

1 – Introduction

Introduction to Machine Learning with R

Plan for this short course

- Overview: R and RStudio
- Data visualization and exploratory analysis
- Supervised Machine Learning
- Classification and Regression

R and RStudio

R has emerged over the last couple decades as a first-class tool for scientific computing tasks, and has been a consistent leader in implementing statistical methodologies for analyzing data. The usefulness of R for data science stems from the large, active, and growing ecosystem of third-party packages.

Hands-On Machine Learning with R (bradleyboehmke.github.io)

Overview

- Data preparation and feature engineering
- Reading and importing data, data exploration and filtering
- Cleaning and Preprocessing
- Exploratory Data Analysis
- Supervised learning
 - Classification: K-nearest neighbors, decision trees, random forest
 - Regression: K-nearest neighbors, linear regression, regression trees, random forest
- Model Evaluation and Model Selection

Outline

- First day:
 - R and R Studio
 - Reading and importing data, exploratory analysis and visualization
- Day two: Regression
- Day three: Classification
- Hands on exercise

R and RStudio

To download R, go to CRAN, the Comprehensive R Archive Network.

www.r-project.org

CRAN is composed of a set of mirror servers distributed around the world and is used to distribute R and R packages.

pick a mirror or use the cloud mirror, <https://cloud.r-project.org>

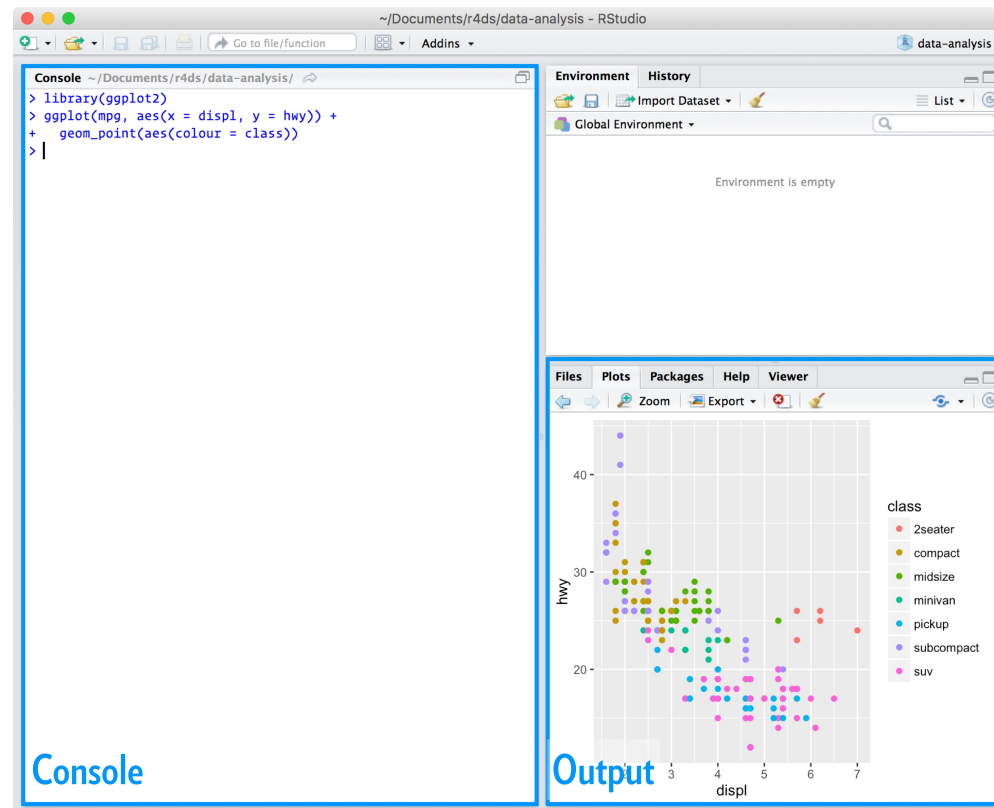
R and Rstudio – Install

RStudio is an integrated development environment, or IDE, for R programming. Download and install it from <http://www.rstudio.com/download>.

RStudio is updated a couple of times a year. When a new version is available, RStudio will let you know.

Rstudio – Overview

When you start RStudio, you'll see two key regions in the interface:



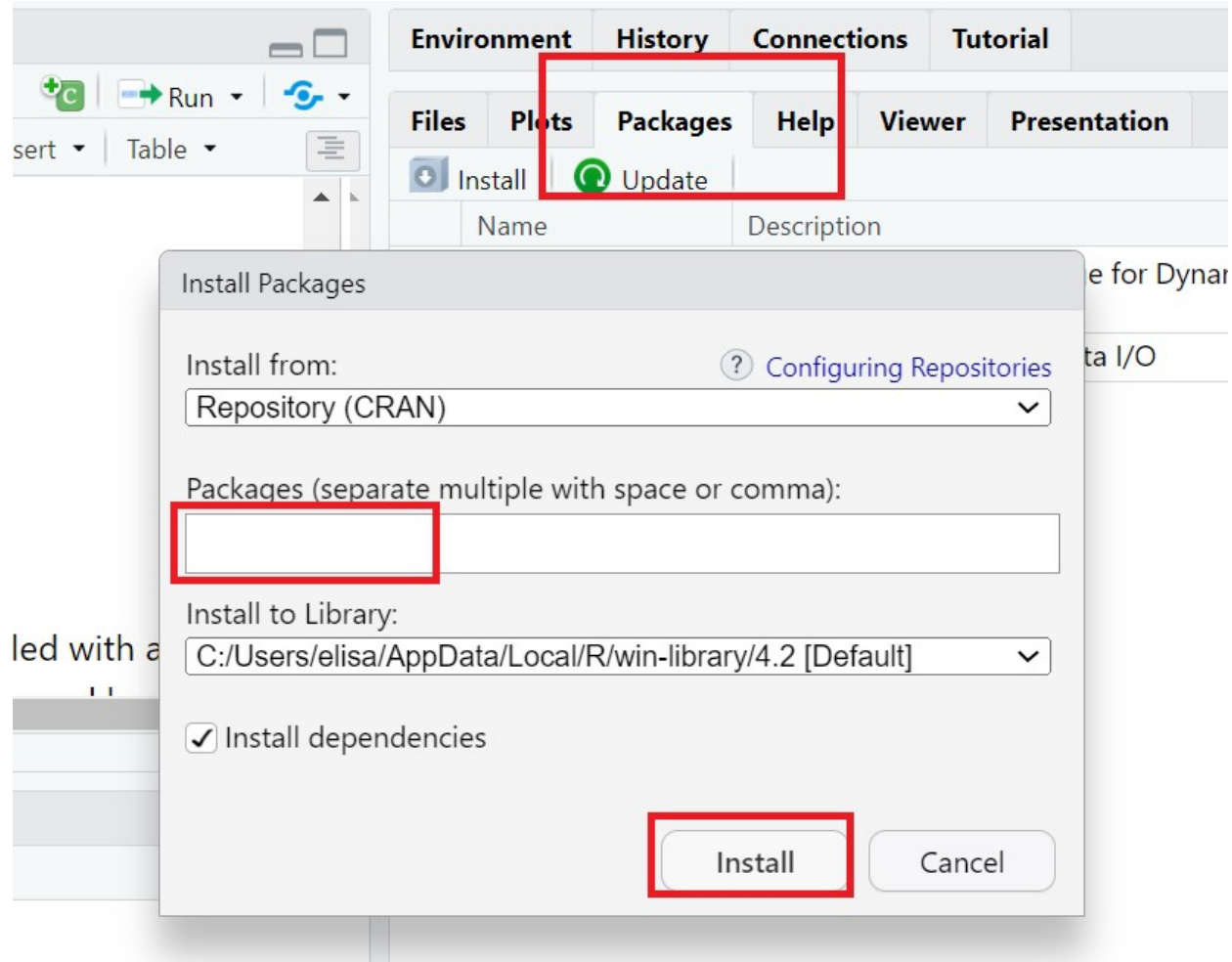
Source: R for Data Science

R Packages

R packages are collections of functions and data sets developed by the community. They increase the power of R by improving existing base R functionalities, or by adding new ones.

