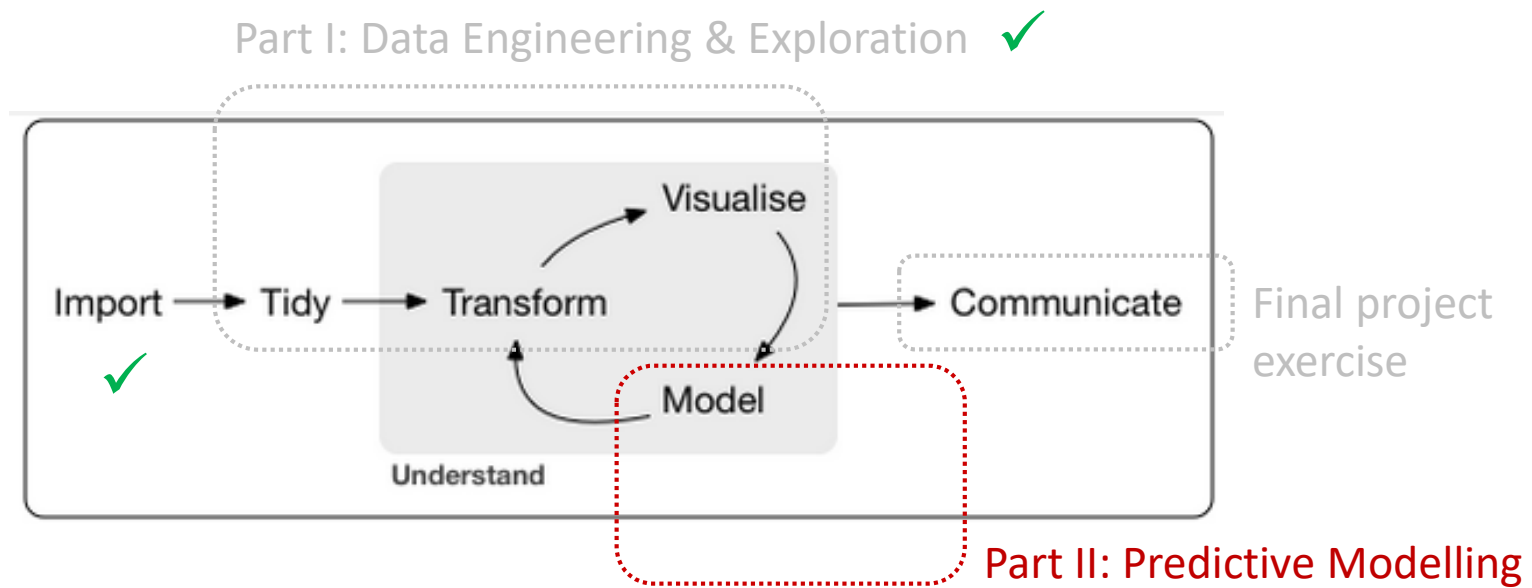


Data Science

Predictive Modelling

Gero Szepannek

Data Science Process



Prediction Model



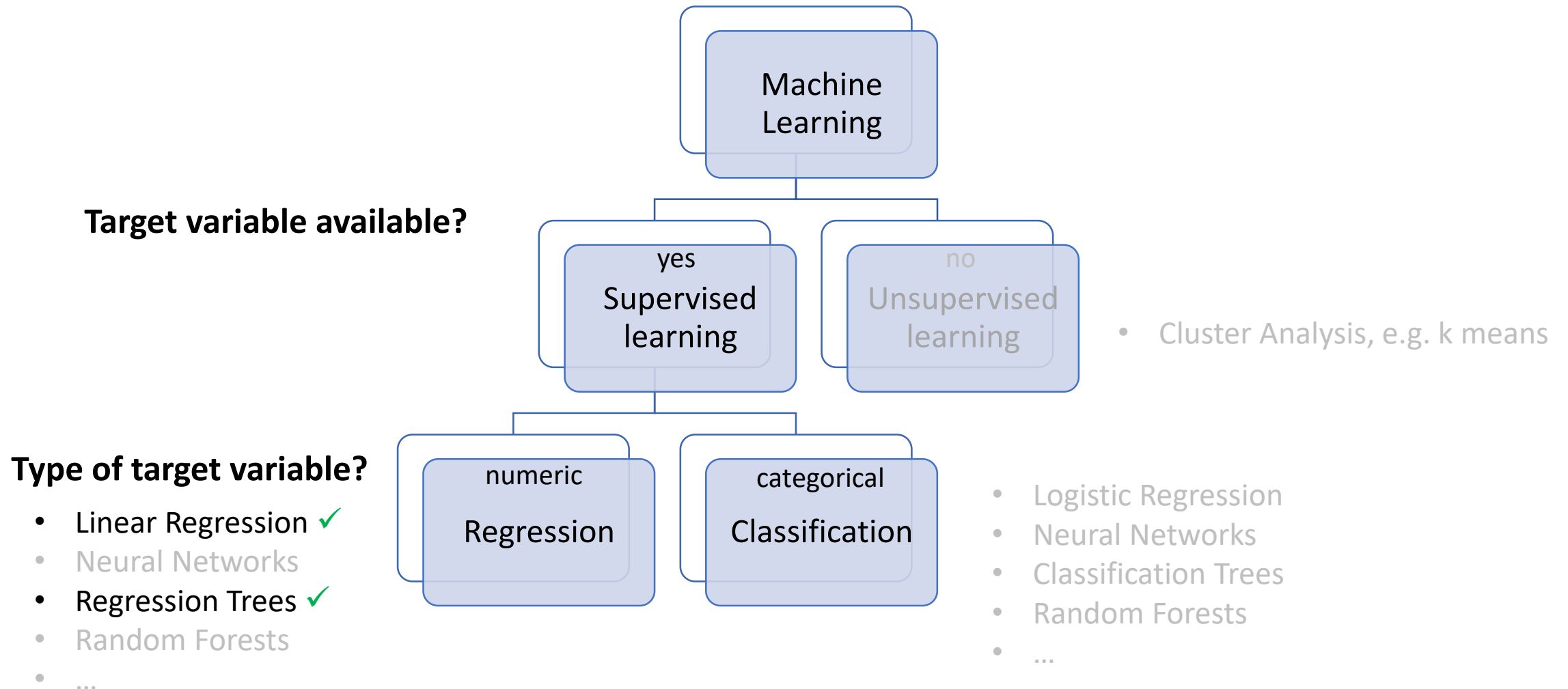
$$y = f(x) + \varepsilon$$

