MACHINE LEARNING WITH TREE-BASED MODELS IN R

Welcome to the course!

Erin LeDell & Gabriela de Queiroz
Machine Learning Scientist & Data Scientist

Tree-based models

- Interpretability + Ease-of-Use + Accuracy
- Make Decisions + Numeric Predictions

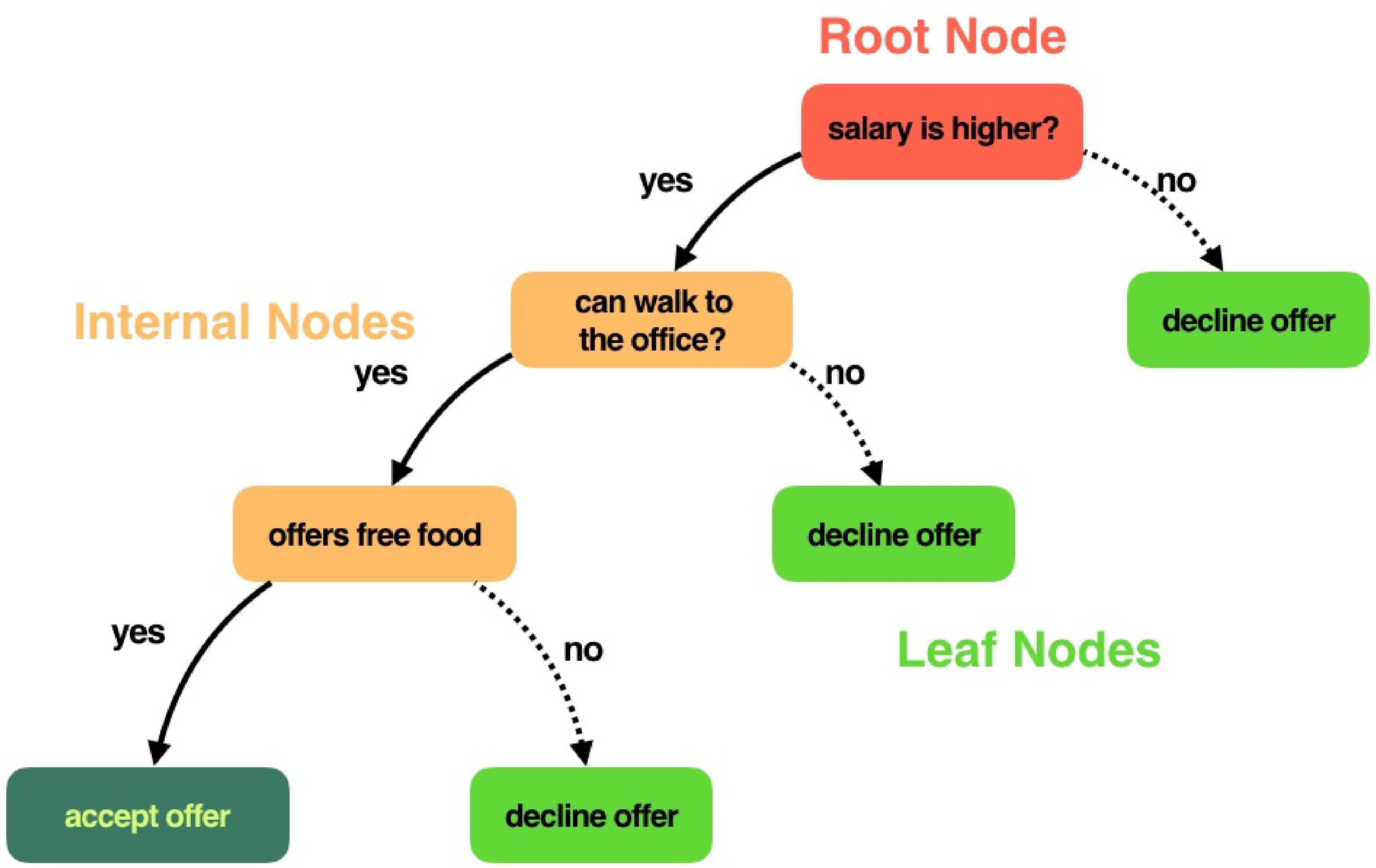
What you'll learn:

- Interpret and explain decisions
- Explore different use cases
- Build and evaluate classification and regression models
- Tune model parameters for optimal performance

We will cover:

- Classification & Regression Trees
- Bagged Trees
- Random Forests
- Boosted Trees (GBM)