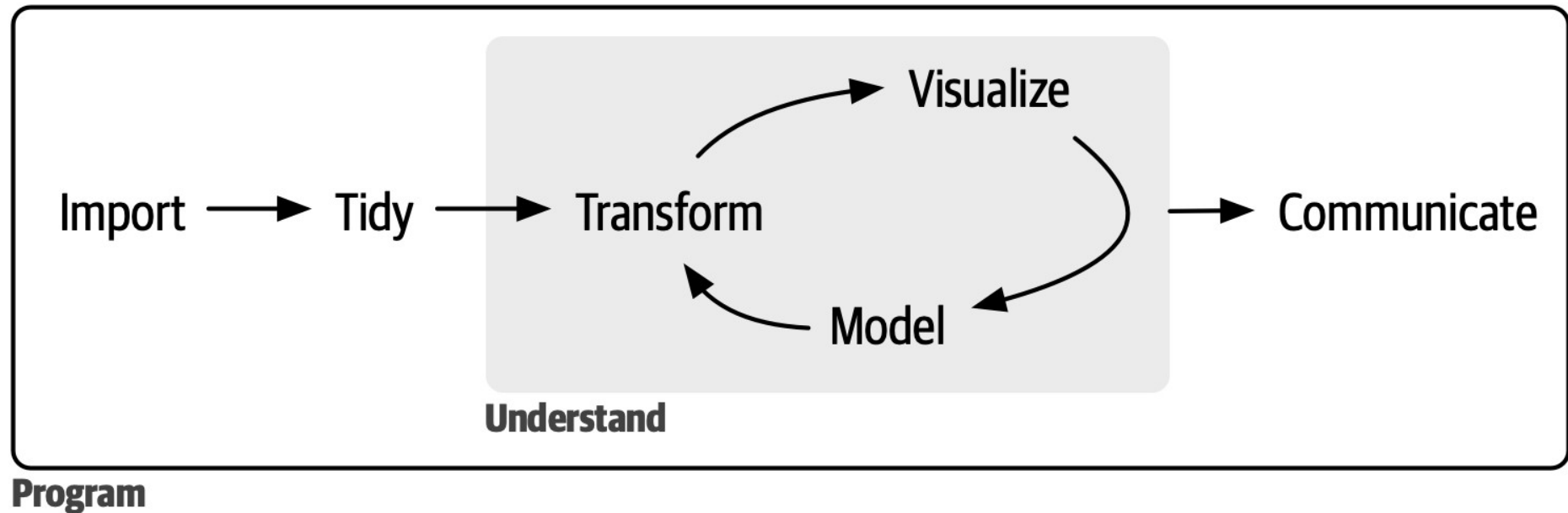
- 10,000 packages published

Install R Packages



tidyverse

- The tidyverse is a collection of packages that can easily be installed with a single “meta”-package, that share a high-level design philosophy and low-level grammar and data structures, so that learning one package makes it easier to learn the next.

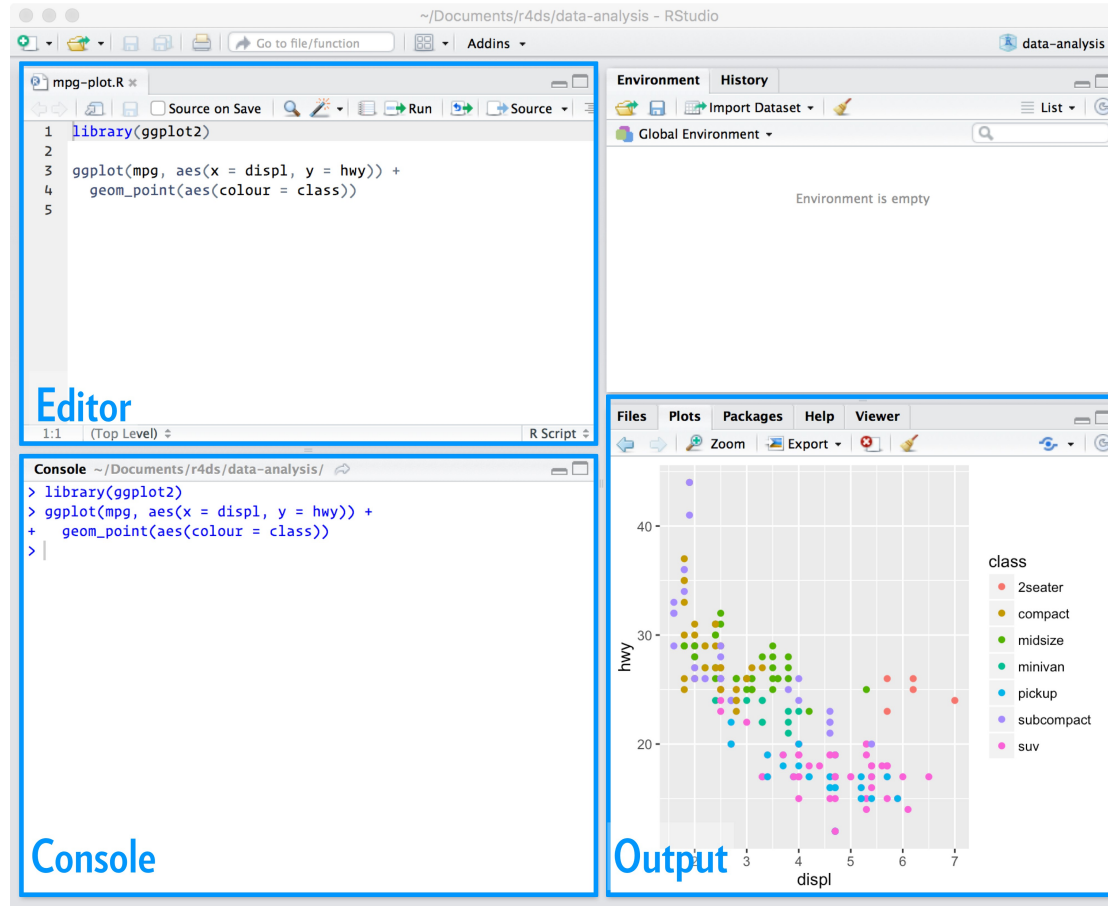


Wickham et al. (2017) model of the data science process

Scripts

- A script is simply a text file containing a set of commands and comments.
- The script can be saved and used later to re-execute the saved commands.
- The script can also be edited so you can execute a modified version of the commands.

Scripts



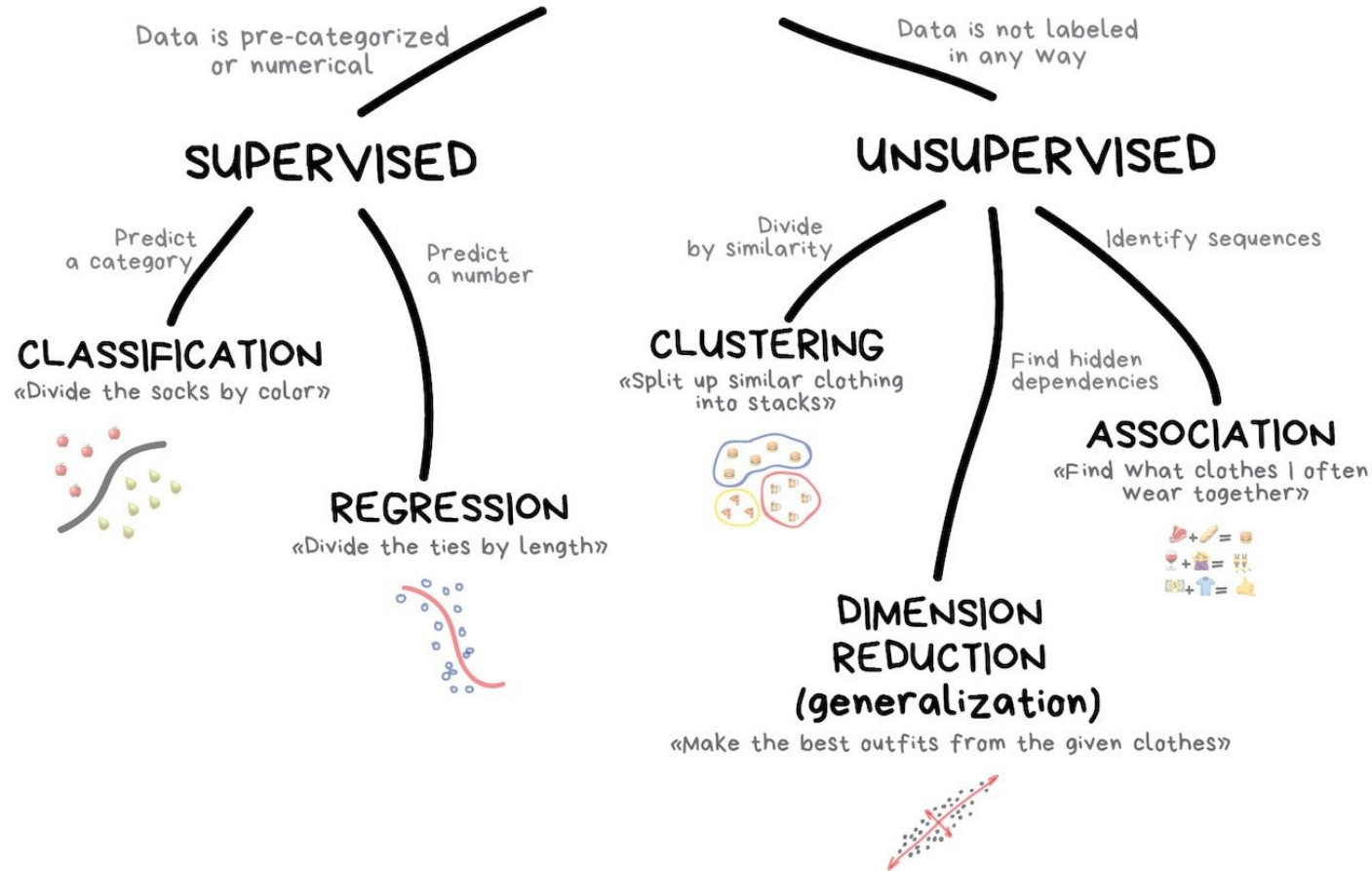
Script example

What is machine learning?



