Module 1
Lesson 1 - What Is Machine Learning?


Machine Learning
Describe Associations
Search For Patterns
Make Predictions

Accuracy = Test Error Rate

Goal: find a model that minimizes test error rate

Module 1

Lesson 2 - Machine Learning and the Bias Variance Trade-Off

Linear Regression

Accuracy=mean squared error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2$$

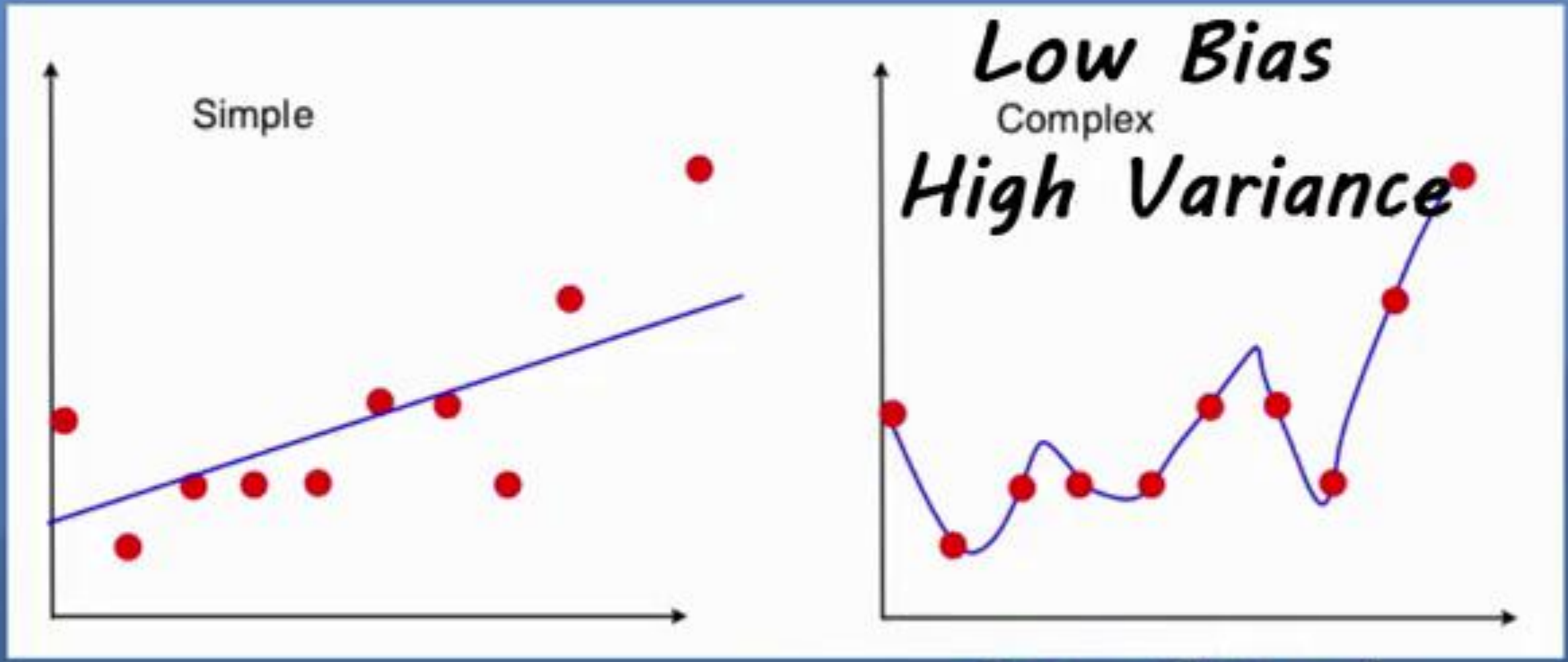
Variance = change in parameter estimates across different data sets

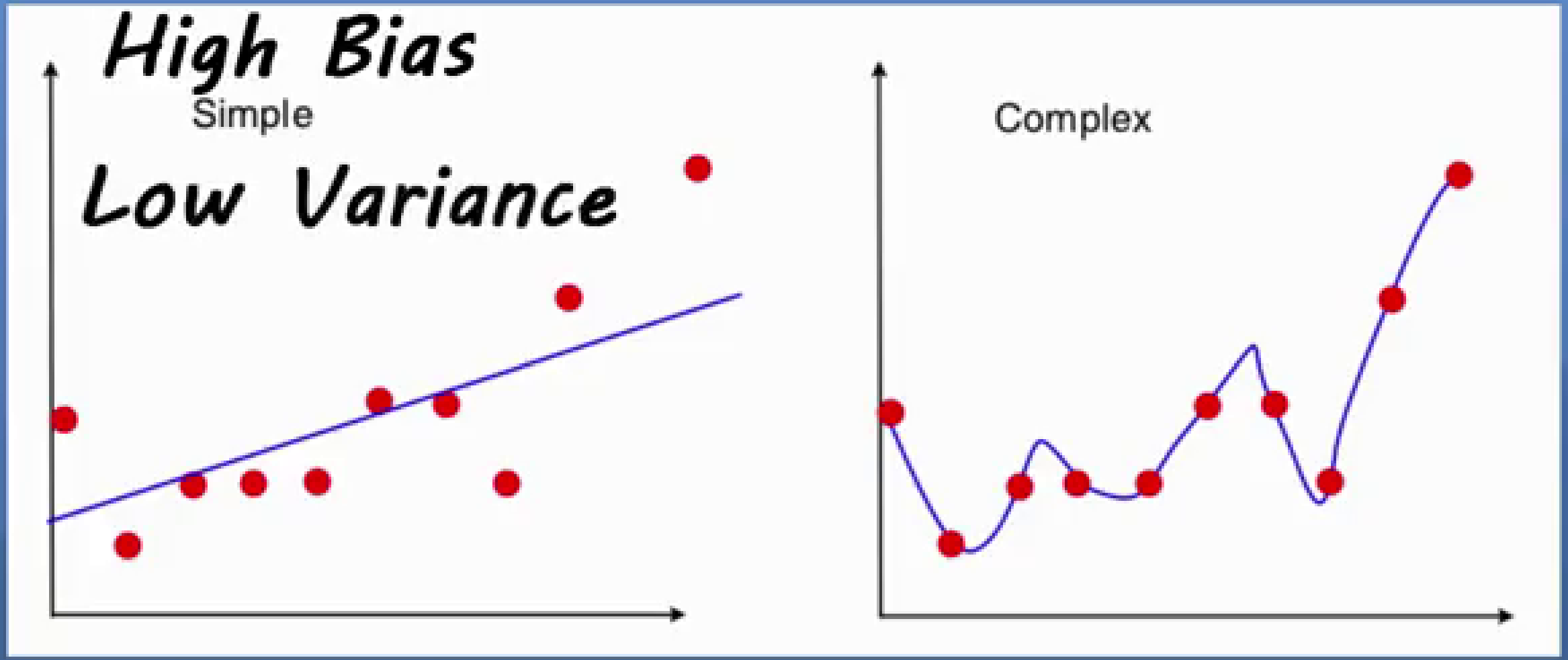
Bias = how far off model estimated values are from true values