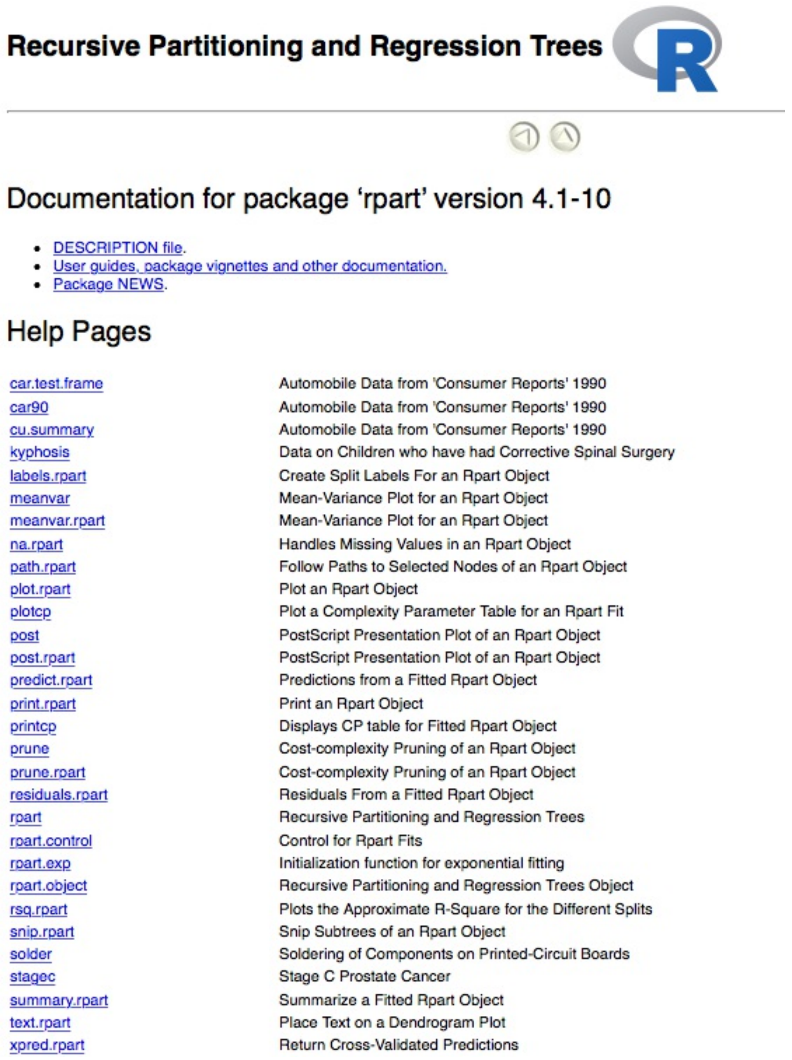
Decision tree terminology: nodes



Training Decision Trees in R

```
> library("rpart")
```

```
> help(package = "rpart")
```



Recursive Partitioning and Regression Trees

Documentation for package ‘rpart’ version 4.1-10

- DESCRIPTION file.
- User guides, package vignettes and other documentation.
- Package NEWS.

Help Pages

car.test.frame	Automobile Data from 'Consumer Reports' 1990
car90	Automobile Data from 'Consumer Reports' 1990
cu.summary	Automobile Data from 'Consumer Reports' 1990
kyphosis	Data on Children who have had Corrective Spinal Surgery
labels.rpart	Create Split Labels For an Rpart Object
meanvar	Mean-Variance Plot for an Rpart Object
meanvar.rpart	Mean-Variance Plot for an Rpart Object
na.rpart	Handles Missing Values in an Rpart Object
path.rpart	Follow Paths to Selected Nodes of an Rpart Object
plot.rpart	Plot an Rpart Object
plotcp	Plot a Complexity Parameter Table for an Rpart Fit
post	PostScript Presentation Plot of an Rpart Object
post.rpart	PostScript Presentation Plot of an Rpart Object
predict.rpart	Predictions from a Fitted Rpart Object
print.rpart	Print an Rpart Object
printcp	Displays CP table for Fitted Rpart Object
prune	Cost-complexity Pruning of an Rpart Object
prune.rpart	Cost-complexity Pruning of an Rpart Object
residuals.rpart	Residuals From a Fitted Rpart Object
rpart	Recursive Partitioning and Regression Trees
rpart.control	Control for Rpart Fits
rpart.exp	Initialization function for exponential fitting
rpart.object	Recursive Partitioning and Regression Trees Object
rsq.rpart	Plots the Approximate R-Square for the Different Splits
snip.rpart	Snip Subtrees of an Rpart Object
solder	Soldering of Components on Printed-Circuit Boards
stagec	Stage C Prostate Cancer
summary.rpart	Summarize a Fitted Rpart Object
text.rpart	Place Text on a Dendrogram Plot
xpred.rpart	Return Cross-Validated Predictions

Training Decision Trees in R

```
> rpart(response ~ ., data = dataset)
```

MACHINE LEARNING WITH TREE-BASED MODELS IN R

Let's practice!

MACHINE LEARNING WITH TREE-BASED MODELS IN R

Introduction to classification trees

Gabriela de Queiroz
Instructor

Advantages

- ✔ Simple to understand, interpret, visualize
- ✔ Can handle both numerical and categorical features (inputs) natively
- ✔ Can handle missing data elegantly
- ✔ Robust to outliers
- ✔ Requires little data preparation
- ✔ Can model non-linearity in the data
- ✔ Can be trained quickly on large datasets