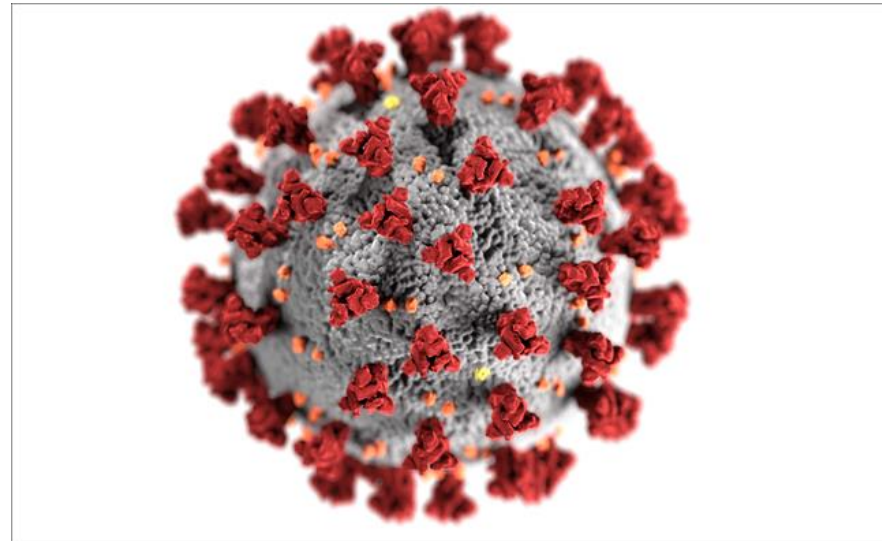


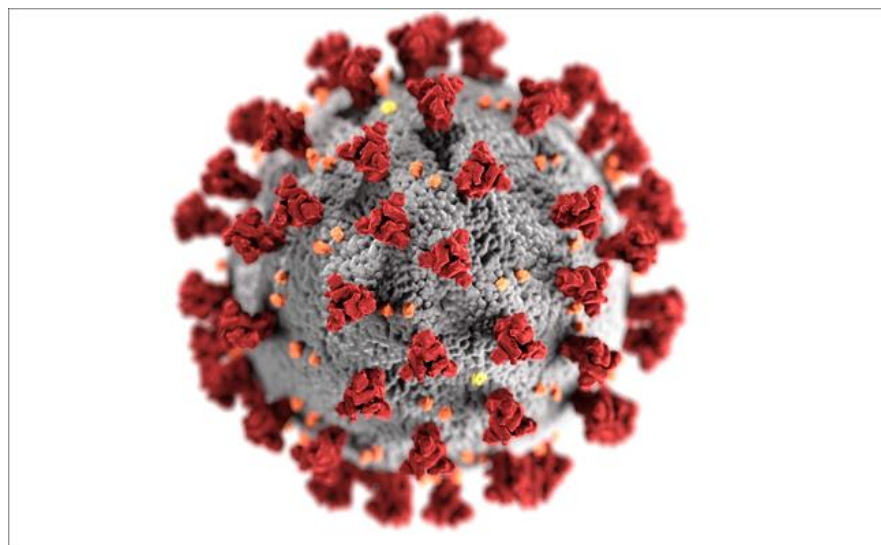
$$\hat{y}_i = \hat{f}(x_i)$$



Regression vs. Classification







amazon

A Point-and-click Framework developed @

