Model Interpretation

Analysis of Mental Health Scores Distribution



The provided density plot serves as a graphical illustration of the distribution of scores across three key mental health variables: Depression, Anxiety, and Loneliness. Through this plot, we gain a deeper understanding of the patterns and variances in the data. Distinctly colored curves delineate the frequency distribution for each mental health condition, providing a clear visual of how scores are dispersed and where they are most concentrated. The x-axis showcases the range of scores obtained from mental health assessments, while the y-axis denotes the density, reflecting the likelihood of scores within the dataset.

Insights Derived from the Density Plot

A comprehensive examination of the density plot reveals diverse insights:

- **Anxiety**: Represented in red, the distribution of Anxiety scores demonstrates a broader spectrum of responses, with a significant prevalence at the lower end of the score range.
- **Depression**: The green curve associated with Depression scores indicates a notable clustering effect around a central range,

suggesting a pattern of symptom commonality among respondents.

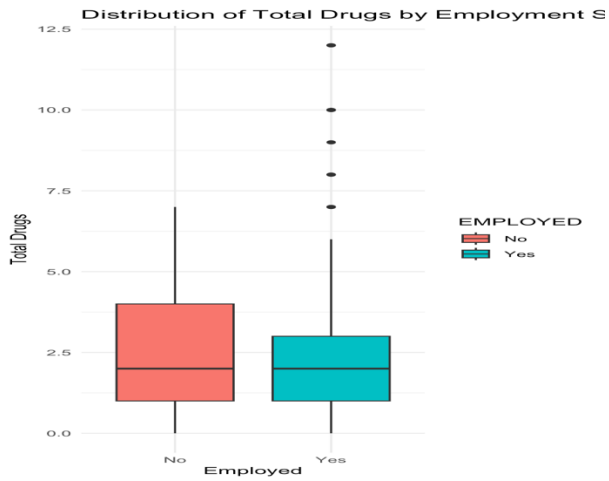
- **Loneliness**: In blue, the Loneliness curve is distinctly narrow and peaked, signifying a high concentration of responses within a more confined score range and fewer instances of extreme scores reported

.

Implications and Considerations for Mental Health Interventions

The density distributions provide valuable indications regarding the prevalence and acuity of these mental health challenges among the sampled population. Notably, the overlap between the distribution's points to the possibility of co-occurrence of symptoms across different conditions. This information is particularly useful for healthcare practitioners and policymakers in identifying intervention priorities. For example, a pronounced peak in moderate-to-severe Depression scores may call for specialized support in this area. Consequently, the analysis facilitates strategic planning and allocation of mental health resources and services to address the specific needs identified within the community.

Distribution of Drug Use Relative to Employment Status



The distribution of total drug use in relation to employment status is effectively visualized through a boxplot. This visualization categorizes individuals into two distinct groups: those who are employed and those who are not. The 'Employed' axis splits the data accordingly, while the 'Total Drugs' axis quantifies drug usage. The red and blue boxes depict the interquartile ranges for the unemployed and employed groups, respectively, and offer a comparative view of the data's spread. Notably, the median — indicated by the line within each box — suggests that the median drug use for the unemployed is higher than for their employed counterparts.

Insights on Drug Use Patterns

The insights garnered from the boxplot are telling: there is a wider variability in drug use among those who are unemployed, as evidenced by the taller box representing this group. Furthermore, the occurrence of outliers, particularly in the unemployed group, points to instances of substantially higher drug use. These outliers are visually represented by the individual points beyond the 'whiskers' of the boxplot and may be indicative of extreme cases that could require more focused investigation or intervention. The data presented in this boxplot underscores the relationship between employment status and drug use, which may prove invaluable for crafting targeted social and health policies.

Limitations of the Analysis

Data-Related Limitations

Sample Size and Generalizability: The conclusions drawn from this analysis are based on the dataset provided. Given its limited size, the results may not be applicable to the wider population. Larger samples would be necessary to ensure the generalizability of the findings.

Representation: The dataset may not accurately represent the diverse demographic variables of the broader population, which could lead to skewed outcomes and potentially unreliable conclusions.

Data Quality: Potential errors, missing values, and biases within the dataset could adversely affect the predictive accuracy of the models. Measures to clean

and preprocess the data were undertaken; however, some issues may persist.