Disadvantages

✖ Large trees can be hard to interpret

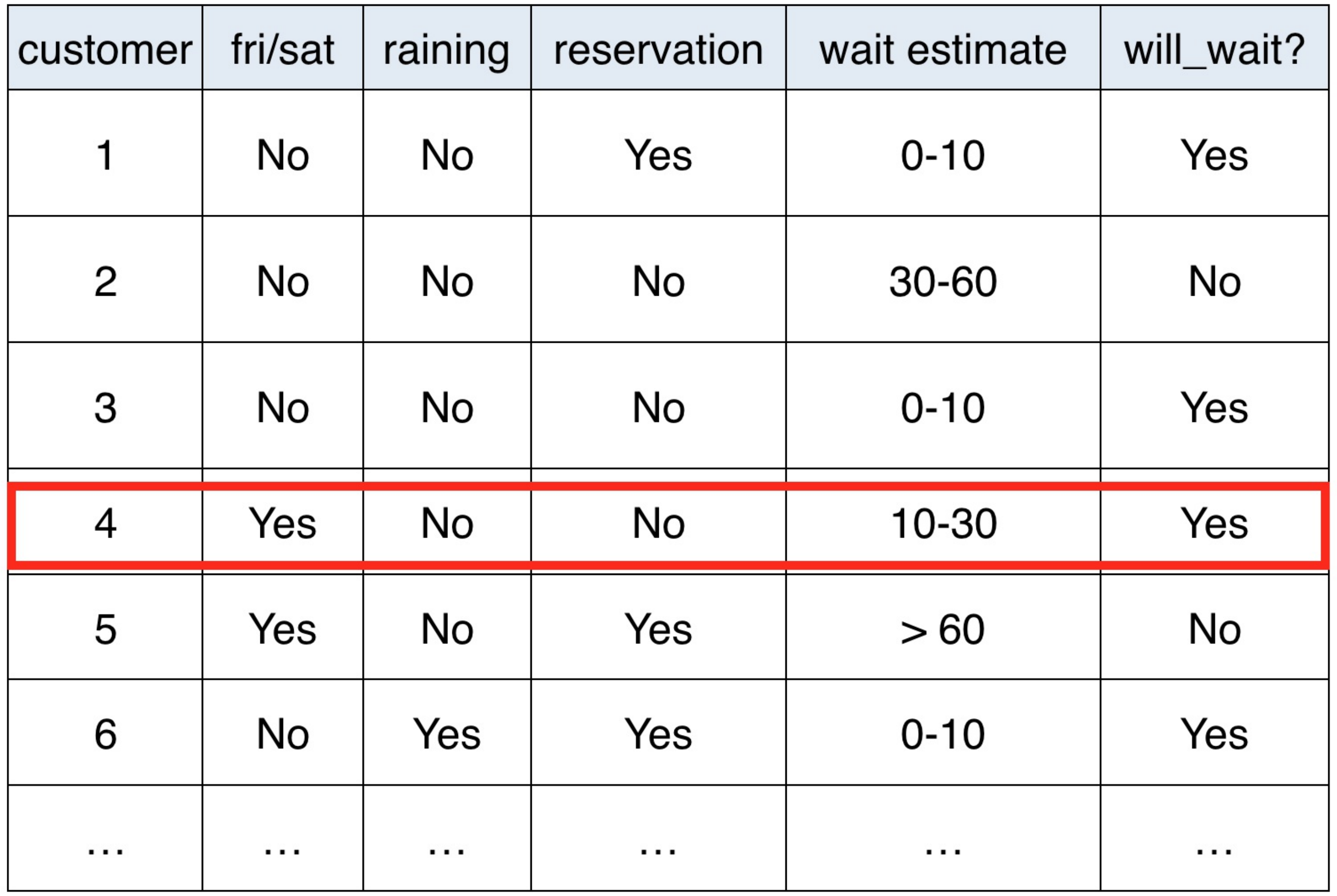
✖ Trees have high variance, which causes model performance to be poor

✖ Trees overfit easily

Will you wait for a table or go elsewhere?

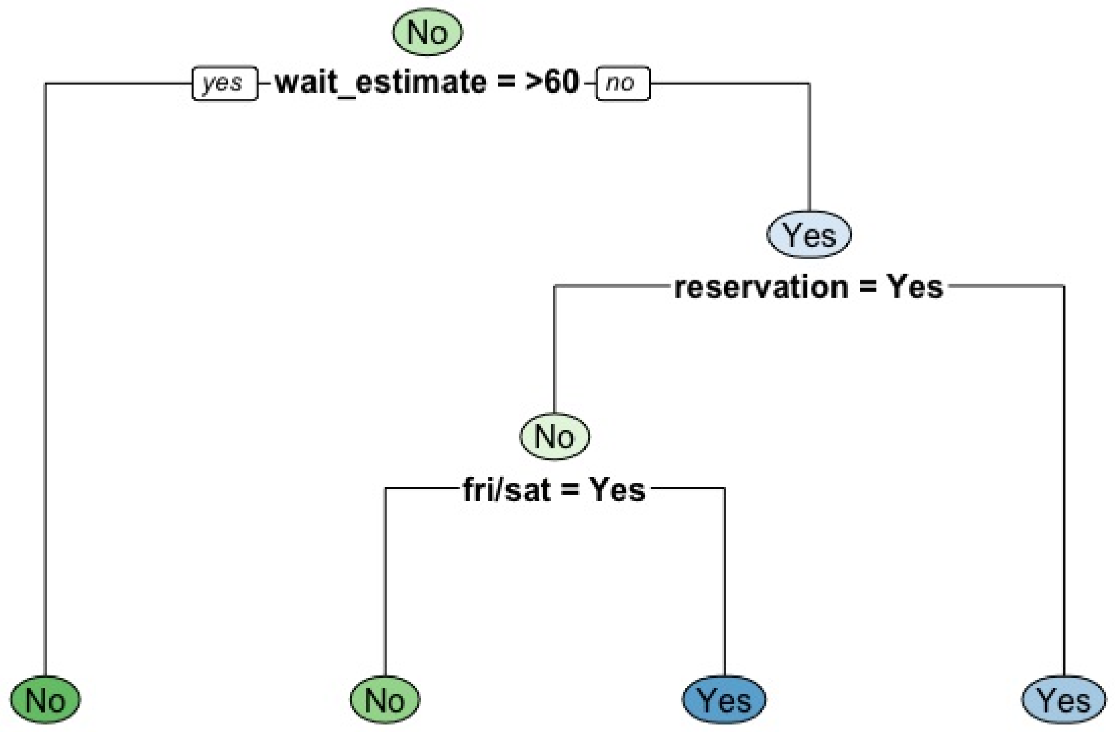
customer	fri/sat	raining	reservation	wait estimate	will_wait?
1	No	No	Yes	0-10	Yes
2	No	No	No	30-60	No
3	No	No	No	0-10	Yes
4	Yes	No	No	10-30	Yes
5	Yes	No	Yes	> 60	No
6	No	Yes	Yes	0-10	Yes
…	…	…	…	…	…