What is machine learning?

CLASSICAL MACHINE LEARNING



How are statistics and machine learning related?

How are they similar? Different?

the “two cultures”

- model first vs. data first
- inference vs. prediction

Supervised learning

A *predictive model* is used for tasks that involve the prediction of a given output using other variables (or features) in the data set.

Or, as stated by Kuhn and Johnson (2013, 26:2), predictive modeling is “...the process of developing a mathematical tool or model that generates an accurate prediction.”

The learning algorithm in a predictive model attempts to discover and model the relationships among the variable being predicted and the other predictor variables.

Examples of predictive modeling

- using home attributes to predict the sales price;
- using employee attributes to predict the likelihood of attrition;
- using patient attributes and symptoms to predict the risk of readmission;
- using production attributes to predict time to market.

Supervised learning

In essence, these tasks all seek to learn from data. To address each scenario, we can use a given set of *variables* to train an algorithm and extract insights.

The supervision refers to the fact that the target values provide a supervisory role, which indicates to the learner the task it needs to learn.

Specifically, given a set of data, the learning algorithm attempts to optimize a function to find the combination of variables values that results in a predicted value that is as close to the actual target output as possible.

Regression or classification

Most supervised learning problems can be bucketed into one of two categories, *regression* or *classification*.

When the objective of our supervised learning is to predict a numeric outcome, we refer to this as a *regression problem*.

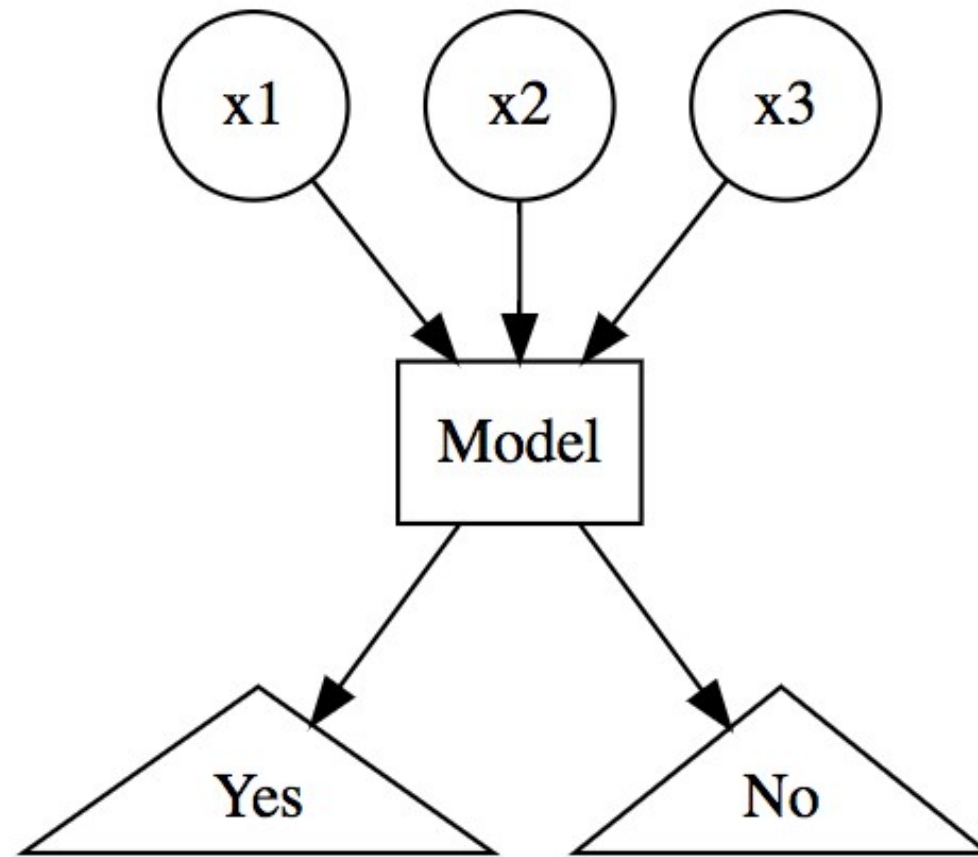
When the objective of our supervised learning is to predict a categorical outcome, we refer to this as a *classification problem*.

Classification

Classification problems most commonly revolve around predicting a binary or multinomial response measure such as:

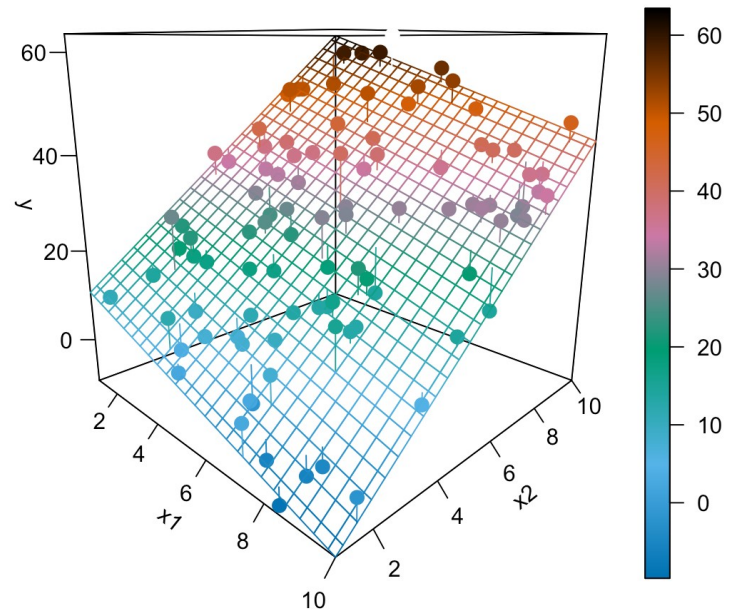
- Did a customer click on our online ad (coded as yes/no or 1/0)?
- Commute choice: car, bike or bus.

