Welcome to the course!

MACHINE LEARNING WITH TREE-BASED MODELS IN R

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

Tree-based models

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

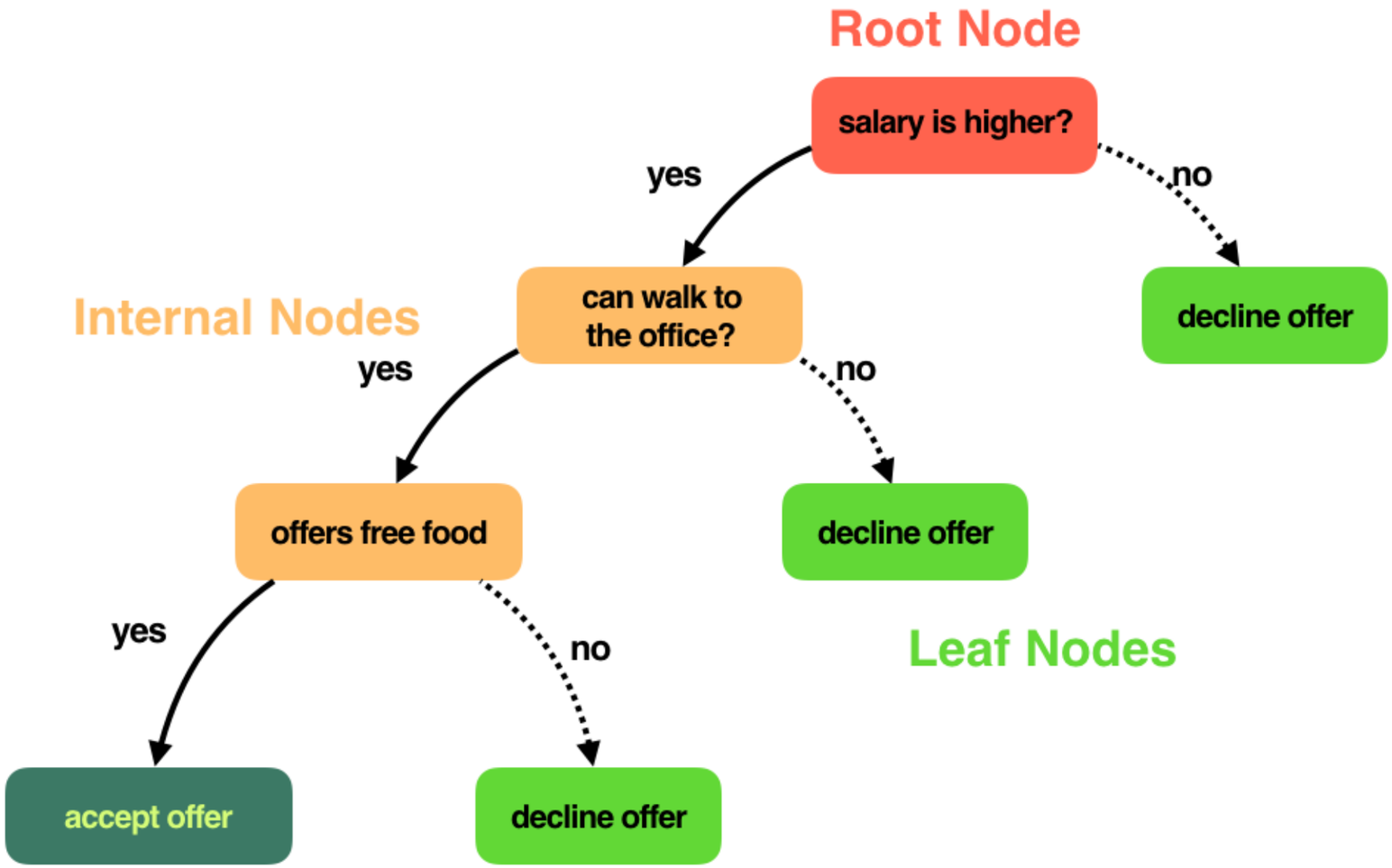
What you'll learn:

- 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

Training Decision Trees in R

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

Let's practice!