Restaurant Example



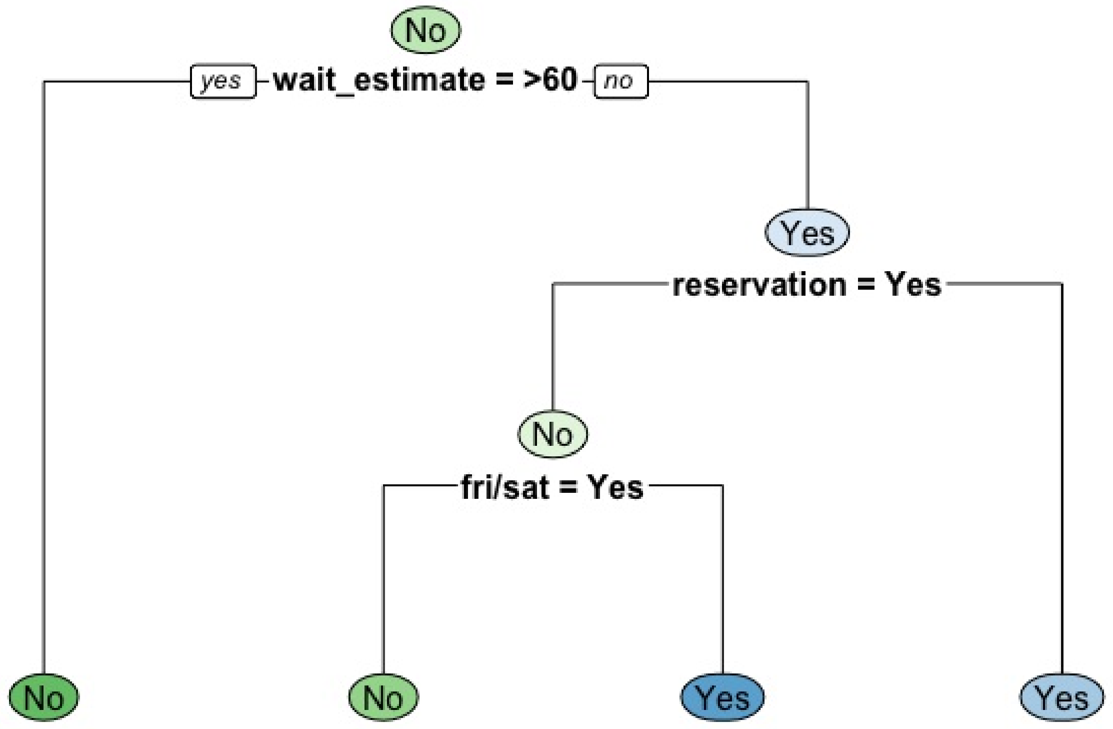
customer	fri/sat	raining	reservation	wait estimate	will_wait?
1	No	No	Yes	0-10	Yes
2	No	No	No	30-60	No
3	No	No	No	0-10	Yes
4	Yes	No	No	10-30	Yes
5	Yes	No	Yes	> 60	No
6	No	Yes	Yes	0-10	Yes
…	…	…	…	…	…