Methodological Limitations

Assumptions: Both LASSO and Random Forest methods carry intrinsic assumptions. For instance, LASSO assumes linear relationships between predictors and response, which might not always exist in complex datasets.
Bias in Feature Selection: The penalization in LASSO regression could lead to the exclusion of significant variables, thus potentially overlooking relevant factors that contribute to the model's response.
Risk of Overfitting: Random Forest models are known for their robustness against overfitting, yet they are not immune. Without careful hyperparameter tuning, these models can overfit to the training data, compromising their performance on unseen data.

Analytical Limitations

Model Interpretability: The simplicity of LASSO allows for greater interpretability, whereas Random Forest models, with their inherent complexity, make it difficult to ascertain clear and direct relationships between variables and outcomes.
Causality: The predictive nature of the models used does not permit the establishment of causation. The significant variables identified serve as indicators of association and require further causal investigation.
Computational Limitations
Resource Intensity: The Random Forest algorithm can be computationally demanding, more so with an increase in the number of trees and the dataset's size, which might limit its application in larger-scale studies.
Scalability: The present code may lack the scalability required to efficiently process larger datasets or more complex modeling scenarios without significant modification.

Code-Specific Limitations

Hardcoded Parameters: The code may contain certain parameters that have been predefined and may not be optimal for other datasets, thus requiring manual adjustments.
Automation Deficiency: There is an absence of automated processes for critical tasks such as hyperparameter tuning and cross-validation, which are vital for the robustness and reliability of the models.

Reporting Limitations

Visualization and Interpretation: The visual aids provided are beneficial but may not fully encapsulate the data's nuances or the model's predictive capabilities. Furthermore, interpretation of the outputs is subject to the analyst's discretion, potentially introducing bias into the reported findings.

Conclusions from the Analysis Using R Code:

The code-based analysis reveals that while both LASSO and Random Forest pinpoint similar influential predictors, their methodologies provide unique interpretations. LASSO is advantageous for its simplicity and ease of interpretation, making it ideal where clear model comprehension is needed. Conversely, Random Forest offers a thorough analysis by considering an extensive range of factors and recognizing intricate patterns, which is particularly useful for datasets anticipated to involve complex variable interactions.
The selection between LASSO and tree-based methods should be informed by the specific objectives of the study, particularly balancing the need for straightforward model interpretation against the requirement to understand complex variable relationships.
For future work on the study of substance use patterns in young adults using regularization and tree-based methods, consider the following potential enhancements:

1. Expansion of Data Sources:

- **Description**: Increase the dataset's scope to encompass a broader range of demographic and psychographic factors, and incorporate longitudinal data to track changes over time and evaluate causality.
- **Objective**: To broaden the data's scope and strengthen the reliability of the predictive models by introducing time-series analysis and capturing a wider array of socio-economic factors.

2. Exploration of Advanced Algorithms:

- **Description**: Investigate the use of sophisticated machine learning algorithms such as deep learning networks and combined model approaches to improve prediction accuracy.
- **Objective**: To uncover more intricate non-linear patterns and interactions that may not be captured by current models and evaluate these advanced models in practical scenarios.

3. Development of a Real-Time Predictive System:

- **Description**: Create a system for real-time prediction that can be used in medical settings or through mobile health apps to identify individuals at risk based on real-time data.
- **Objective**: To offer immediate and customized intervention options, aiming to reduce initial drug use and its escalation among young adults.

4. Cross-Cultural Validation Studies:

- **Description**: Extend validation studies to include diverse populations and regions to confirm the models' applicability on a global scale.
- **Objective**: To assess whether the predictive models are effective worldwide and to pinpoint unique or shared predictors across different cultures.

5. Testing of Psychosocial Interventions:

- **Description**: Examine the effects of specific interventions, such as cognitive behavioral therapy and peer support, on the predictors of drug use identified.
- **Objective**: To assess and refine intervention strategies based on predictive data, enhancing preventive public health measures.

6. Advancement of Computational Techniques:

- **Description**: Employ advanced computational methods, including hybrid models that merge characteristics of both LASSO and Random Forest, or adaptive boosting to enhance performance.
- **Objective**: To enhance prediction precision and efficiency, particularly for handling complex, voluminous datasets.

7. Focus on Ethics and Privacy:

- **Description**: Establish guidelines for managing ethical and privacy issues related to sensitive data handling in the research.
- **Objective**: To maintain high ethical standards and respect for participant privacy, ensuring trust and participation compliance in the research.