MACHINE LEARNING WITH TREE-BASED MODELS IN R

Introduction to classification trees

MACHINE LEARNING WITH TREE-BASED MODELS IN R

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

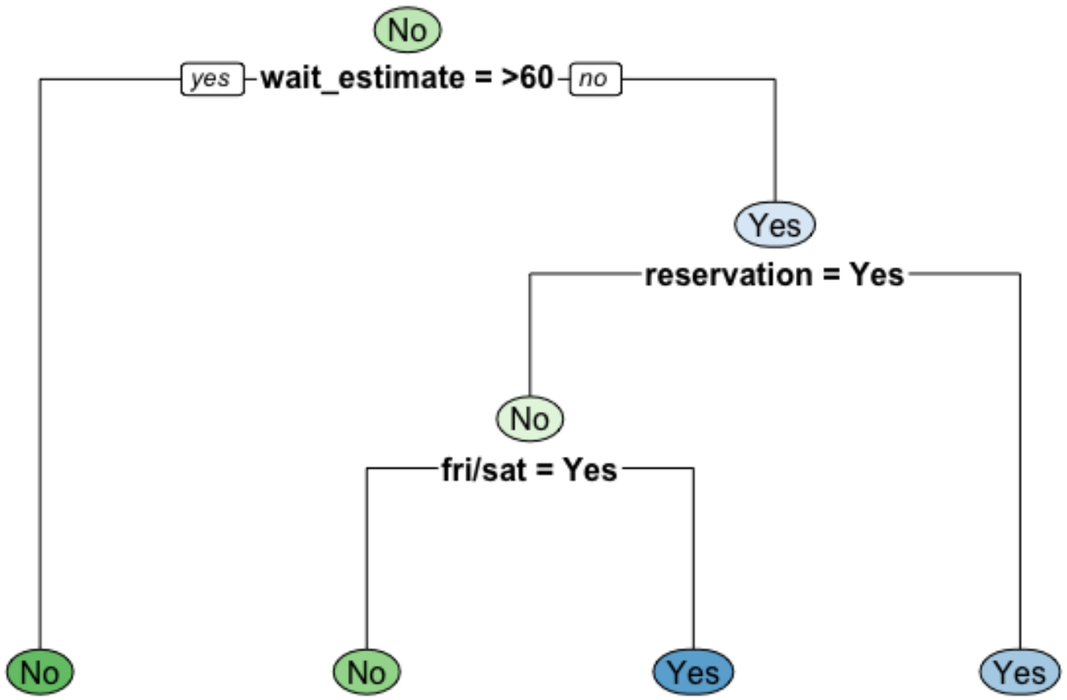
Will you wait for a table or go elsewhere?

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

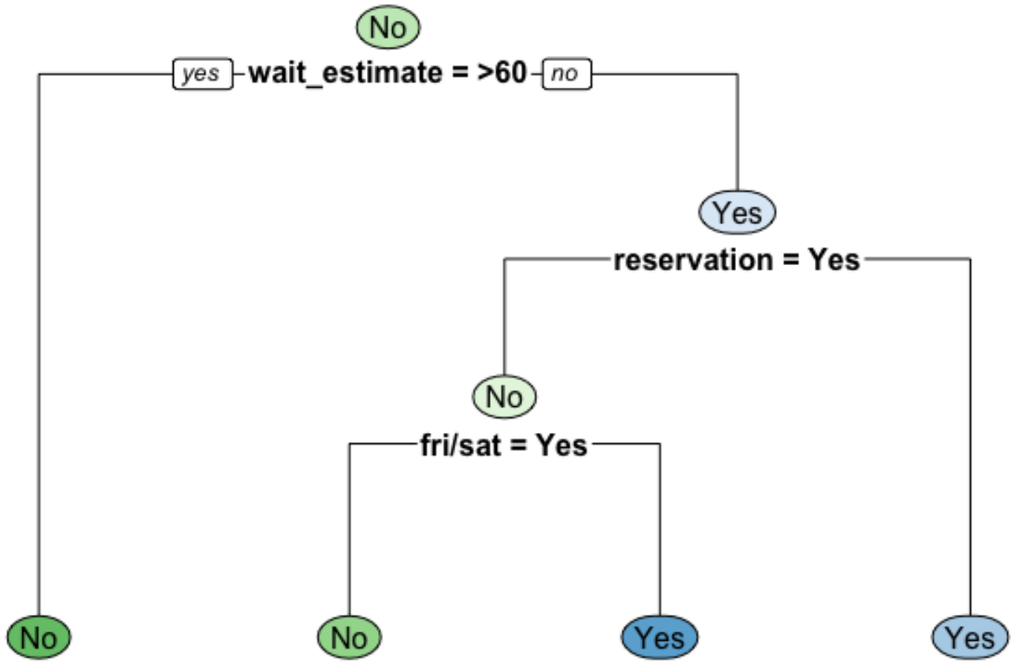
customer	fri/sat	raining	reservation	wait estimate	will_wait?
1	No	No	Yes	0-10	Yes
2	No	No	No	30-60	No
3	No	No	No	0-10	Yes
4	Yes	No	No	10-30	Yes
5	Yes	No	Yes	> 60	No
6	No	Yes	Yes	0-10	Yes
…	…	…	…	…	…