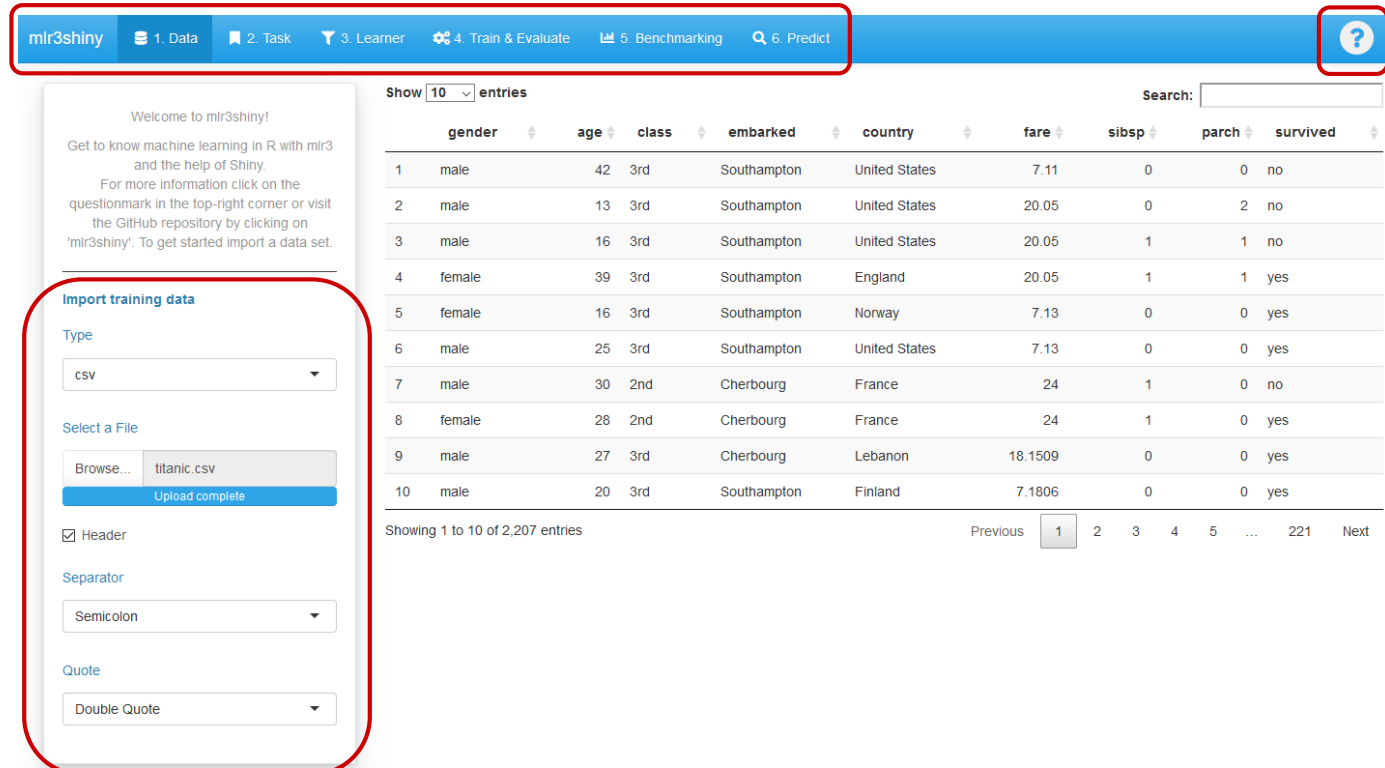
```
# install
```

```
remotes::install_github("https://github.com/LamaTe/mlr3shiny.git")
```

```
# start
```

```
library(mlr3shiny)
```

```
launchMlr3Shiny()
```

