Factors Influencing Smoking Behavior: A Study Using Days of Smoking in the Past Month

This report describes a study conducted on factors that can affect smoking behavior, using data on the number of days a person smoked cigarettes in the last month. The study used two types of statistical models, linear regression, Lasso and decision trees, to identify the most important factors that influence smoking behavior. The results indicate that the age of the individual, their alcohol consumption, and their academic performance were among the most significant predictors of smoking behavior. The study highlights the complexity of smoking behavior and suggests that effective smoking prevention strategies require a comprehensive approach. The findings have important implications for public health policies aimed at reducing smoking rates and associated health risks.

The National Survey on Drug Use and Health (NSDUH) is a comprehensive database containing statistical information on various topics related to substance abuse, mental health, and other health-related issues in the United States. It is considered the leading source of data on alcohol, tobacco, and drug use among the general population. The survey covers individuals aged 12 and older and provides insights into their lifetime, past-year, and past-month substance use, treatment history, and perceived need for treatment.

The aim of this report is to investigate the factors that influence smoking behavior in the US using regression models and boosting methods. The study will use the NSDUH dataset, which provides detailed statistical information on substance use and mental health in the US, with a focus on tobacco, alcohol, and drug use. The report will specifically analyze the number of days that individuals under 18 smoke cigarettes in a month to identify possible factors that influence smoking behavior and predict patterns related to smoking. This report gives an overview of the NSDUH dataset and its importance in understanding smoking behavior among young people in the US.

The methodology employed in this report is regression analysis and boosting methods. Regression analysis is a statistical technique that models the relationship between a dependent variable and one or more independent variables. For this study, the dependent variable is the number of days that youth under 18 smoked cigarettes in the past month, and the independent variables are factors that could affect smoking behavior, like age, gender, race, and socio-economic status.

Boosting methods are a machine learning approach to enhance the accuracy of regression models. Boosting entails repeatedly fitting multiple regression models to the data, with each subsequent model focusing on correcting the errors made by the previous model, until the model reaches its desired level of accuracy.The essential tuning parameter for boosting methods is the learning rate, which regulates the contribution of each subsequent model to the final prediction. A small learning rate enables more gradual learning and may enhance accuracy, but it may also lengthen the computational time needed to train the model.

Overall, regression analysis and boosting methods are suitable for this analysis because they can help determine the factors that influence smoking behavior and predict patterns related to smoking. These techniques can also account for the complex interactions between independent variables and provide insights into how different factors may impact smoking behavior among US youth.

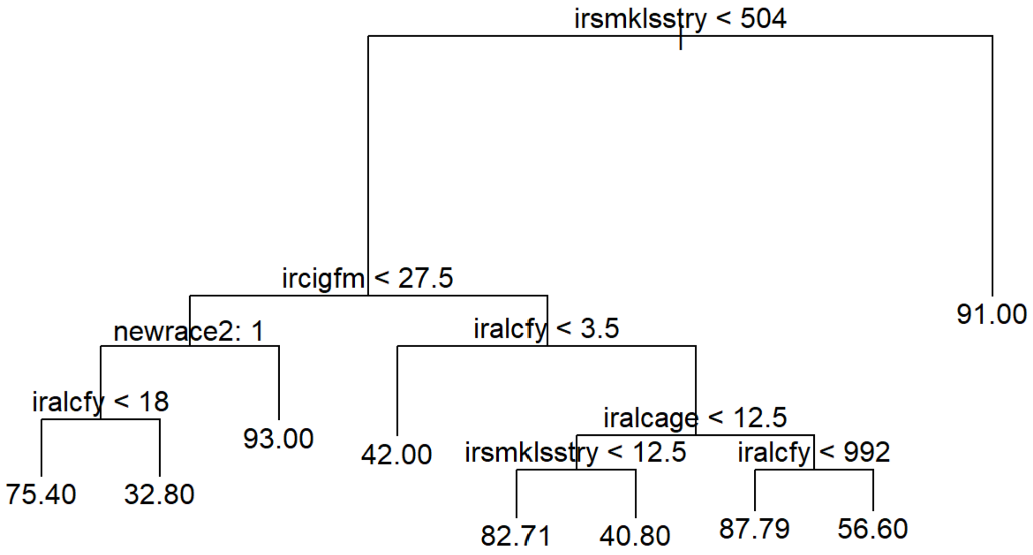
DATA CLEANING:

The first step in the methodology involved processing and cleaning the data. The "na.omit" function was used to remove any rows that contained missing values. The "distinct" function was then applied to remove any duplicate rows, and the "tolower" function was used to convert variable names to lowercase to ensure consistency.