Decision Tree in R

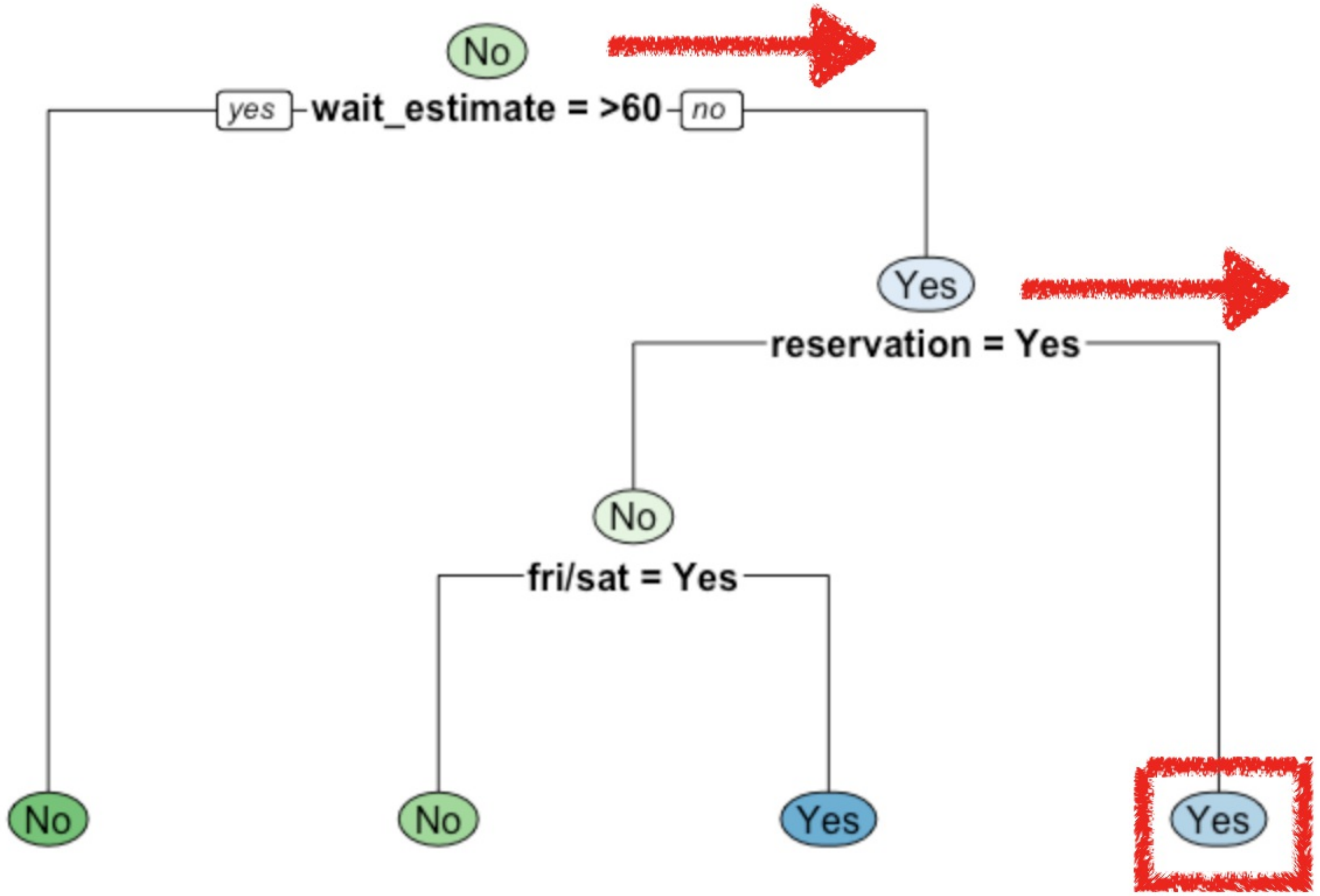


Prediction example

- The wait estimate is 20 minutes, no reservation was made, and it is Wednesday



Example


No
yes
wait_estimate = >60
no
Yes
reservation = Yes
No
fri/sat = Yes
No
No
Yes
Yes

MACHINE LEARNING WITH TREE-BASED MODELS IN R

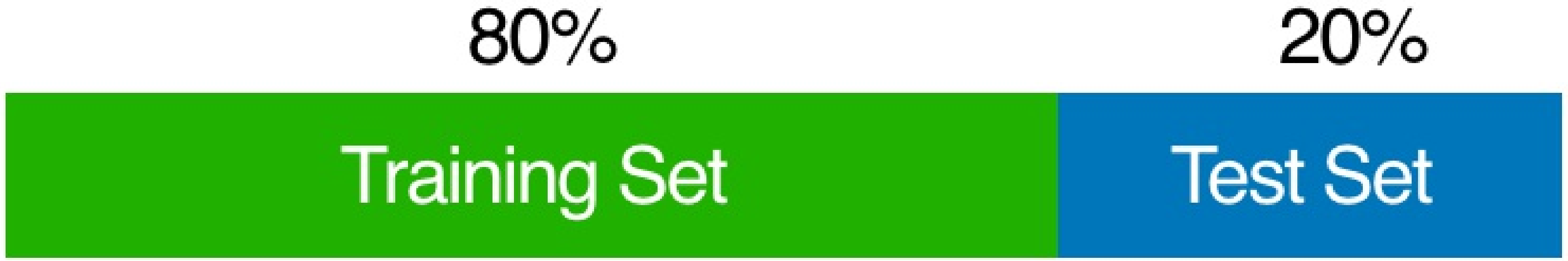
Let's practice!

MACHINE LEARNING WITH TREE-BASED MODELS IN R

Overview of the modeling process

Gabriela de Queiroz
Instructor

Train/Test Split


80%

20%

Training Set

Test Set

Train/test split in R

```
# total number of rows in the restaurant data frame
n <- nrow(restaurant)
```

```
# number of rows for the training set (80% of the dataset)
n_train <- round(0.80 * n)
```

```
# create a vector of indices which is an 80% random sample
set.seed(123) # set a random seed for reproducibility
train_indices <- sample(1:n, n_train)
```

```
# subset the data frame to training indices only
restaurant_train <- restaurant[train_indices, ]

# exclude the training indices to create the test set
restaurant_test <- restaurant[-train_indices, ]
```

Train a Classification Tree

```
# train the model to predict the binary response, "will_wait"

restaurant_model <- rpart(formula = will_wait ~.,
                          data = restaurant_train,
                          method = "class")
```

formula: response variable ~ predictor variables

MACHINE LEARNING WITH TREE-BASED MODELS IN R

Let's practice!

MACHINE LEARNING WITH TREE-BASED MODELS IN R

Evaluate Model Performance

Gabriela de Queiroz
Instructor

Predicting class labels for test data

```
> predict(model, test_dataset)
```

```
> predict(model, test_dataset, type = ___)
```

```
class_prediction <- predict(object = restaurant_model, # model object
                            newdata = restaurant_test, # test dataset
                            type = "class")  # return classification labels
```

