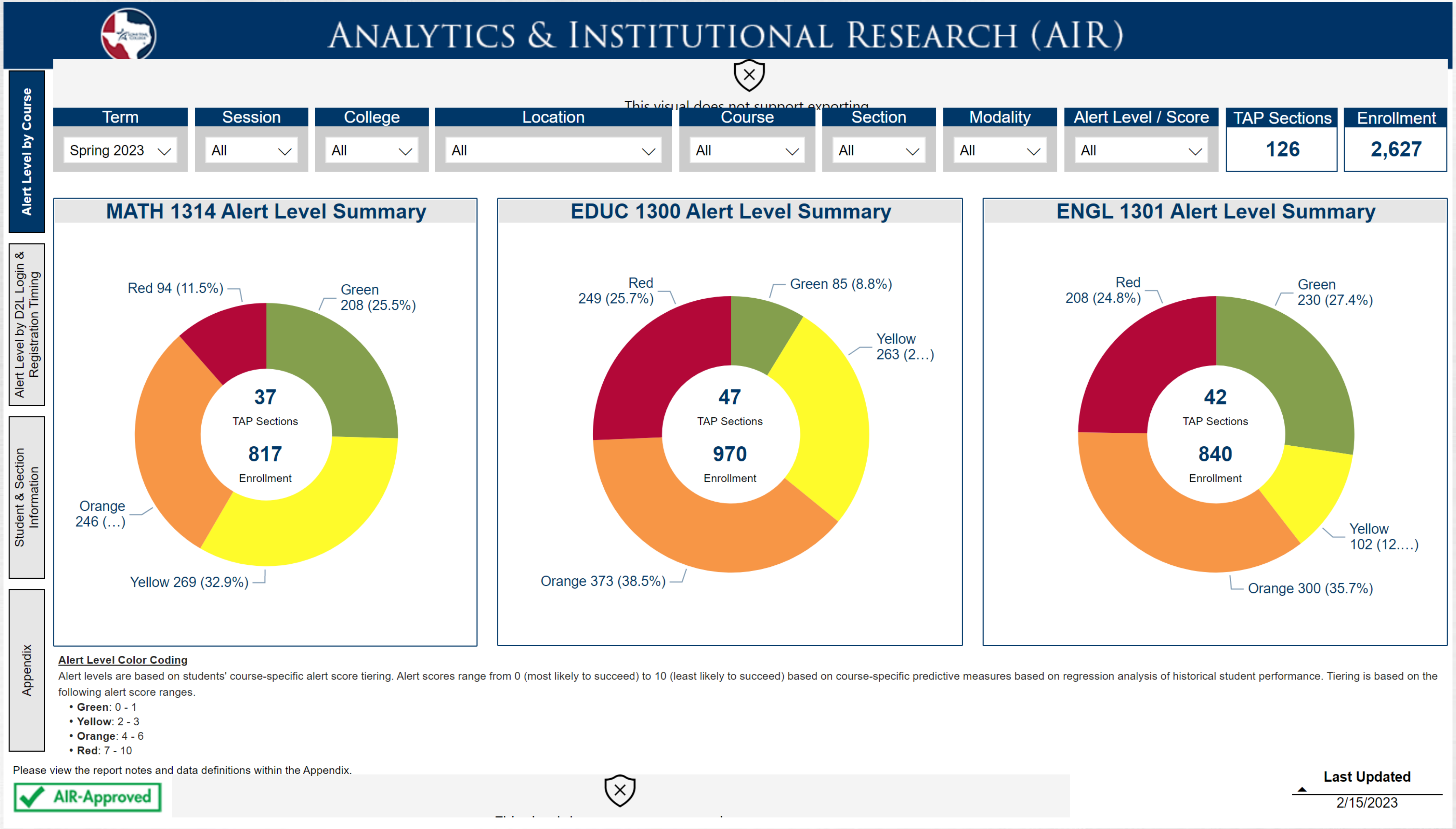
Enhancing Predictive Models for Student Success: A Multidimensional Approach

Jae Hak Jung, Ph. D. Director, Institutional Research, Lone Star College
(JaeHak.H.Jung@lonestar.edu)
Kwanghee Jung, Ph.D., Associate Professor, Texas Tech University
Jaehoon Lee, Ph.D., Associate Professor, Texas Tech University

Evolution of LSC Early Alert System

- Initiated by LSC Leadership and faculty request to identify at-risk students proactively
- Asked to create an Early Alert model aimed at accurately predicting classroom performance and potential dropouts
- Performed an in-depth logistic regression analysis to find significant predictors contributing to academic success
- Beta Early Alert Power BI dashboard developed, incorporating these predictors for real-time monitoring

Snapshot of the LSC Early Alert Power BI Dashboard



ANALYTICS & INSTITUTIONAL RESEARCH (AIR)

This visual does not support exporting.

Alert Level by Course | Alert Level by D2L Login & Registration Timing | Student & Section Information | Appendix

Term	Session	College	Location	Course	Section	Modality	Alert Level / Score	TAP Sections	Enrollment
Spring 2023	All	All	All	All	All	All	All	126	2,627

Alert Level Color Coding

Alert levels are based on students' course-specific alert score tiering. Alert scores range from 0 (most likely to succeed) to 10 (least likely to succeed) based on course-specific predictive measures based on regression analysis of historical student performance. Tiering is based on the following alert score ranges.

- **Green**: 0 - 1
- **Yellow**: 2 - 3
- **Orange**: 4 - 6
- **Red**: 7 - 10

Please view the report notes and data definitions within the Appendix.

AIR-Approved

Last Updated
2/15/2023

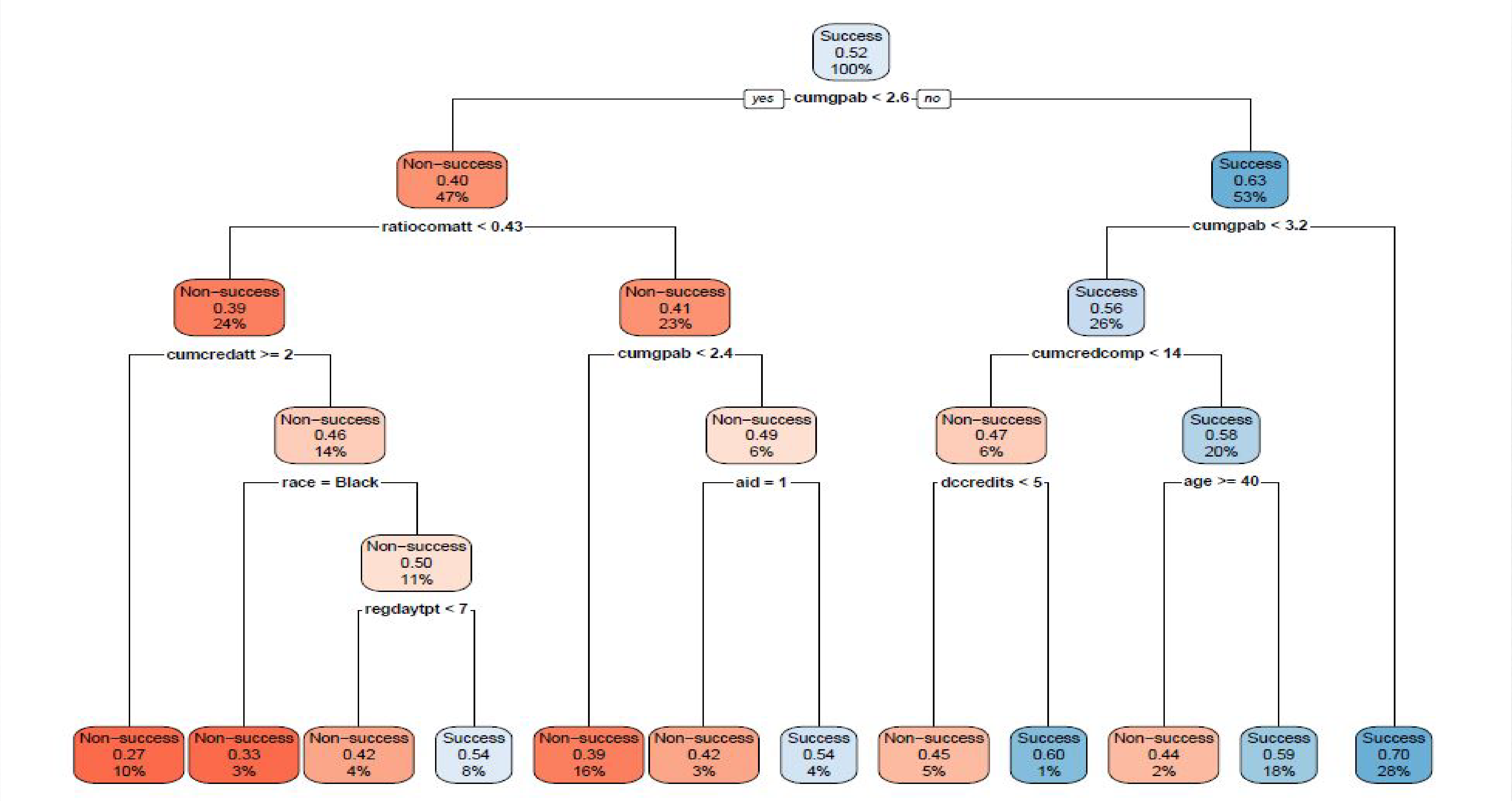
Overcoming Limitations and Exploring Implications