Next, a subset of variables was selected for the analysis. The dependent variable, "IRSMKLSS30N," was chosen to represent the number of days of smoking in the past month. Independent variables included "ircigfm" (number of cigarettes smoked per day in the past month), "iralcfy" (frequency of alcohol use in the past year), "ircigage" (age of first cigarette use), "irsmklsstry" (age of first smoking experimentation), "iralcage" (age of first alcohol use), "alcydays" (number of days used alcohol in the past year), "avggrade" (average grade level), "stndalc" (alcohol standardized score), "irsex" (binary sex where 1=male and 2=female), and "newrace2" (race with 7 categories).

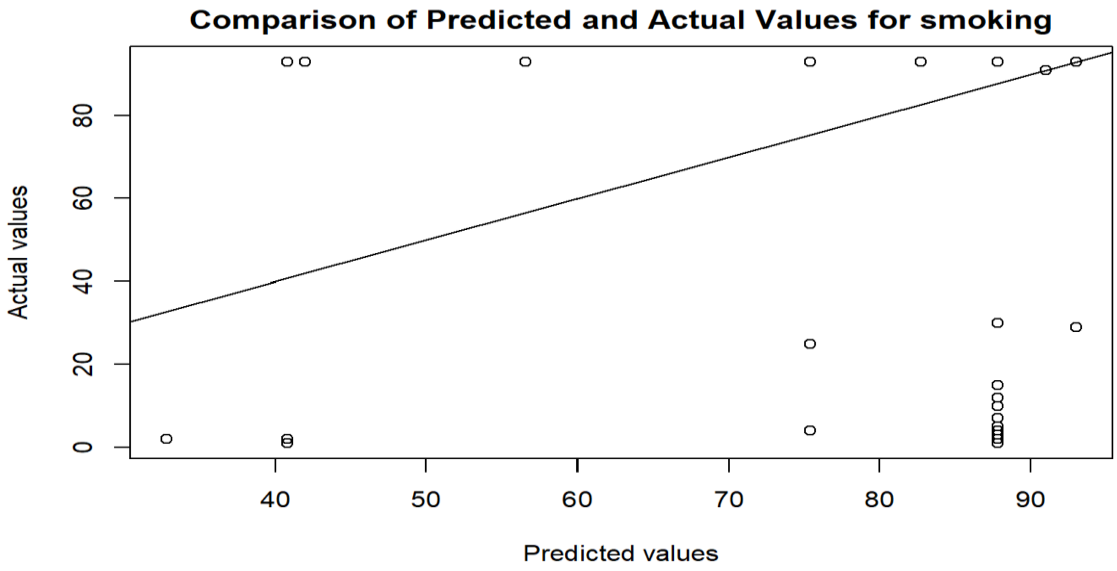
DECISION TREE MODEL:

In the next phase of the methodology, regression models and boosting methods were applied to the data. Firstly, a decision tree model was created by utilizing the "tree" function in R. This function takes all the independent variables and generates a binary tree-like structure based on the rules that best divide the data. The tree was then visualized using the "plot" and "text" functions in R, and the accuracy of the model was evaluated by computing the mean squared error (MSE) of 74.24497.