Classification

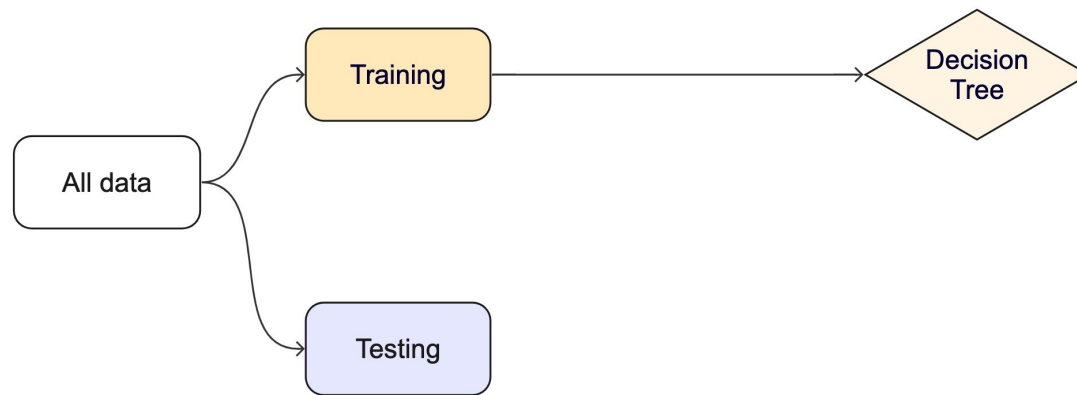


Regression problems

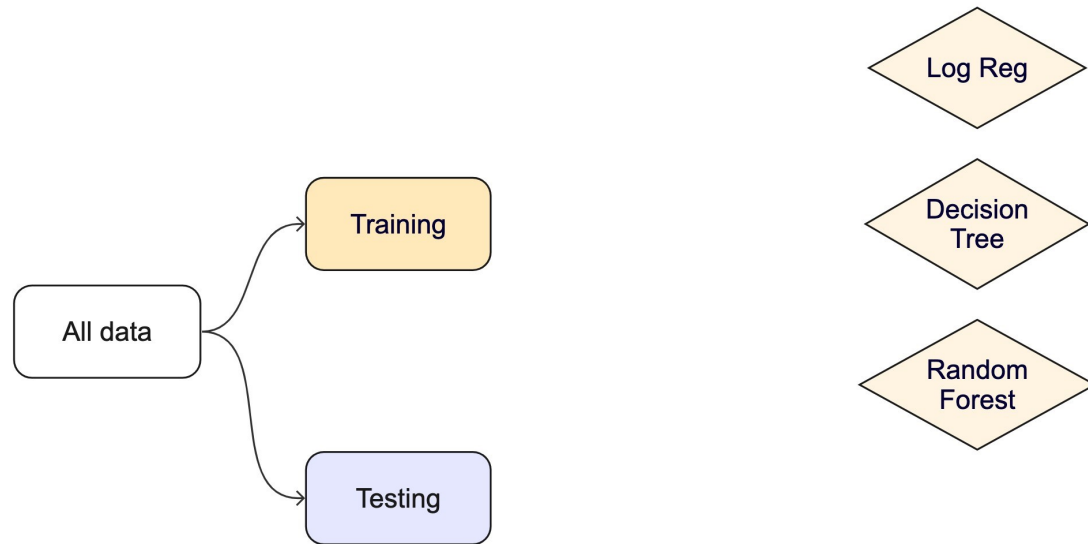
Predict a numeric outcome



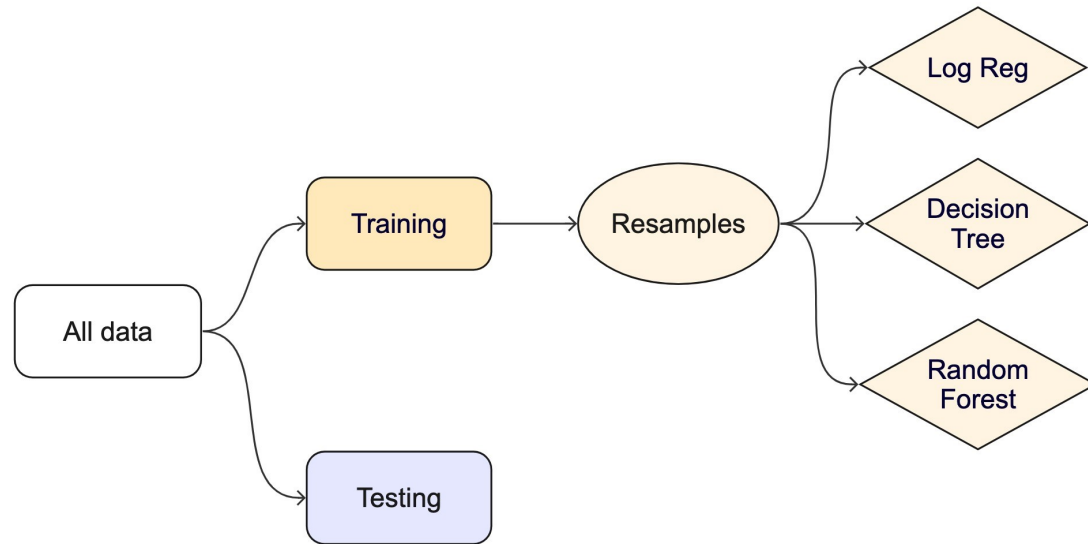
Flowchart



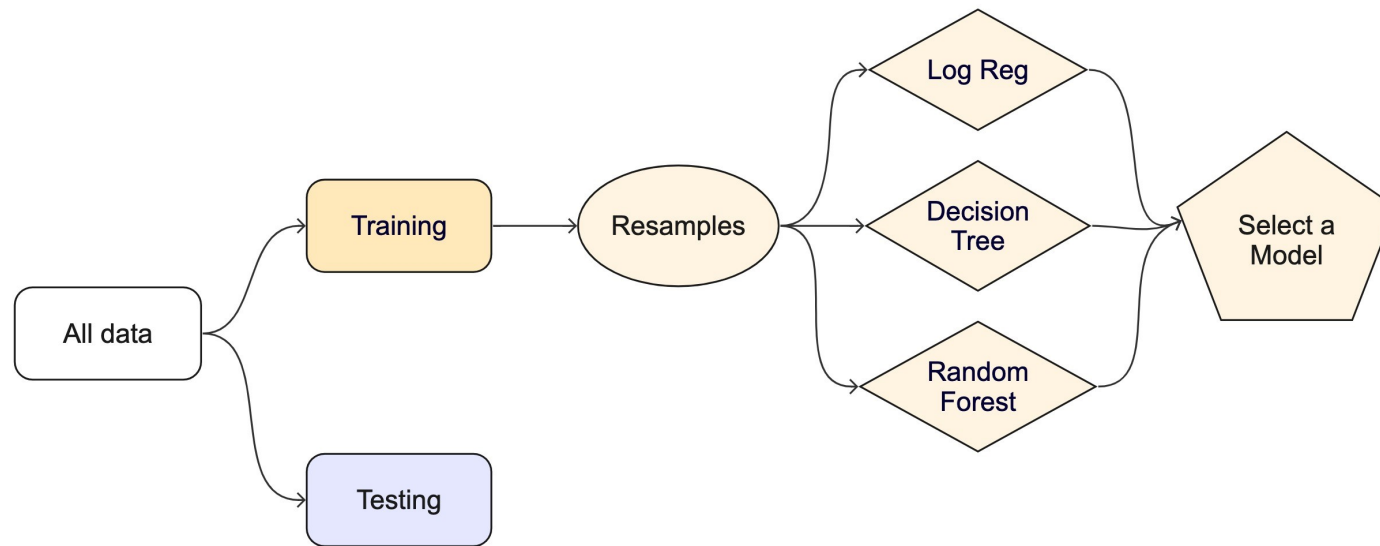
Flowchart



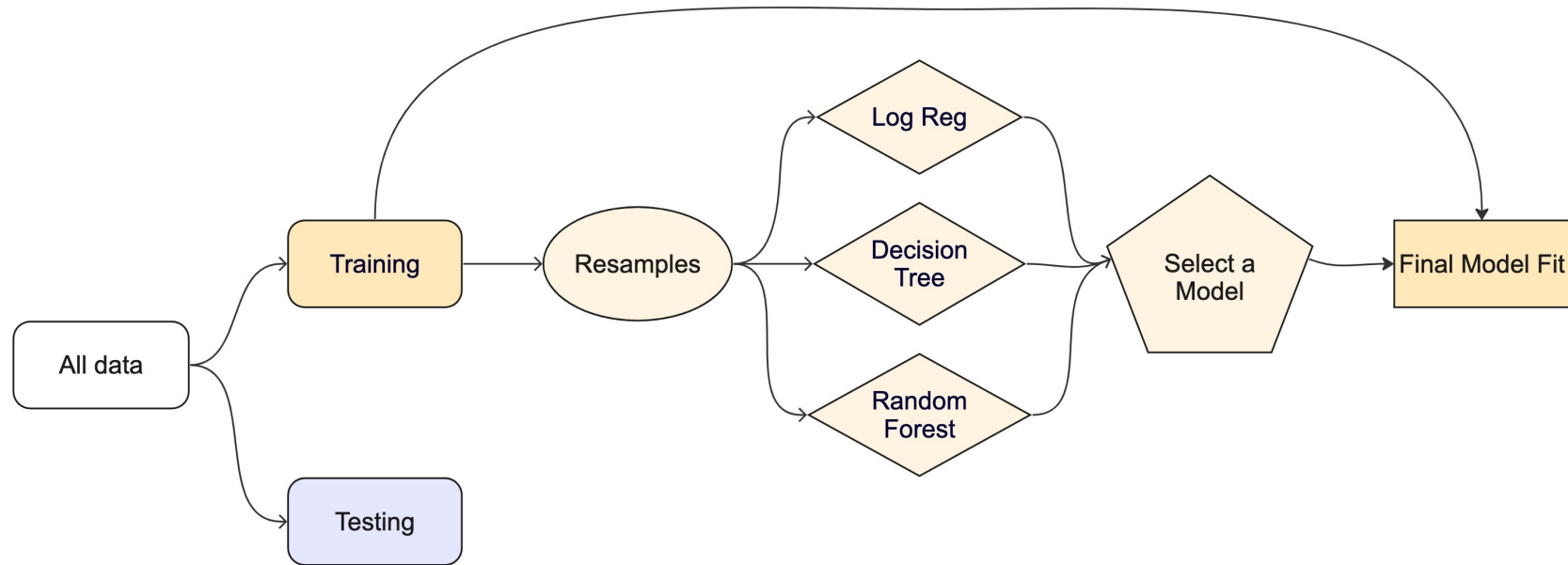
Flowchart



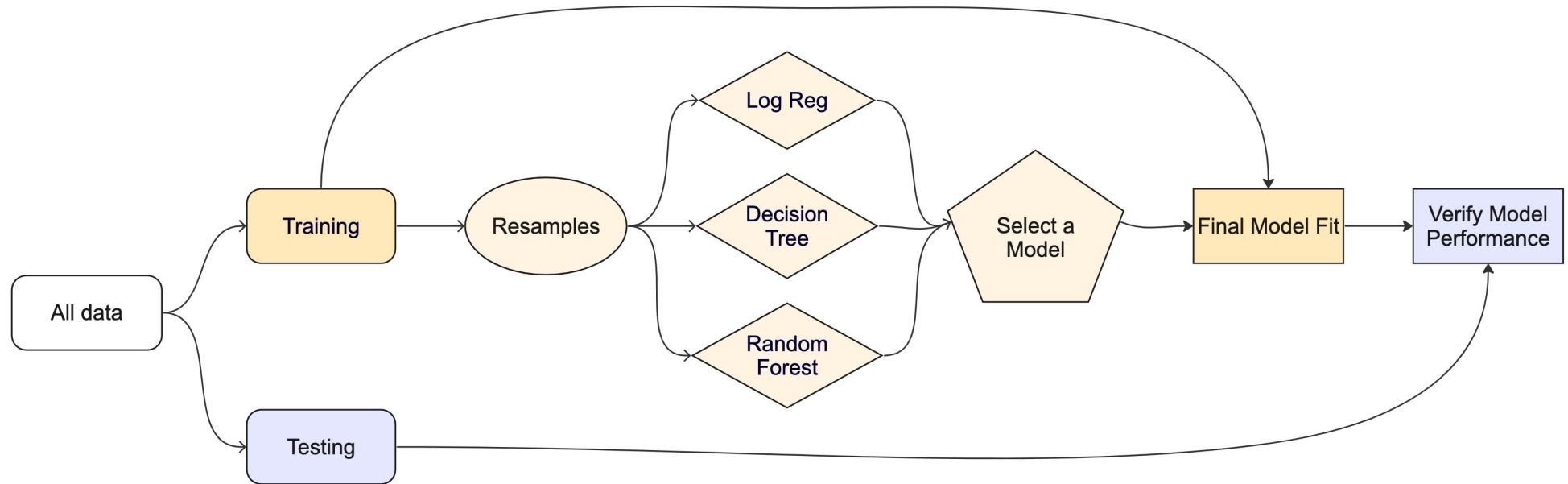
Flowchart



Flowchart



Flowchart




Data splitting and spending

For machine learning, we typically split data into training and test sets:

- The **training set** is used to estimate model parameters.
- The **test set** is used to find an independent assessment of model performance.

Do not  use the test set during training.

Data splitting

- Spending too much data in **training** prevents us from computing a good assessment of predictive **performance**.
- Spending too much data in **testing** prevents us from computing a good estimate of model **parameters**.
- The testing data is precious 
- How much data in training vs testing?
 - R function uses a good default, but this depends on your specific goal/data. We will talk about more powerful ways of splitting, like stratification, later.

What is `set.seed()`?