Decision Tree in R

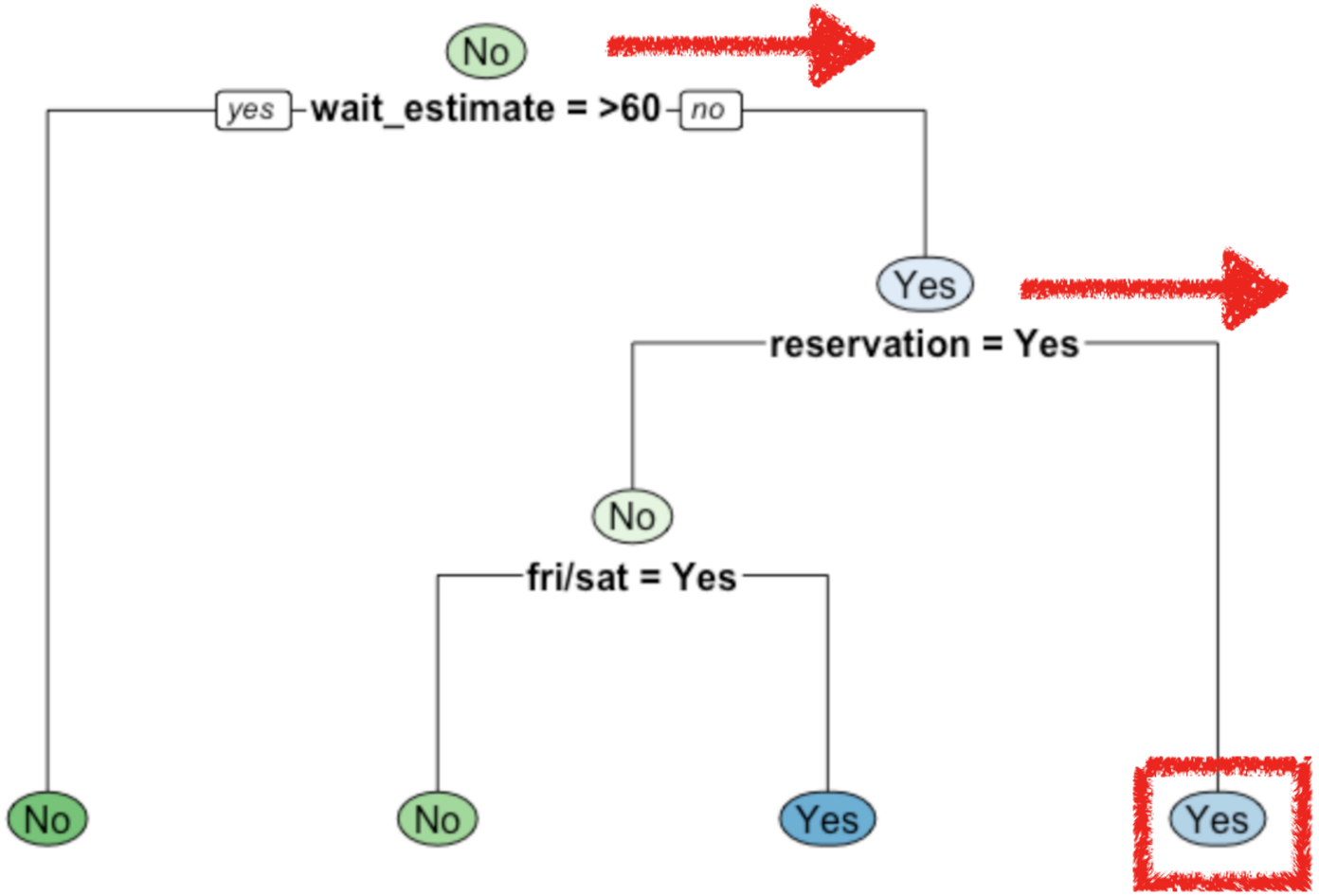


Prediction example

- 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

Let's practice!