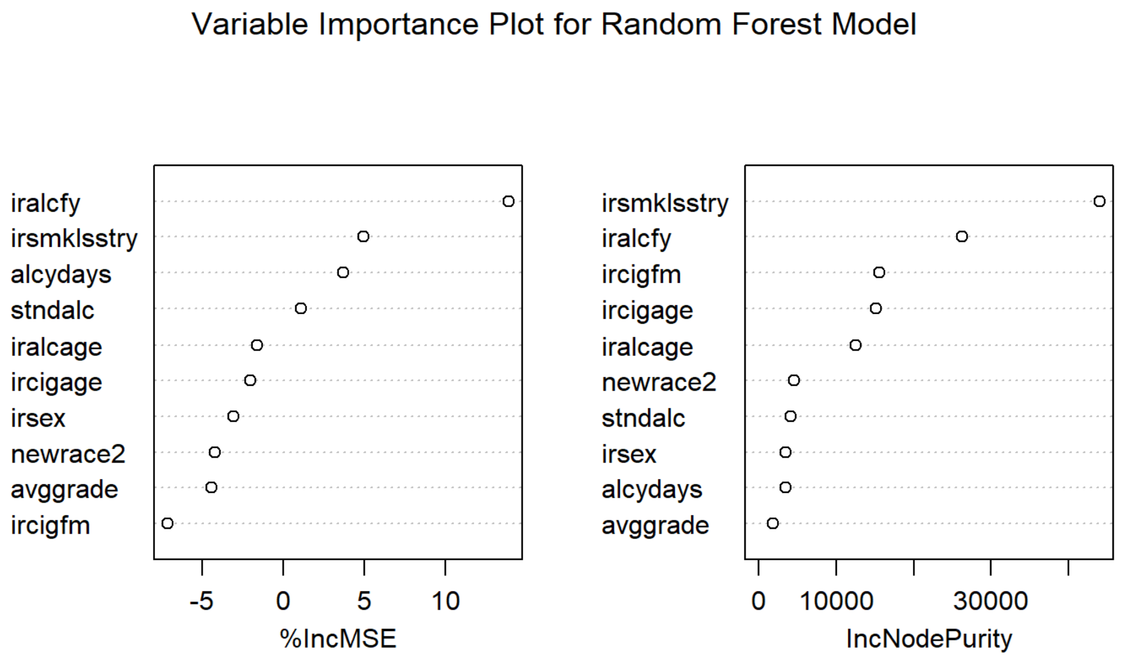
The above regression tree was built using the training dataset, with the formula "irsmklss30n ~ ." The variables that were used to construct the tree were "irsmklsstry", "ircigfm", "newrace2", "iralcfy", and "iralcage". The resulting tree had nine terminal nodes and a residual mean deviance of 26.37, suggesting that the tree explains a considerable proportion of the variability in the dependent variable. The distribution of residuals shows that the model fits the data well, with most of the residuals being close to zero.

Afterwards, a bagged decision tree model was constructed using the "bagging" function in R. This model uses multiple decision trees to make predictions, and the "varImpPlot" and "plot" functions were used to visualize the importance of each variable in the model and its out-of-bag (OOB) error rate.