8. Development of Interactive Visualization Tools:

- **Description**: Develop interactive platforms and visualization tools that allow for dynamic exploration of data and results by researchers and practitioners.
- **Objective**: To facilitate the use and understanding of research outcomes, enabling informed decision-making based on dynamic data analysis.

Conclusion

These proposed enhancements aim to extend the findings of the current study, utilizing new technologies and approaches to deepen our understanding of substance use trends. The ultimate goal is to create more effective, adaptable public health strategies that respond to changes in young adult behaviors and societal conditions.

Conclusions

Comparative Analysis of LASSO and Tree-based

Methods from Computational Results

The examination of outcomes from LASSO and Random Forest methods using the supplied R code provides valuable insights into their distinct analytical approaches. Here is an assessment of their similarities and differences based on the analysis:

Similarities:

1. Identification of Critical Predictors: Both methodologies identified the use of Alcohol and Tobacco as critical predictors for total drug consumption. This agreement across methods underscores the significance of these variables in analyzing drug usage patterns.Variables related to illegal drug use were also highlighted as important by both models, confirming their direct correlation with overall drug consumption.
2. Effectiveness in Prediction: Each method successfully pinpointed variables that heavily influence drug use, demonstrating their efficacy in predictive modeling within the domain of substance use.

Differences:

1. Complexity and Clarity of Models:

LASSO Regression: Offers a cleaner, more interpretable model by zeroing out many coefficients, which simplifies understanding by clearly identifying which predictors matter.

Random Forest:This method doesn't simplify the predictor set but rather assesses the importance of each, resulting in a model that accommodates a broader spectrum of variables, including potential. Interactions and non-linear dynamics not captured by LASSO.

2. Handling of Variable Importance and Complexity:

Random Forest: This model excels in recognizing complex interactions and non-linear relationships between predictors and outcomes, reflecting its ability to detect detailed patterns.

LASSO: Tends to focus on linear associations and may not capture complex interdependencies between predictors as effectively, though it excels in highlighting significant, independent predictors.

3. Approach to Regularization:

LASSO: Implements a regularization penalty that significantly adjusts the values of coefficients, reducing many to zero to prevent overfitting and boost generalizability.

Random Forest: Utilizes a bagging approach rather than regularization to diminish variance and bolster the model's resilience against overfitting.

References

1. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32. This seminal paper describes the Random Forest algorithm, detailing its operation and providing a basis for its use in predictive modeling and variable importance evaluation.
2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York. This book provides a comprehensive introduction to the field of statistical learning, essential for understanding advanced techniques like LASSO and Ridge regression.
3. Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1), 55-67. This article introduces Ridge regression, discussing its benefits in handling multicollinearity and its application in regression scenarios where predictor variables are highly correlated.
4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. This text is valuable for readers looking to apply statistical learning techniques in R, including methodologies like Random Forests, and is

suitable for real-world applications.
5. Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, 4(3), 279-282. This paper
6. discusses the importance of effect size in the context of medical research, encouraging researchers to focus on the magnitude of effects rather than just statistical significance.
7. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267-288. Tibshirani introduces the LASSO (Least Absolute Shrinkage and Selection Operator) technique for regression analysis, explaining the method's ability to enhance prediction accuracy and interpretability.

These references provide a solid foundation for your report, supporting the advanced statistical techniques and models discussed in your research on substance use patterns in young adults.