- Addressed the complexities of applying logistic regression analysis within the Early Alert Power BI environment and the interpretive challenges encountered
- Issues with subjective selection of predictor thresholds and multidimensional data interpretation
- Acknowledged difficulties in accurate student categorization and fulfilling the assumptions, especially with skewed datasets

Recap of Insights from the Last TAIR Presentation

- Reflect on the previous TAIR conference where we showcased the construction of a predictive model for college student success using the CART(Classification and Regression Trees) method
- Discuss how this method was utilized to interpret patterns and aid in the prediction of student outcomes

Visual Recap: Last Year's Decision Tree Model



Feedback Integration and Future Research Directions

- Addressed the exclusion of non-cognitive factors, high school performance, and college readiness in predictions
- Highlighted the importance of validating the CART method against other ML algorithms
- Stressed the need to look beyond course success to holistic outcomes like graduation and persistence rates

Advancing Research with Methodological Innovations

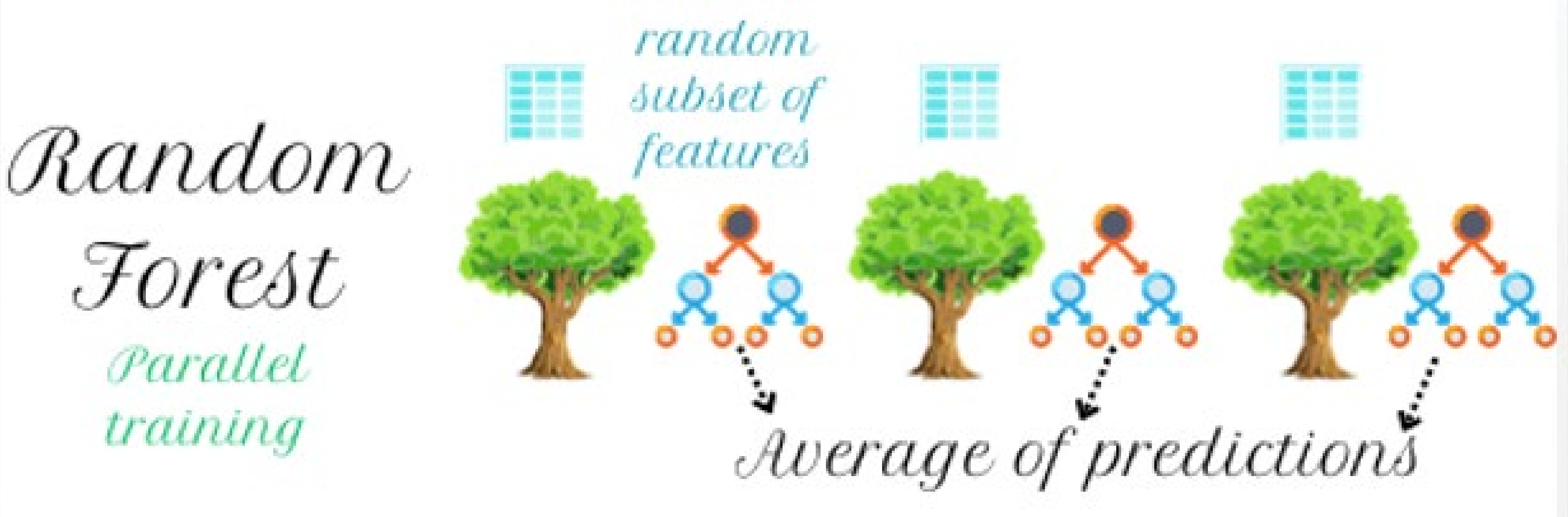
- Expose on expanding our analytical horizon by adopting the XGBoost & Random Forest algorithm alongside CART, enabling a comprehensive comparison of their predictive efficiencies
- Clarify our commitment to utilizing broader success metrics, such as graduation and persistence rates, reflecting a shift towards more holistic educational success indicators in line with emerging funding models for community colleges
- Upcoming development of an enhanced Power BI Interactive dashboard to facilitate early intervention for at-risk students