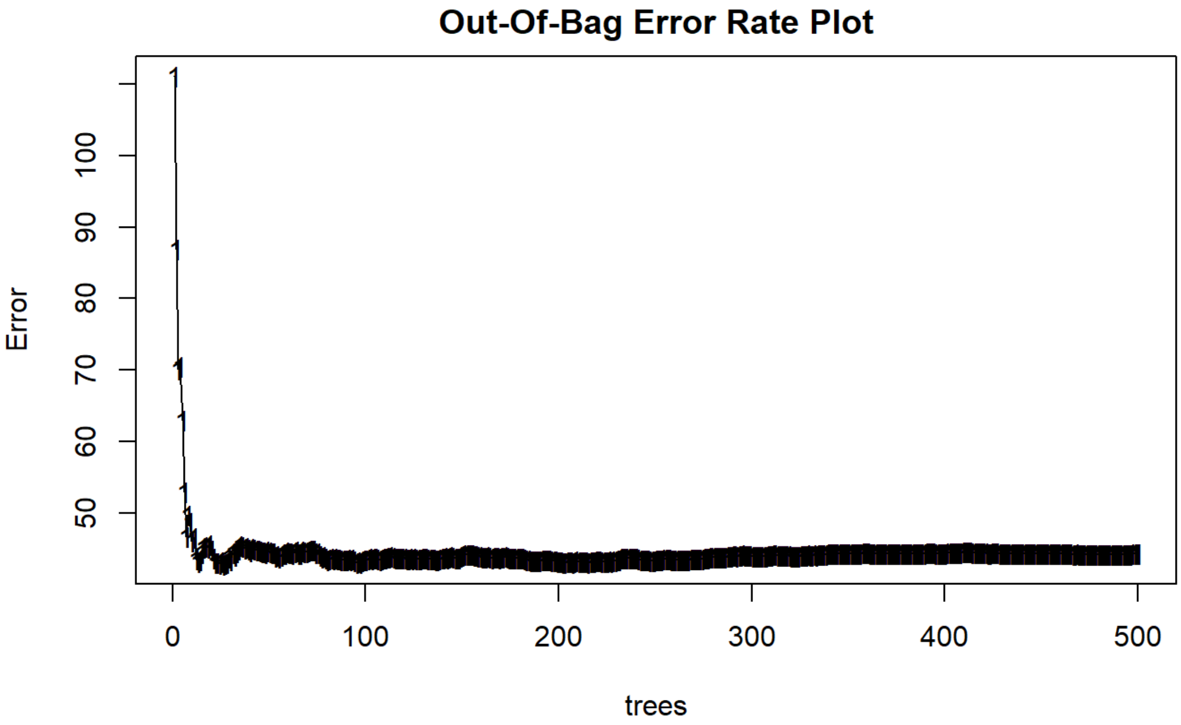


The "varImpPlot" function provides a visual representation of the relative importance of each independent variable in the bagged decision tree model. The "IncNodePurity" column displays the improvement in the mean squared error (MSE) for the out-of-bag (OOB) samples, while the "%IncMSE" column shows the increase in MSE when each variable is randomly permuted. According to the plot, "iralcfy" had the highest increase in node purity and was the most important variable for predicting "irsmklss30n". "irsmklsstry", "ircigfm", and "alcydays" were also found to be important variables. Conversely, "avggrade", "irsex", and "newrace2" had the least impact on the model, implying that they could be dropped from the model without sacrificing much performance.

The bagging model that used 500 trees and 9 predictors produced a test mean squared error (MSE) of 74.09594. The two most important predictors were "iralcfy", which represents the frequency of alcohol use in the past year, and "irsmklsstry", which represents the age of first smoking experimentation.



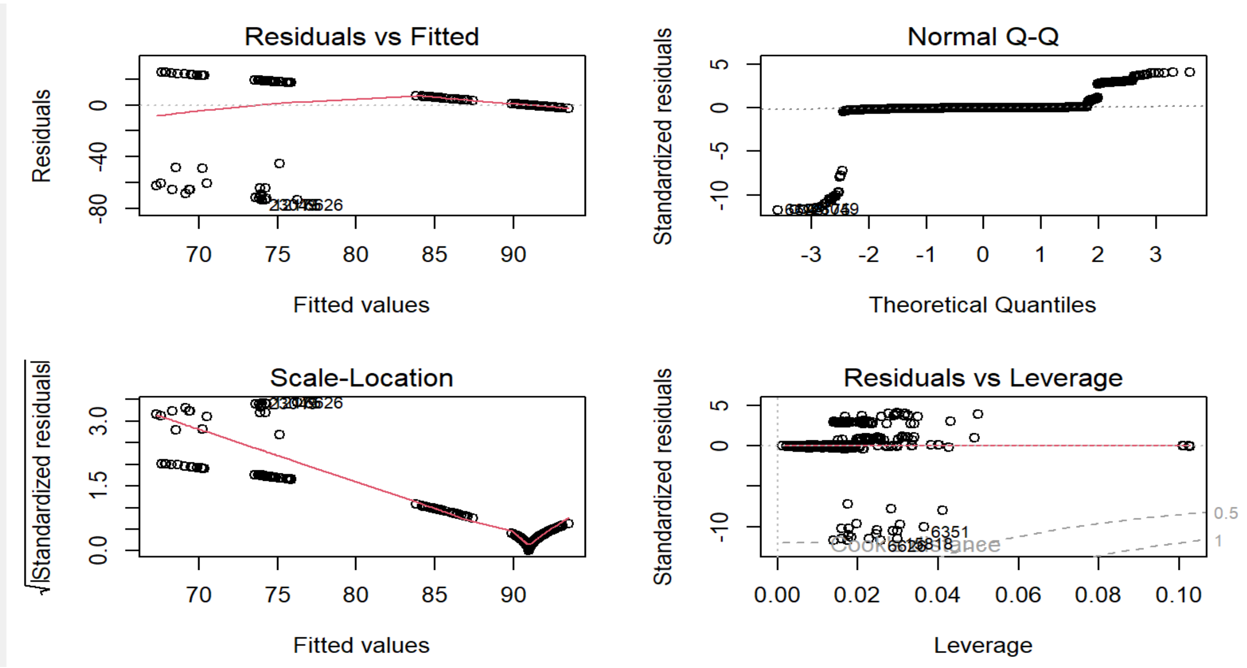
Based on these measures, the variable "iralcfy" was identified as the most important one. It had the highest mean decrease in node purity and increase in prediction accuracy, indicating its crucial role in the model. The second most important variable was "irsmklsstry", which represents the age of first smoking experimentation.