Low Variance ✓
Low Bias ✓

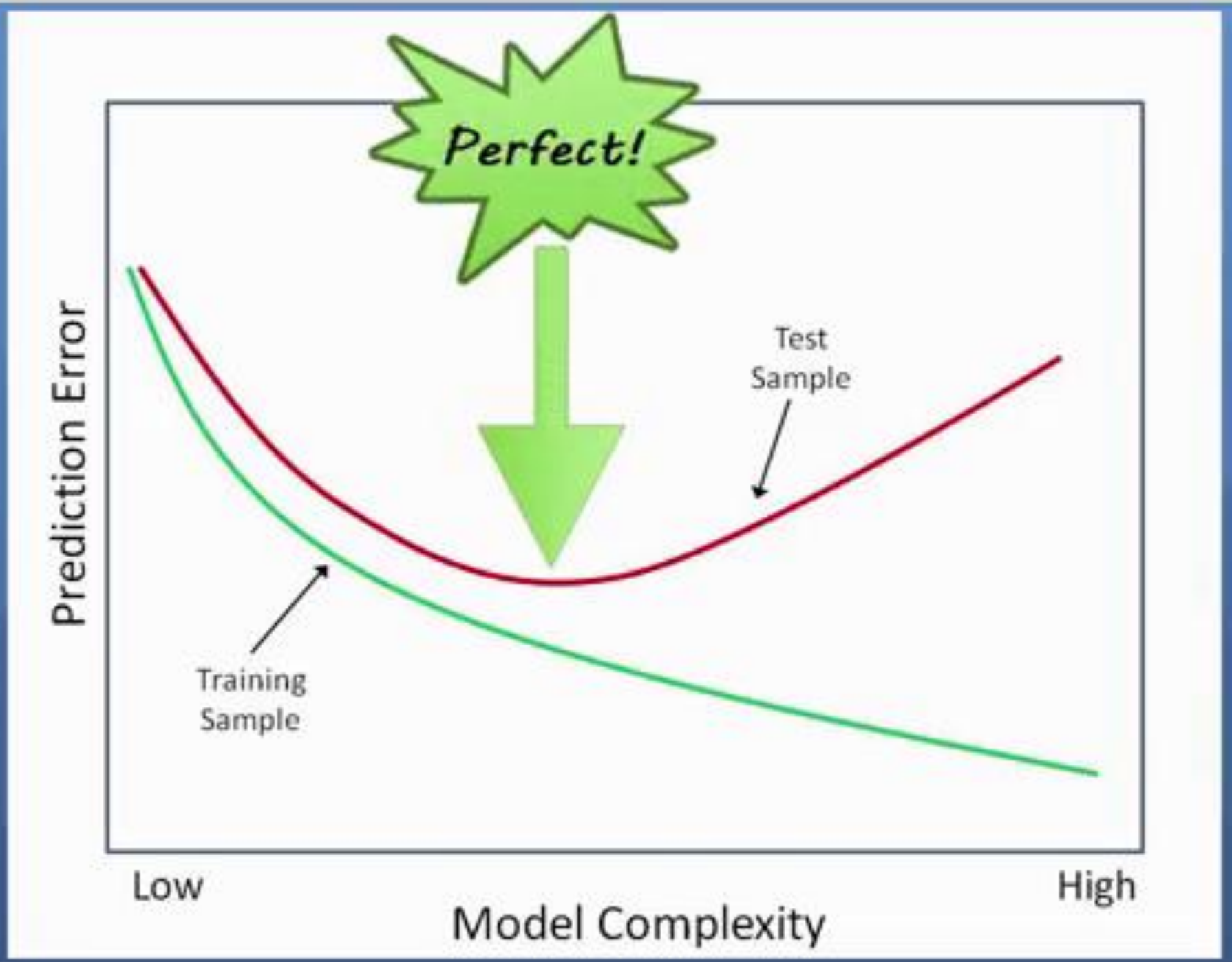

Complexity = High Variance/Low Bias


Complexity = Low Variance/High Bias


Simple
Low Bias
Complex
High Variance
Overfitted



Underfitted


Perfect!
Test Sample
Training Sample
Prediction Error
Low
High
Model Complexity

Logistic Regression

Accuracy = How well a model correctly classifies observations

Confusion Matrix

Test Sample Nicotine Dependence Classification

Actual

Predicted	Yes	No	Total
Yes	184	91	275
No	32	685	717
Total	216	776	992

Confusion Matrix

Test Sample Nicotine Dependence Classification

		Actual		
		Yes	No	Total
Predicted	Yes	184 ✓	91 ✗	275
	No	32 ✗	685 ✓	717
	Total	216	776	992

Confusion Matrix

Test Sample Nicotine Dependence Classification

Predicted \ Actual	Yes	No	Total
Yes	184	91	275
No	32	685	717
Total	216	776	992

91 + 32=123 incorrectly classified

Confusion Matrix

Test Sample Nicotine Dependence Classification

Predicted \ Actual	Yes	No	Total
Yes	184	91	275
No	32	685	717
Total	216	776	992

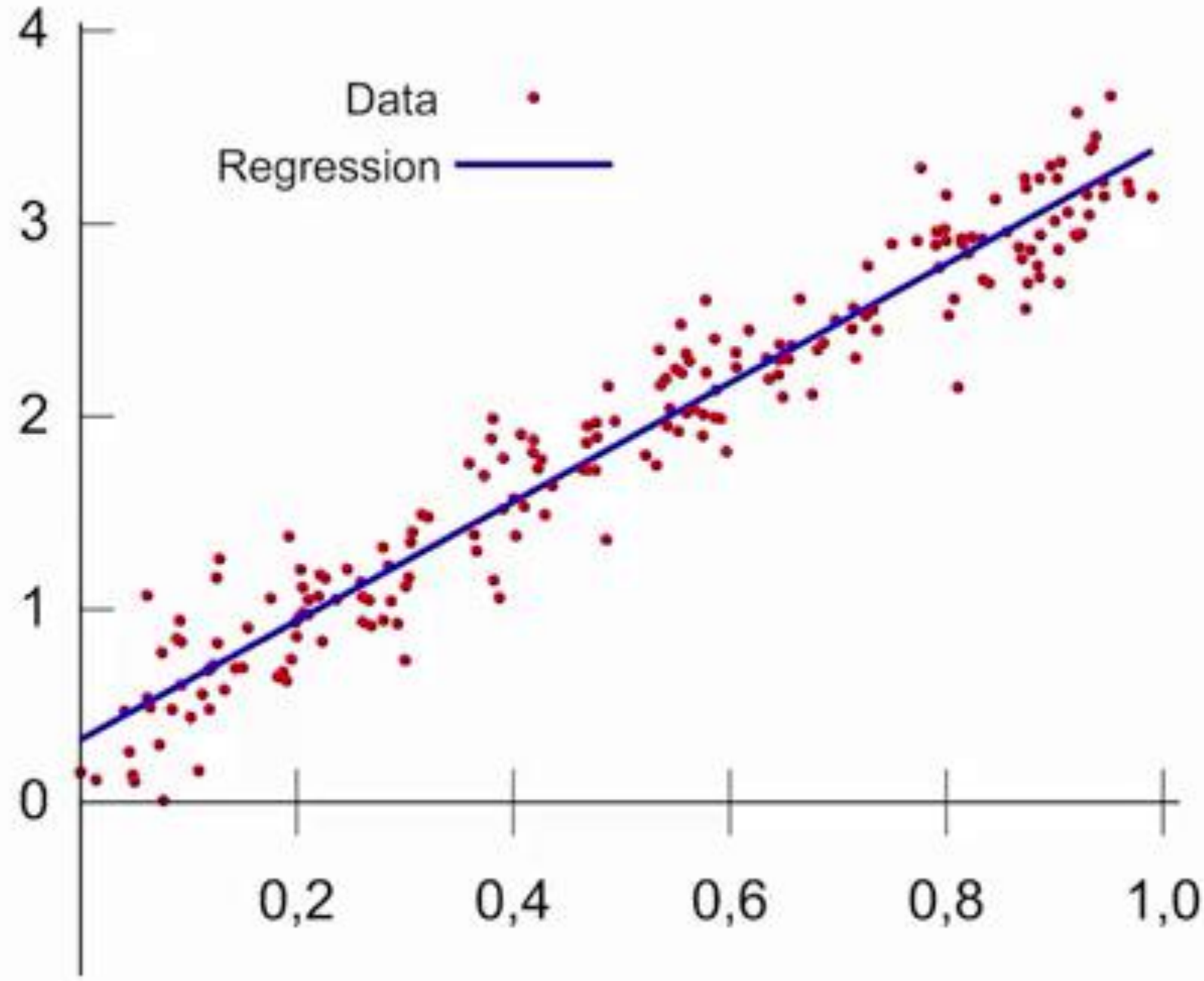
Test error rate = % misclassified = 12%

WESLEYAN
UNIVERSITY

Decision Trees

What is a Decision Tree?

