DATA 5322

Practical Homework 1: Decision Trees

Mark Daza

Overview of Youth Drug Use Dataset

- The survey is conducted annually by the Substance Abuse and Mental Health Services Administration to assess behavioral health in the U.S.
- Provides nationally representative data for individuals aged 12+
- Covers tobacco, alcohol, and drug use; mental health; and treatment access
  - Lifetime, past-year, and past-month substance use
  - Age of first use and treatment history
  - Mental health metrics like depression and suicidal ideation
- It includes additional information, such as demographics, income, and factors like family conflict and participation in prevention programs.

Preliminary Cleaning

- The data set has been pre-cleaned to include responses from youth under 18 based on the responses observed in the SCHFELT variable.

(YESCHFLT)
SCHFELT Len : 1 RC-HOW YTH FELT: ABOUT GOING TO SCHOOL IN PST YR

	Freq	Pct
. = Unknown/Aged 18+ (Otherwise)	27393	83.28
1 = Liked A Lot/Kind of Liked (YESCHFLT=1,2)	3981	12.10
2 = Didn't Like Very Much/Hated (YESCHFLT=3,4)	1519	4.62

- During the preliminary cleaning, values that were NAs were excluded, and those who answered the youth experience questions (ages 12-18) were retained.

Ensuring Completeness in Youth Response Data (12-18)

- The strategy previously used to ensure that data included responses from youth (ages 12-18) will be applied to the columns containing youth experience responses. While the previous strategy was appropriate, there are still NA values in this category of responses. To ensure we can generate the appropriate models, we need to remove those missing values.

```
YOSELL2      YOSTOLE2     YOATTAK2    PRPKCIG2    PRMJEVR2     PRMJMO      PRALDLY2    YFLPKCG2    YFLTMRJ2    YFLMJMO
1   :  130   1   :  427   1   : 590   1   :9822   1   :8540   1   :8837   1   :9530   1   :9660   1   :8166   1   :8188
2   :10406   2   :10094   2   :9930   2   : 643   2   :1920   2   :1624   2   : 940   2   : 811   2   :2306   2   :2281
NA's:   25   NA's:   40   NA's:  41   NA's:  96   NA's: 101   NA's: 100   NA's:  91   NA's:  90   NA's:  89   NA's:  92



YFLADLY2    FRDPCIG2    FRDMEVR2    FRDMJMON    FRDADLY2    TALKPROB    PRTALK3     PRBSOLV2    PREVIOL2    PRVDRGO2
1   :9349   1   :9510   1   :8110   1   :8222   1   :9277   1   : 644   1   :5585   1   :1945   1   : 801   1   : 761
2   :1119   2   : 905   2   :2299   2   :2187   2   :1137   2   :9581   2   :4777   2   :8330   2   :9618   2   :9698
NA's:  93   NA's: 146   NA's: 152   NA's: 152   NA's: 147   NA's: 336   NA's: 199   NA's: 286   NA's: 142   NA's: 102
```

Handling ‘Never Used’ and Missing Values in Use Frequency and Age Variables

- Columns recording frequency of use contain codes with the same (or consistent) meaning across all variables.
- Observations coded as 91 or 991 in **frequency-of-use variables**, denoting ‘NEVER USED,’ may be recoded to 0, as all indicate an absence of use and are analytically equivalent.

IRALCFY - ALCOHOL FREQUENCY PAST YEAR - IMPUTATION REVISED

RANGE = 1 - 365

991 = NEVER USED ALCOHOL

993 = DID NOT USE ALCOHOL PAST YEAR

- Values of 991 in **age-of-first-use** variables cannot be recoded to 0, as doing so would misrepresent the respondent’s actual age at first use. These variables will be used as positive indicators of substance use; responses coded as 991 will be excluded where applicable.

IRCIGAGE - CIGARETTE AGE OF FIRST USE - IMPUTATION REVISED

RANGE = 1 - 55

991 = NEVER USED

Goals for Assignment 1

Generate tree models and provide one example for each of the following types of problems:

- Binary classification
  - ❑ MRJFLAG - has or has not used marijuana
  - ✓I converted MRJFLAG responses into factors for most ensemble methods, excluding Boosting.
- Multi-class classification
  - ❑ Alcohol user in the past year - never, seldom, and frequent
  - ✓ Generated multi-class outcome variable based on alcohol use in the past year
- Regression
  - ❑ Number of days per year a person has used alcohol

Theoretical Background: Decision Tree Methods (Updated)

Trees can make predictions for regression or classification.