Machine Learning Algorithms

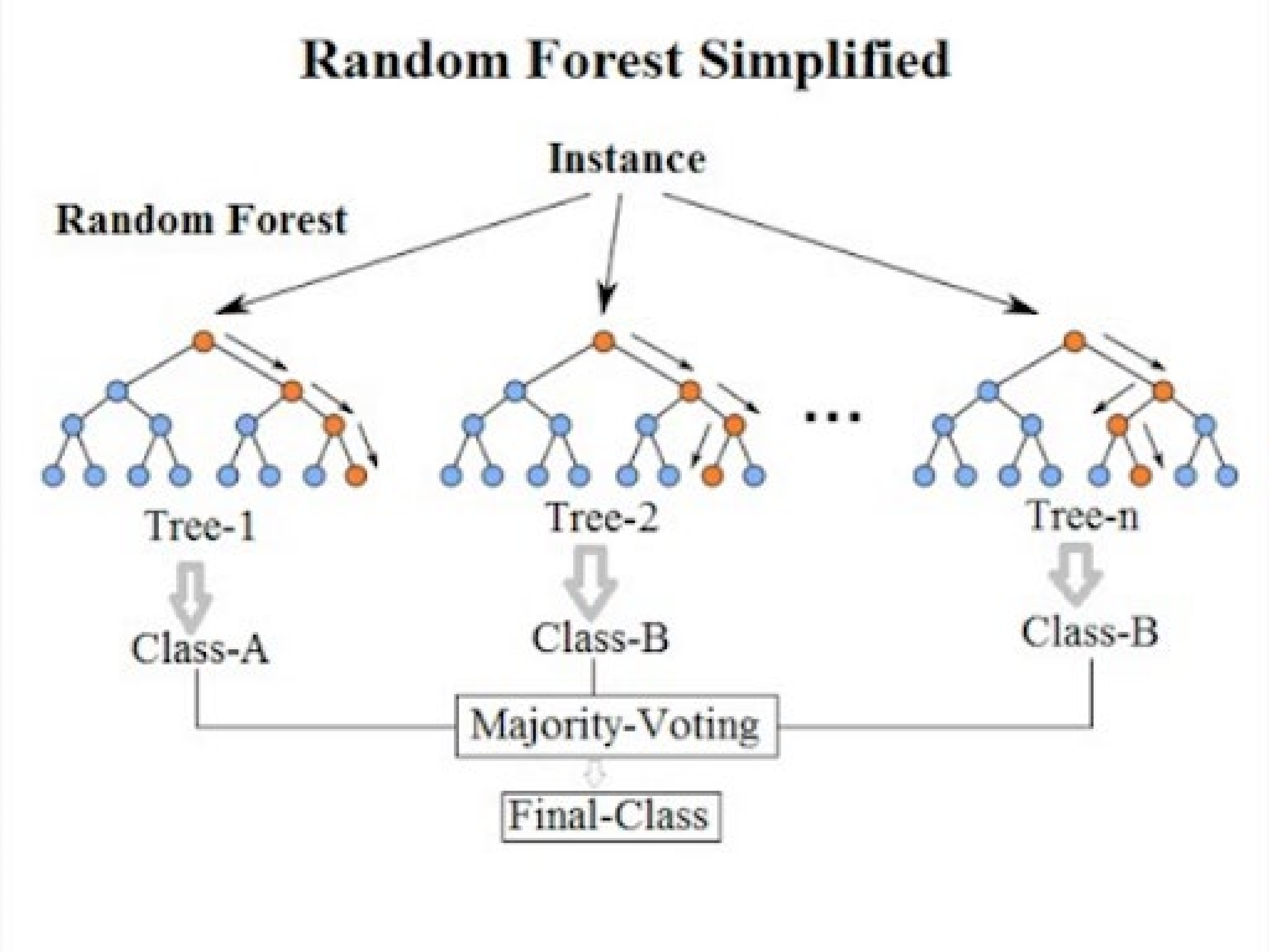
- **CART (Classification and Regression Trees)**
  - ✓ This is a fundamental machine learning method that builds a decision tree to make predictions
  - ✓ It's akin to asking a series of yes/no questions to infer the answer, which in our context is the likelihood of a student's persistence or dropout. It’s known for its simplicity and interpretability

Machine Learning Algorithms

- **Random Forest**
  - ✓ This method creates a ‘forest’ of decision trees
  - ✓ It's akin to assembling a committee where each member (tree) casts a vote, and the majority determines the prediction
  - ✓ Random Forests are great for increasing accuracy without the risk of overfitting, making them more reliable for complex decision-making

