



Supporting Data-Driven Decision Making: The Classification and Regression Tree (CART) Algorithm

José Noel Caraballo


Irmannette Torres-Lugo

University of Puerto Rico at Cayey



Outline

- Institutional research and data-driven decision making
- Adequacy of the UPR's admission criteria
- Recursive partitioning – Basic ideas
- Tree construction – A simple example
 - Splitting predictors
 - Prunning
 - Misclassification error
- Predicting time to degree completion
 - Findings
- Final comments
- Questions?



Institutional Research and Data-Driven Decision Making

- Central Role of Institutional Research and Assessment: Providing the foundations for data-driven decision making.
- Quantitative multivariate analysis uncovering evidence regarding the relationships amongst variables is frequently used as a mechanism for refining institutional policies (Toutkoushian, 2007).
- To undergraduate institutions, policies regarding first-time degree seeking students are of particular interest especially when uniform admission criteria have been established for systemic admissions for diverse campuses, each with their own peculiarities and service groups.



Adequacy of the UPR's Admission Criteria

- Twelve years after the implementation of revised admission criteria, the central administration has requested all campuses to analyze the adequacy of the criteria for predicting their students' success, in an effort to understand the predictive potential of the current process and uncover the relationship between multiple variables.
- Recursive partitioning has been identified as the method for analyzing the relationship amongst the variables gathered through the student application forms and the student achievement information which form part of institution's database.



Recursive Partitioning

- Procedure whereby a given set of data is partitioned into increasingly homogeneous subsets.
- Typically, the result of the application of this procedure is presented as a dendrogram, or inverted decision tree.
- Can be used as an alternative for regression or discriminant analysis.

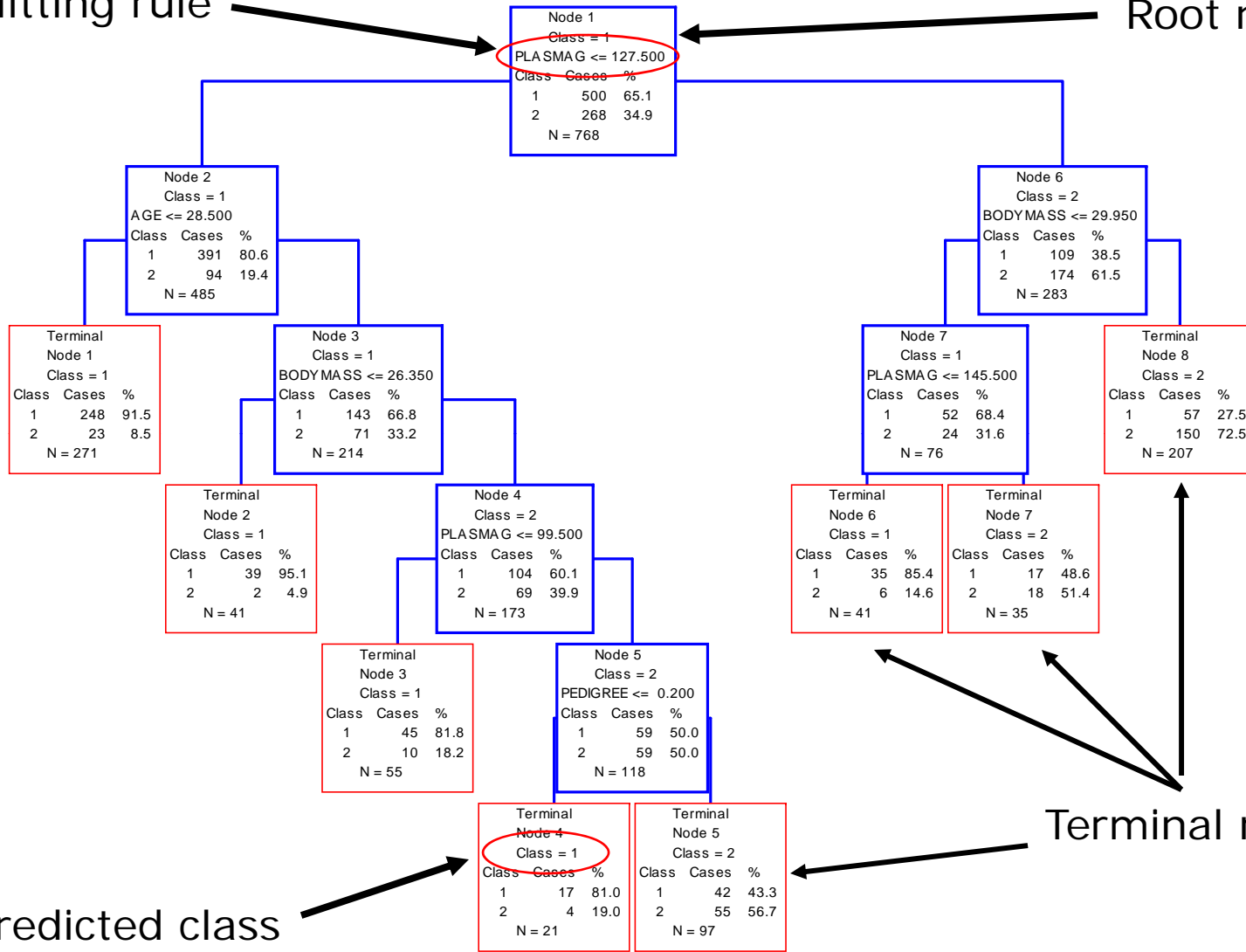


Characteristics

- Non parametric
- Flexible
 - Classification
 - Regression
- In general, predictions are as good or better than those obtained by discriminant analysis or multiple regression
- Higher-order interactions