Decision Trees

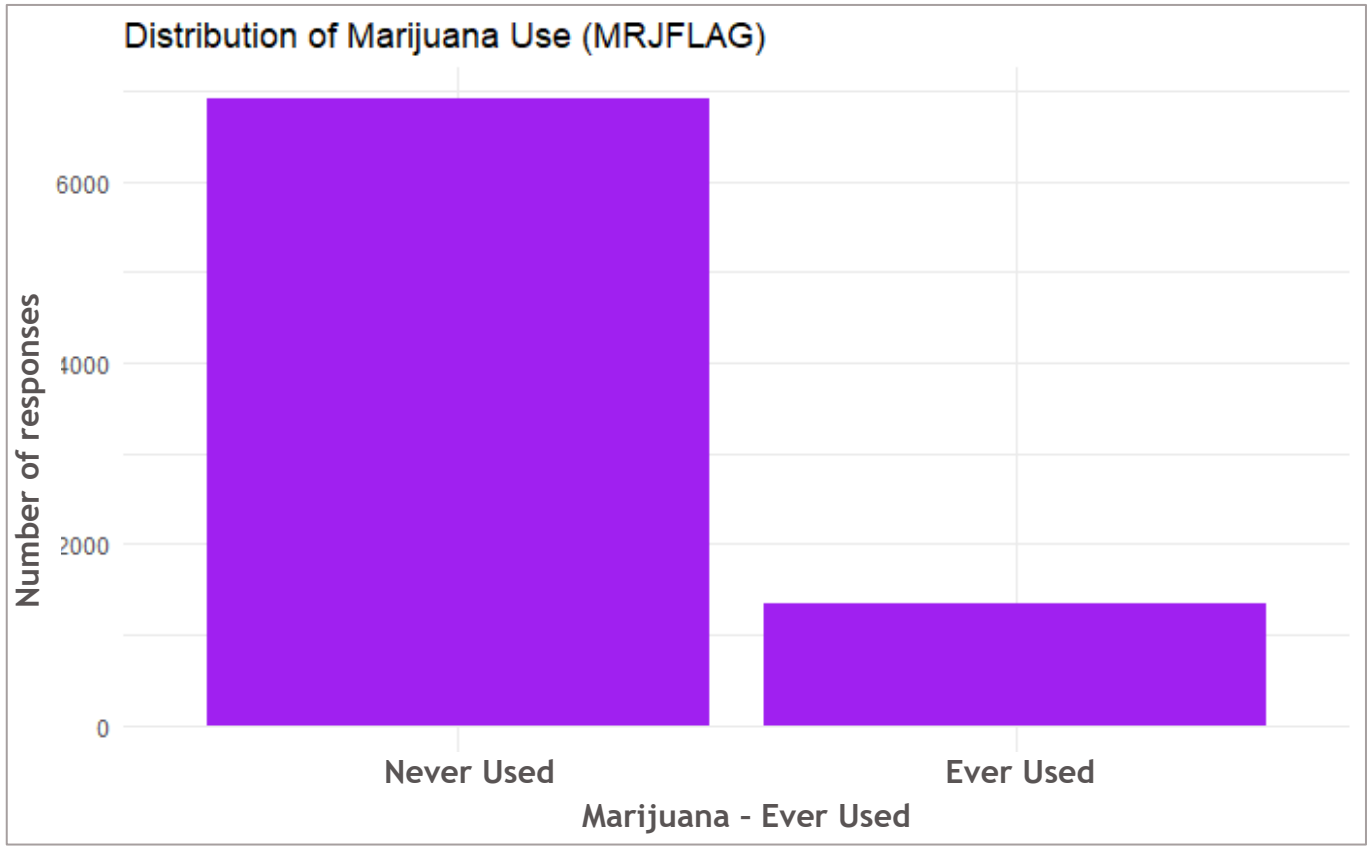
- Simple and interpretable models that mimic human decision-making. However, they can overfit the data!

Optimizing Trees

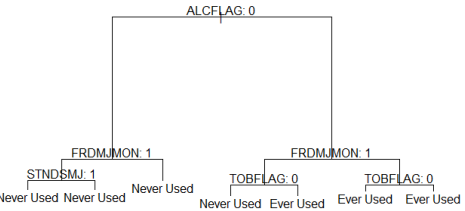
- Pruning
  - Reduces the size of a tree to prevent overfitting and improve generalization.
- Bagging
  - Builds many decision trees on different random samples (with replacement).
  - Final prediction is an average (regression) or majority vote (classification).
- Random Forest
  - A form of bagging where each tree split considers a random subset (m) of predictors.
  - Additional parameters, such as ‘m and the number of trees built, can be optimized.
- Boosting
  - The idea of building many small trees, converting many weak trees into a strong model.
  - Also has tuning parameters like number of trees, shrinkage, learning rate (λ), and depth(d)

Binary Classification - Can We Predict Whether A Youth Has Ever Used Marijuana? (Updated)