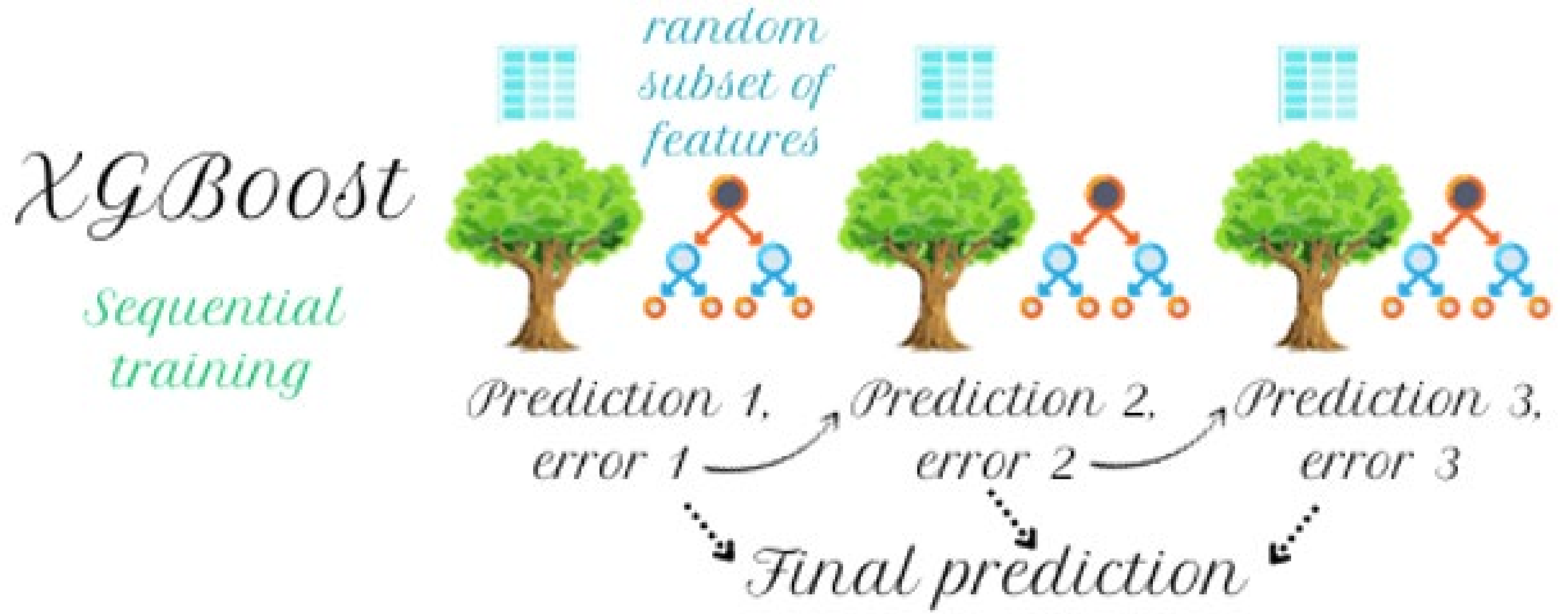


Random Forest

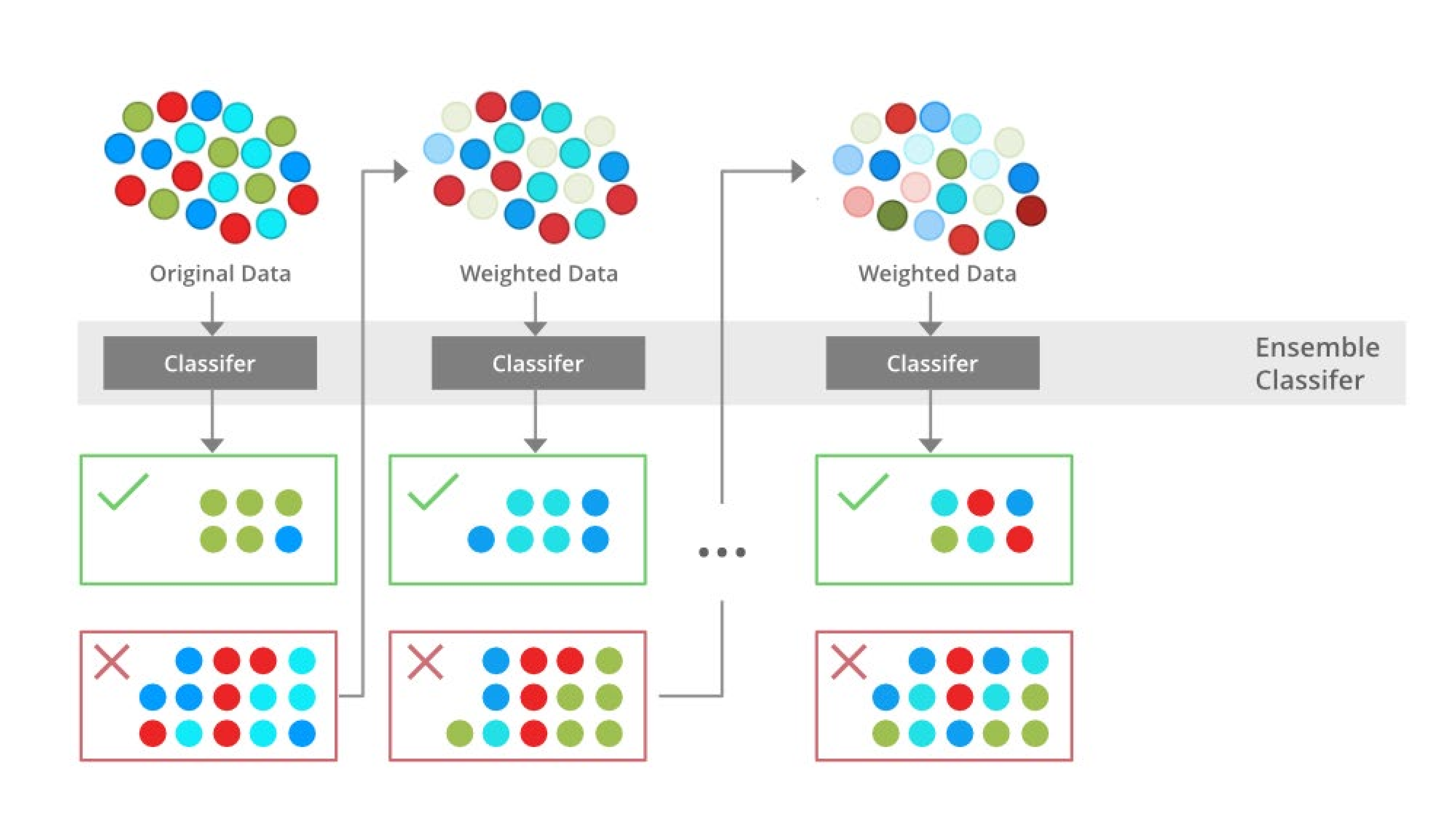


Machine Learning Algorithms

- **XGBoost (Extreme Gradient Boosting)**:
  - ✓ XGBoost builds trees one at a time, where each new tree helps to correct errors made by previously trained trees
  - ✓ With its high performance and speed, it is particularly useful for large datasets and challenging machine learning problems



XGBoost


Original Data
Weighted Data
Weighted Data
Classifer
Classifer
Classifer
Ensemble Classifer

Comparisons

Aspect	CART	Random Forest	XGBoost
Model Complexity	Simple, one tree	Complex, multiple trees	Complex, multiple boosted trees
Interpretability	High (single tree structure)	Medium (due to multiple trees)	Medium-Low (due to boosting and many trees)
Predictive Accuracy	Generally good baseline accuracy	Higher accuracy due to ensemble method	Often highest accuracy due to model sophistication
Use of Feature Information	Direct use of features to split nodes	Combines feature information across trees	Utilizes feature information iteratively for boosting

Lone Star College Student Data: Spring 2023 Cohort

- Enrollment Overview
  - ✓ Total Students Enrolled: 4,633
- Success Metrics
  - ✓ Persistence and Graduation in Fall 2023
  - ✓ Our definition of student success includes both persistence to the next semester and graduation from Spring 2023 to Fall 2023 for students who enrolled in Spring 2023

Predictors of student success

1) Cumulative GPA before Spring 2023
2) Term GPA in Fall 2022
3) Community College Survey of Student Engagement (CCSSE)
    - **Active and Collaborative Learning (ACTCOLL):** the extent to which students participate in class, interact with other students, and extend learning outside of the classroom.
    - **Student Effort (STUEFF):** time on task, preparation, and use of student services.
    - **Academic Challenge (ACCHALL):** The academic challenge benchmark measures the extent to which students engage in challenging mental activities, such as evaluation and synthesis, as well as the quantity and rigor of their academic work.
    - **Student-Faculty Interaction (STUFAC):** the extent to which students and faculty communicate about academic performance, career plans, and course content and assignments.
    - **Support for Learners (SUPPORT)** students’ perceptions of their colleges and assess their use of advising and counseling services
4) Full-time/Part-time in SP23

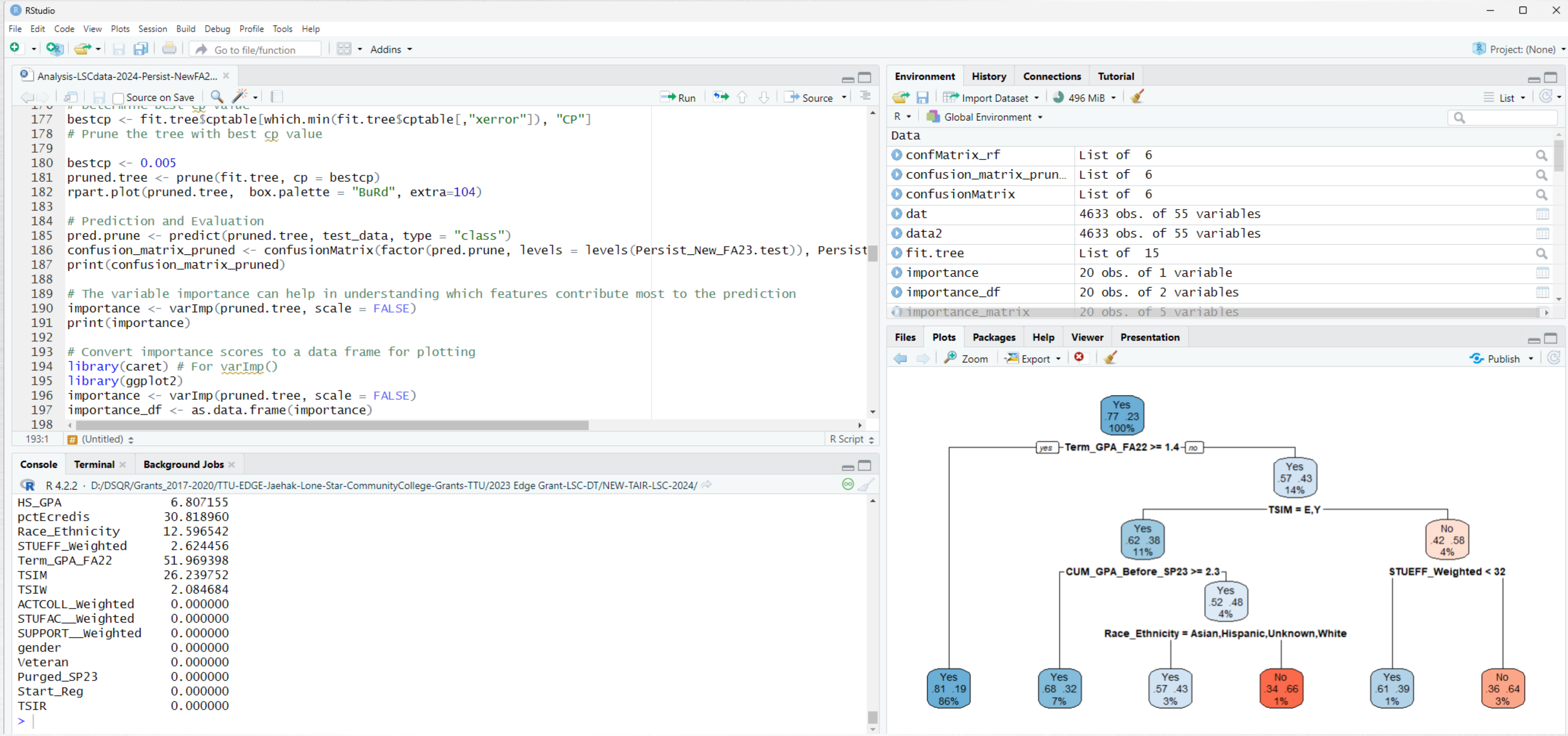
Predictors of student success

5) Gender
6) Age
7) Race/Ethnicity
8) Veteran Status
9) Ratio between Credits Earned and Credits Attempted
10) High School GPA
11) Financial aid
12) College Readiness (TSIM, TSIR, TSIW)
13) How much earlier the student registered in SP23
14) Purged (non-payment) Experience in SP23