MACHINE LEARNING WITH TREE-BASED MODELS IN R

Overview of the modeling process

MACHINE LEARNING WITH TREE-BASED MODELS IN R

Gabriela de Queiroz
Instructor

Train/Test Split

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)
```

```
# Set a random seed for reproducibility
set.seed(123)

# Create a vector of indices which is an 80% random sample
train_indices <- sample(1:n, n_train)
```

Train/test split in R

```
# 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_wai

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

formula: `response variable` ~ `predictor variables`

Let's practice!

MACHINE LEARNING WITH TREE-BASED MODELS IN R

Evaluate Model Performance

MACHINE LEARNING WITH TREE-BASED MODELS IN R

Gabriela de Queiroz
Instructor

Predicting class labels for test data

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

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

```
class_pred <- predict(object = restaurant_model,
                      newdata = restaurant_test,
                      type = "class")
```

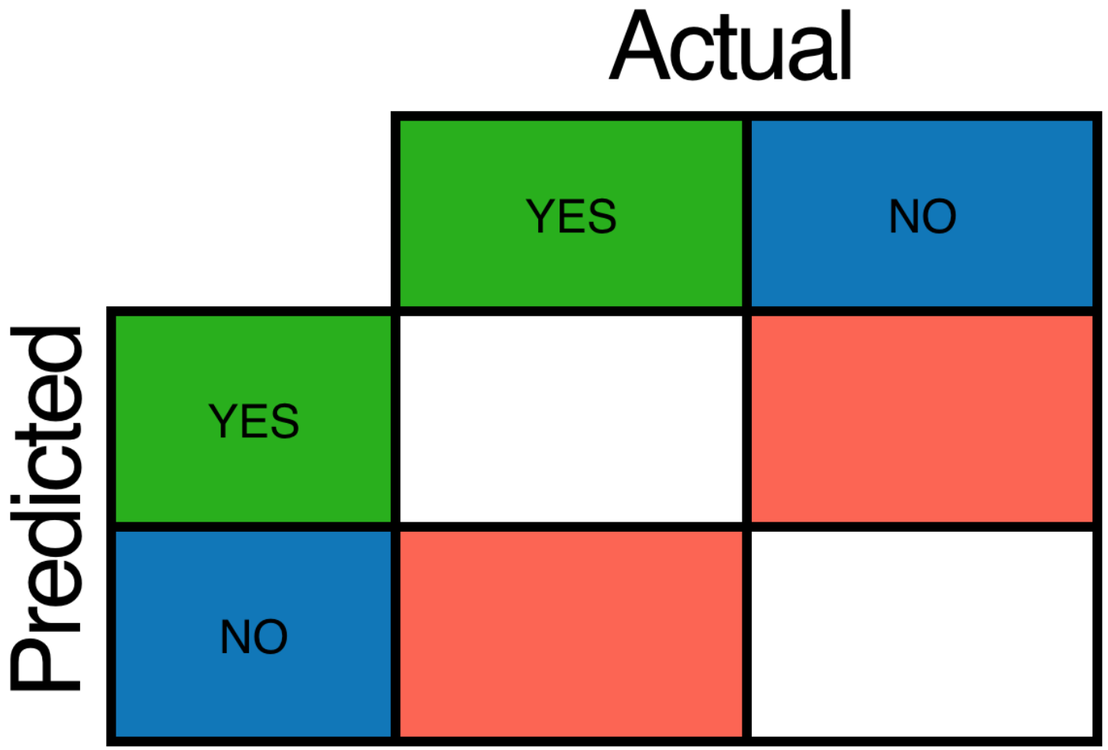
Evaluation Metrics for Binary Classification