To create that split of the data, R generates “pseudo-random” numbers: while they are made to behave like random numbers, their generation is deterministic given a “seed”.

This allows us to reproduce results by setting that seed.

Which seed you pick doesn't matter, as long as you don't try a bunch of seeds and pick the one that gives you the best performance.

Evaluating models

- Look at the predictions
- Metrics

REGRESSION

- Mean absolute error (MAE)
- Root mean squared error (RMSE)
- R-Squared R^2 or
- Adjusted R^2

CLASSIFICATION

- Accuracy
- Precision
- AUC (Area under the curve)
- Recall
- F1

Confusion Matrix – Classification

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

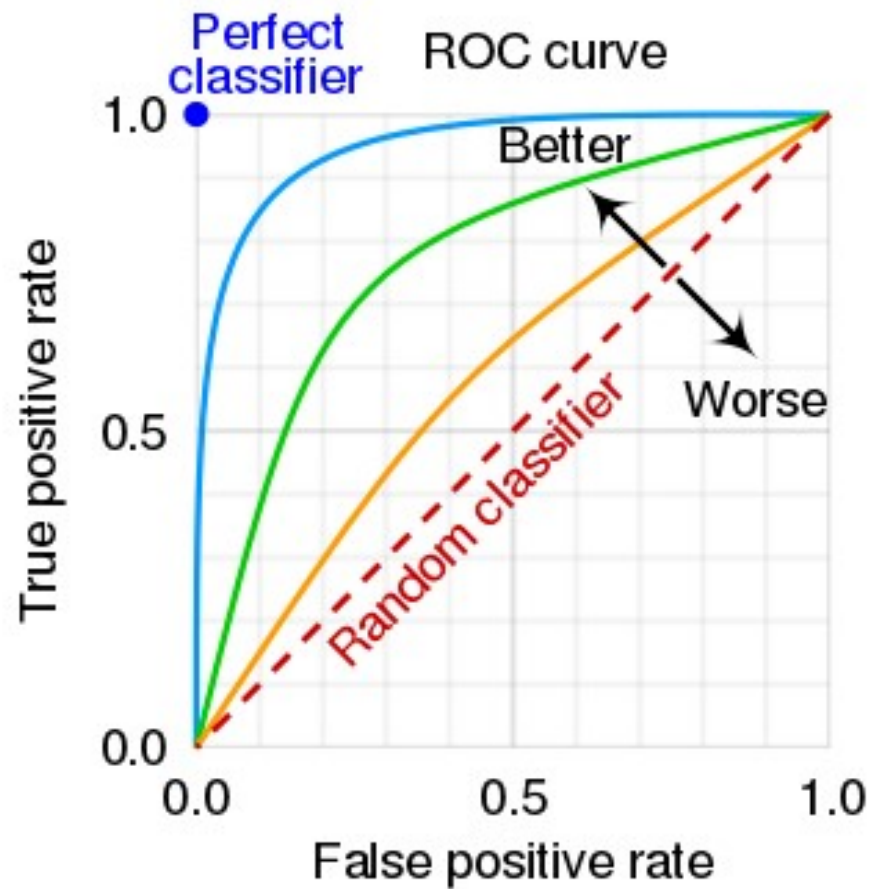
<https://h2o.ai/wiki/confusion-matrix/>

ROC curves

To make an ROC (receiver operator characteristic) curve, we:

- calculate the sensitivity (true positive rate) and specificity true negative rate) for all possible thresholds
- plot false positive rate (x-axis) versus true positive rate (y-axis)
- . . .

ROC Curves



https://en.wikipedia.org/wiki/Receiver_operating_characteristic

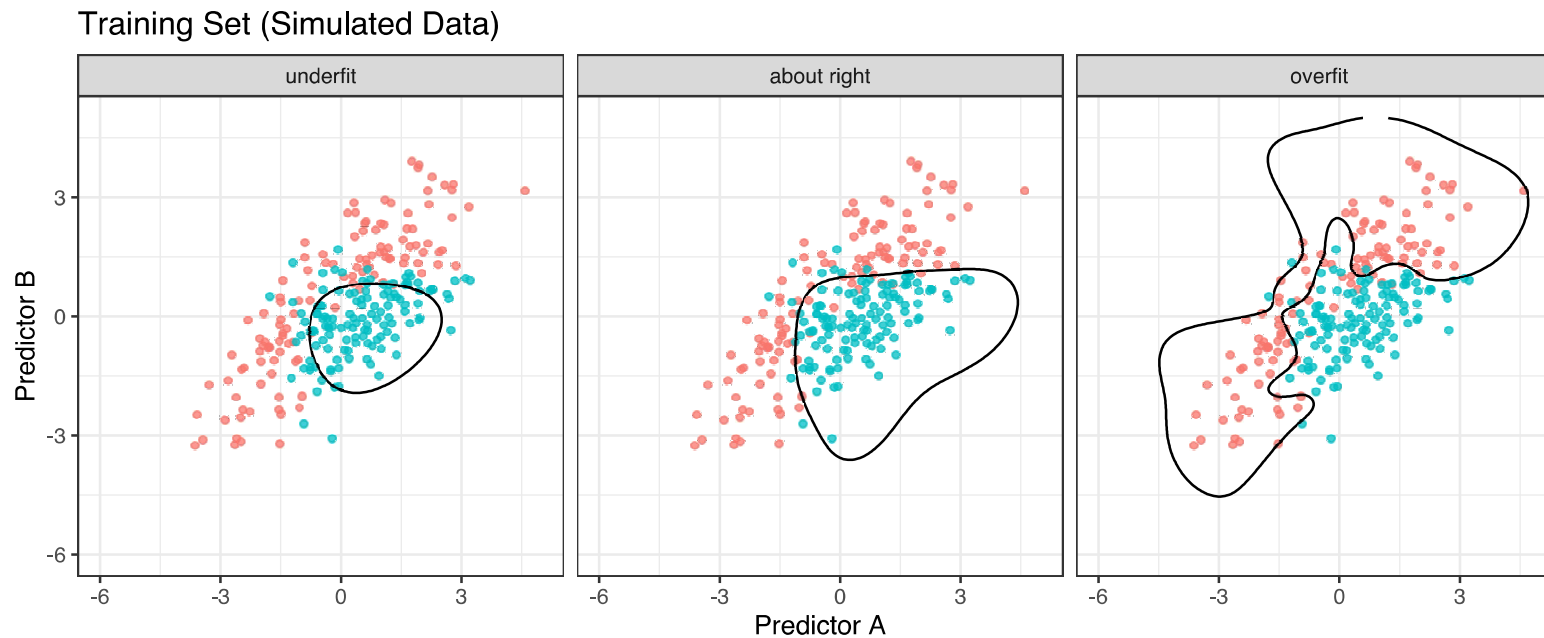
ROC curve

We can use the area under the ROC curve as a classification metric:

- ROC AUC = 1 😄
- ROC AUC = 1/2 😭

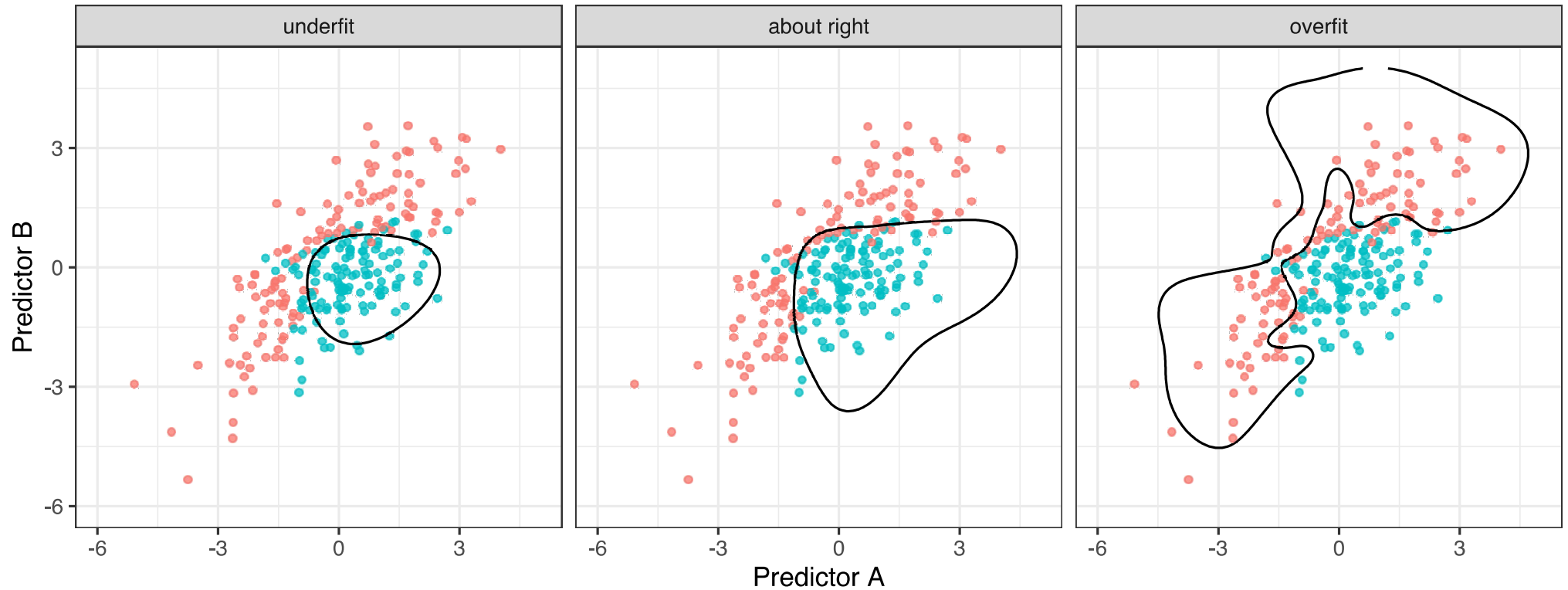
OVERFITTING

Overfitting is the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit to additional data or predict future observations reliably.



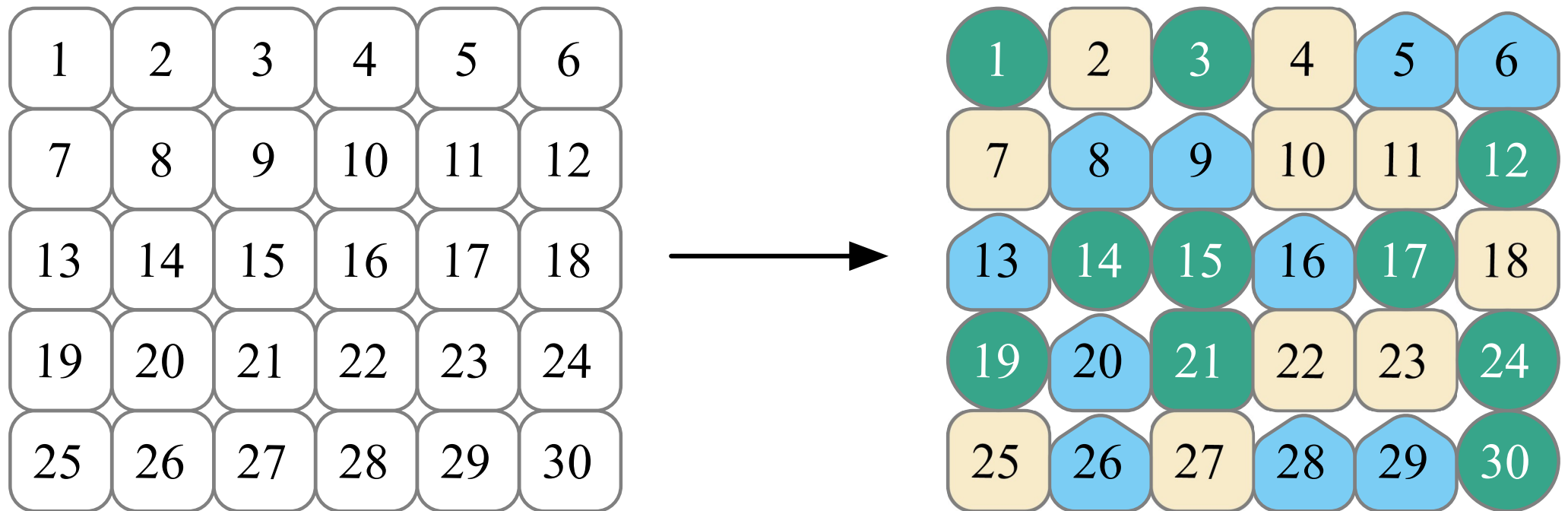
Overfitting

Test Set (Simulated Data)