Splitting rule

Root node



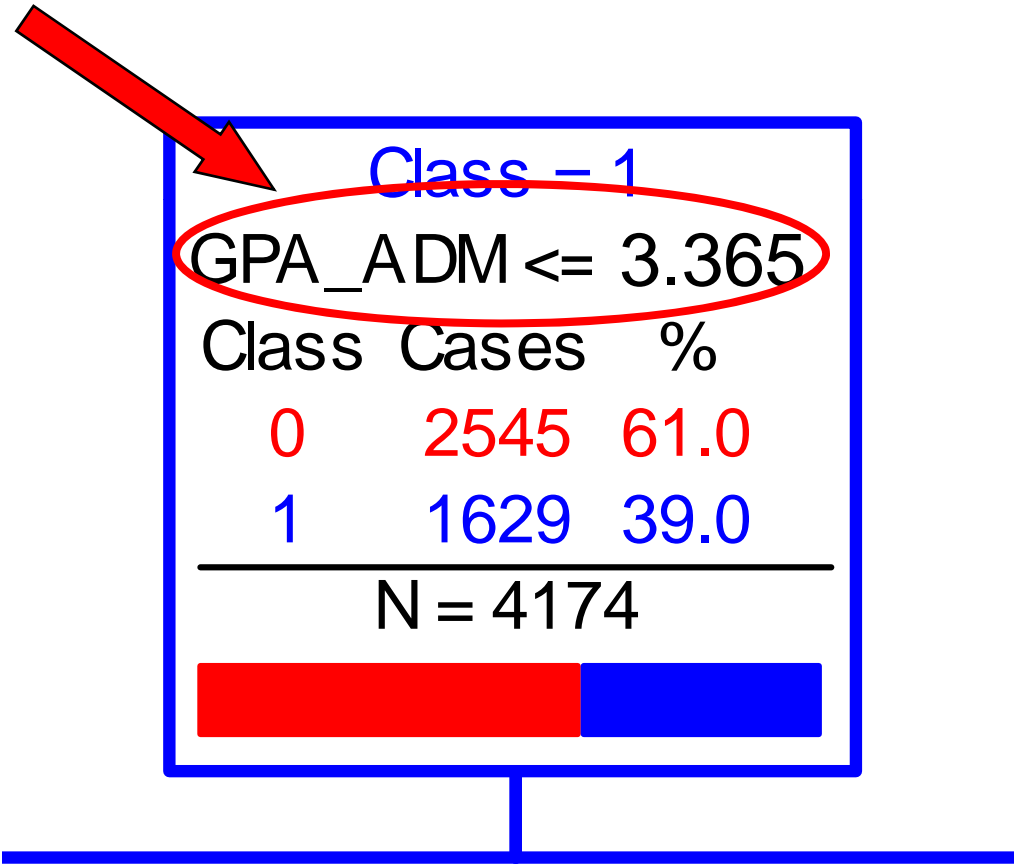
Predicted class membership

Terminal nodes

Regression Tree (a.k.a. dendogram)

Root Node

Splitting rule



Terminal Node

Predicted class membership



Terminal Node 4		
Class	Cases	%
1	17	81.0
2	4	19.0
N = 21		

Class = 1

19.0



Misclassification error



How a tree is constructed

- The basic idea is to partition the data set into homogeneous subgroups (in terms of the dependent variable):
 - The root node is split into two subnodes.
 - Each subnode is split into subnodes.
 - Process continues until some criteria is met (e.g., node is homogeneous; $n \leq 5$)
 - The tree is pruned to obtain the one with minimum classification error.
- What characterizes the CART algorithm is that all splits are binary.



Number of possible splits

- The CART algorithm will try all possible splits of all independent variables and select the one with the ‘best’ split (i.e., the one that reduced more the heterogeneity of the data.)
- For a continuous variable with n distinct values, there are $n-1$ possible splits. Each of the $n-1$ splits are performed at a point x_{jn} that is midway between two consecutive ordered values $X_j^{(i)}$ and $X_j^{(i+1)}$.
- For a discrete variable with J categories there are $2^{(J-1)}-1$ possible splits.



Example: Categorical dependent variable

- Prediction of color (red, blue, yellow)
- Predictors: size (big, small) and shape (circle, square, triangle)
- $n = 15$

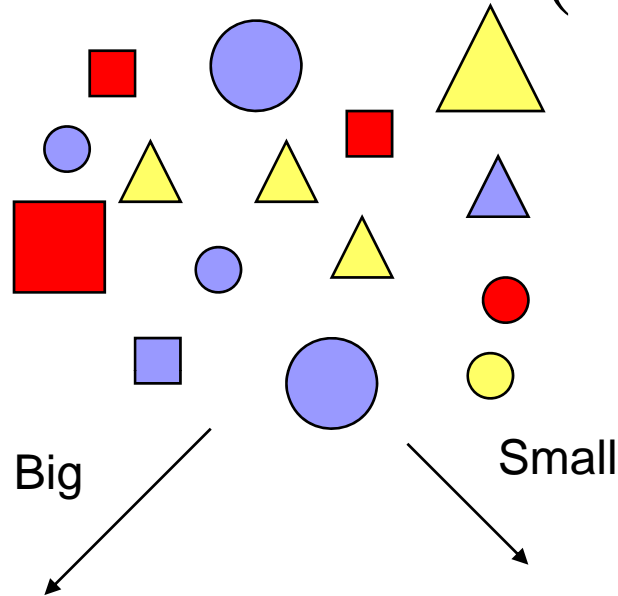


Size

(color, size, shape)

Number of splits:

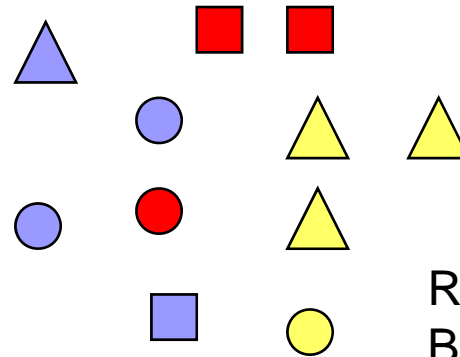
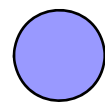
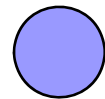
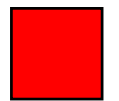
$$2^{k-1} - 1 = 2^{2-1} - 1 = 1$$