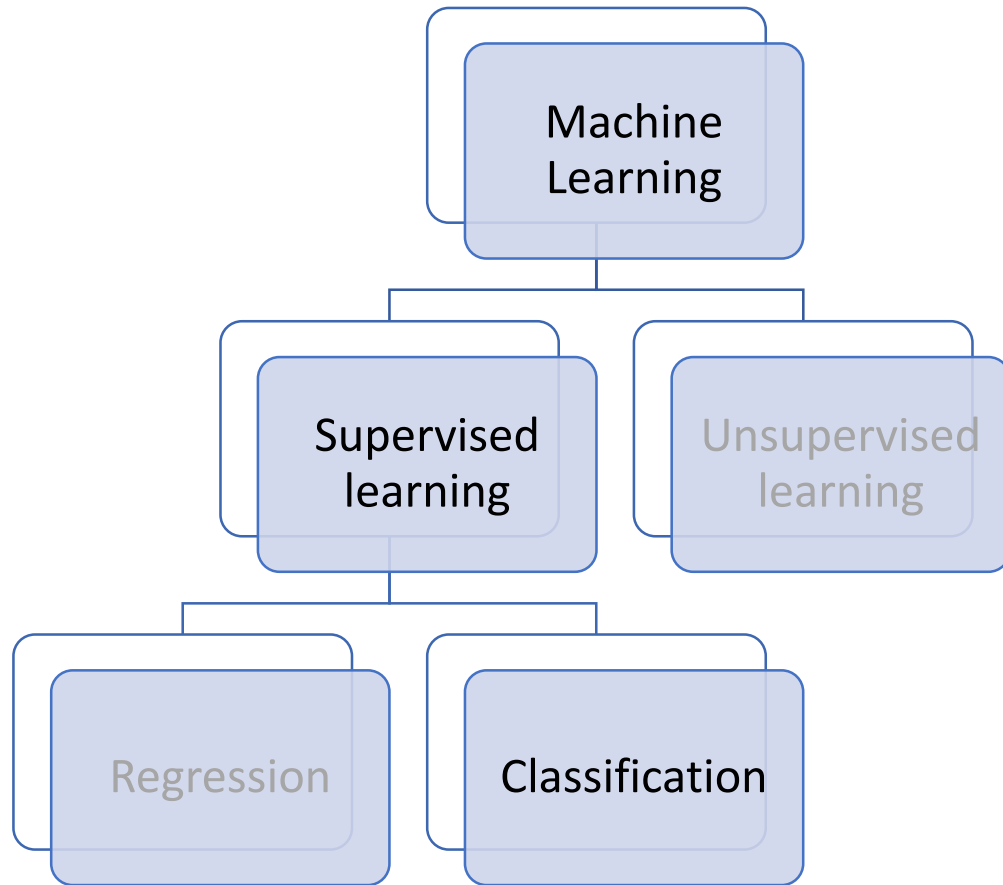


The screenshot shows the mlr3shiny web interface. The top navigation bar includes tabs for '1. Data', '2. Task', '3. Learner', '4. Train & Evaluate', '5. Benchmarking', and '6. Predict'. A red box highlights the '1. Data' tab and a question mark icon in the top right corner. The main content area displays a table of 10 entries from the Titanic dataset, with columns for gender, age, class, embarked, country, fare, sibsp, parch, and survived. A red box highlights the 'Import training data' section on the left, which includes a 'Type' dropdown set to 'CSV', a 'Select a File' section with a 'Browse...' button and a 'titanic.csv' file listed, and an 'Upload complete' button. Below this, there are checkboxes for 'Header' and 'Separator' (set to 'Semicolon'), and a 'Quote' dropdown set to 'Double Quote'.

| | gender | age | class | embarked | country | fare | sibsp | parch | survived |
|----|--------|-----|-------|-------------|---------------|---------|-------|-------|----------|
| 1 | male | 42 | 3rd | Southampton | United States | 7.11 | 0 | 0 | no |
| 2 | male | 13 | 3rd | Southampton | United States | 20.05 | 0 | 2 | no |
| 3 | male | 16 | 3rd | Southampton | United States | 20.05 | 1 | 1 | no |
| 4 | female | 39 | 3rd | Southampton | England | 20.05 | 1 | 1 | yes |
| 5 | female | 16 | 3rd | Southampton | Norway | 7.13 | 0 | 0 | yes |
| 6 | male | 25 | 3rd | Southampton | United States | 7.13 | 0 | 0 | yes |
| 7 | male | 30 | 2nd | Cherbourg | France | 24 | 1 | 0 | no |
| 8 | female | 28 | 2nd | Cherbourg | France | 24 | 1 | 0 | yes |
| 9 | male | 27 | 3rd | Cherbourg | Lebanon | 18.1509 | 0 | 0 | yes |
| 10 | male | 20 | 3rd | Southampton | Finland | 7.1806 | 0 | 0 | yes |

Classification

\hat{y}_i



$$\hat{y}_i = \hat{f}(x_i)$$

1. In classification y_i is categorical.
2. ...most relevant practical case: **binary classification** $\rightarrow y_i \in \{0,1\}$