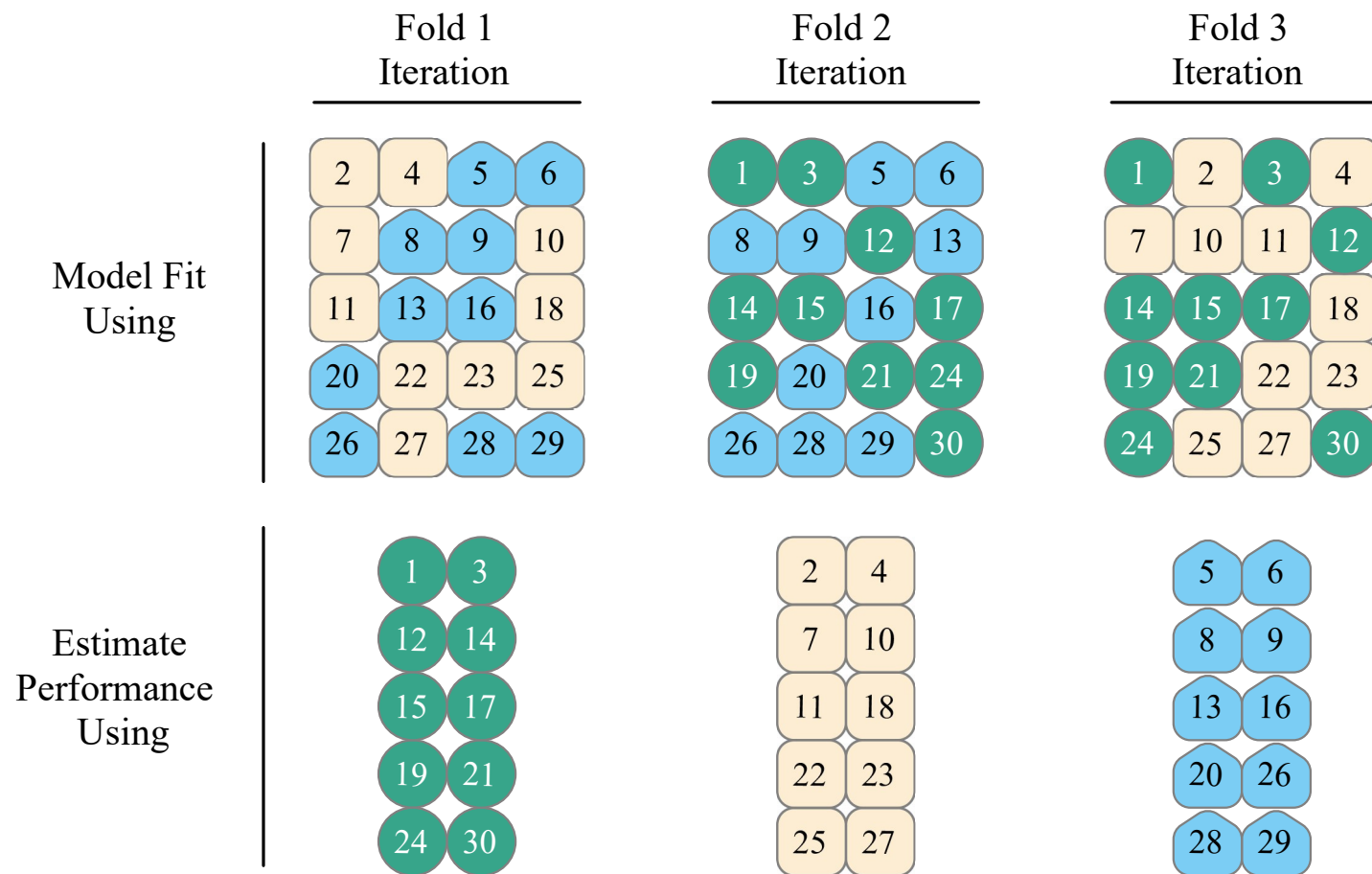


How can we use the *training* data to compare and evaluate different models?

Cross-validation



Cross-validation



Alternate resampling schemes

Bootstrapping

	Bootstrap Iteration 1	Bootstrap Iteration 2	Bootstrap Iteration 3																																																																																										
Model Fit Using	<table><tr><td>1</td><td>1</td><td>4</td><td>7</td><td>8</td><td>8</td></tr><tr><td>10</td><td>13</td><td>13</td><td>13</td><td>14</td><td>15</td></tr><tr><td>16</td><td>16</td><td>16</td><td>17</td><td>19</td><td>19</td></tr><tr><td>21</td><td>22</td><td>23</td><td>23</td><td>24</td><td>23</td></tr><tr><td>25</td><td>25</td><td>25</td><td>27</td><td>28</td><td>29</td></tr></table>	1	1	4	7	8	8	10	13	13	13	14	15	16	16	16	17	19	19	21	22	23	23	24	23	25	25	25	27	28	29	<table><tr><td>2</td><td>2</td><td>3</td><td>3</td><td>3</td><td>4</td></tr><tr><td>4</td><td>4</td><td>6</td><td>6</td><td>7</td><td>10</td></tr><tr><td>11</td><td>12</td><td>12</td><td>14</td><td>14</td><td>15</td></tr><tr><td>17</td><td>17</td><td>18</td><td>21</td><td>22</td><td>22</td></tr><tr><td>23</td><td>23</td><td>28</td><td>27</td><td>28</td><td>30</td></tr></table>	2	2	3	3	3	4	4	4	6	6	7	10	11	12	12	14	14	15	17	17	18	21	22	22	23	23	28	27	28	30	<table><tr><td>2</td><td>2</td><td>3</td><td>3</td><td>4</td><td>5</td></tr><tr><td>5</td><td>5</td><td>6</td><td>7</td><td>10</td><td>11</td></tr><tr><td>12</td><td>15</td><td>16</td><td>18</td><td>18</td><td>19</td></tr><tr><td>19</td><td>20</td><td>20</td><td>20</td><td>21</td><td>21</td></tr><tr><td>21</td><td>21</td><td>22</td><td>22</td><td>29</td><td>30</td></tr></table>	2	2	3	3	4	5	5	5	6	7	10	11	12	15	16	18	18	19	19	20	20	20	21	21	21	21	22	22	29	30
	1	1	4	7	8	8																																																																																							
	10	13	13	13	14	15																																																																																							
	16	16	16	17	19	19																																																																																							
	21	22	23	23	24	23																																																																																							
25	25	25	27	28	29																																																																																								
2	2	3	3	3	4																																																																																								
4	4	6	6	7	10																																																																																								
11	12	12	14	14	15																																																																																								
17	17	18	21	22	22																																																																																								
23	23	28	27	28	30																																																																																								
2	2	3	3	4	5																																																																																								
5	5	6	7	10	11																																																																																								
12	15	16	18	18	19																																																																																								
19	20	20	20	21	21																																																																																								
21	21	22	22	29	30																																																																																								
Estimate Performance Using	<table><tr><td>2</td><td>3</td><td>5</td><td>6</td><td>9</td><td>11</td></tr><tr><td>12</td><td>18</td><td>20</td><td>24</td><td>26</td><td>28</td></tr><tr><td></td><td></td><td>30</td><td></td><td></td><td></td></tr></table>	2	3	5	6	9	11	12	18	20	24	26	28			30				<table><tr><td>1</td><td>5</td><td>8</td><td>9</td><td>13</td><td>16</td></tr><tr><td>19</td><td>20</td><td>24</td><td>26</td><td>29</td><td></td></tr></table>	1	5	8	9	13	16	19	20	24	26	29		<table><tr><td>1</td><td>8</td><td>9</td><td>13</td><td>14</td><td>17</td></tr><tr><td>23</td><td>24</td><td>25</td><td>26</td><td>27</td><td>28</td></tr></table>	1	8	9	13	14	17	23	24	25	26	27	28																																																
	2	3	5	6	9	11																																																																																							
	12	18	20	24	26	28																																																																																							
		30																																																																																											
1	5	8	9	13	16																																																																																								
19	20	24	26	29																																																																																									
1	8	9	13	14	17																																																																																								
23	24	25	26	27	28																																																																																								

A validation set is just another type of resampling.

Machine Learning Model Classes

Training a machine learning model using the `caret` package is deceptively simple in RStudio. We simply use the `train` function, specify our outcome variable and data set, and specify the model we would like to apply via the argument `method =`

However, there are over 200 available models in the `caret` package.

Each model will have a different composition and different requirements, which we refer to as **tuning parameters**.

The main classes of machine learning

- Linear Discriminant Analysis
- Decision Trees
- Regression Trees
- Random Forests
- Support Vector Machines
- K Nearest Neighbors
- Multiple linear regression
 - We won't go into all the mathematics behind these models. Rather, we will focus on the R code required to apply several of these to our data.

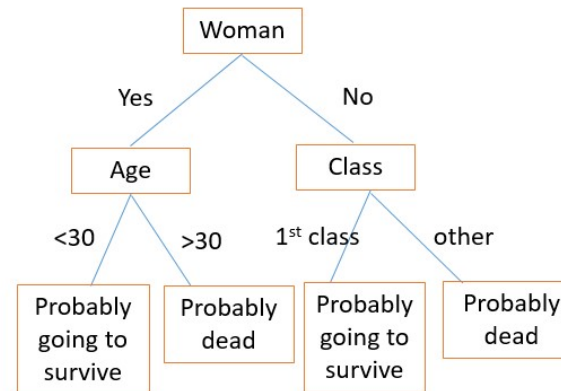
Decision trees

The simplest type of tree model is a **Decision Tree model**. The easiest way to describe a decision tree model is probably to show one.

A normal tree



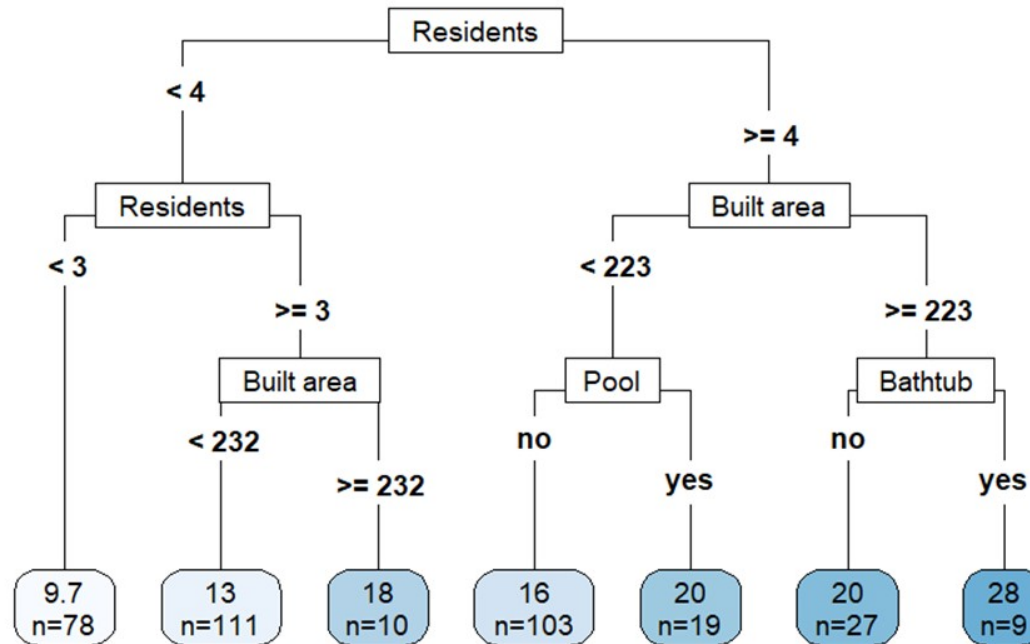
A decision tree



<https://www.kaggle.com/code/akashchola/decision-tree-for-classification-regression>