R packages and functions

Algorithm	R Package	Main Function	Auxiliary Functions & Methods
CART	rpart	rpart()	printcp(), plotcp(), prune(), rpart.plot()
XGBoost	xgboost	xgboost(), xgb.train()	xgb.DMatrix(), xgb.importance(), xgb.plot.importance(), xgb.plot.tree(), xgb.dump()
Random Forest	randomForest	randomForest()	importance(), varImpPlot(), randomForest::getTree()

RStudio Interface



```r
# Determine best cp value
bestcp <- fit.tree$cptable[which.min(fit.tree$cptable[,"xerror"]), "CP"]
# Prune the tree with best cp value

bestcp <- 0.005
pruned.tree <- prune(fit.tree, cp = bestcp)
rpart.plot(pruned.tree,  box.palette = "BuRd", extra=104)

# Prediction and Evaluation
pred.prune <- predict(pruned.tree, test_data, type = "class")
confusion_matrix_pruned <- confusionMatrix(factor(pred.prune, levels = levels(Persist_New_FA23.test)), Persist
print(confusion_matrix_pruned)

# The variable importance can help in understanding which features contribute most to the prediction
importance <- varImp(pruned.tree, scale = FALSE)
print(importance)

# Convert importance scores to a data frame for plotting
library(caret) # For varImp()
library(ggplot2)
importance <- varImp(pruned.tree, scale = FALSE)
importance_df <- as.data.frame(importance)
```

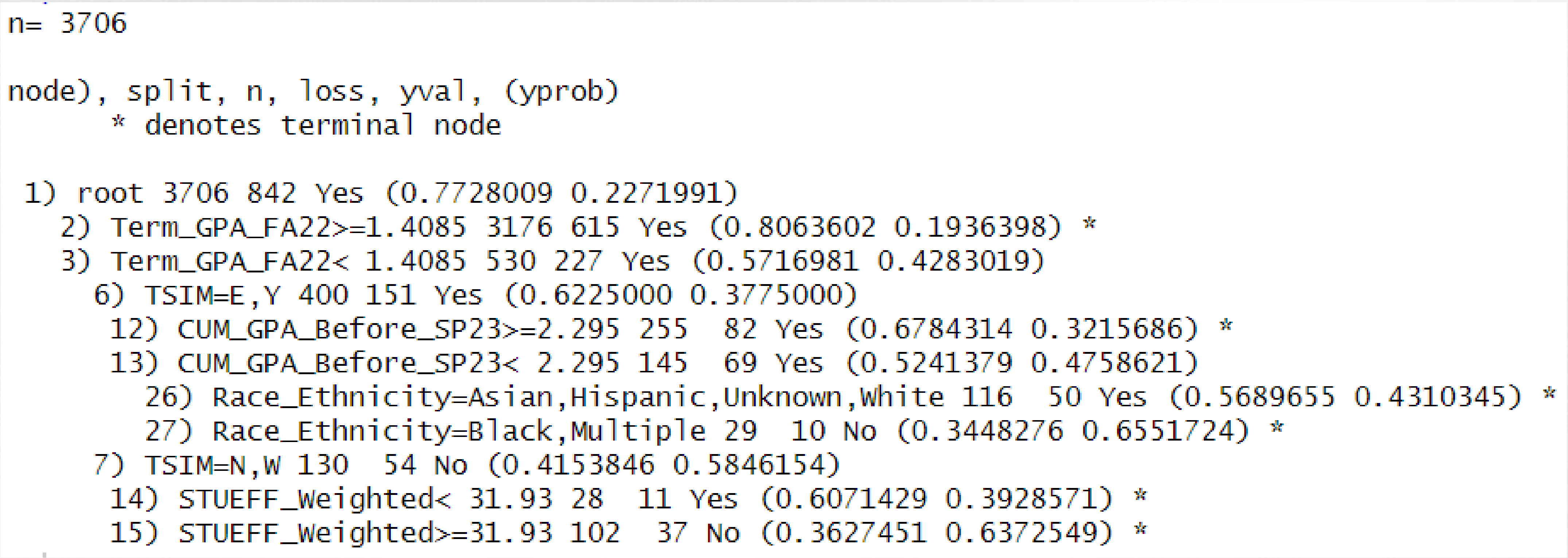
```
HS_GPA               6.807155
pctEcredis          30.818960
Race_Ethnicity      12.596542
STUEFF_Weighted      2.624456
Term_GPA_FA22       51.969398
TSIM                26.239752
TSIW                 2.084684
ACTCOLL_Weighted     0.000000
STUFAC__Weighted     0.000000
SUPPORT__Weighted    0.000000
gender               0.000000
Veteran              0.000000
Purged_SP23          0.000000
Start_Reg            0.000000
TSIR                 0.000000
>
```

Data

confMatrix_rf	List of 6
confusion_matrix_prun...	List of 6
confusionMatrix	List of 6
dat	4633 obs. of 55 variables
data2	4633 obs. of 55 variables
fit.tree	List of 15
importance	20 obs. of 1 variable
importance_df	20 obs. of 2 variables
importance_matrix	20 obs. of 5 variables