Classification

mlr3shiny

1. Data

2. Task

3. Learner

4. Train & Evaluate

5. Benchmarking

6. Predict

Define a Task

Select Data Backend

imported training data

Task ID

my_task

Task Target

survived

Create Task

Task Overview

Supervised Task:

twoclass classif

Task ID:

my_task

Data:

9 Variables with 2207 Observations

Target:

survived

Positive Class:

no

Features:

| id | type |
|----------|---------|
| age | numeric |
| class | factor |
| country | factor |
| embarked | factor |

Task Processing

Set Positive Class:

Nothing selected

Change

Drop Features:

Nothing selected

Drop

Classification Algorithms

The screenshot displays the mlr3shiny web interface. The top navigation bar includes tabs for '1. Data', '2. Task', '3. Learner' (highlighted with a red box), '4. Train & Evaluate', '5. Benchmarking', and '6. Predict'. On the left, the 'Select a machine-learning algorithm' panel shows a dropdown menu with 'decision tree' selected (also highlighted with a red box). The dropdown list includes 'decision tree', 'random forest', 'support vector machine', and 'logistic regression'. A 'Go' button is next to the dropdown. The main content area is divided into two sections: 'Learner Information' and 'Learner Parameters'. The 'Learner Information' section shows the selected algorithm as 'Decision Tree', the current predict type as 'response', and supported predict types as 'response, prob'. The 'Learner Parameters' section shows parameters for 'minsplit' (Lower: 1, Upper: Inf, value: 20), 'cp' (Lower: 0, Upper: 1, value: 0,01), and 'maxdepth' (Lower: 1, Upper: 30, value: 30). A 'Change Parameters' button is located below these parameters. At the bottom, there is a 'Change Predict Type' section with a dropdown menu set to 'response' and a 'Change' button.

mlr3shiny 1. Data 2. Task 3. Learner 4. Train & Evaluate 5. Benchmarking 6. Predict ?