- Accuracy
- Confusion Matrix
- Log-loss
- AUC

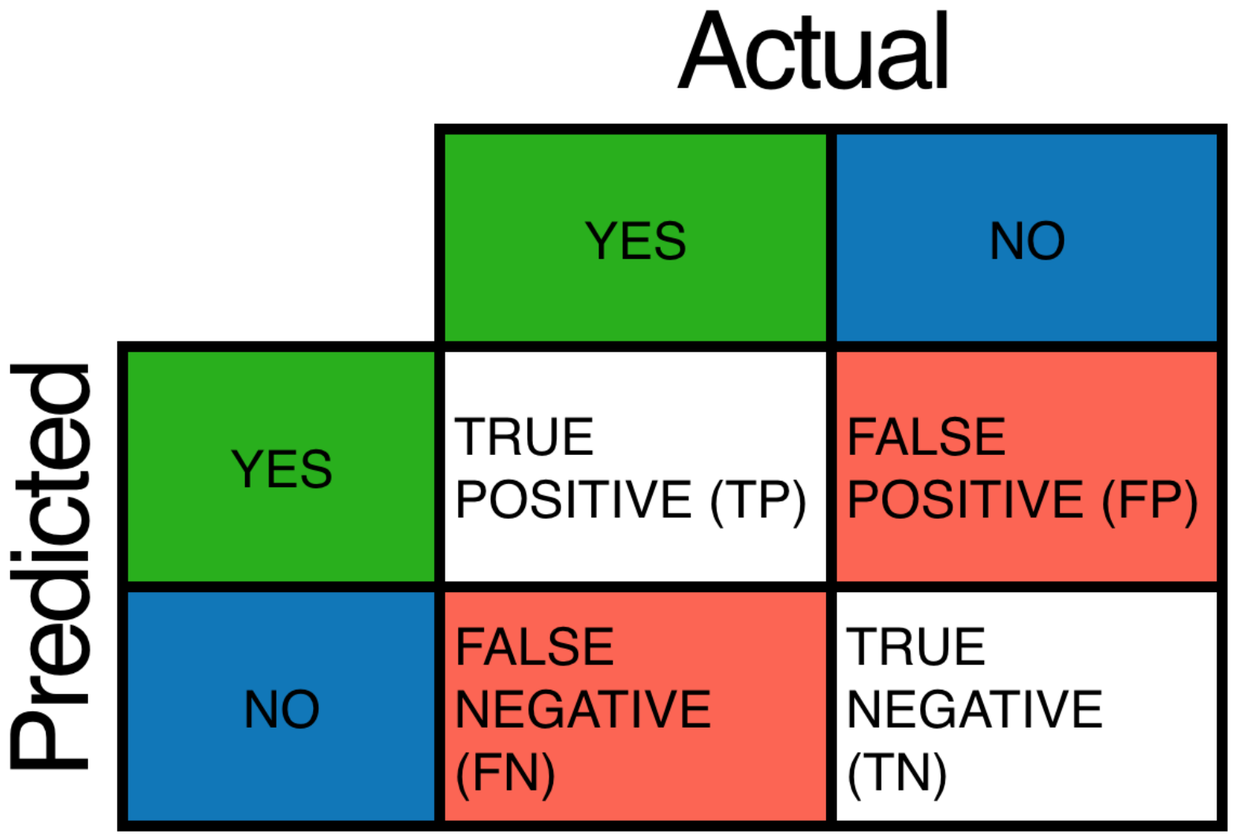
Accuracy

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

Confusion Matrix



Confusion Matrix



Confusion Matrix

```
library(caret)

# Calculate the confusion matrix for the test set
confusionMatrix(data = class_pred,
                reference = restaurant_test$will_wait)
```

Let's practice!