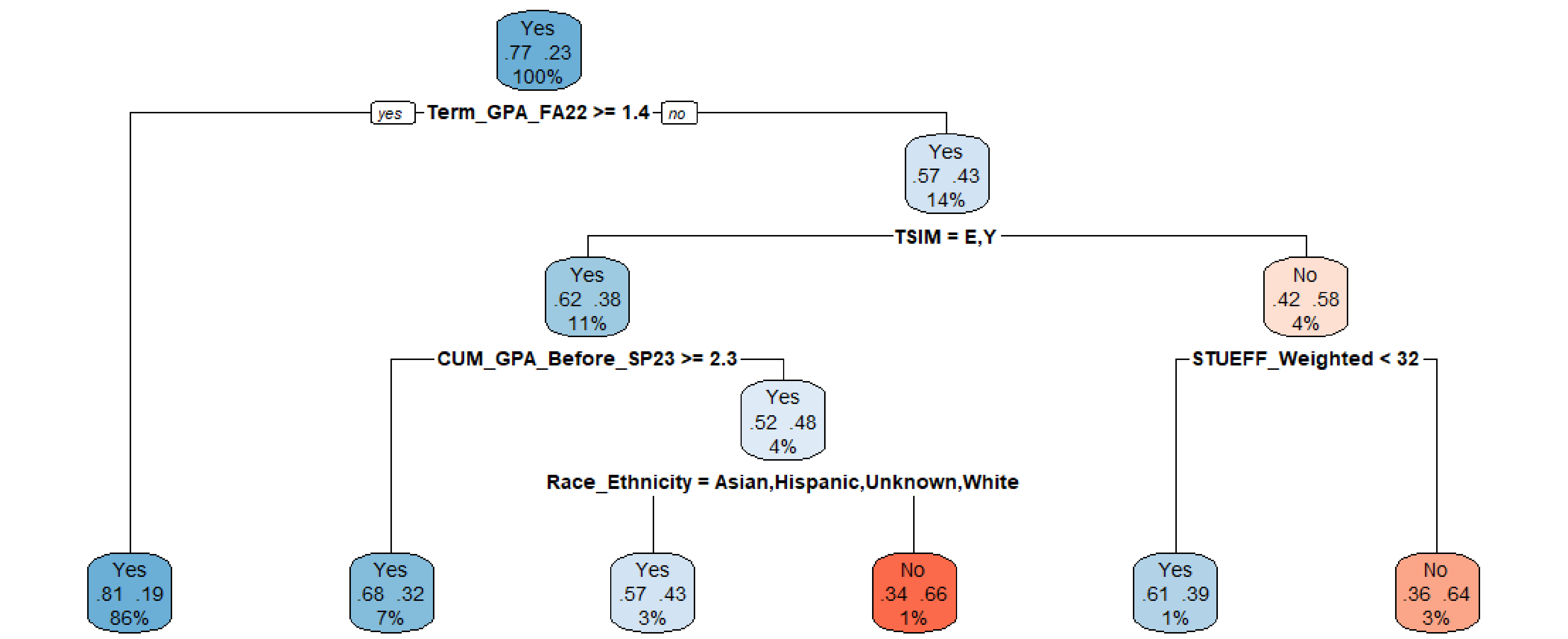
Results

Decision Tree (CART): Persistence Fall 2023

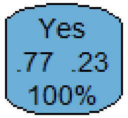


```
n= 3706

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 3706 842 Yes (0.7728009 0.2271991)
   2) Term_GPA_FA22>=1.4085 3176 615 Yes (0.8063602 0.1936398) *
   3) Term_GPA_FA22< 1.4085 530 227 Yes (0.5716981 0.4283019)
     6) TSIM=E,Y 400 151 Yes (0.6225000 0.3775000)
      12) CUM_GPA_Before_SP23>=2.295 255  82 Yes (0.6784314 0.3215686) *
      13) CUM_GPA_Before_SP23< 2.295 145  69 Yes (0.5241379 0.4758621)
        26) Race_Ethnicity=Asian,Hispanic,Unknown,White 116  50 Yes (0.5689655 0.4310345) *
        27) Race_Ethnicity=Black,Multiple 29  10 No (0.3448276 0.6551724) *
     7) TSIM=N,W 130  54 No (0.4153846 0.5846154)
      14) STUEFF_Weighted< 31.93 28  11 Yes (0.6071429 0.3928571) *
      15) STUEFF_Weighted>=31.93 102  37 No (0.3627451 0.6372549) *
```

CART: Persistence Fall 2023



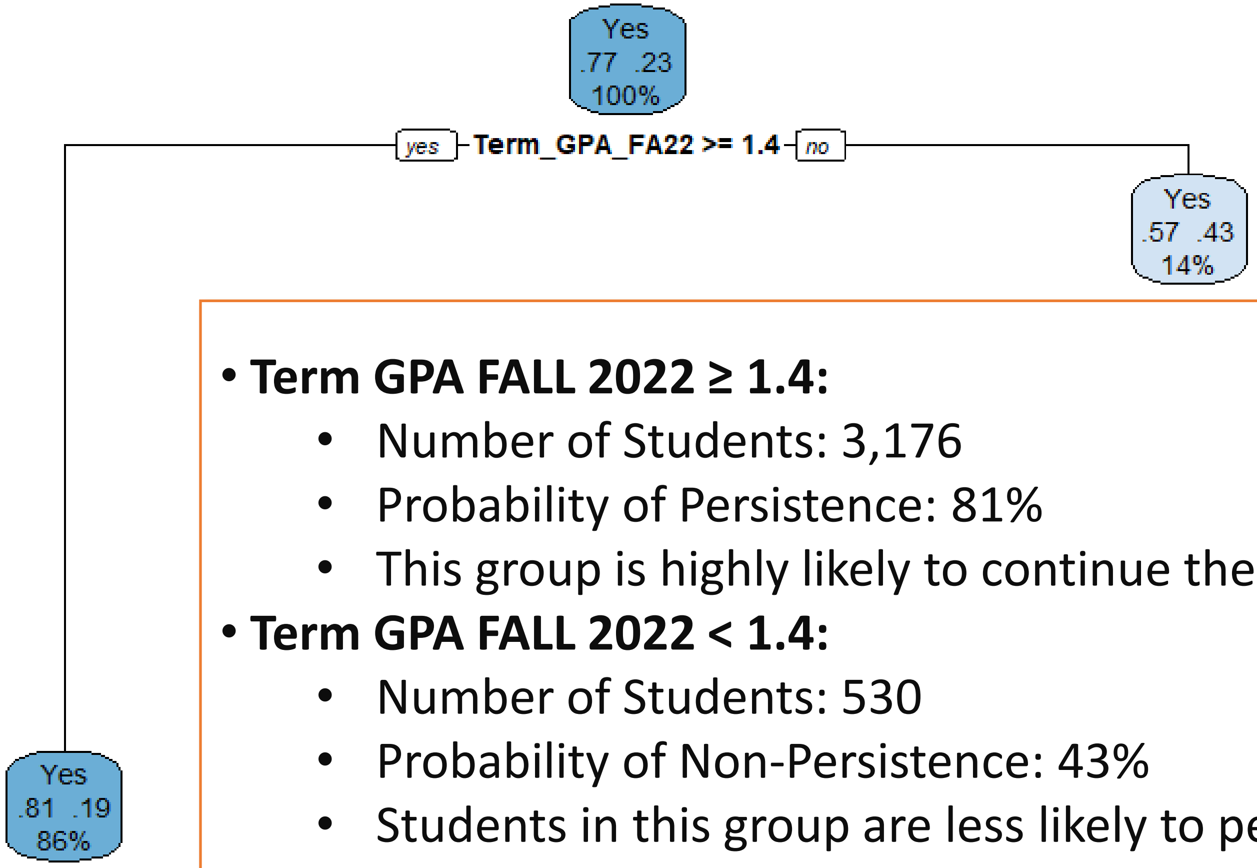
CART: Persistence Fall 2023



Root Node (Node 1):

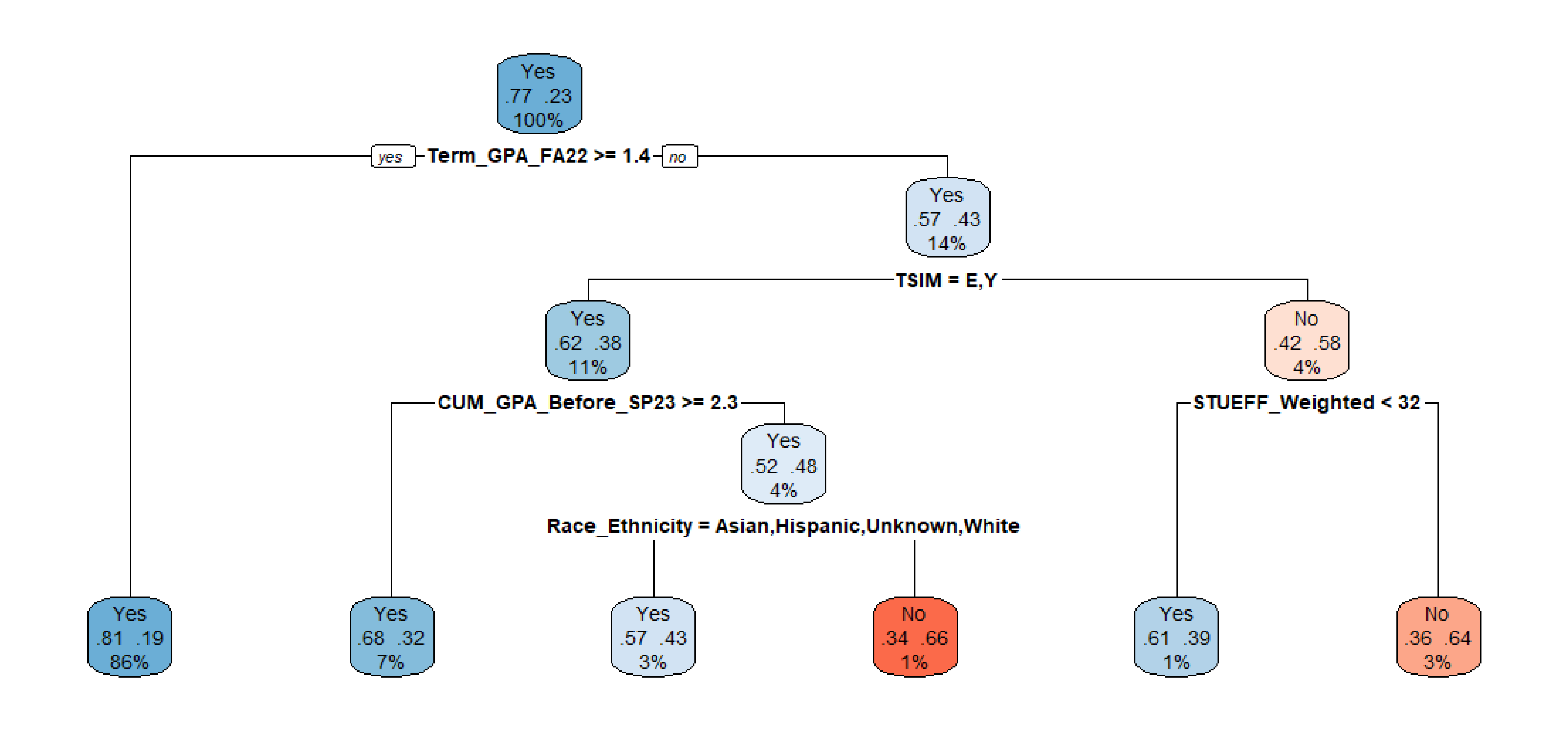
- This is the starting point of the tree, encompassing all 3,706 students (training data)
- The probability of a student persisting is 77.28%, while the probability of not persisting is 22.72%.