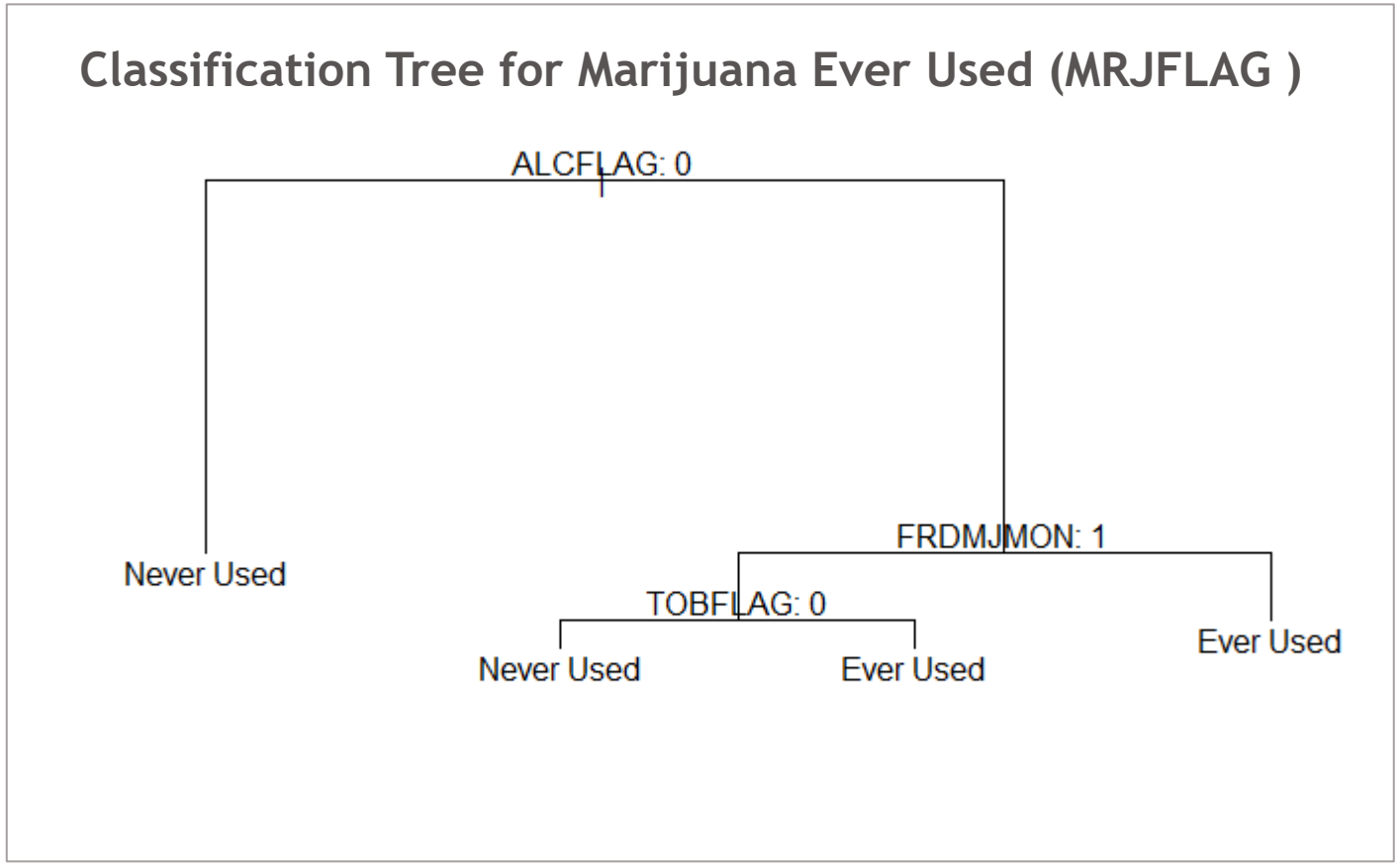
- **Distribution**
  - Majority of youth in the dataset report never using marijuana.
- **Predictors Used**
  - Drug use history, family/peer influences, school performance, demographics, and risk factors (59 predictors)
- **Approach**
  - Trained classification models using decision trees, bagging, random forests, and boosting.



Decision Tree - Initial Model (Updated)



- **Model**
  - Basic classification tree trained with 50% of youth responses.
- **Top Predictor**
  - The most important variable was Alcohol Ever Used (ALCFLAG)
- **Interpretation**
  - Among those who have used alcohol, having friends who use marijuana or having used tobacco themselves makes it much more likely that they've used marijuana too.