with Professor Lisa Dierker


Data
Regression
4
3
2
1
0
0,2
0,4
0,6
0,8
1,0

Supervised Prediction

linear regression

pattern recognition

discriminant analysis

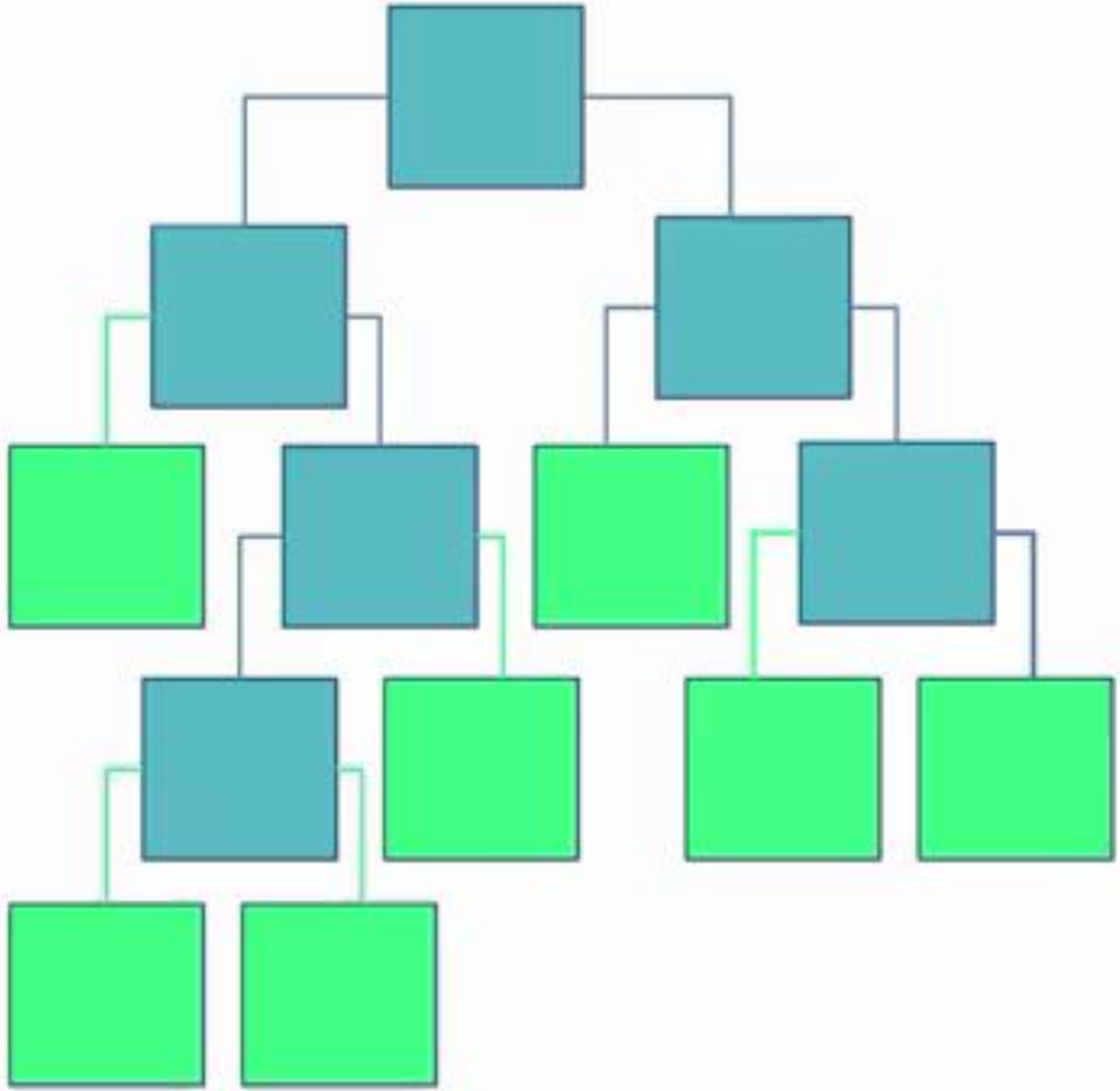
multivariate function estimation

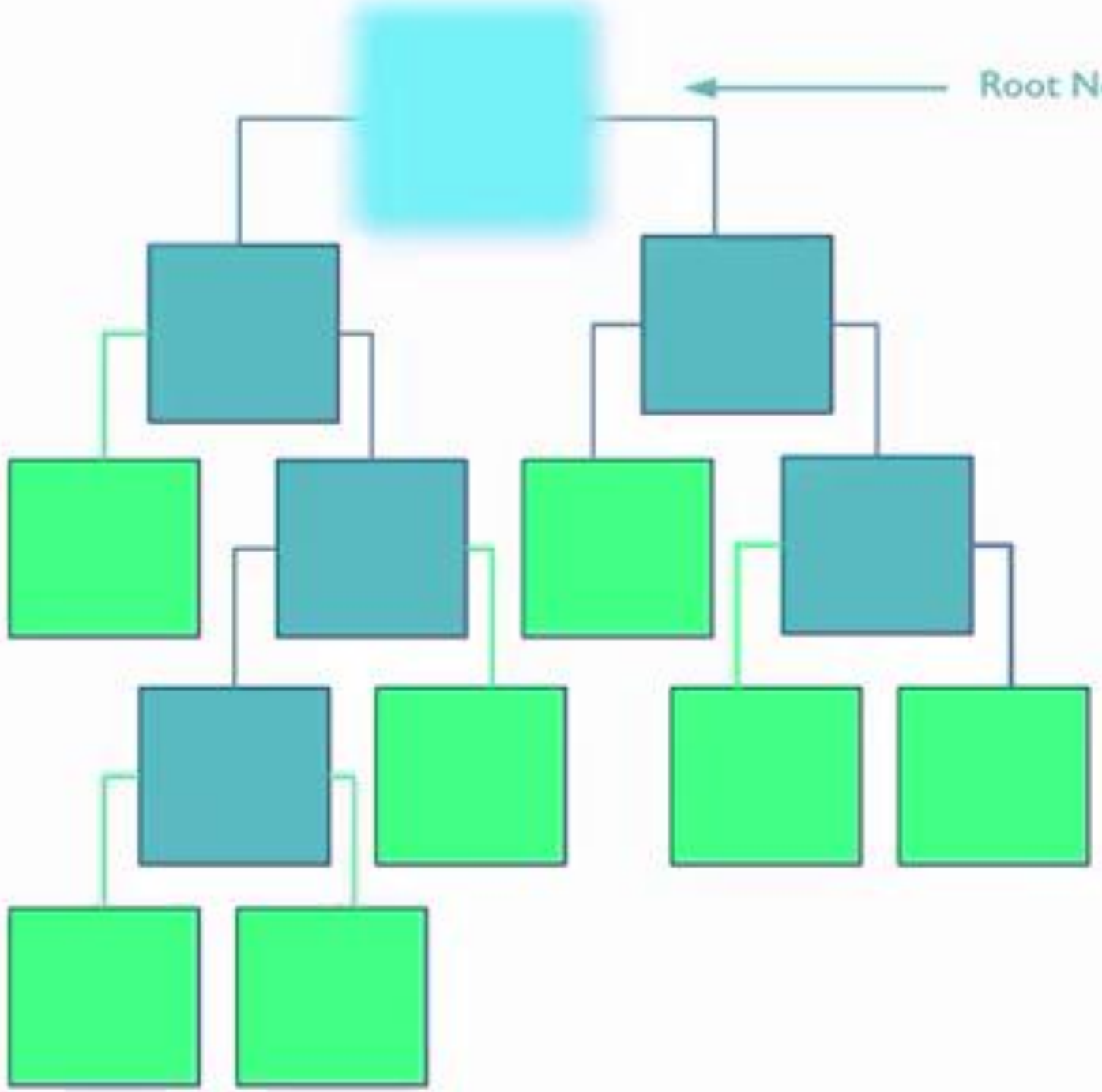
supervised machine learning techniques

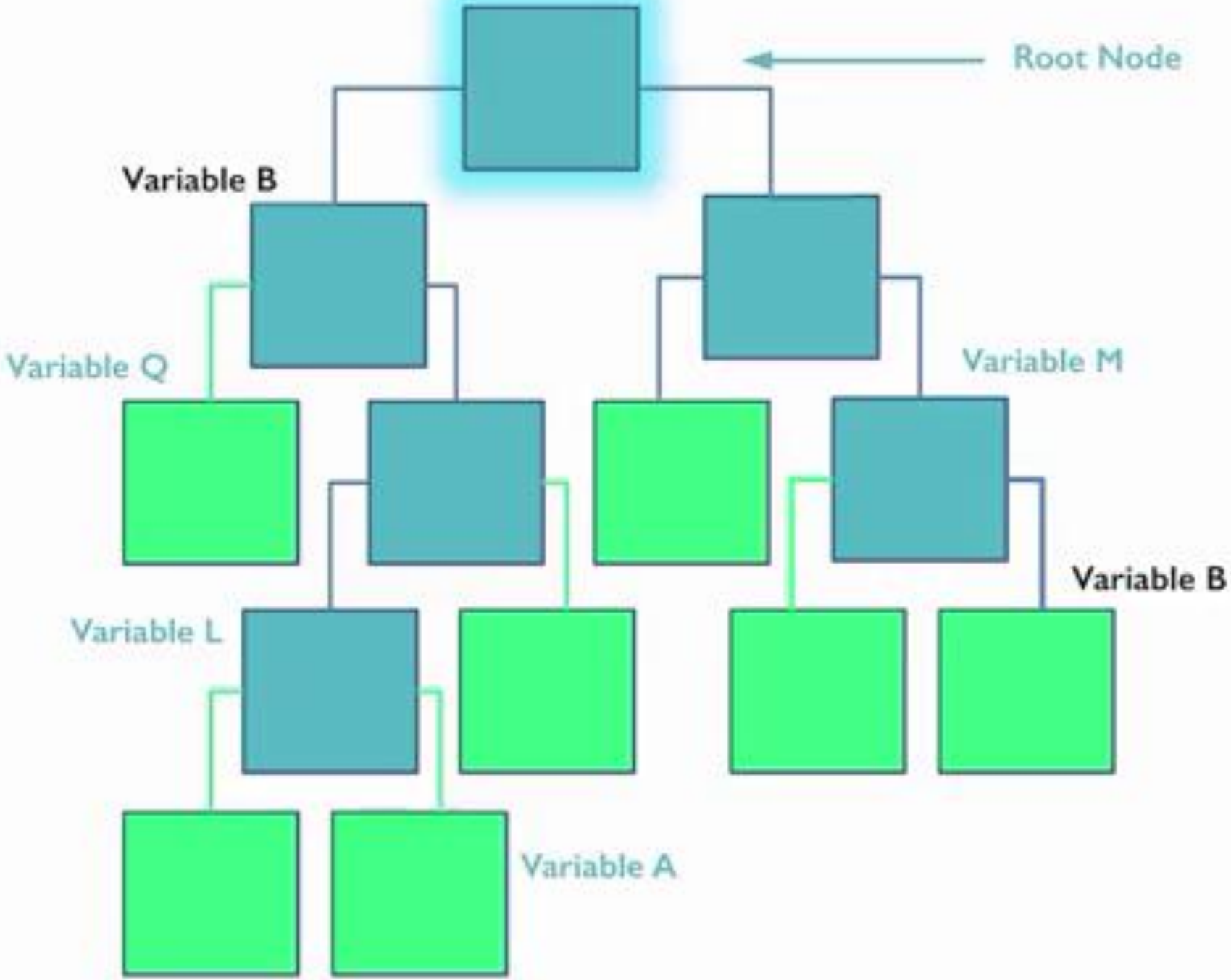
Supervised Prediction

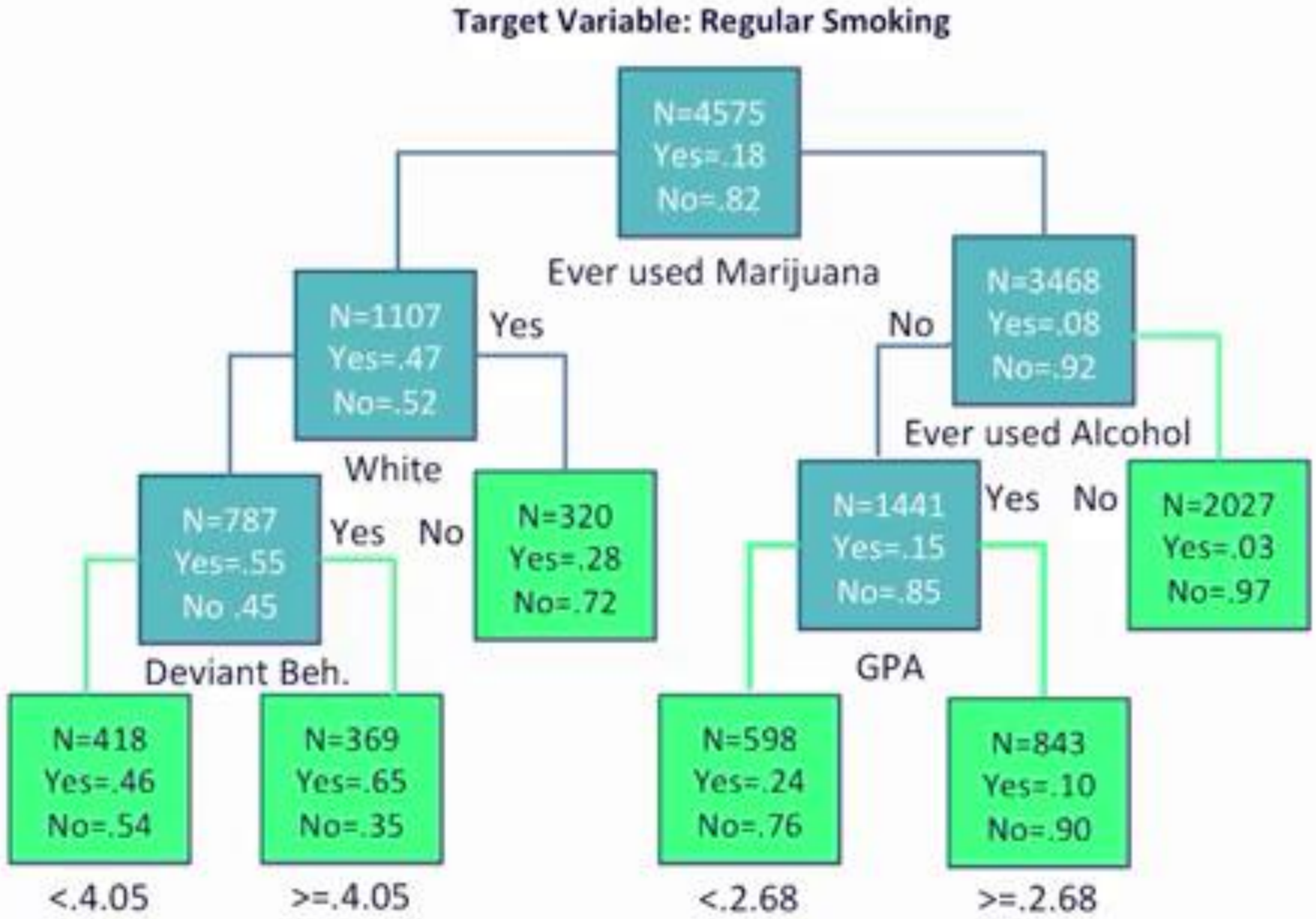
explanatory variables

response variable




Root Node


Root Node
Variable B
Variable Q
Variable M
Variable B
Variable L
Variable A


Target Variable: Regular Smoking
N=4575
Yes=.18
No=.82
Ever used Marijuana
Yes
No
N=1107
Yes=.47
No=.52
N=3468
Yes=.08
No=.92
White
Ever used Alcohol
Yes
No
N=787
Yes=.55
No .45
N=320
Yes=.28
No=.72
N=1441
Yes=.15
No=.85
N=2027
Yes=.03
No=.97
Deviant Beh.
GPA
N=418
Yes=.46
No=.54
N=369
Yes=.65
No=.35
N=598
Yes=.24
No=.76
N=843
Yes=.10
No=.90
<.4.05
>=.4.05
<.2.68
>=.2.68

24 possible explanatory variables

target variable - regular smoking

target variable - regular smoking

"Have you (ever) smoked cigarettes regularly, that is, at least one cigarette every day for 30 days?"