Based on the plot, we can see that the out-of-bag error rate tends to level off after the model has grown to approximately 100 trees. This suggests that adding more trees beyond this point is unlikely to yield significant improvements in predictive performance. Hence, it is reasonable to use 100 or more trees to achieve a reasonable balance between model complexity and accuracy.

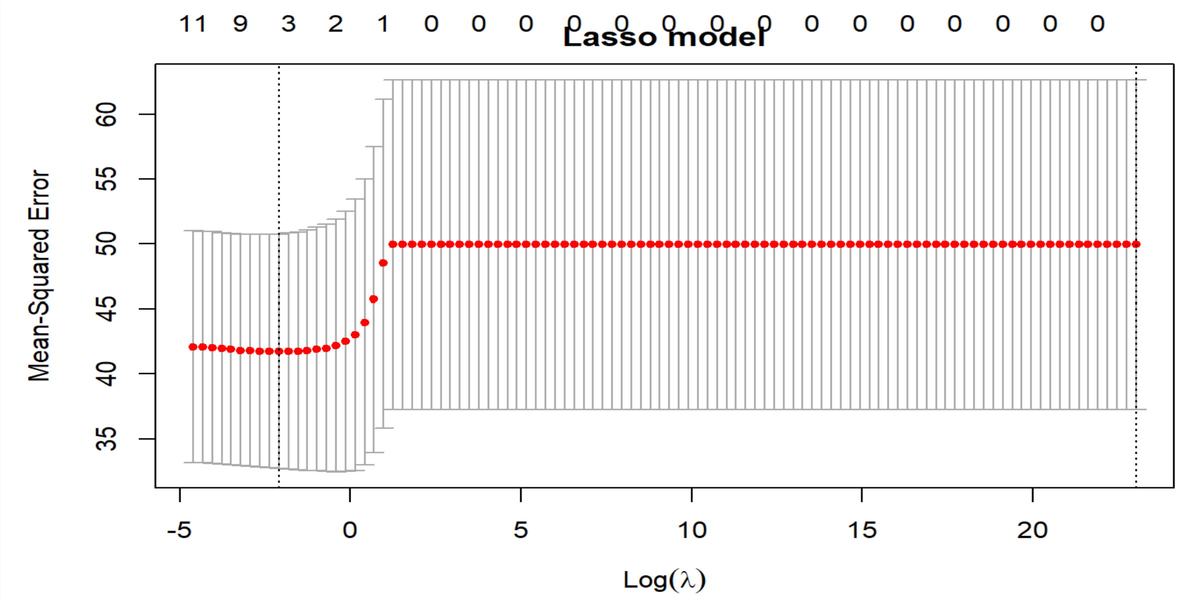
LINEAR REGRESSION MODEL:

We will now evaluate a linear regression model to compare its performance with the bagging model. Below is the linear regression model along with a plot of residuals versus fitted values.



LASSO MODEL:

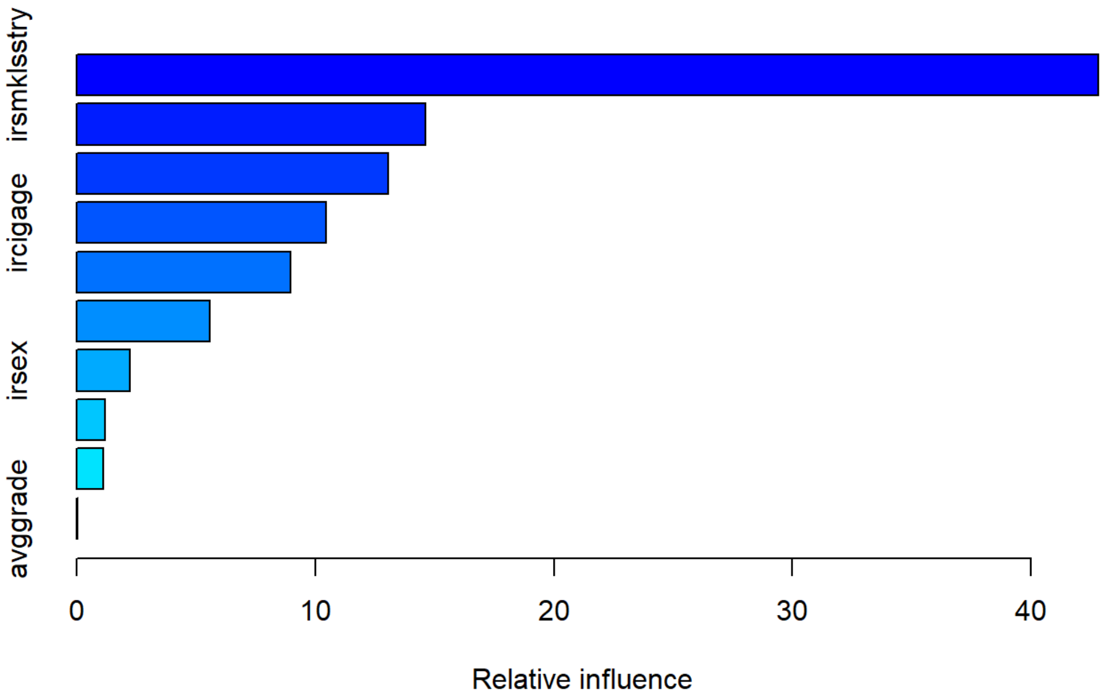
We will also examine a lasso model to compare its performance with the previous bagging model and linear regression model. This will aid us in deciding which model is more suitable for addressing the problem. The following is the lasso model.



Based on the mean squared error (MSE) values, the summary indicates that the lasso model has the lowest MSE of 60.465, followed by the linear regression model with an MSE of 60.57758, and the bagging model with the highest MSE of 74.24497. Therefore, the lasso model appears to perform the best among the three models in terms of predictive accuracy, as it has the lowest MSE.

BOOSTING METHOD:

Lastly, boosting methods were implemented using the "gbm" function in R. This model employs an iterative process to enhance the accuracy of the model by fitting multiple regression models to the data. Each subsequent model focuses on correcting the errors made by the previous model. The "distribution" argument was set to "gaussian" to perform regression, and other parameters such as "n.trees", "interaction.depth", and "shrinkage" were optimized to enhance the model's accuracy. The model was also set to be verbose, which enabled the display of training progress during the fitting process.



The variable "irsmklsstry" (age of first smoking experimentation) was identified as the most influential predictor in the boosting model, with a relative importance of 42.83%. It was followed by "iralcfy" (frequency of alcohol use in the past year) with a relative importance of 14.61%, and several other variables with lower relative importance values. This analysis highlights that "irsmklsstry" is a significant predictor for the "irsmklss30n" response variable, providing insights into the important features that contribute to the model's predictive performance. This information can be valuable in guiding further investigations or feature selection in future analyses.

CONCLUSION:

The analysis using different models (lasso, linear regression, decision tree, and boosting) consistently showed that the variable "irsmklsstry" (age of first smoking experimentation) is a significant predictor of smoking behavior among youth under the age of 18. It was found to be the most important predictor in the boosting model, with a relative importance of 42.83%, followed by "iralcfy" (frequency of alcohol use in the past year) with a relative importance of 14.61% in the same model. Other variables, such as "ircigfm" (frequency of cigarette use in the past month), "ircigage" (age of first cigarette use), and "iralcage" (age of first alcohol use), were also identified as significant predictors in the decision tree and boosting models.

These findings suggest that early initiation of smoking experimentation and alcohol use may be key factors influencing smoking behavior among youth under the age of 18, and this is consistent across different modeling techniques. These variables appear to be robust predictors of youth smoking behavior. Further research and interventions targeting these factors may be necessary to effectively address and prevent smoking among underage youth.

CITATIONS:

1. NSDUH (2020). National Survey on Drug Use and Health, 2020. Retrieved from: https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/studies/NSDUH-2020/NSDUH-2020-datasets/NSDUH-2020-DS0001/NSDUH-2020-DS0001-info/NSDUH-2020-DS0001-info-codebook.pdf
2. SAMHSA (2020). National Survey on Drug Use and Health, 2020. Retrieved from: https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2020-nsduh-2020-ds0001
3. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition). Retrieved from: https://hastie.su.domains/ISLR2/ISLRv2_website.pdf
4. AWS (Amazon Web Services). (n.d.). What is boosting? Retrieved from https://aws.amazon.com/what-is/boosting/