Accuracy: 0.897

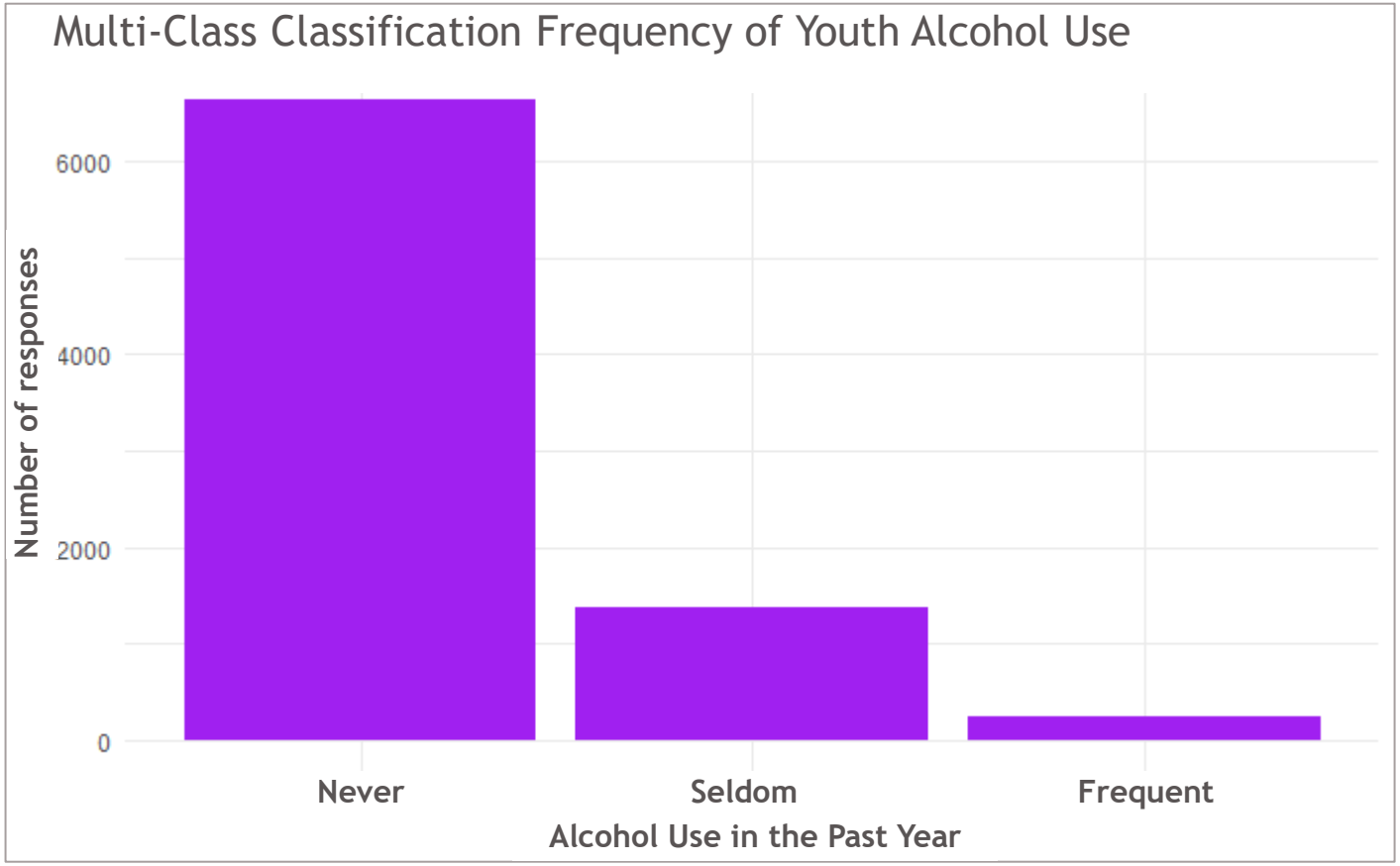
Decision Tree – Improving Accuracy with Ensemble Trees (Updated)

Method	Baggings	Random Forest	Boosting
Accuracy	0.8923	0.8977	0.8904
Classification Error (Ever Used)	0.38	0.37	0.424
Top 5 Predictors	1. Alcohol Ever Used 2. How Friends Feel Abt Marijuana 3. Any Tobacco Ever Used 4. Students in Yth Grade Use Marijuana 5. Race/Hispanicity	1. Alcohol Ever Used 2. Any Tobacco Ever Used 3. How Friends Feel Abt Marijuana 4. Students in Yth Grade Use Marijuana 5. Race/Hispanicity	1. Alcohol Ever Used 2. Any Tobacco Ever Used 3. Students in Yth Grade Use Marijuana 4. Race/Hispanicity 5. How Friends Feel Abt Marijuana

- Boosting had the highest classification error for Ever Used, despite comparable overall accuracy.
- Alcohol Ever Used (ALCFLAG) was consistently the strongest predictor across all models.
- Random Forest had the lowest classification error and highest accuracy.

Multi-Class Classification - Can We Differentiate Between Never, Seldom, And Frequent Alcohol Use? (Updated)

- **Distribution**
  - Most youth reported never using alcohol in the past year
- **Predictors Used**
  - Included drug use history, school performance, peer/family influences, demographics, and other behavioral factors (59 predictors total)
- **Approach**
  - We trained multi-class classification models using bagging and random forests.