24 possible explanatory variables

associated with regular smoking behavior in adolescents

Binary Categorical Variables

Gender
Hispanic
White
Black
Native American
Asian

Substance use measured with individual questions

Alcohol
Marijuana
Cocaine
Inhalants

Additional Categorical Variables

Availability of Cigarettes
Parents on Public Assistance
Expelled from School

Quantitative Variables

Age
Alcohol Problems
Deviant Behavior
Violent Behavior
Depression
Self Esteem
Parental Presence
Parental Activities
Family Connectedness
School Connectedness
Grade Point Average

For more complete details on how these variables were constructed see:

Dierker, et al., 2004 paper from Prevention Science

SAS program: "Decision Trees Data Management"

WESLEYAN
UNIVERSITY

Decision Trees

What is the Process of "Growing" a Decision Tree?

with Professor Lisa Dierker

Growing the Tree

Binary splits maximize correct classification

All cut-points are tested

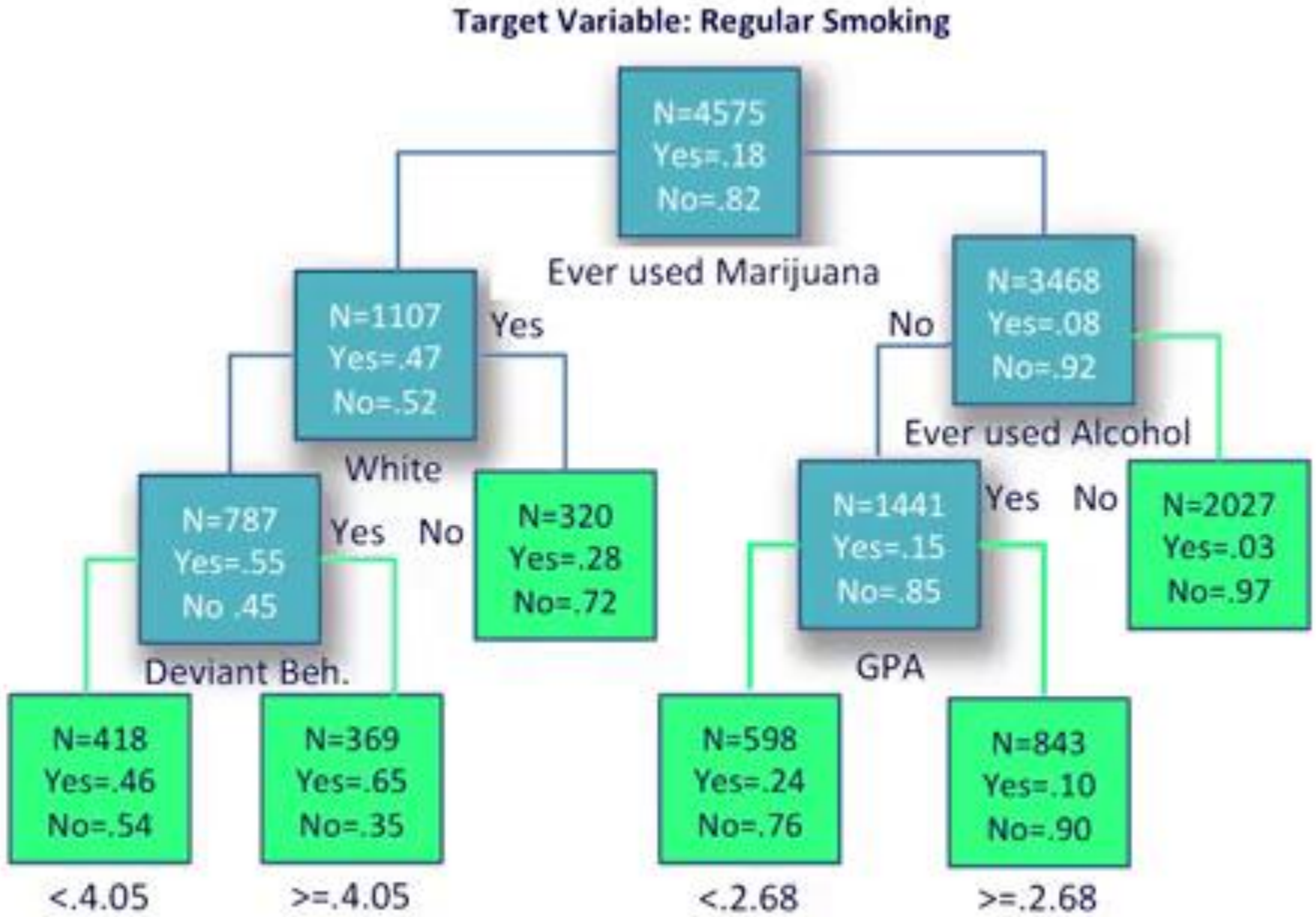
Subgroups showing similar outcomes are generated