CART: Persistence Fall 2023

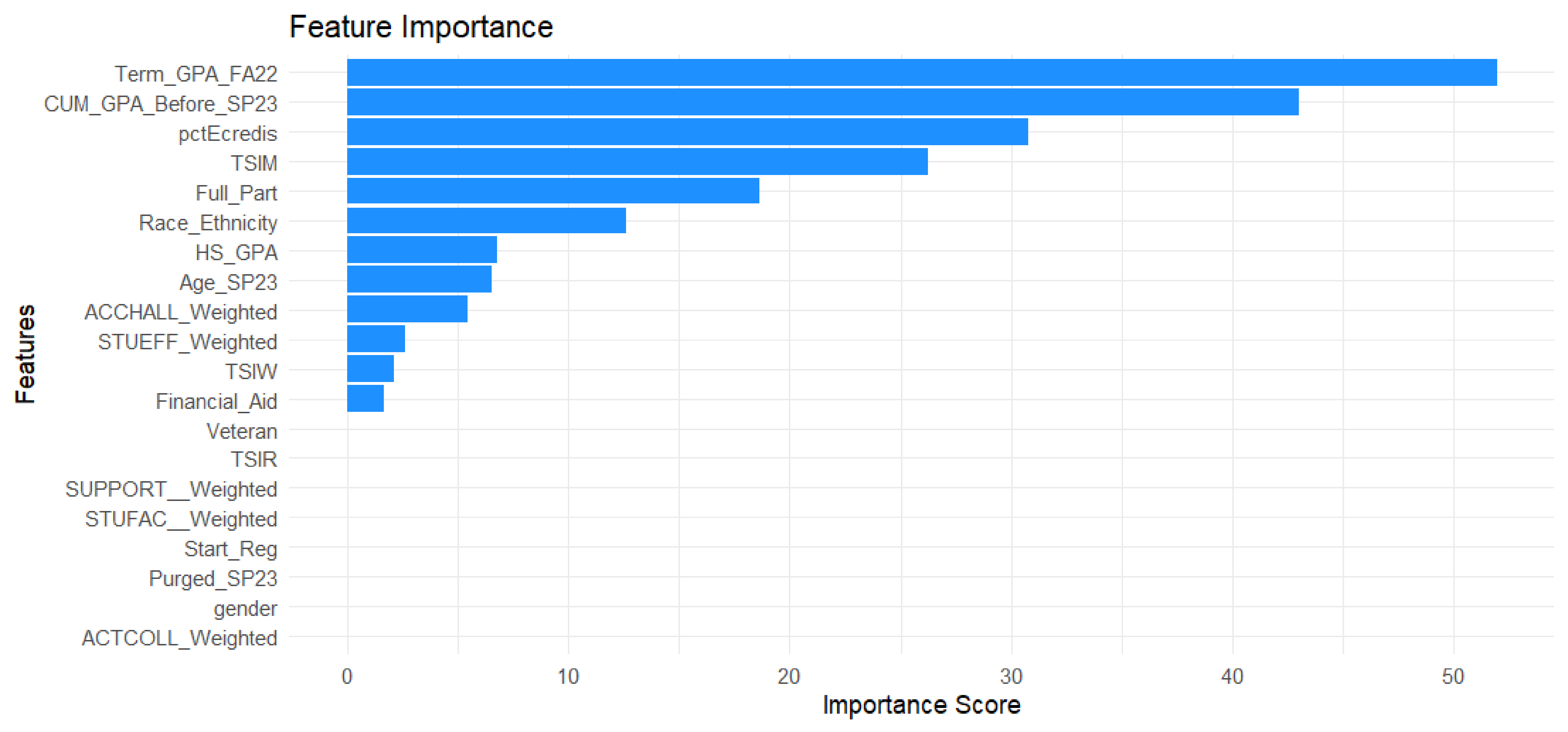


- **Term GPA FALL 2022 ≥ 1.4:**
  - Number of Students: 3,176
  - Probability of Persistence: 81%
  - This group is highly likely to continue their studies (Terminal node)
- **Term GPA FALL 2022 < 1.4:**
  - Number of Students: 530
  - Probability of Non-Persistence: 43%
  - Students in this group are less likely to persist, leading to further branching based on other factors