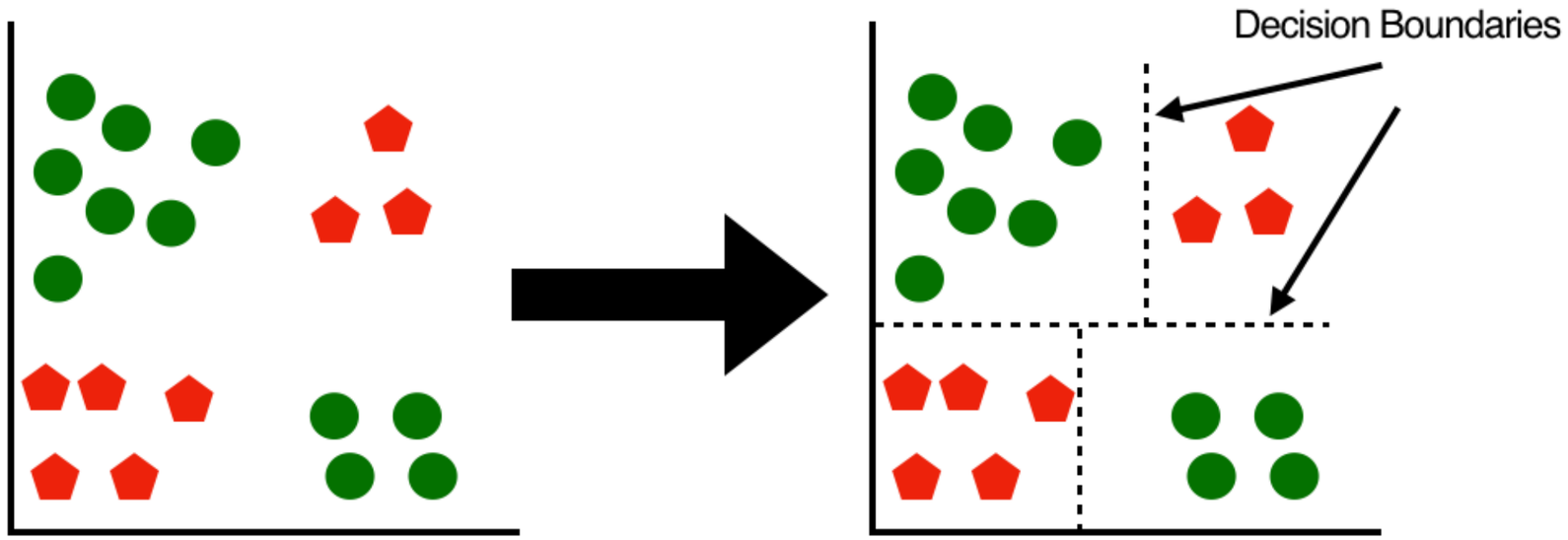
MACHINE LEARNING WITH TREE-BASED MODELS IN R

Use of splitting criterion in trees

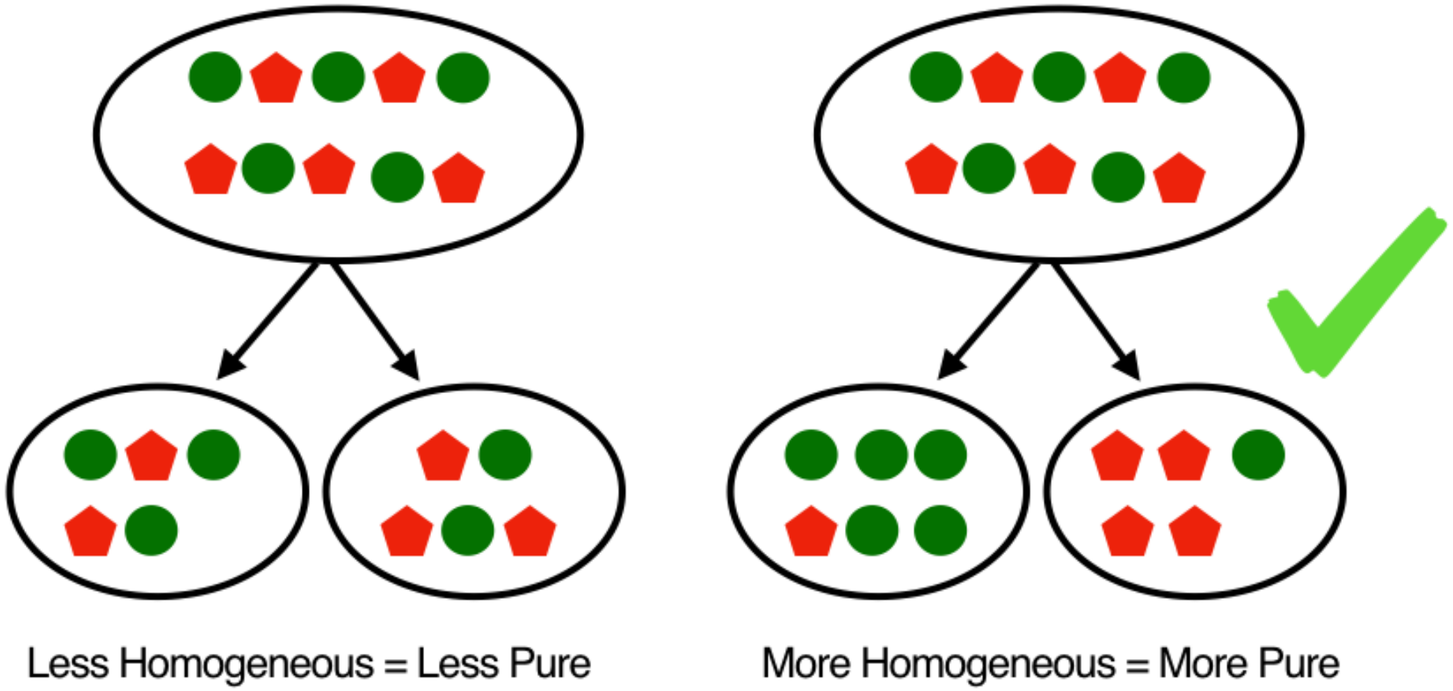
MACHINE LEARNING WITH TREE-BASED MODELS IN R