#Substance Use Patterns in Young Adults: A Data-Driven Investigation Using Regularization Techniques and Tree-Based Methods.(R-Code)

```
library(dplyr)
library(ggplot2)
data <- read.csv("/Users/balajikolusu/Downloads/data_math678.csv")
head(data)
str(data)

# --------------------------------PRE-PROCESSING--------------------------------

# Handling missing values
data_clean <- na.omit(data)

# Filling missing numeric data with the mean of the column
numeric_columns <- sapply(data_clean, is.numeric)
data_clean[numeric_columns] <- lapply(data_clean[numeric_columns], function(x) ifelse(is.na(x), mean(x, na.rm = TRUE), x))

# Encoding categorical variables into dummy variables
categorical_columns <- c("GENDER", "RACE", "US_BORN", "EMPLOYED", "SCHOOL", "THERAPY", "ALCOHOL", "TOBACCO",
                         "ECSTASY", "KETAMINE", "METHAMPHETAMINE", "MARIJUANA", "COCAINE_POWEDER", "GHB",
                         "COCAINE_CRACK", "HEROIN", "MUSHROOM", "LSD", "RX")
data_clean <- data_clean %>%
  mutate(across(all_of(categorical_columns), as.factor)) %>%
  model.matrix(~.-1, data = .) %>%
  as.data.frame()
str(data_clean)
 summary(data)

# --------------------------------METHOD 1: glmnet--------------------------------

library(glmnet)
data_glmnet <- data_clean
# Extracting the response variable
response <- data_glmnet$TOTALDRUGS
# Remove the response variable from the predictor dataset
data_glmnet$TOTALDRUGS <- NULL

# Convert to matrix format for glmnet
predictors <- as.matrix(data_glmnet)

# Setting up the glmnet model for LASSO regression (alpha = 1)
set.seed(123)
cv_lasso_model <- cv.glmnet(predictors, response, alpha = 1) # alpha = 1 for LASSO
cv_ridge_model <- cv.glmnet(predictors, response, alpha = 0) # alpha = 0 for Ridge

# Plotting the cross-validation results to see the performance
plot(cv_lasso_model)
plot(cv_ridge_model)

# Identify the lambda that gives the minimum mean cross-validated error
best_lambda_lasso <- cv_lasso_model$lambda.min
best_lambda_ridge <- cv_ridge_model$lambda.min
best_model_lasso <- glmnet(predictors, response, alpha = 1, lambda = best_lambda_lasso)
best_model_ridge <- glmnet(predictors, response, alpha = 0, lambda = best_lambda_ridge)

# Print the coefficients of the model
print(coef(best_model_lasso))

# 30 x 1 sparse Matrix of class "dgCMatrix"
#                              s0
# (Intercept)               0.1055241
# X                         .
# AGE                       .
# GENDERfemale              .
# GENDERmale                .
# RACEBLACK                 .
#RACEMIXED RACE            .
#RACEWHITE                 .
# US_BORNYes                .
# EMPLOYEDYes               .
# SCHOOLYes                 .
# THERAPYYes                .
# ILLEGALTOTALILLEGALDRUGS1 0.9821027
# ILLEGTOTALILLEGALDRUGS2   .
# ALCOHOLYes                0.9348273
# TOBACCOYes                0.9061479
# ECSTASYYes                .
# KETAMINEYes               .
# METHAMPHETAMINEYes        .
# MARIJUANAYes              .
# COCAINE_POWEDERYes        .
# GHBYes                    .
# COCAINE_CRACKYes          .
# HEROINYes                 .
# MUSHROOMYes               .
# LSDYes                    .
# RXYes                     .
# DEPRESSION                .
# ANXIETY                   .
# LONELINESS                .

print(coef(best_model_ridge))

# 30 x 1 sparse Matrix of class "dgCMatrix"
# s0
# (Intercept)               -0.2259396041
```

```
# X                         -0.0006429291
# AGE                        0.0006294003
# GENDERfemale              -0.0280468813
# GENDERmale                 0.0282115451
# RACEBLACK                  0.0093471692
# RACEMIXED.RACE             0.0028085434
# RACEWHITE                  0.0006104133
# US_BORNYes                 0.0293912099
# EMPLOYEDYes                0.0081754342
# SCHOOLYes                 -0.0163418927
# THERAPYYes                 0.0192084442
# ILLEGALTOTALILLEGALDRUGS1  0.1775974092
# ILLEGTOTALILLEGALDRUGS2    0.1641478734
# ALCOHOLYes                 0.8908836661
# TOBACCOYes                 0.8732955443
# ECSTASYYes                 0.5849378692
# KETAMINEYes                0.6319159302
# METHAMPHETAMINEYes         0.6002392228
# MARIJUANAYes               0.7387923877
# COCAINE_POWEDERYes         0.6027194123
# GHBYes                     0.5575379513
# COCAINE_CRACKYes           0.6344908943
# HEROINYes                  0.6885167850
# MUSHROOMYes                0.5893154889
# LSDYes                     0.5916894442
# RXYes                      0.5786050026
# DEPRESSION                 0.0042086983
# ANXIETY                    0.0146360467
# LONELINESS                -0.0159347180

# view non-zero coefficients
print(coef(best_model_lasso)[coef(best_model_lasso) != 0])