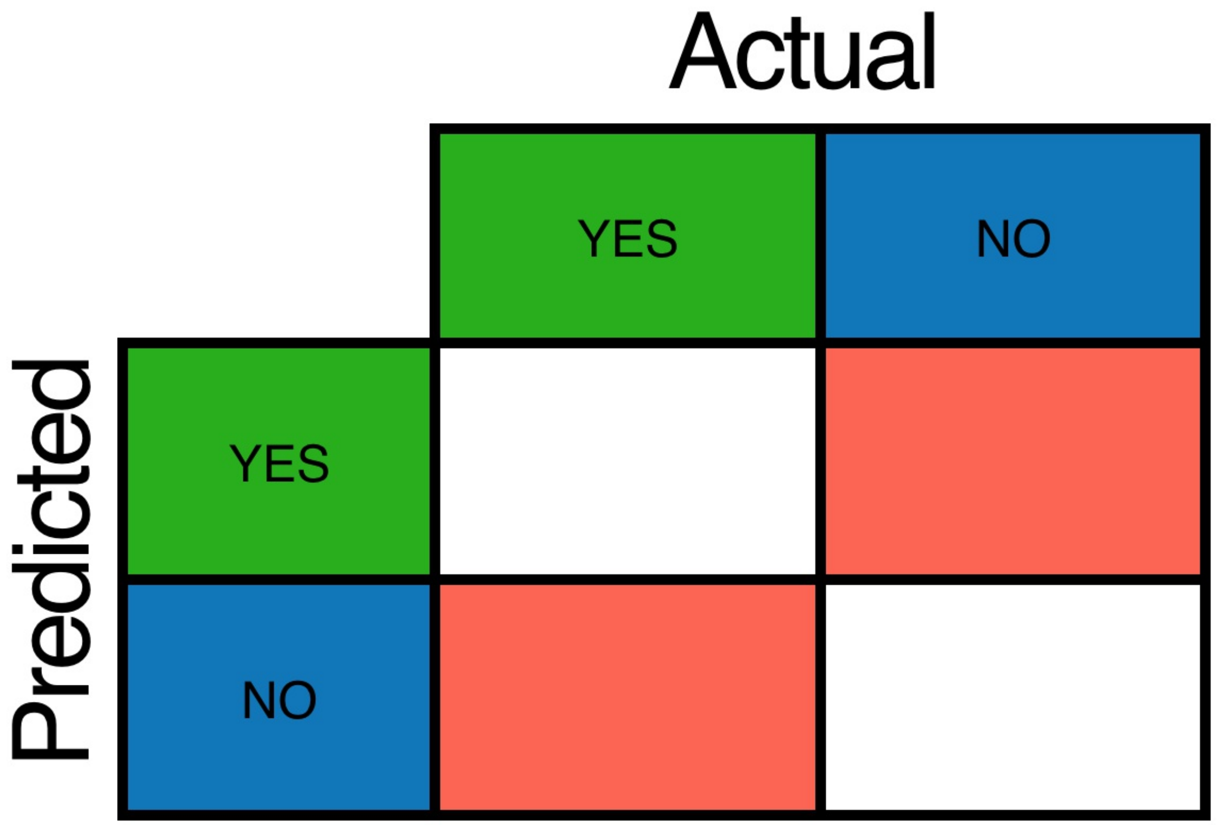
Evaluation Metrics for Binary Classification

- Accuracy
- Confusion Matrix
- Log-loss
- AUC

Accuracy

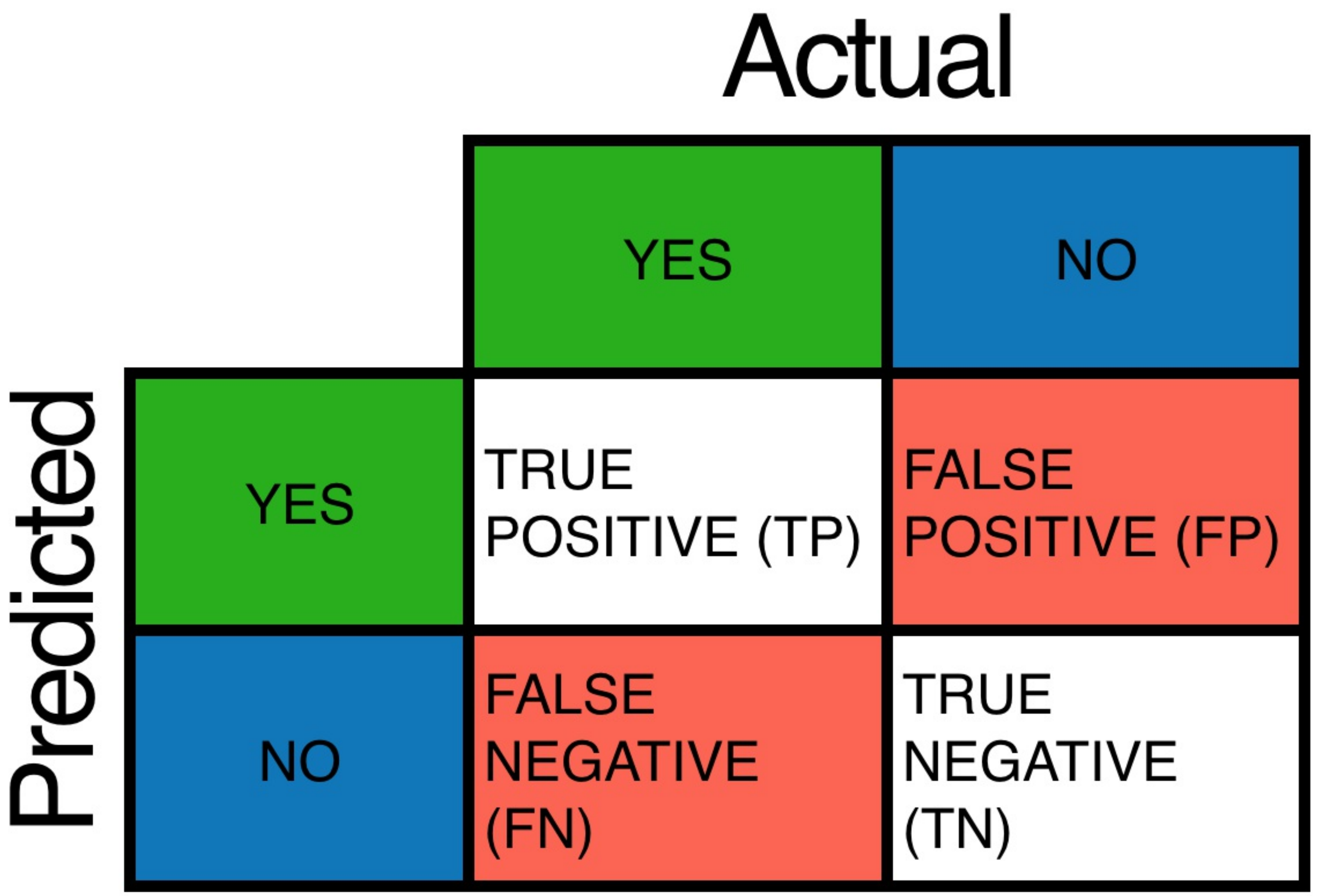
$$accuracy = \frac{\text{n of correct predictions}}{\text{n of total data points}}$$