Gabriela de Queiroz
Instructor

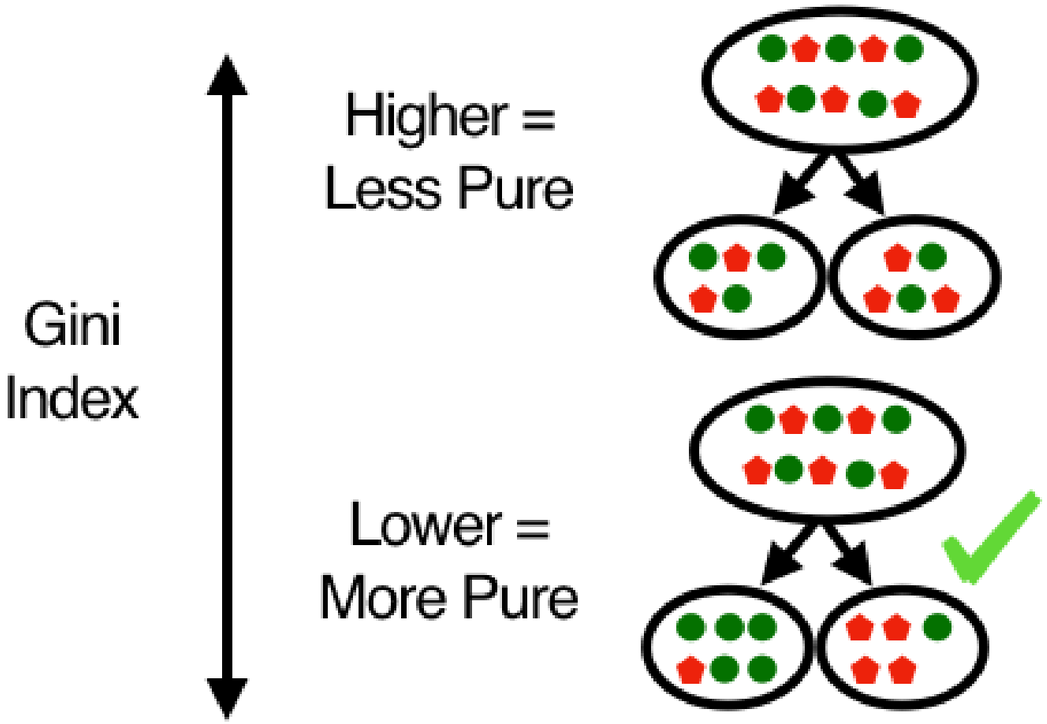
Split the data into "pure" regions



How to determine the best split?



Impurity Measure - Gini Index



Let's practice!

MACHINE LEARNING WITH TREE-BASED MODELS IN R