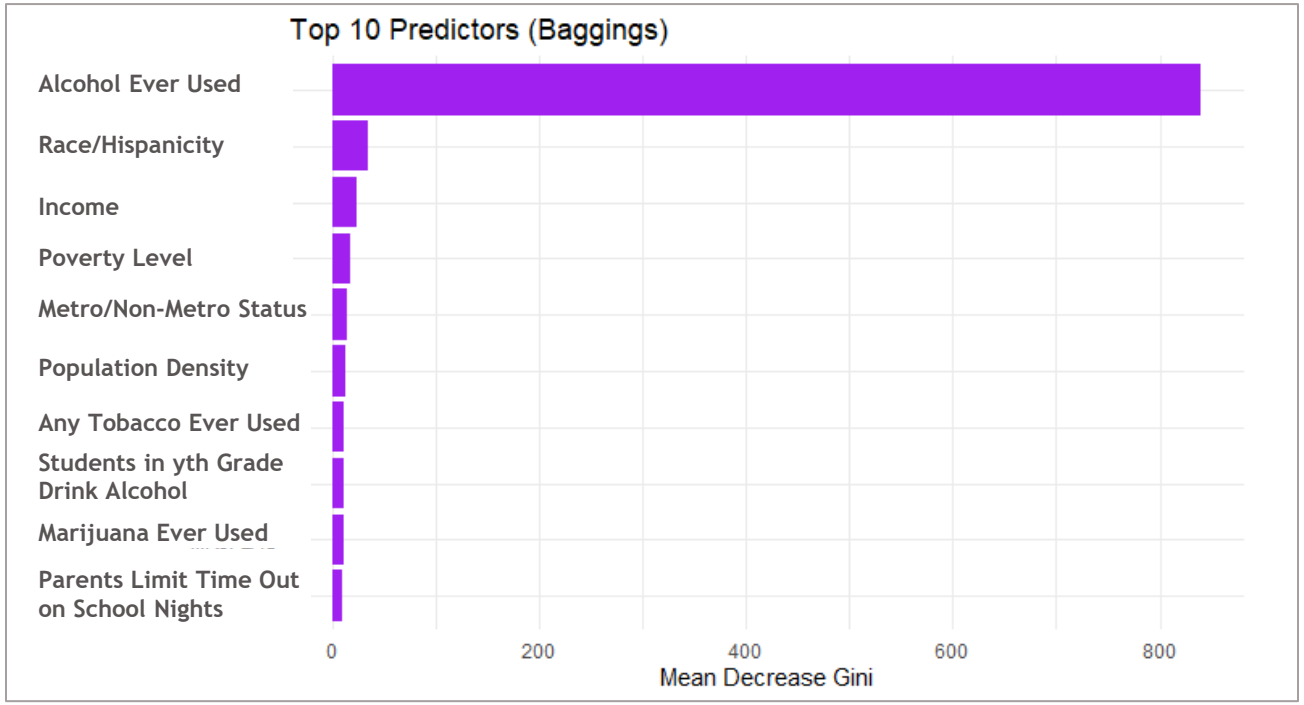


Bagging Model - Predicting Alcohol Use Frequency (Updated)

- **Performance**
  - Accuracy: 91.2%
  - Confusion matrix shows strong prediction for “Never” class, moderate for others
- **Interpretation**
  - The model performed best at identifying non-users. This reflects imbalanced distribution.
  - Alcohol Ever Used (ALCFLAG) greatly improves the model’s performance. Therefore, it was kept as a predictor.

Confusion Matrix

	Frequent	Never	Seldom	Class Error
Frequent	10	7	100	0.915
Never	2	3147	184	0.056
Seldom	3	27	646	0.044