# [1] 0.1055241 0.9821027 0.9348273 0.9061479
print(coef(best_model_ridge)[coef(best_model_ridge) != 0])
# [1] -0.2259396041 -0.0006429291  0.0006294003 -0.0280468813  0.0282115451  0.0093471692
# [7]  0.0028085434  0.0006104133  0.0293912099  0.0081754342 -0.0163418927  0.0192084442
# [13]  0.1775974092  0.1641478734  0.8908836661  0.8732955443  0.5849378692  0.6319159302
# [19]  0.6002392228  0.7387923877  0.6027194123  0.5575379513  0.6344908943  0.6885167850
# [25]  0.5893154889  0.5916894442  0.5786050026  0.0042086983  0.0146360467 -0.0159347180

# --------------------------------METHOD 2: vip---------------------------------

library(randomForest)
library(vip)

data_vip <- data_clean

# Clean up column names to avoid spaces and special characters
names(data_vip) <- make.names(names(data_vip))

# Fitting a Random Forest model
rf_model <- randomForest(TOTALDRUGS ~ ., data = data_vip, ntree = 500, importance = TRUE)

# Plotting variable importance
vi <- vip(rf_model, num_features = 10)
plot(vi)

# Importance score generated by randomForest
importance(rf_model)


#                            %IncMSE IncNodePurity
#X                          11.59651542    20.8523250
#AGE                        -0.71719124     3.3610060
#GENDERfemale                3.66131644     0.9329935
#GENDERmale                  2.54530961     1.1088823
#RACEBLACK                  -0.50541057     0.7953879
#RACEMIXED.RACE              0.35642577     0.8547220
#RACEWHITE                  -0.07555131     1.1363432
#US_BORNYes                  4.53464348     1.3514203
#EMPLOYEDYes                 2.17623463     1.4449428
#SCHOOLYes                  -0.06087980     0.6546473
#THERAPYYes                 -0.61975128     0.7009120
#ILLEGALTOTALILLEGALDRUGS1 20.91610665   236.7092252
#ILLEGTOTALILLEGALDRUGS2   19.52724934   264.6365508
#ALCOHOLYes                26.01072677    35.0890983
#TOBACCOYes                34.21158857    15.9608705
#ECSTASYYes                 6.10165509    32.2706349
#KETAMINEYes                0.90283349     2.9457813
#METHAMPHETAMINEYes         2.44582640     5.0058209
#MARIJUANAYes              16.25062777    29.8781866
#COCAINE_POWEDERYes         7.73264301    53.2511810
# GHBYes                    1.41929187     3.0679982
# COCAINE_CRACKYes          1.91026913     8.8919438
# HEROINYes                 2.59288522     7.7921750
# MUSHROOMYes               5.80552857    32.3446140
# SDYes                     3.07784616    15.8681280
# RXYes                     3.56592491     4.2908171
# DEPRESSION               10.18361622    95.1610426
# ANXIETY                  24.39519140   153.9924997
# LONELINESS                5.82760985     8.0290280

# Model summary
print(rf_model)

# Type of random forest: regression
# Number of trees: 500
```

```
# No. of variables tried at each split: 9
#
# Mean of squared residuals: 0.1234435
# % Var explained: 97.5

# Check for and handle missing values
data <- na.omit(data)

# Convert relevant columns to factors if they are not already
data$EMPLOYED <- as.factor(data$EMPLOYED)
data$SCHOOL <- as.factor(data$SCHOOL)

# Visualization 1: Boxplot of Total Drugs by Employment Status
ggplot(data, aes(x=EMPLOYED, y=TOTALDRUGS, fill=EMPLOYED)) +
  geom_boxplot() +
  labs(title="Distribution of Total Drugs by Employment Status", x="Employed", y="Total Drugs") +
  theme_minimal()

# Visualization 2: Boxplot of Total Drugs by School Attendance
ggplot(data, aes(x=SCHOOL, y=TOTALDRUGS, fill=SCHOOL)) +
  geom_boxplot() +
  labs(title="Distribution of Total Drugs by School Attendance", x="Attends School", y="Total Drugs") +
  theme_minimal()

# Visualization 3: Density Plots for Depression, Anxiety, and Loneliness
data %>%
  select(DEPRESSION, ANXIETY, LONELINESS) %>%
  pivot_longer(cols = everything(), names_to = "Mental_Health_Variable", values_to = "Score") %>%
  ggplot(aes(x=Score, fill=Mental_Health_Variable)) +
  geom_density(alpha=0.7) +
  labs(title="Density Plots for Depression, Anxiety, and Loneliness", x="Score", y="Density") +
  theme_minimal()


# ------------------------------------------------------------------------------
```