Big

Small

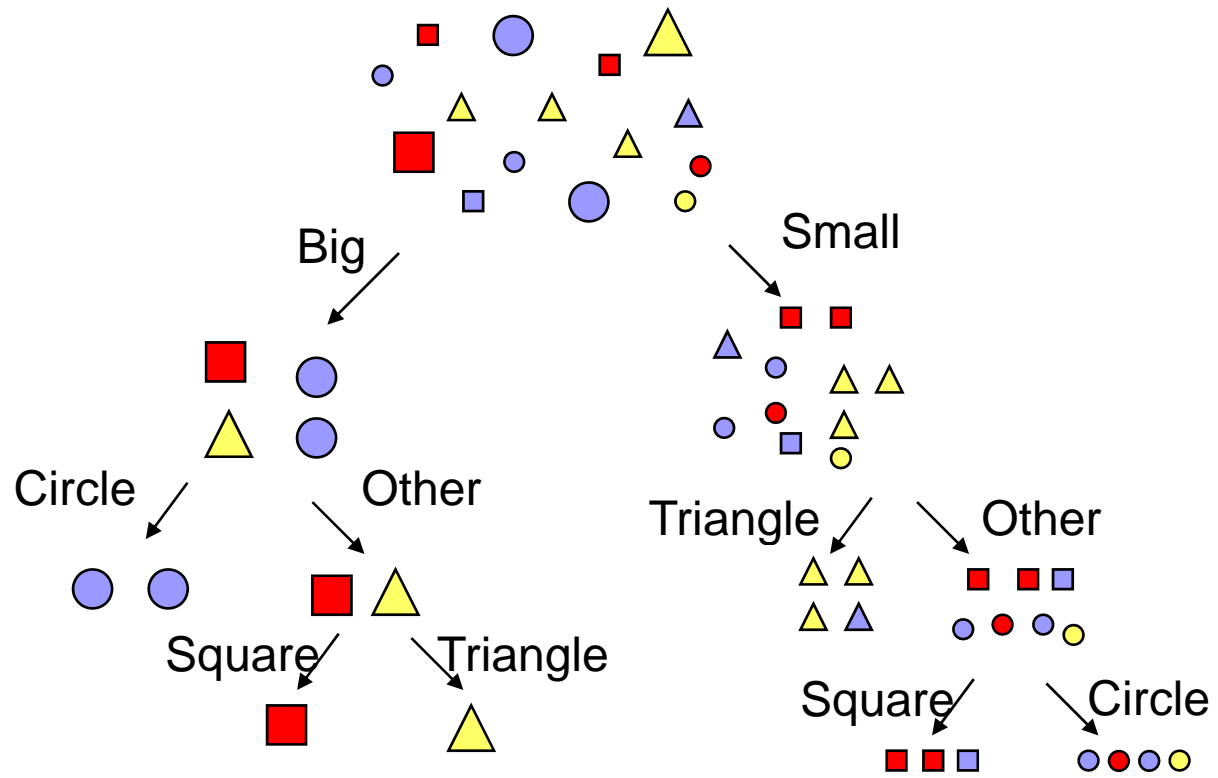
Red = 1
Blue = 2
Yellow = 1



Red = 3
Blue = 4
Yellow = 4



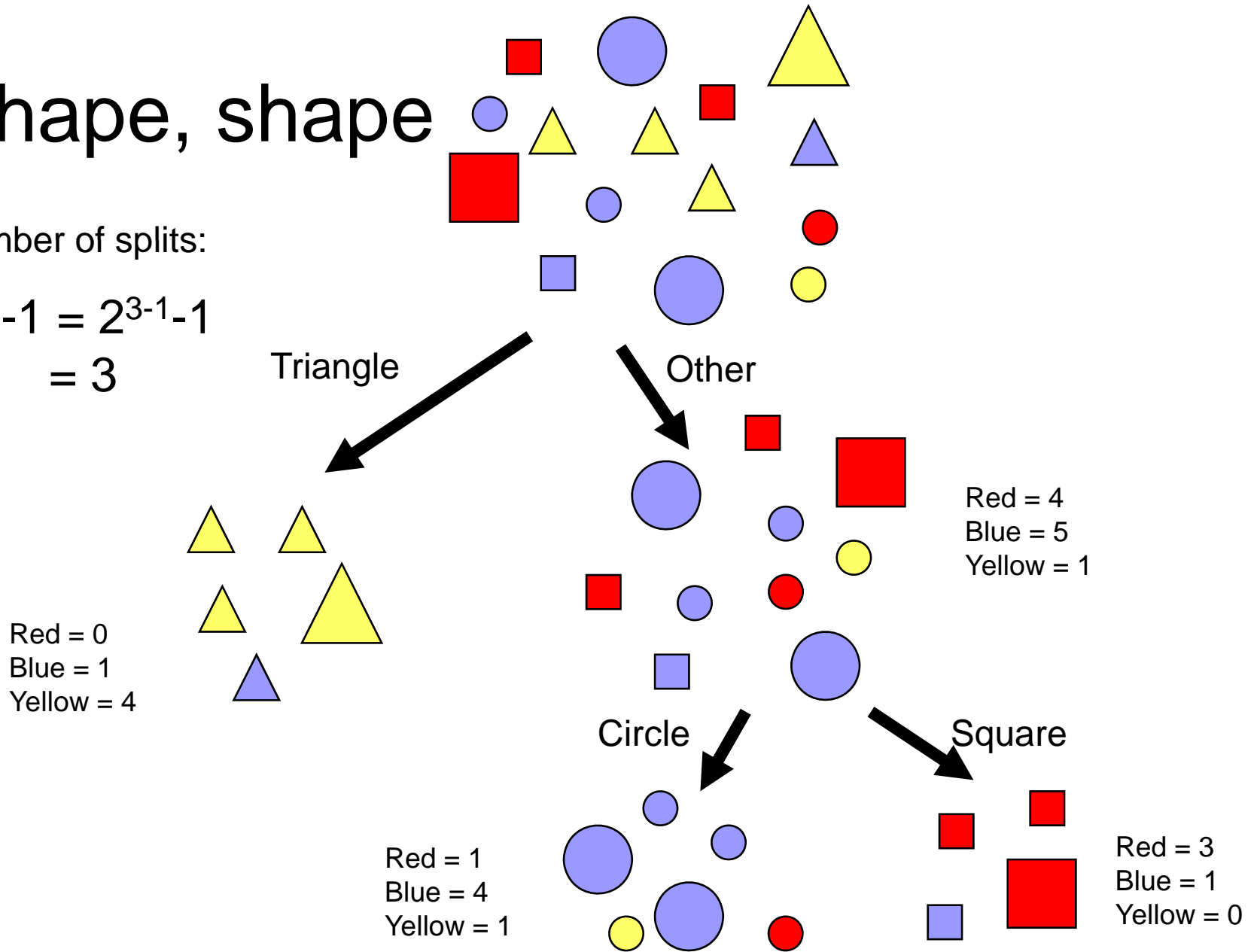
Size



Shape, shape

Number of splits:

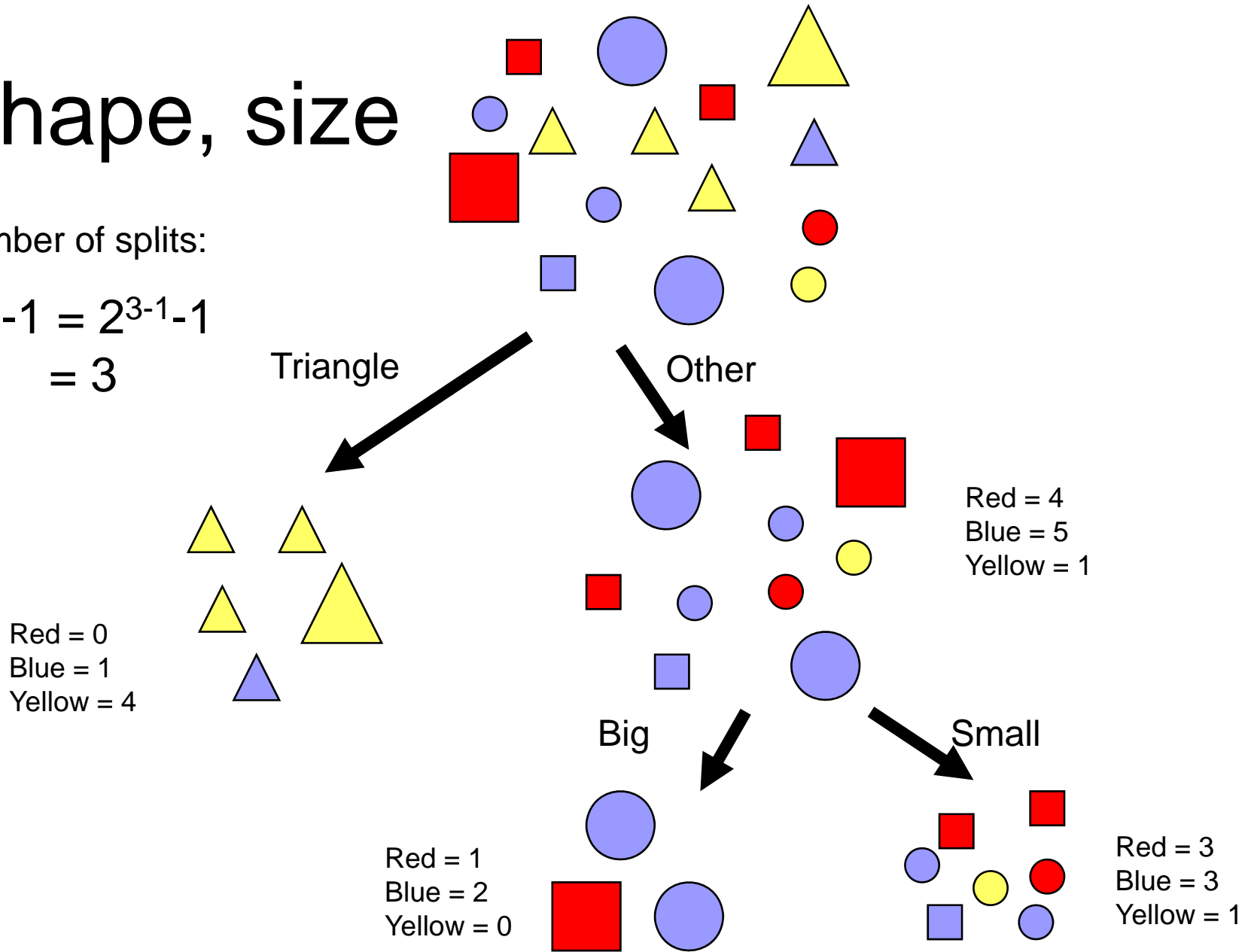
$$2^{k-1} - 1 = 2^{3-1} - 1 = 3$$



Shape, size

Number of splits:

$$2^{k-1} - 1 = 2^{3-1} - 1 = 3$$





CART

- Splitting in CART is based on the minimization of an impurity function
- Continues until data in a node is homogeneous, or too few cases are available.
- This 'maximal' tree is in turn pruned until a tree is obtained with the minimum error rate.



Regression Trees

- Measuring heterogeneity for continuous dependent variable
 - Least squares
 - Least absolute deviation



Classification Trees

- Measuring heterogeneity for a categorical dependent variable:
- We need a function $\Phi(p_1, p_2, \dots, p_k)$ with the following characteristics:
 - If $p_1 = p_2 = \dots p_k$ then Φ is a maximum.
 - If $p_j = 1$ and $p_i = 0, \forall i \neq j$ then $\Phi = 0$
- CART provides different splitting rules, two of the most commonly used are
 - GINI
 - Twoing



GINI splitting criteria

- Measure of heterogeneity (impurity) at a node:

$$i(t) = 1 - \sum_{j=1}^J p^2(j | t)$$

- Select as the splitting variable the one that minimizes:

$$\Delta i = i(t) - p_1 i(1) - p_2 i(2)$$

$p(j|t)$ is the proportion of class j in node t .

Tree construction example

Y_1	X_1	X_2	X_3	Y_1	X_1	X_2	X_3
1	1	1	10	3	1	2	9
2	1	2	7	3	1	2	10
1	2	1	5	2	1	1	7
2	2	1	7	2	2	1	9
3	3	2	8	1	3	2	6
3	3	2	8	3	3	2	9
3	2	1	8	3	3	1	8
3	2	1	9	2	2	1	6
1	4	1	6	1	4	2	5
1	1	1	6	1	4	2	4

Categorical Continuous



Impurity at the root node:

$$Y = \{1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3\}$$

$$p(1 | r) = \frac{7}{20} \qquad p(2 | r) = \frac{5}{20}$$

$$p(3 | r) = \frac{8}{30}$$

$$\begin{aligned} i(r) &= 1 - \sum_{j=1}^3 p^2(j | r) = 1 - p^2(1 | r) - p^2(2 | r) - p^2(3 | r) \\ &= 1 - \left(\frac{7}{20}\right)^2 - \left(\frac{5}{20}\right)^2 - \left(\frac{8}{20}\right)^2 = 0.40 \end{aligned}$$



Splitting a categorical variable: X_1

- With 4 values there are $2^{4-1}-1=7$ possible splits:

1-234

2-134

3-124

4-123

12-34

13-24

14-23



Splitting a categorical variable: X_2


- With 2 values there are $2^{2-1}-1=1$ possible splits: 1-2

For $X_2=1$, Y is

$\{1,1,1,1,2,2,2,2,3,3,3\}$

For $X_2=2$, Y is


$\{1,1,1,2,3,3,3,3,3\}$



For $X_2=1$, $Y=\{1,1,1,1,2,2,2,2,3,3,3\}$

$$p(1|1) = \frac{4}{11} \quad p(2|1) = \frac{4}{11} \quad p(3|1) = \frac{4}{11}$$

$$\begin{aligned} i(2) &= 1 - \sum_{j=1}^3 p^2(j|2) \\ &= 1 - p^2(1|2) - p^2(2|2) - p^2(3|2) \\ &= 1 - \left(\frac{4}{11}\right)^2 - \left(\frac{4}{11}\right)^2 - \left(\frac{3}{11}\right)^2 \\ &\approx 0.66 \end{aligned}$$



For $X_2=2$, $Y=\{1,1,1,2,3,3,3,3,3\}$

$$p(1|2) = \frac{3}{9} \quad p(2|2) = \frac{1}{9} \quad p(3|2) = \frac{5}{9}$$

$$i(3) = 1 - \sum_{j=1}^3 p^2(j|3)$$

$$= 1 - p^2(1|3) - p^2(2|3) - p^2(3|3)$$

$$= 1 - \left(\frac{3}{9}\right)^2 - \left(\frac{1}{9}\right)^2 - \left(\frac{5}{9}\right)^2$$

$$\approx 0.57$$



So, the change in “impurity” is:

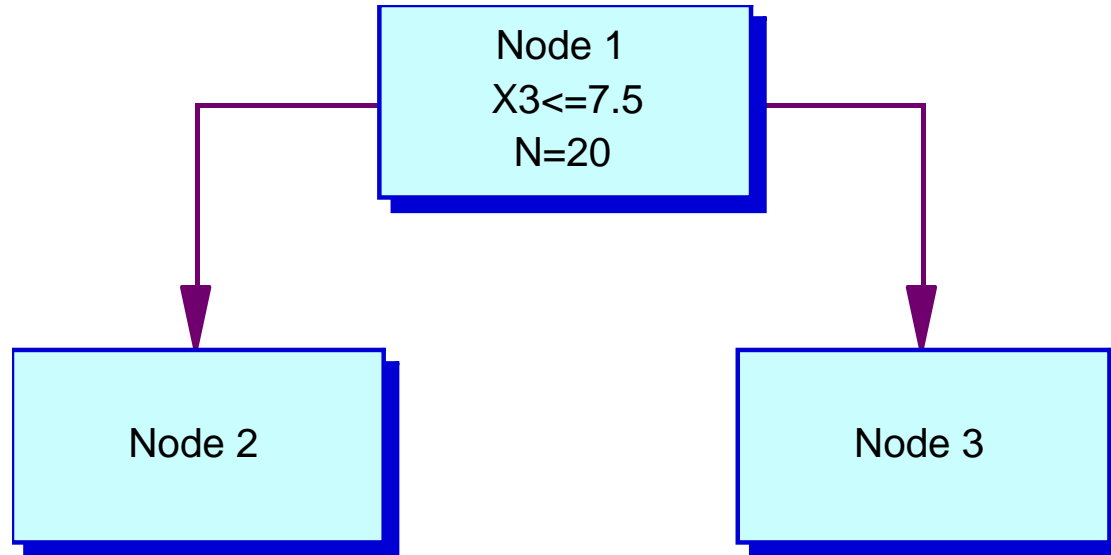
$$\begin{aligned}\Delta i &= i(r) - p_1 i(1) - p_2 i(2) \\ &= 0.40 - \left(\frac{11}{20}\right)(0.6011) - \left(\frac{9}{11}\right)(0.5679) \\ &= 0.036\end{aligned}$$