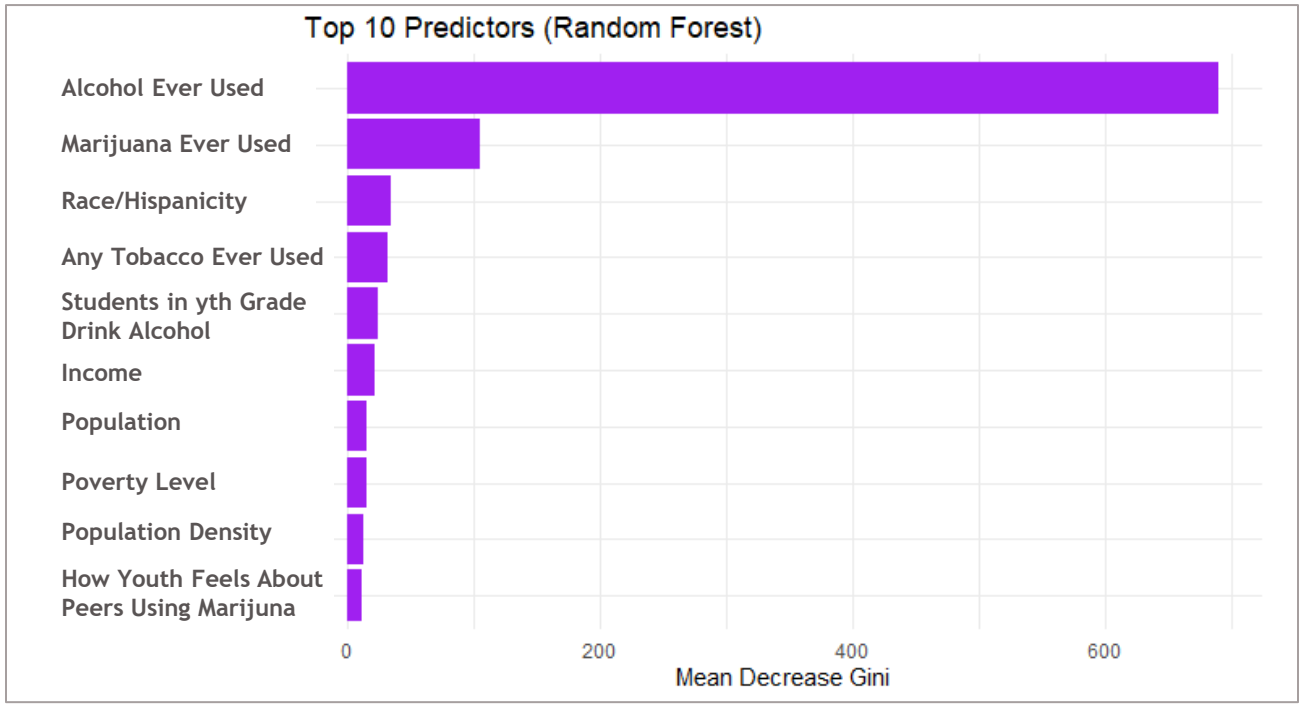


Random Forest Model - Predicting Alcohol Use Frequency (Updated)

- **Performance**
  - Accuracy: 91.08%
  - Again, the confusion matrix shows strong prediction for "Never" class, moderate for others
- **Interpretation**
  - ALCFLAG (Alcohol Ever Used) is the most important predictor for alcohol use across multiple frequency categories.
  - Random Forest shows similar strengths and weaknesses as Bagging, especially favoring the majority class.