Regression Trees

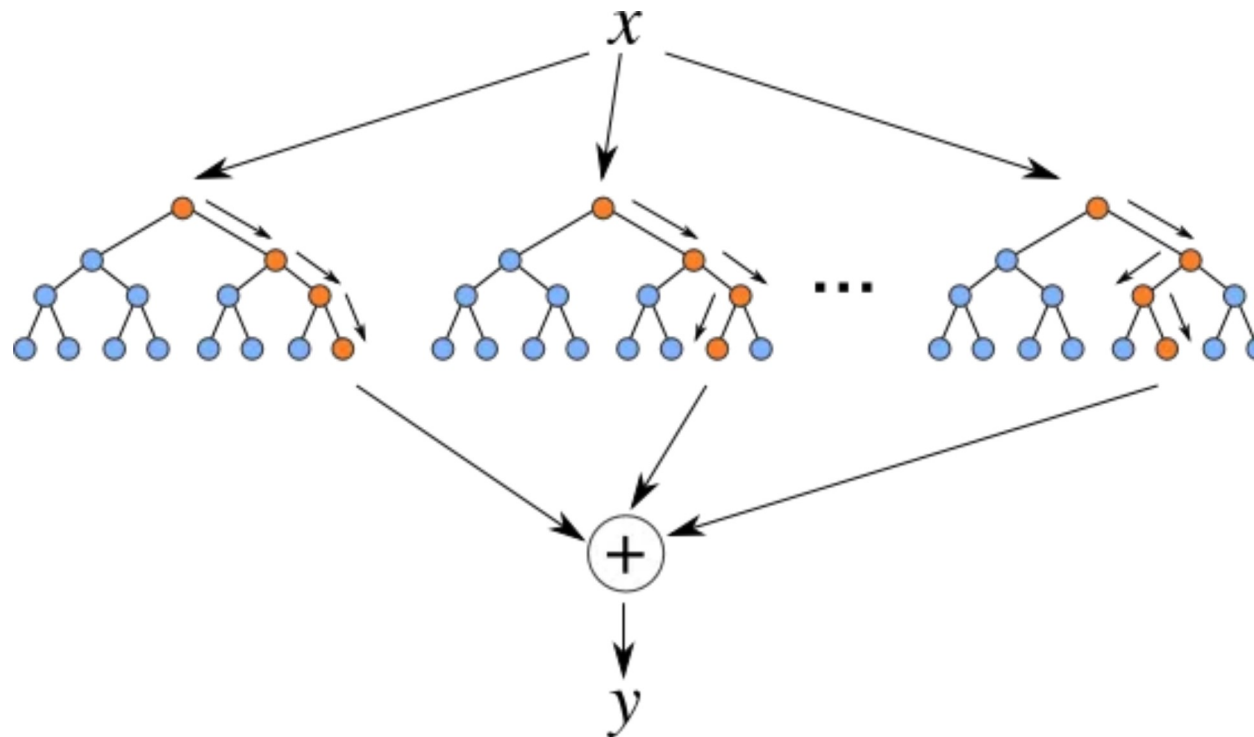
A regression tree is like a decision tree, except that the regression model fitting process involved is more sophisticated.



<https://www.sciencedirect.com/science/article/abs/pii/S2210670722004991>

Random Forest

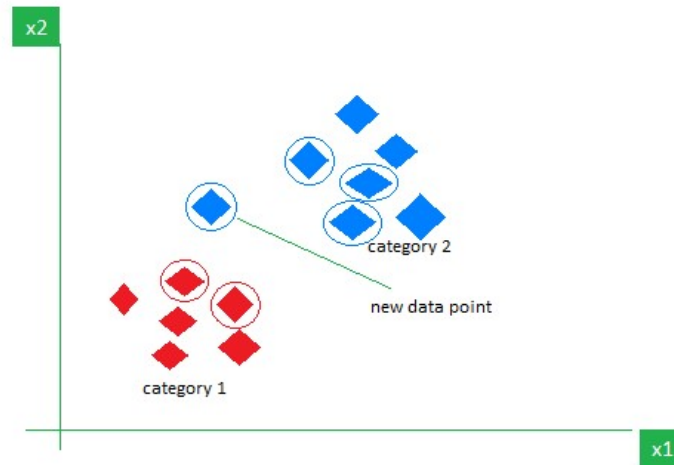
Random forest models combine multiple decision trees to achieve better results than any single decision tree considered could offer.



<https://blog.toadworld.com/2018/08/31/random-forest-machine-learning-in-r-python-and-sql-part-1>

K Nearest Neighbor KNN

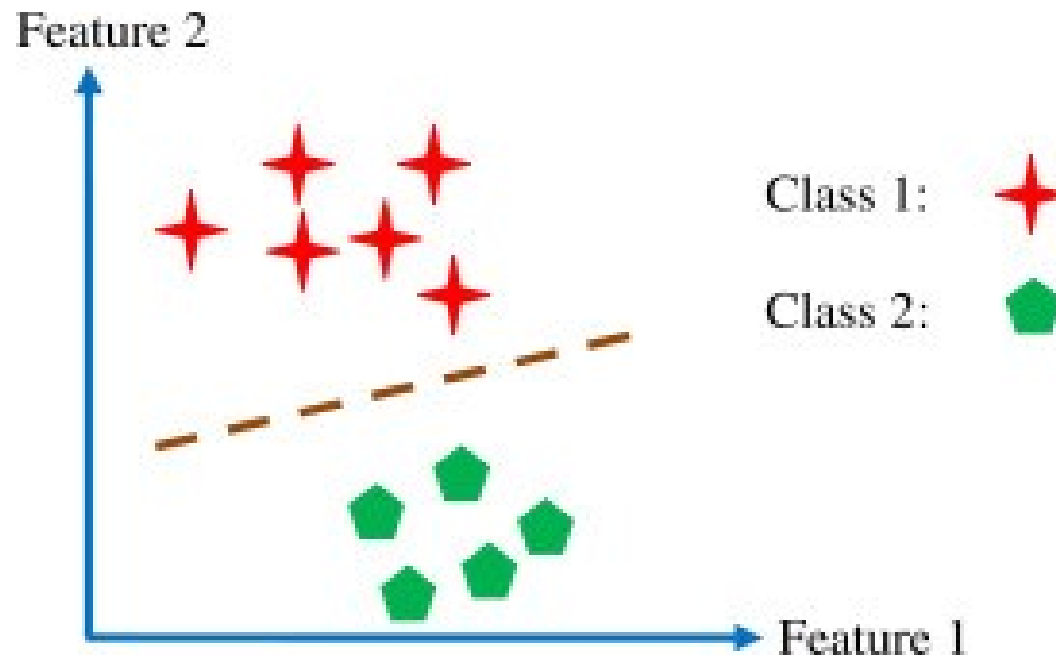
Nearest Neighbour are another class of non-linear models, that assess distances between observations, grouping nearby observations together - a bit like k-means clustering.



<https://www.geeksforgeeks.org/k-nn-classifier-in-r-programming/>

Linear discriminant analysis LDA

A linear discriminant analysis (LDA) model is a type of linear model that uses Bayes' theorem to classify new observations based on characteristics of the outcome variable classes.



<https://www.sciencedirect.com/topics/computer-science/linear-discriminant>

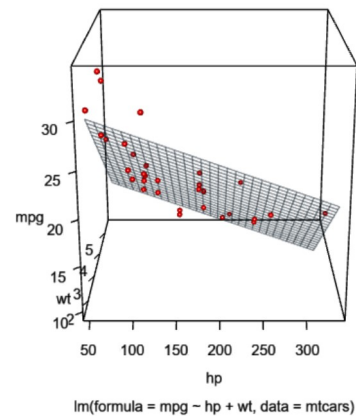
Support vector machine SVM

A support vector machine (SVM) is a type of non-linear model that operates similarly to LDA models, with a focus on clearly separating outcome variable classes.

By Original: Alisneaky Vector: Zirguezi - Own work based on: Kernel Machine.png, CC BY-SA 4.0

Multiple linear regression

Linear regression is a linear approach for modelling the relationship between a quantitative response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.



<https://rpubs.com/cardiomoon/474707>

References

- <https://workshops.tidymodels.org/>
 - Licensed under Creative Commons Attribution CC BY-SA 4.0.
- R for Data Science (2e) (hadley.nz)
- <https://www.tidymodels.org/>

<https://www.tidymodels.org/learn/>

This material is made with [Quarto](#).