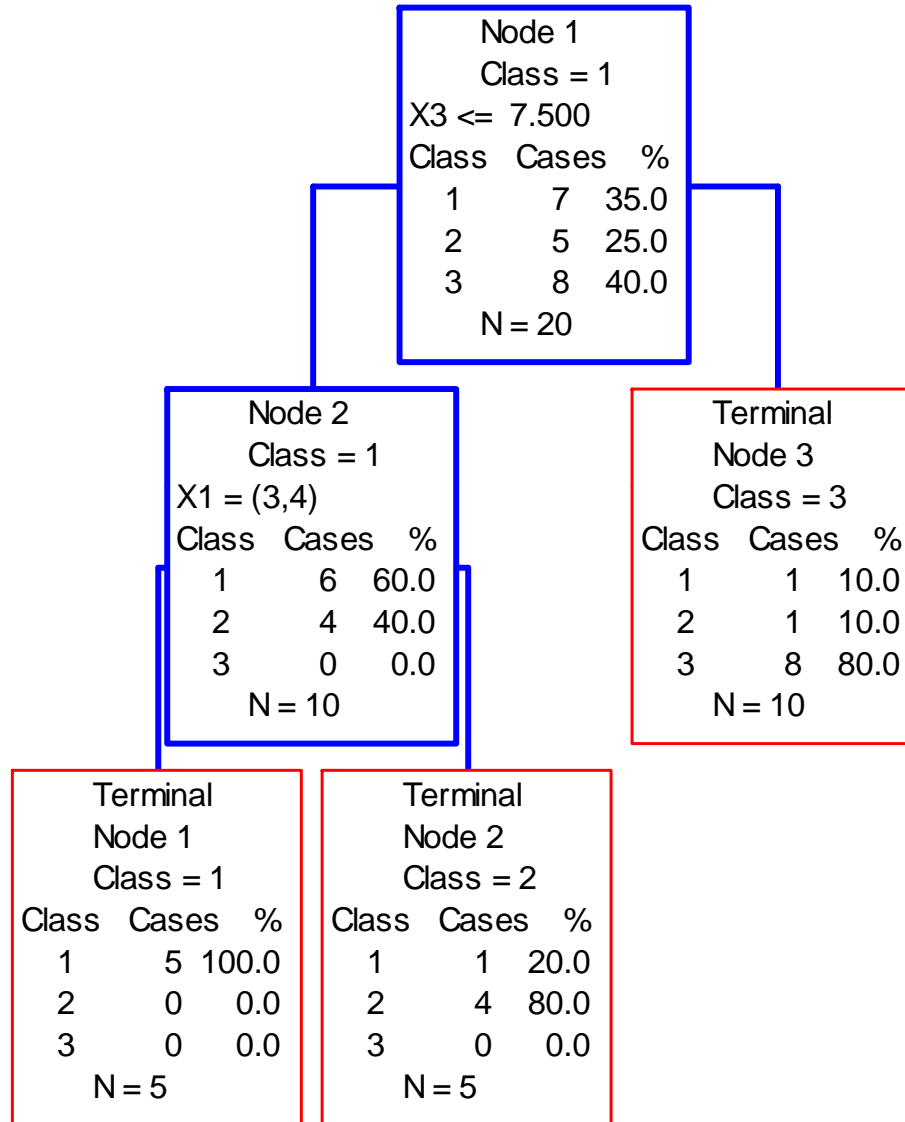


Splitting Variable	Split	Δi
X_1	1-234	0.015
	2-134	0.043
	3-124	0.082
	4-123	0.149
	12-34	0.087
	13-24	0.065
	14-23	0.061
X_2	1-2	0.036
X_3	≤ 4.5	0.034
	≤ 5.5	0.161
	≤ 6.5	0.231
	≤ 7.5	0.395
	≤ 8.5	0.048
	≤ 9.5	0.011

Splitting variable and split value \longrightarrow

Biggest decrease in impurity \longleftarrow







The splitting process is continued until:

- A node is homogeneous.
- A certain criteria is met (e.g., $n \leq 5$).
- The problem with this process is that you will get a tree that “overfits” the data.
- Solution: Prunning



Pruning

1. A “maximal” tree is grown.
2. A set of subtrees is obtained by cutting down branches (pruning).
3. For each tree the misclassification rate is computed.
4. The “best” tree is selected based on the misclassification error.



Misclassification Error: Naive

- Proportion of misclassified cases.

$$\begin{aligned} m.e. &= \frac{0 + 1 + 2}{20} \\ &\approx 0.15 \\ &= 15\% \end{aligned}$$



Misclassification Error

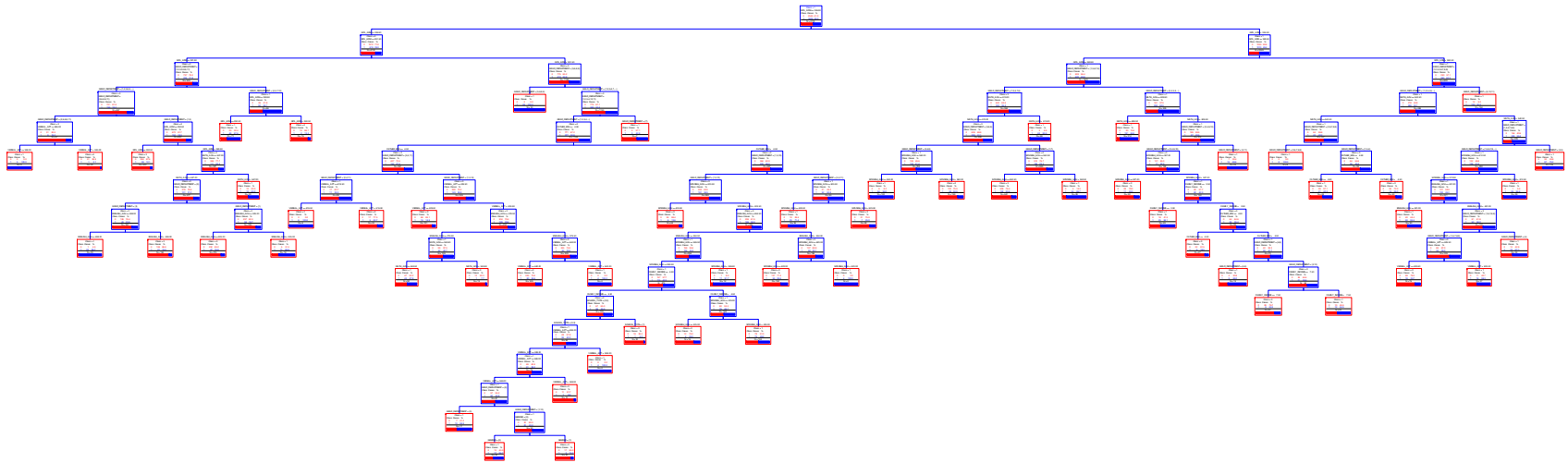
- A more accurate value is obtained using:
 - A test sample
 - Crossvalidation



Example: Predicting Time to Degree Completion (N = 5,240)