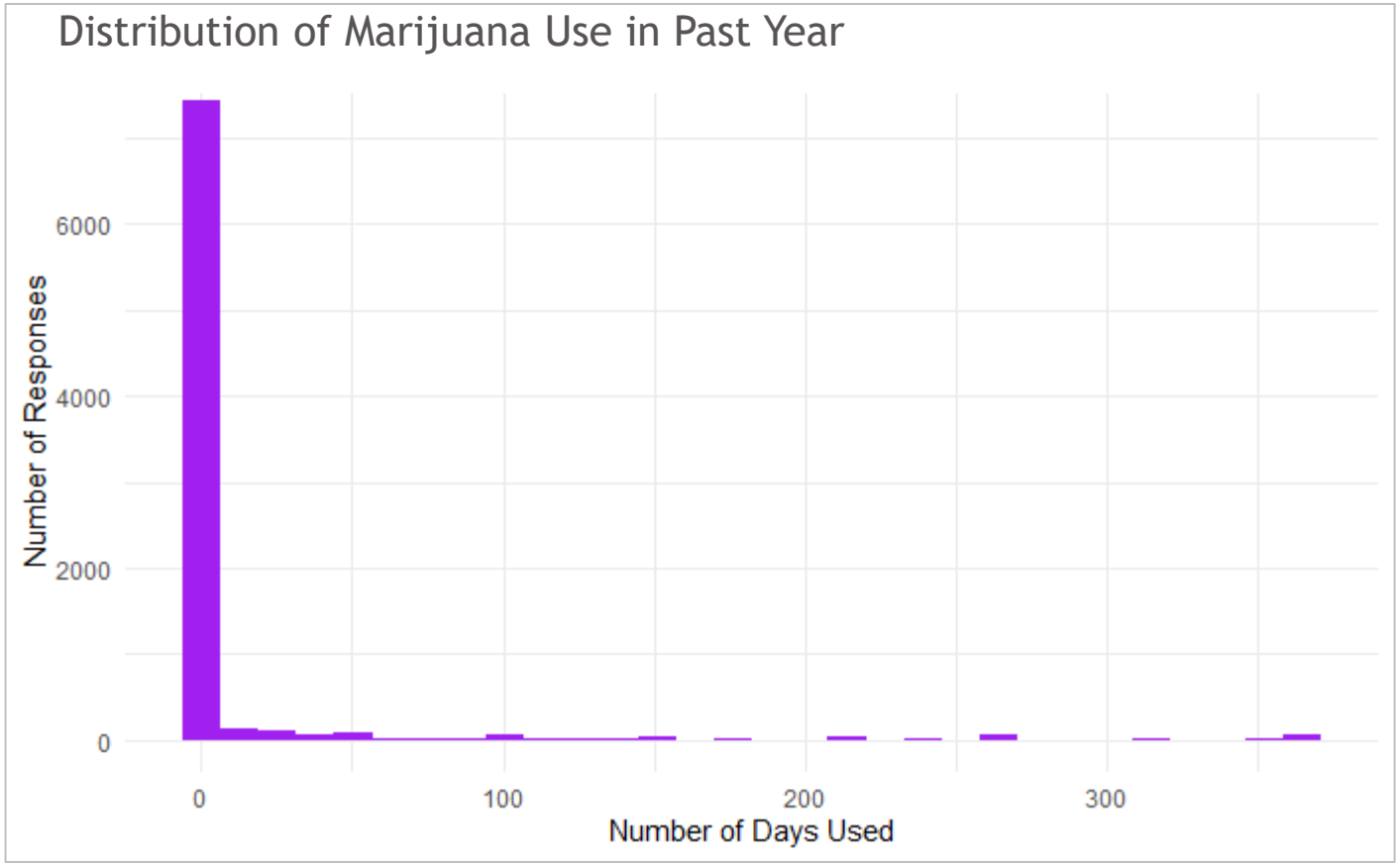
Confusion Matrix

	Frequent	Never	Seldom	Class Error
Frequent	8	4	105	0.932
Never	0	3146	187	0.056
Seldom	4	26	646	0.044



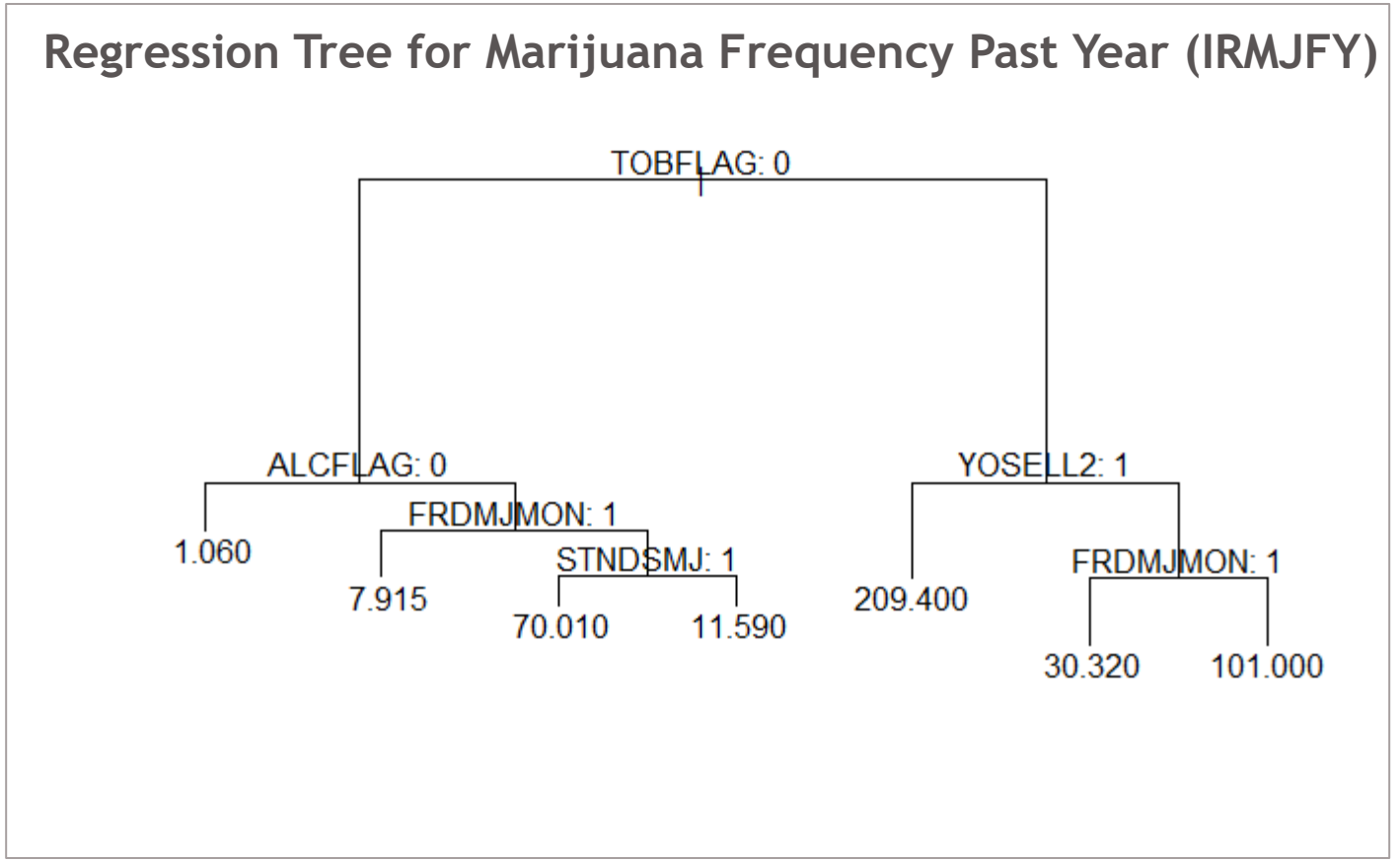
Regression – Can We Predict How Often Youth Use Marijuana?