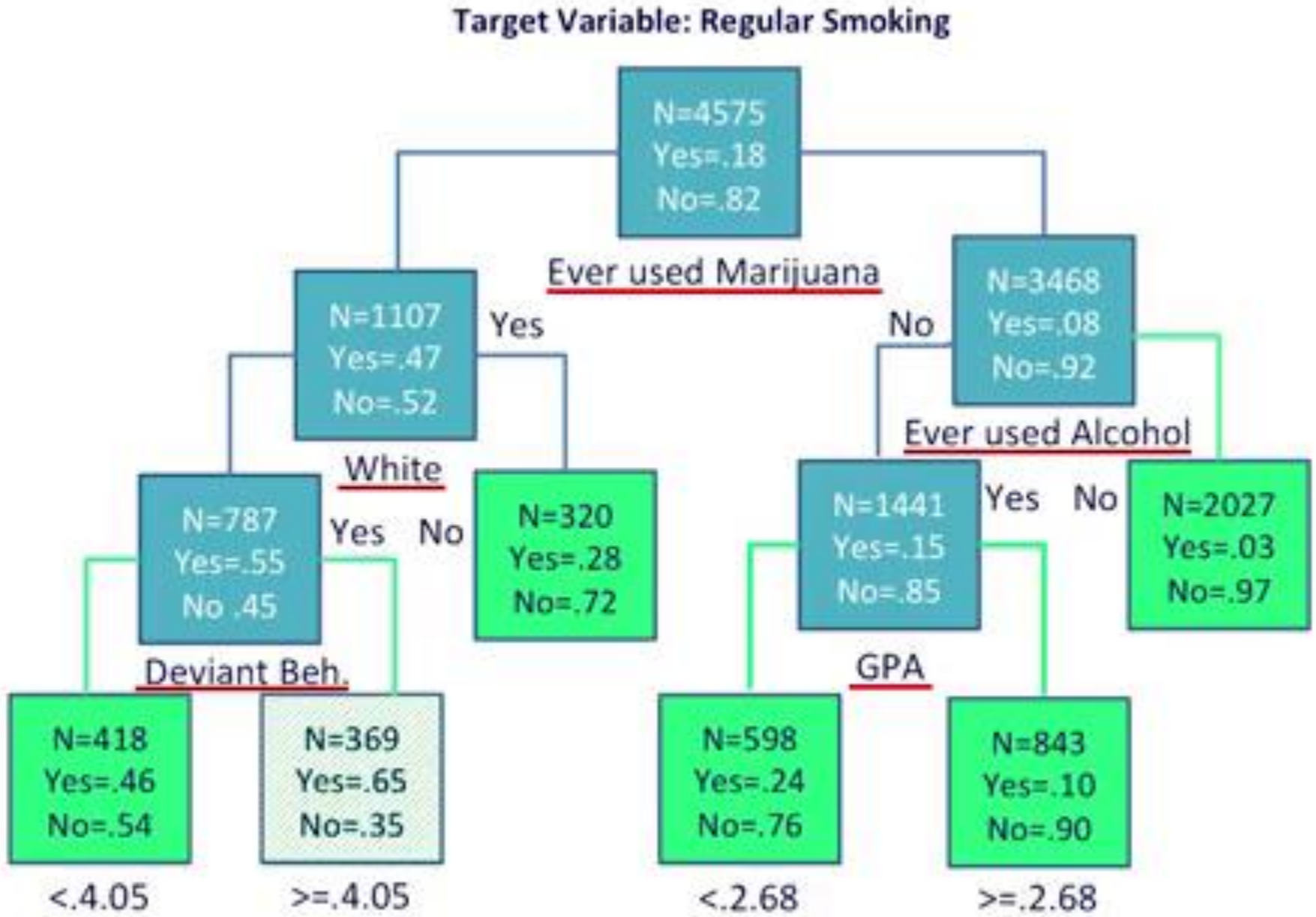
Validating the Tree

Cross-validation guards againts overfit

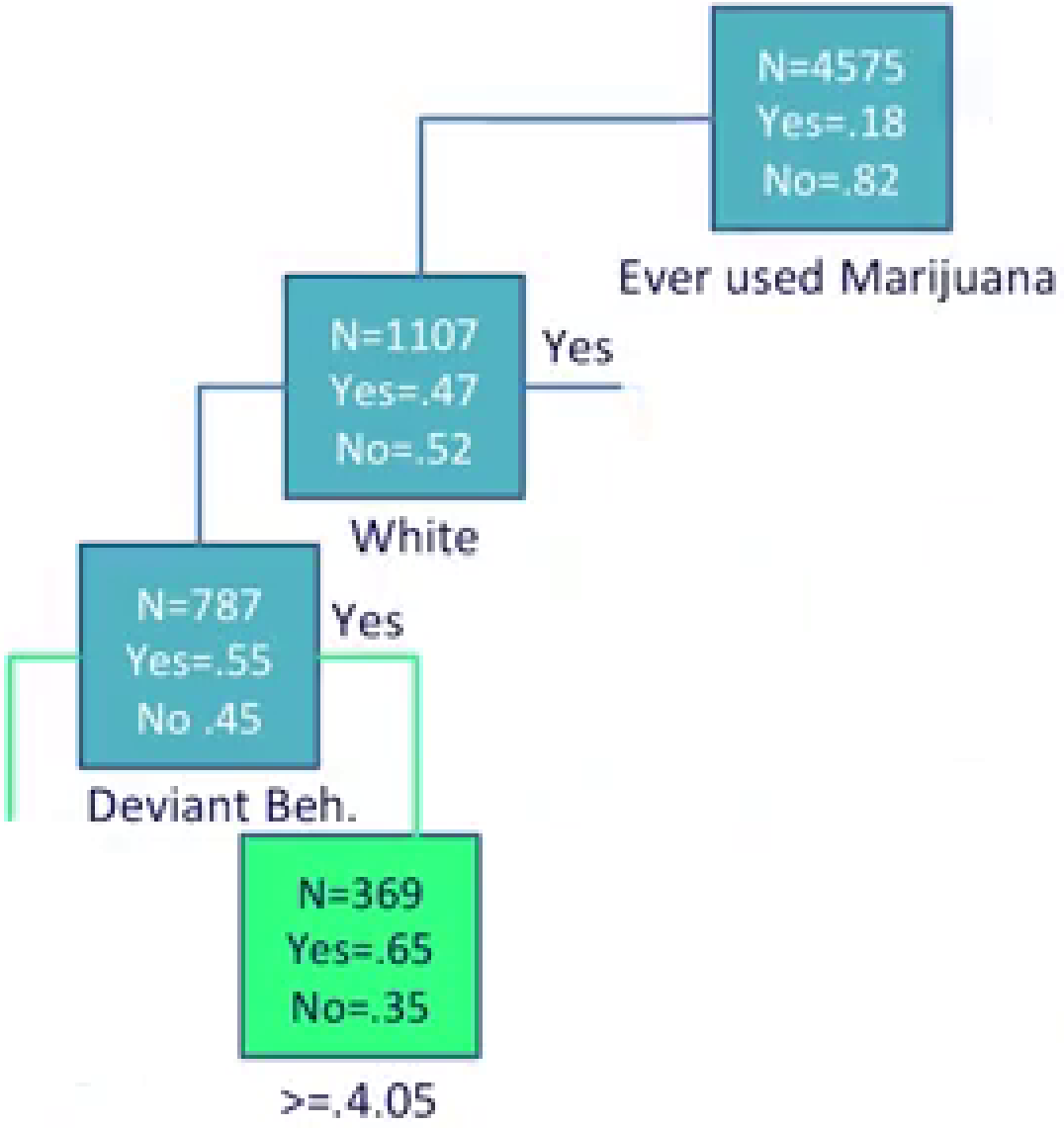
A random subset is tested and only "branches" that improve the classification are retained

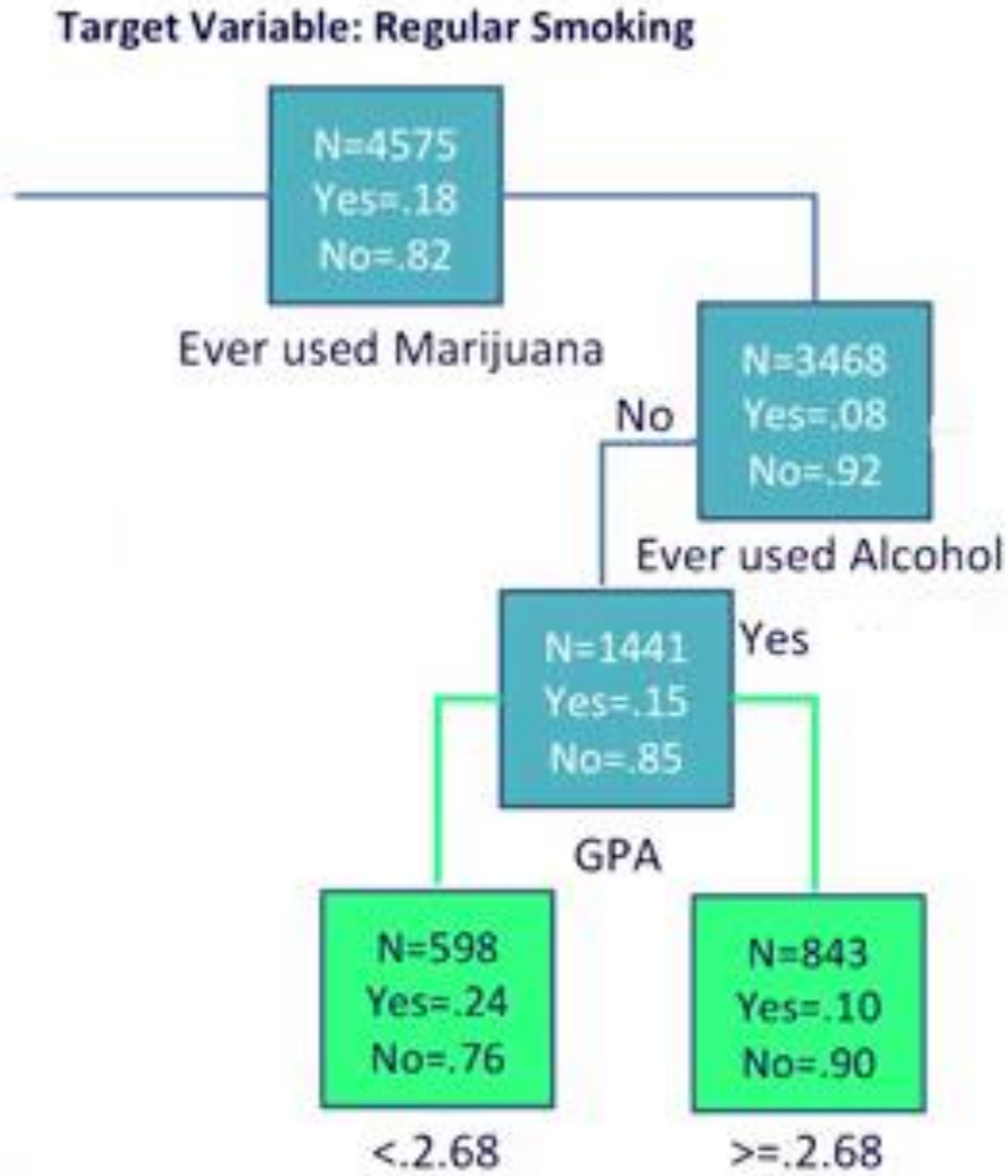
Selected sub-tree is the lowest probability of misclassification