Confusion Matrix



		Actual	
		YES	NO
Predicted	YES		🟥
	NO	🟥	

Confusion Matrix



		Actual	
		YES	NO
Predicted	YES	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)
	NO	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)

Confusion Matrix

```
library(caret)

# calculate the confusion matrix for the test set
confusionMatrix(data = class_prediction,                # predicted classes
                reference = restaurant_test$will_wait)  # actual classes
```

MACHINE LEARNING WITH TREE-BASED MODELS IN R

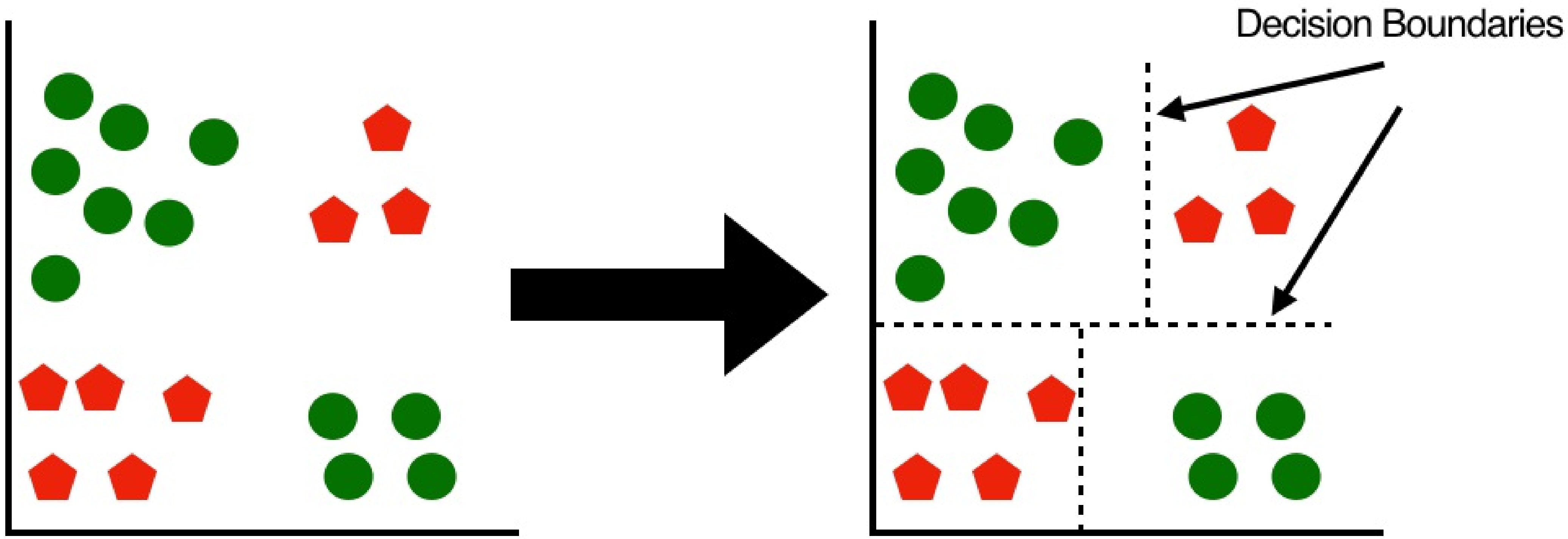
Let's practice!