- **Distribution**
  - Most youth reported never using marijuana in the past year
- **Predictors Used**
  - Included drug use history, school performance, peer/family influences, demographics, and other behavioral factors (59 predictors total)
- **Approach**
  - We trained regression models using decision trees, bagging, random forests, and boosting



Decision Tree - Initial Model (Updated)

- **Model**
  - Basic classification tree trained with 50% of youth responses.
- **Top Predictor**
  - The most important variable was Tabacco Ever Used (TOBFLAG)
- **Interpretation**
  - Youth who reported using tobacco had significantly higher predicted marijuana use



Test MSE: 1986.22

Decision Tree – Improving Accuracy with Ensemble Trees (Updated)

Method	Baggings	Random Forest	Boosting
Test MSE	1953.839	1750.569	1722.246
Top 5 Predictors	1. Any Tobacco Ever Used 2. Youth Sold Illegal Drugs 3. Race/Hispanicity 4. How Friends Feel Abt Marijuana 5. Alcohol Ever Used	1. Any Tobacco Ever Used 2. Alcohol Ever Used 3. Youth Sold Illegal Drugs 4. Race/Hispanicity 5. How Friends Feel Abt Marijuana	1. Any Tobacco Ever Used 2. Alcohol Ever Used 3. Youth Sold Illegal Drugs 4. How Youth Think Parents Feel Abt Marijuana 5. How Friends Feel Abt Marijuana

- Boosting had the lowest MSE, suggesting slightly better predictive performance.
- Tobacco Ever Used (TOBFLAG) was the top predictor across all models.
- Peer factors like Youth Sold Illegal Drugs and How Friends Feel Abt Marijuana also consistently ranked high, suggesting peer behavior influences the number of days of marijuana use in youth.

Discussion (Updated)

Binary Classification (Marijuana Ever Used)

- Top predictor: Alcohol Ever Used (ALCFLAG) was the strongest predictor across all models.
- Best performing model: Random Forest had the highest accuracy (89.77%) and lowest classification error (37%).
- Additional improvements to the model, such as tuning of alpha and mtry, could help improve ensemble methods.

Multi-Class Classification (Alcohol Use Frequency)

- Top predictor: Again, Alcohol Ever Used (ALCFLAG) appeared to be the strongest predictor.
- Baggings and Random Forest both had strong prediction for “Never” class.
- Imbalanced class sizes, especially “Frequent”, had high classification errors. Models can be tuned, but the class imbalance will likely continue having an impact on classification errors.

Regression (Number of Days Marijuana Was Used)

- Top predictor: Tobacco Ever Used (TOBFLAG) was the top predictor across all the models.
- Best performing model: Boosting had the lowest test MSE (1722.246), while pruned tree had the highest (1986.22).
- Improvements to models could be made by reducing the number of predictors used.

Conclusion

Drug Use is Interconnected

- In all our models, drug use was highly interconnected. This suggests that there is a co-use in youth data.

Real Use Scenarios

- Our models could be used to identify at-risk youth populations. However, further optimization is recommended to identify underlying predictors that are not related to drug use.

Limitations

- Class imbalance was the greatest source of error in our models. Focusing strictly on the youth population currently using drugs might uncover additional predictors.

Thank you

References

[1]Ariana Mendible, Lecture Notes for DATA5322, Seatle University, Spring 2025.

- **Decision Trees.pdf**
- **Decision Tree Ensembles.pdf**

[2]Ariana Mendible, Labs for DATA5322, Seatle University, Spring 2025.

- **Lab Ch8-1.Rmd**
- **Lab Ch8-2.Rmd**

[3] Kassambara, A. ggplot Colors: Best Tricks You Will Love. Datanovia, 2018. Available at: https://www.datanovia.com/en/blog/ggplot-colors-best-tricks-you-will-love/. Accessed April 14, 2025

[4] Ridgeway, G. gbm: Generalized Boosted Regression Models. R package version 2.1.8. Available at: https://cran.r-project.org/web/packages/gbm/gbm.pdf. Accessed April 14, 2025.

[5] Liaw, A., & Wiener, M. randomForest: Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.7-1. Available at: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf. Accessed April 14, 2025.