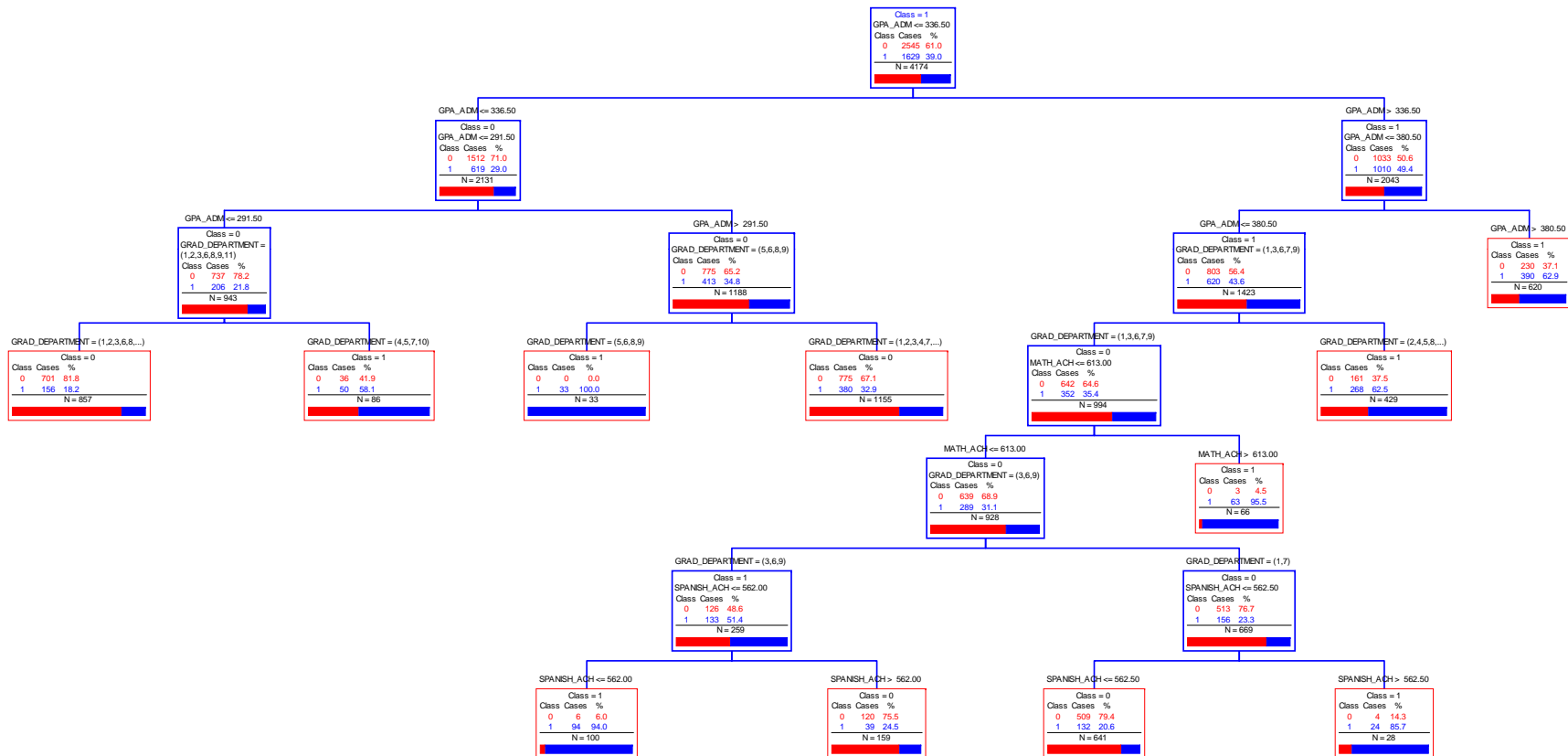
- Time to degree:
 - 6.5 years or less
 - More than 6.5 years
- Predictors
 - High School GPA
 - Math Achievement (CEEB)
 - Math Aptitude (CEEB)
 - Verbal Aptitude (CEEB)
 - English Achievement (CEEB)
 - Spanish Achievement (CEEB)
- Predictors (Cont.)
 - Father's education
 - Mother's education
 - Family income
 - Type of school (private, public)
 - Actual program of study