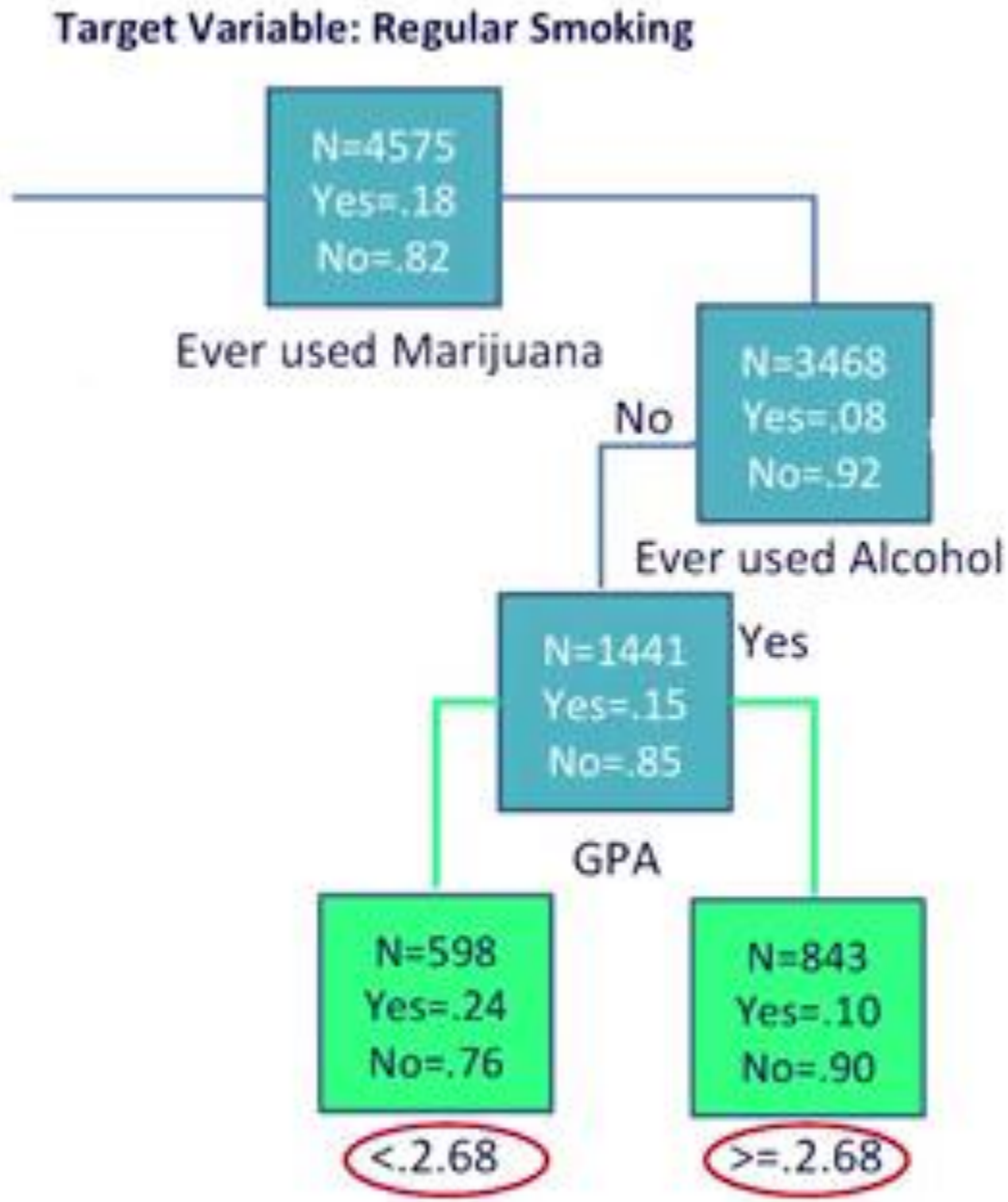

Target Variable: Regular Smoking
N=4575
Yes=.18
No=.82
Ever used Marijuana
Yes
No
N=1107
Yes=.47
No=.52
N=3468
Yes=.08
No=.92
White
Ever used Alcohol
N=787
Yes=.55
No .45
Yes
No
N=320
Yes=.28
No=.72
N=1441
Yes=.15
No=.85
Yes
No
N=2027
Yes=.03
No=.97
Deviant Beh.
GPA
N=418
Yes=.46
No=.54
N=369
Yes=.65
No=.35
N=598
Yes=.24
No=.76
N=843
Yes=.10
No=.90
<.4.05
>=.4.05
<.2.68
>=.2.68


Target Variable: Regular Smoking
N=4575
Yes=.18
No=.82
Ever used Marijuana
Yes
No
N=1107
Yes=.47
No=.52
N=3468
Yes=.08
No=.92
White
Ever used Alcohol
Yes
No
N=787
Yes=.55
No .45
N=320
Yes=.28
No=.72
N=1441
Yes=.15
No=.85
N=2027
Yes=.03
No=.97
Deviant Beh.
GPA
N=418
Yes=.46
No=.54
N=369
Yes=.65
No=.35
N=598
Yes=.24
No=.76
N=843
Yes=.10
No=.90
<.4.05
>=.4.05
<.2.68
>=.2.68