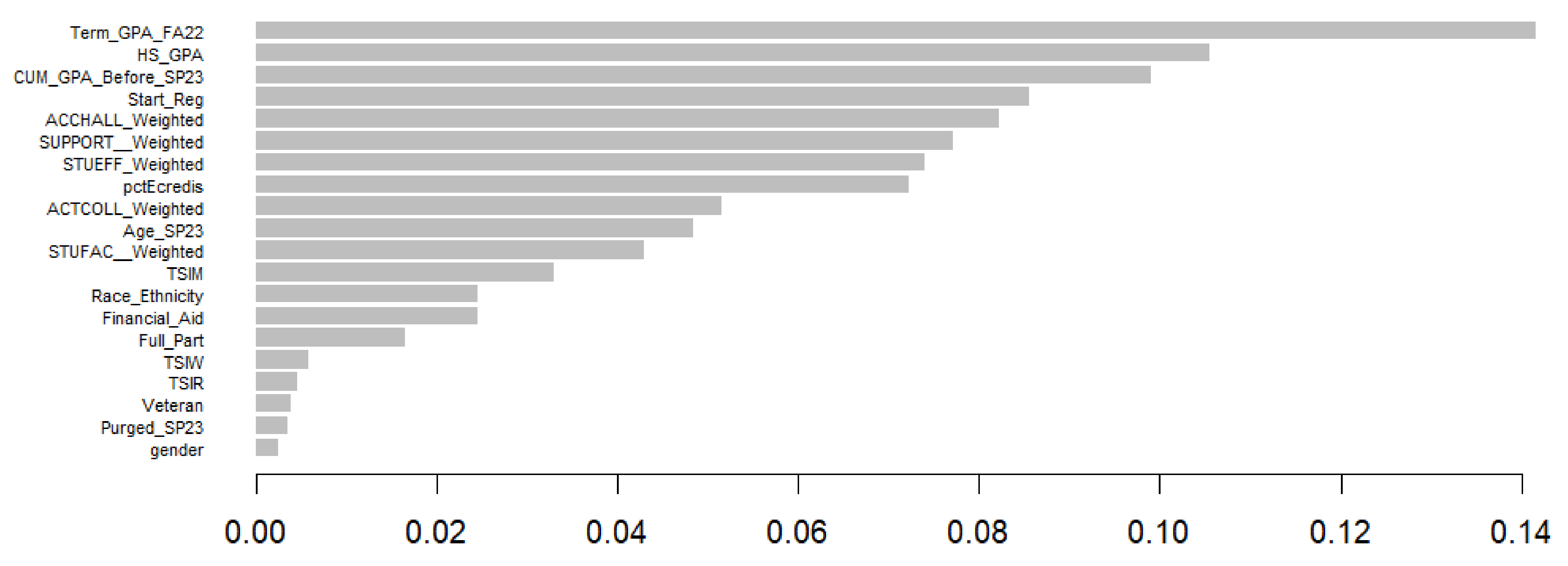
CART: Persistence Fall 2023



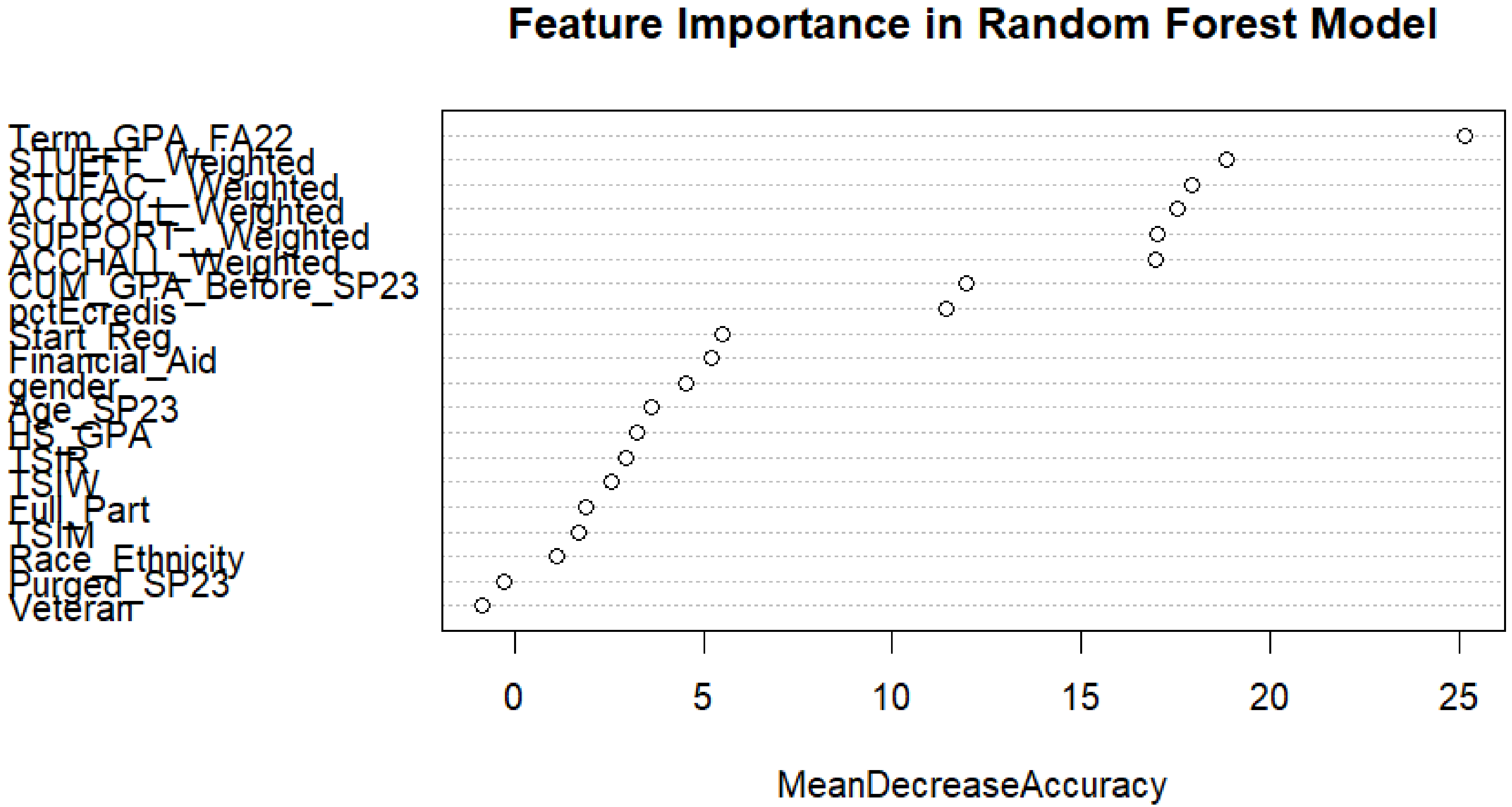
CART: Persistence Fall 2023



XGBoost: Persistence Fall 2023



Random Forest: Persistence Fall 2023



Summary: Persistence Fall 2023

Metric/Model	CART	XGBoost	Random Forest
Accuracy	77.45%	76.59%	77.87%
Sensitivity	93.55%	93.82%	96.50%
Specificity	16.33%	11.22%	7.14%
Top Variables	- Term GPA in FA22 - Cumulative GPA before SP23 - Ratio between Credits Earned and Credits Attempted	- Term GPA in FA22 - High School GPA - Cumulative GPA before SP23	- Term GPA in FA22 - CCSSE subscales - Cumulative GPA before SP23
AUC	0.5747	0.559	0.6008

Conclusion: Evaluation of Machine-Learning Algorithms

- **Accuracy Comparison**
  No significant difference in accuracy was observed among the three machine-learning algorithms
- **Key Predictive Variables**
  - ✓**Most Important Predictor Across All Models:**
    GPA from the Previous Semester
  - ✓**Secondary Important Variables (Varied by Model):**
    Cumulative GPA
    High School GPA
    Registration Timing for the Course
    CCSSE Subscale Scores

Implications: Machine-Learning Algorithms

- **CART Model Performance**
  - ✓ Comparable in accuracy and sensitivity to other algorithms
  - ✓ Consistency in key variables predicting persistence across models
  - ✓ Validated for use due to its intuitive explanation and ease of application in Power BI
- **Practical Implications**
  - ✓ The CART model's user-friendly nature supports broader acceptance and application
  - ✓ Its compatibility with analytical tools, like Power BI, enhances practicality in educational and predictive settings

Future Directions

- **Expansion of Study Scope**
  - **New Student Cohorts**: Extend research to include newly enrolled students to diversify insights and validate findings across broader demographics
  - **Separate Analyses**: Conduct distinct studies for graduation rates and student transfer patterns to uncover specific predictors and trends
  - **Integrate additional significant predictors** to improve the model’s accuracy and predictive power, ensuring more precise and actionable insights

Implications of Machine-Learning Integration

- **Power BI and Azure Machine Learning Integration**
  - ✓ Develop a Power BI report that seamlessly integrates with Azure Machine Learning models and datasets
  - ✓ This integration aims to enhance the reporting and analysis framework, enabling more sophisticated insights derived from machine learning predictions
- **Benefits**
  - ✓ Leverage advanced analytics to uncover deeper insights into student success factors and educational trends
  - ✓ Facilitate the sharing of complex findings in an accessible, interactive format, enhancing decision-making processes for educational administrators and stakeholders

Thank you

Any questions?

Appendix: Sensitivity vs. Specificity

	Persistence (Condition Positive)	Non-Persistence (Condition Negative)
Persistence (Prediction Positive)	**True Positive (TP)**	False Positive (FP)
Non-Persistence (Prediction Negative)	False Negative (FN)	**True Negative (TN)**

- **Sensitivity (True Positive Rate) = TP / (TP + FN)**
- **Specificity (True Negative Rate) = TN / (TN + FP)**