Classification tree



Misclassification error (independent sample) = 23.4%

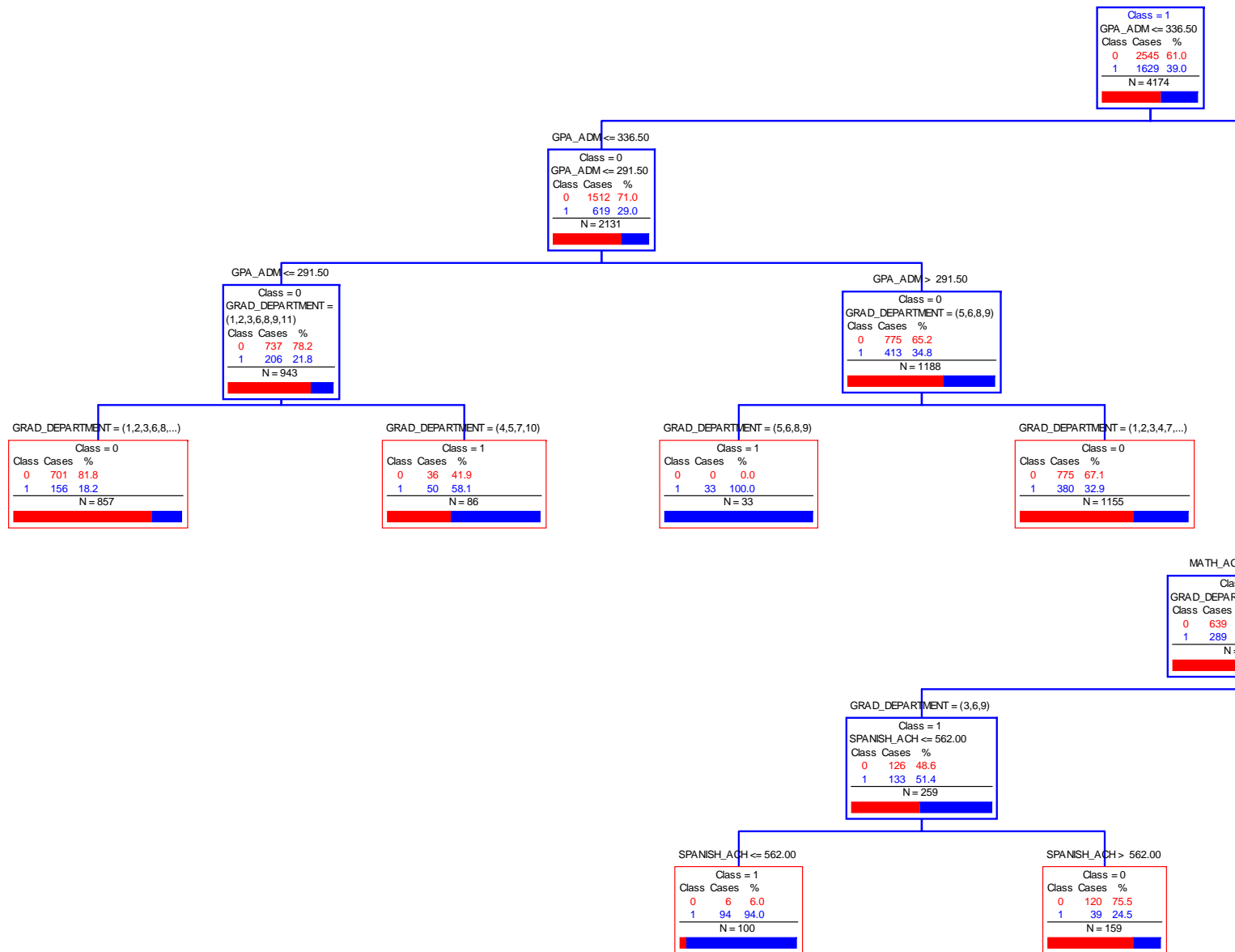
Classification tree

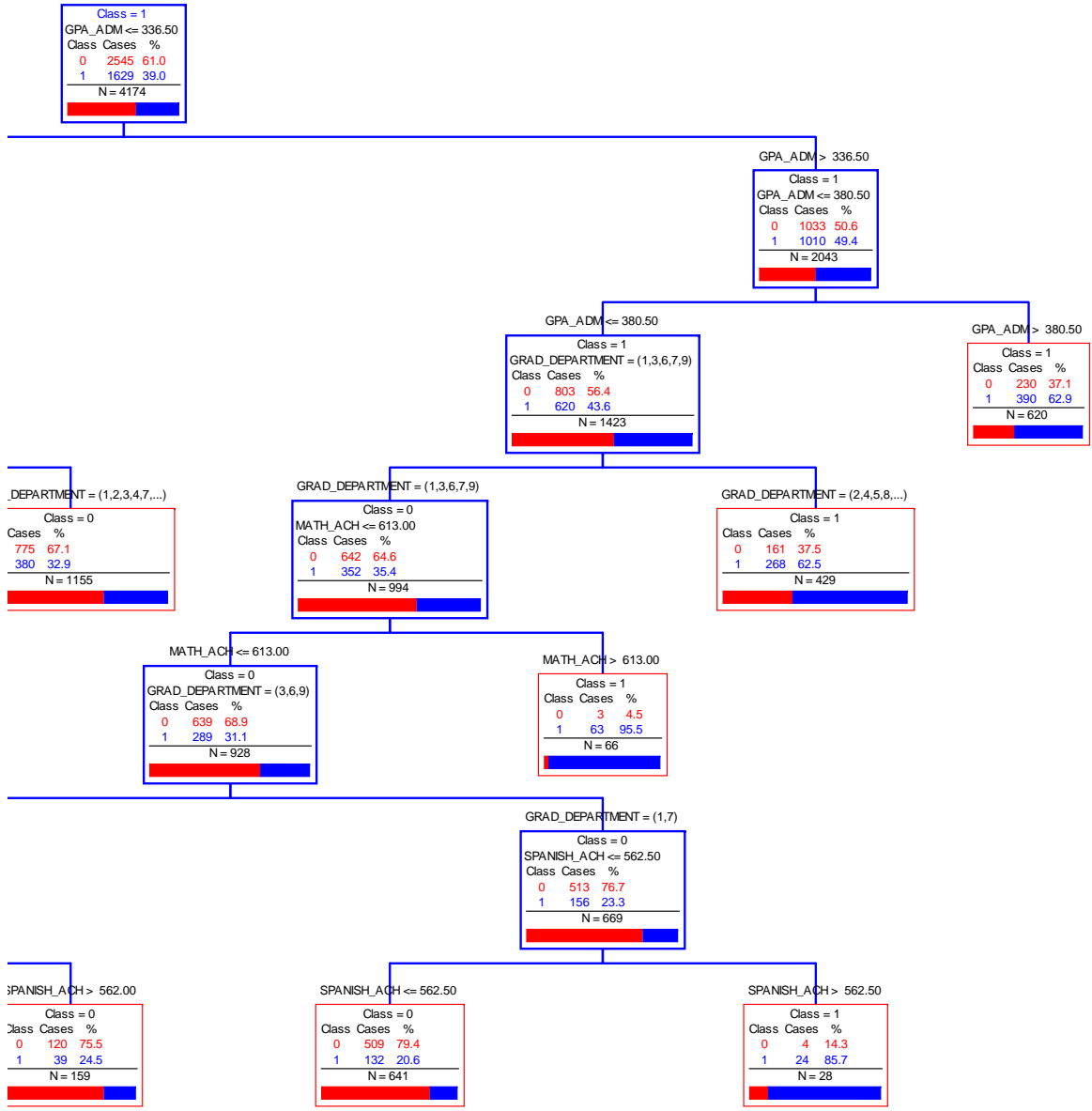


Root node

Class = 1		
GPA_ADM <= 3.365		
Class	Cases	%
0	2545	61.0
1	1629	39.0
<hr/>		
N = 4174		









Variable Importance

Variable	Score	
DEPARTMENT	100.00	
GPA_ADM	92.08	
SPANISH_ACH	85.81	
MATH_ACH	77.13	
VERBAL_APT	39.79	
ENGLISH_ACH	38.63	
MATH_APT	29.64	
FATHER_EDU	10.70	
FAMILY_INCOME	8.80	
SCHOOL_TYPE	5.73	
GENDER	2.57	
MOTHER_EDU	2.07	
FIRST_GENERATION	2.06	



Some findings

- Overall 49% of students complete degree in 150% time or less.
- If Admission GPA > 3.80 , 63% will complete degree.
- If Admission GPA is between 3.37 and 3.80, 43.6% will complete degree.
 - and department is Business Adm, Social Sciences, Hispanic Studies, English and Mathematics 62.5% will complete degree.
- If Admission GPA < 3.37 , 71% will not complete degree.
- If Admission GPA is between 2.92 and 3.36, and student Department is
 - Physical Ed, Hispanic Studies, English, and Mathematics, 100% (n = 33) will complete degree.
 - In other departments 67% (n = 33) will not complete degree.
- If Admission GPA < 2.92 , 78.2% will not complete degree.



Value of Higher Order Interactions

- **CART Analysis:**

- can be used as an alternative to regression, discriminant analysis and other parametric methods for prediction problems.
- can be used as a complement to these analyses providing a better understanding of the interaction between variables.

- **Recursive partitioning allows us to obtain predictions which are as good as, or better, those obtained by discriminant analysis or multiple regression.**



Value of Higher Order Interactions

- Higher-order interactions provide stronger empirical evidence for developing institutional policies regarding student admissions.
- Offers Institutional Research and Assessment Offices an alternative model for identifying the underlying variables, and their interactions, affecting today's academic environment.



Value of Recursive Partitioning to Institutional Research

- Higher-order interactions provide for designing high quality intervention plans to increase retention and graduation rates, and promote academic success.
- CART analyses allows researchers to identify different “at-risk” scenarios, leading to the assignment of responsibility to those institutional units who can tend to these populations.



Value of Recursive Partitioning to Institutional Stakeholders

- The use of Recursive Partitioning offers the administration, faculty, counselors, and other stakeholders data that will help identify groups, such as at-risk students, so that specific interventions can be developed in order to help them succeed in college.
- Institutional stakeholders get a better grasp at the importance that institutional research has both to the institution as a whole and their individual units.



Questions?



Contact Information:

José Noel Caraballo

jcaraballo@cayey.upr.edu

Irmannette Torres-Lugo

irmannette.torres@upr.edu