MACHINE LEARNING WITH TREE-BASED MODELS IN R

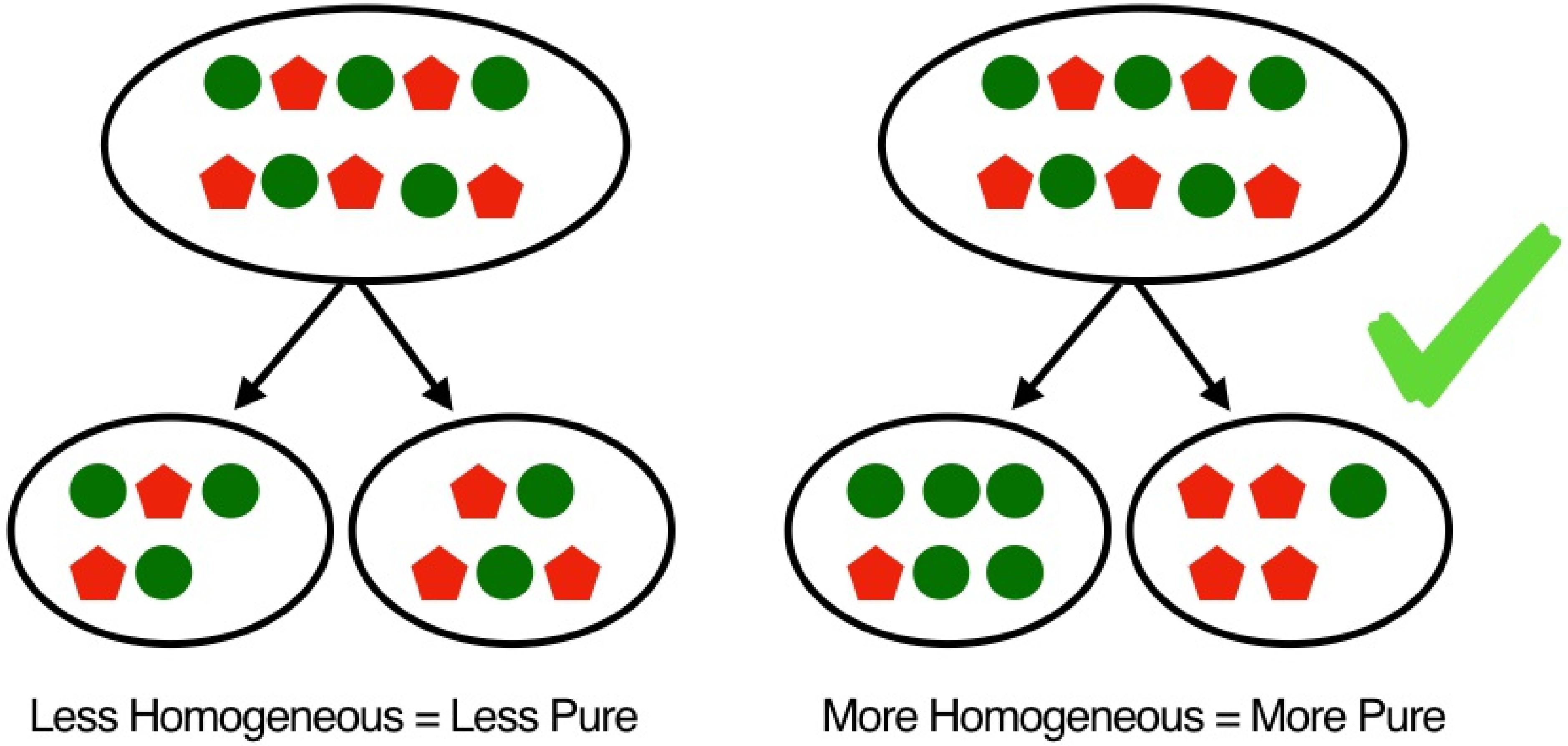
Use of splitting criterion in trees

Gabriela de Queiroz
Instructor

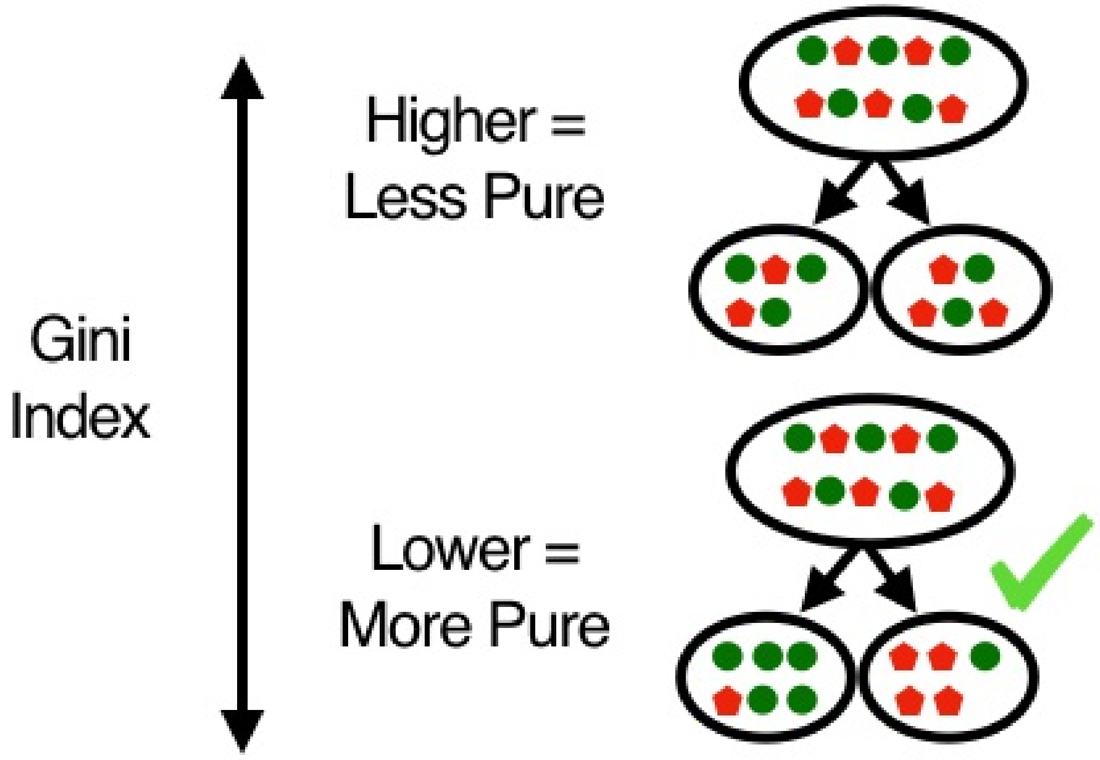
Split the data into "pure" regions



How to determine the best split?



Impurity Measure - Gini Index



MACHINE LEARNING WITH TREE-BASED MODELS IN R

Let's practice!