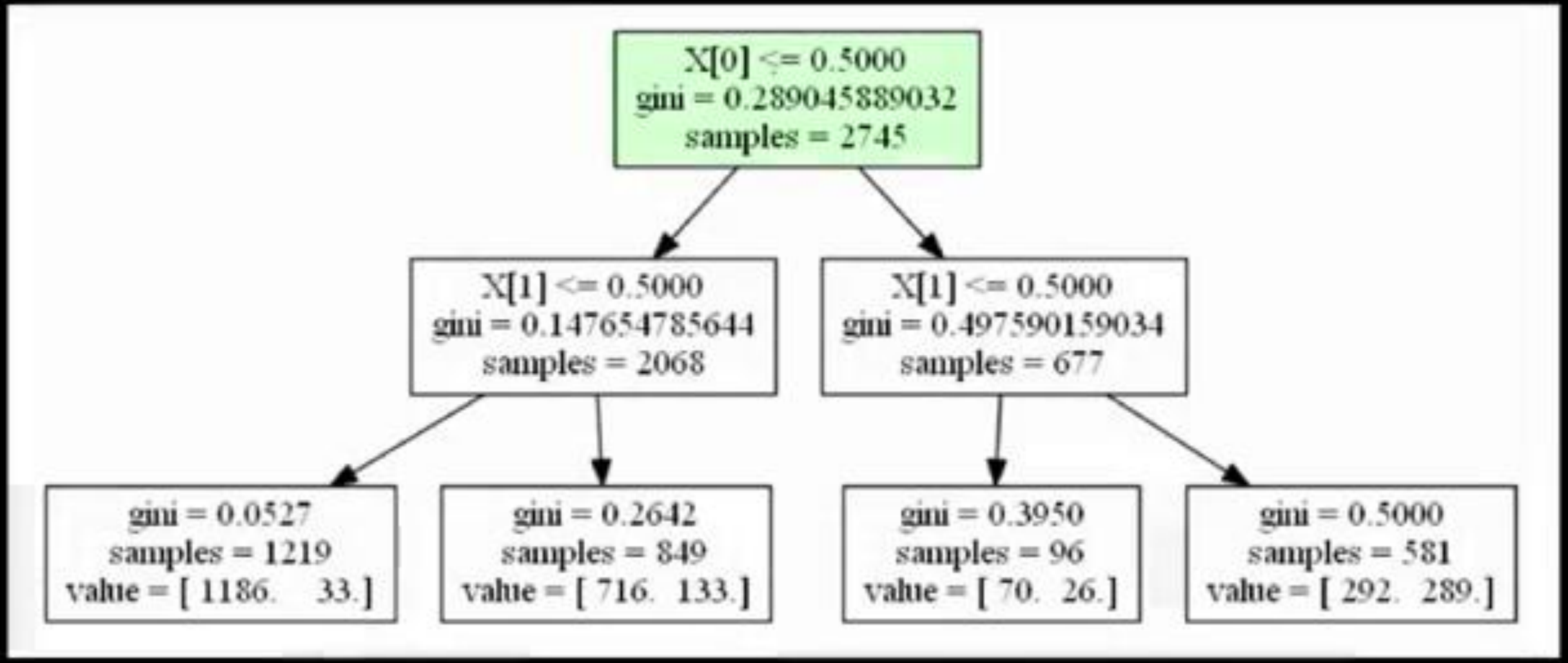
Target Variable: Regular Smoking

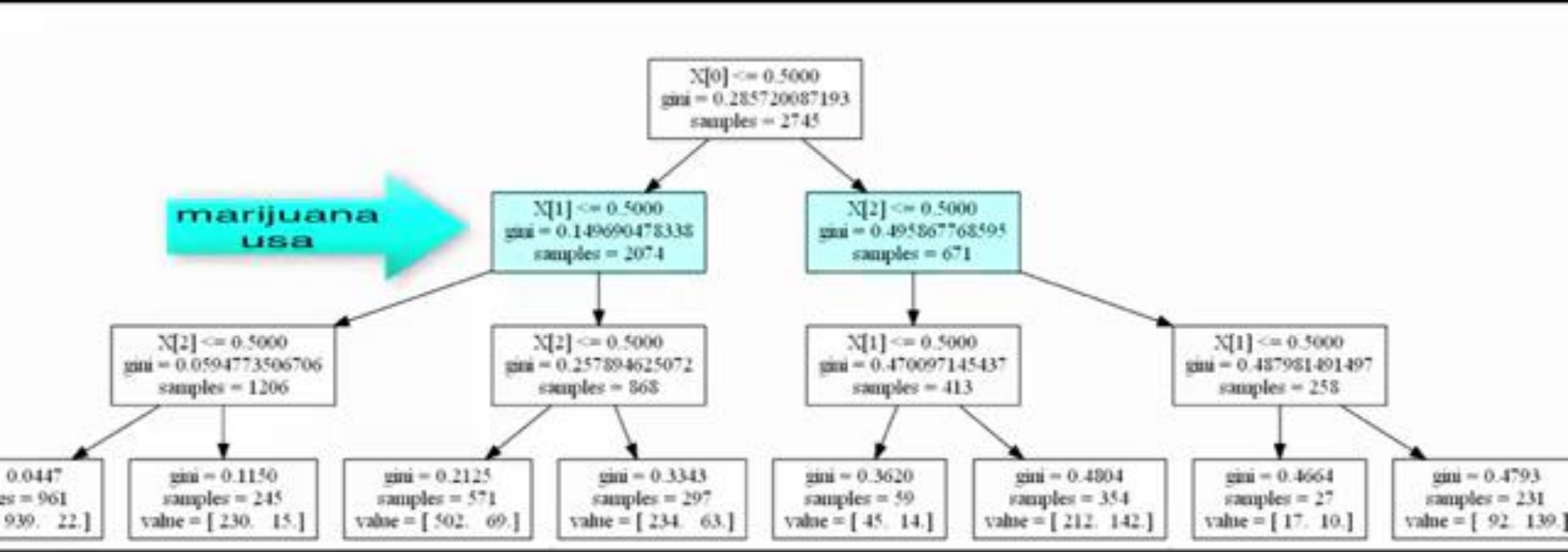



Target Variable: Regular Smoking
N=4575
Yes=.18
No=.82
Ever used Marijuana
No
N=3468
Yes=.08
No=.92
Ever used Alcohol
Yes
N=1441
Yes=.15
No=.85
GPA
N=598
Yes=.24
No=.76
<.2.68
N=843
Yes=.10
No=.90
>=.2.68


Target Variable: Regular Smoking
N=4575
Yes=.18
No=.82
Ever used Marijuana
No
N=3468
Yes=.08
No=.92
Ever used Alcohol
Yes
N=1441
Yes=.15
No=.85
GPA
N=598
Yes=.24
No=.76
<.2.68
N=843
Yes=.10
No=.90
>=.2.68

Module 1
Lesson 5 - Building a Decision Tree in Python


X[0] <= 0.5000
gini = 0.289045889032
samples = 2745
X[1] <= 0.5000
gini = 0.147654785644
samples = 2068
X[1] <= 0.5000
gini = 0.497590159034
samples = 677
gini = 0.0527
samples = 1219
value = [1186. 33.]
gini = 0.2642
samples = 849
value = [716. 133.]
gini = 0.3950
samples = 96
value = [70. 26.]
gini = 0.5000
samples = 581
value = [292. 289.]


X[0] <= 0.5000
gini = 0.285720087193
samples = 2745
marijuana
usa
X[1] <= 0.5000
gini = 0.149690478338
samples = 2074
X[2] <= 0.5000
gini = 0.495867768595
samples = 671
X[2] <= 0.5000
gini = 0.0594773506706
samples = 1206
X[2] <= 0.5000
gini = 0.257894625072
samples = 868
X[1] <= 0.5000
gini = 0.470097145437
samples = 413
X[1] <= 0.5000
gini = 0.487981491497
samples = 258
0.0447
es = 961
939. 22.]
gini = 0.1150
samples = 245
value = [230. 15.]
gini = 0.2125
samples = 571
value = [502. 69.]
gini = 0.3343
samples = 297
value = [234. 63.]
gini = 0.3620
samples = 59
value = [45. 14.]
gini = 0.4804
samples = 354
value = [212. 142.]
gini = 0.4664
samples = 27
value = [17. 10.]
gini = 0.4793
samples = 231
value = [92. 139.]

Strengths of Decision Trees

- Can select from among a large number of variables those and their interactions that are most important in determining the target or response variable to be explained.
- They are easy to interpret and visualize, especially when the tree is small.
- Can handle large data sets well and can predict both binary, categorical target variables (shown in our example) and also quantitative target variables (known as regression trees).

Limitations: Small changes in the data can lead to different splits and this can undermine the interpretability of the model.

Also decision trees are not very reproducible on future data!