Select a machine-learning algorithm

Learner 1

decision tree decision tree random forest support vector machine logistic regression Go

Learner 1

Learner Information

| | | | |
|--------------------------|----------------|--------------------------|--|
| Algorithm: | Decision Tree | Properties: | importance, missings, multiclass, selected_features, twoclass, weights |
| Current Predict Type: | response | | |
| Supported Predict Types: | response, prob | Supported Feature Types: | logical, integer, numeric, factor, ordered |

Learner Parameters

| | | | |
|----------|----------|------------|------|
| minsplit | Lower: 1 | Upper: Inf | 20 |
| cp | Lower: 0 | Upper: 1 | 0,01 |
| maxdepth | Lower: 1 | Upper: 30 | 30 |

Change Parameters

Change Predict Type response Change

Flavours of Classification Models

\hat{y}_i

$$\hat{y}_i = \hat{f}(x_i)$$

1. In classification y_i is categorical.
2. ...most relevant practical case: binary classification $\rightarrow y_i \in \{0,1\}$
3. Predicting **class labels**: $\hat{y}_i \in \{0,1\}$
4. Predicting **posterior probabilities**: $\hat{P}(Y_i = 1) \in [0,1]$



What is a natural rule for prediction of class labels in order to minimize the probability of making an error?

$$\hat{y}_i = \begin{cases} 1 & \dots \text{if } \hat{P}(Y_i = 1) \geq \textit{cut off} \\ 0 & \dots \text{if } \hat{P}(Y_i = 1) < \textit{cut off} \end{cases}$$

Classification Algorithms

mlr3shiny

1. Data

2. Task

3. Learner

4. Train & Evaluate

5. Benchmarking

6. Predict

?

Select a machine-learning algorithm

Learner 1

decision tree

Go

+ Add Learner

Learner 1

Learner Information

| | | | |
|--------------------------|----------------|--------------------------|--|
| Algorithm: | Decision Tree | Properties: | importance, missings, multiclass, selected_features, twoclass, weights |
| Current Predict Type: | response | | |
| Supported Predict Types: | response, prob | Supported Feature Types: | logical, integer, numeric, factor, ordered |

Learner Parameters

| | | | |
|----------|----------|------------|------|
| minsplit | Lower: 1 | Upper: Inf | 20 |
| cp | Lower: 0 | Upper: 1 | 0,01 |
| maxdepth | Lower: 1 | Upper: 30 | 30 |

Change Parameters

Change Predict Type

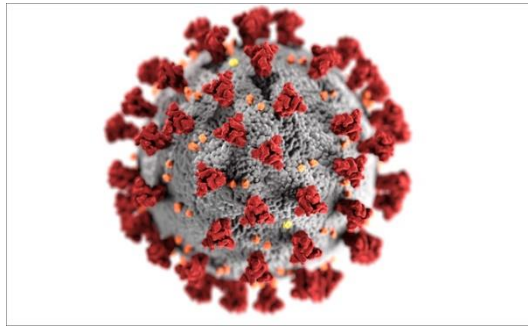
response

response

prob

Change

Performance Measures for Classification

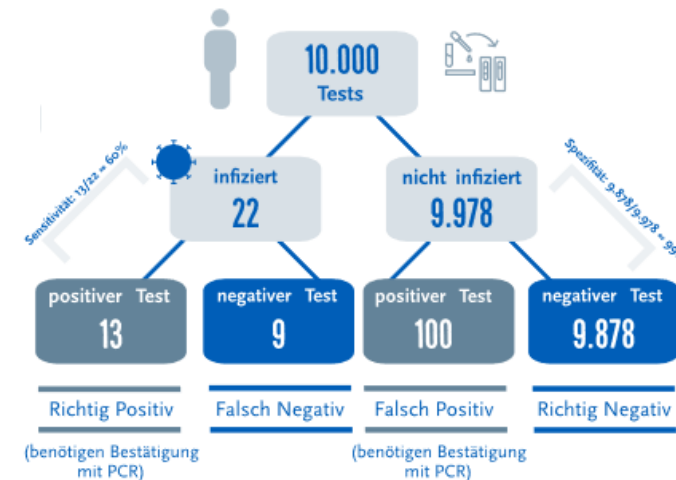


1. Accuracy
2. Sensitivity
3. Specificity
4. Precision

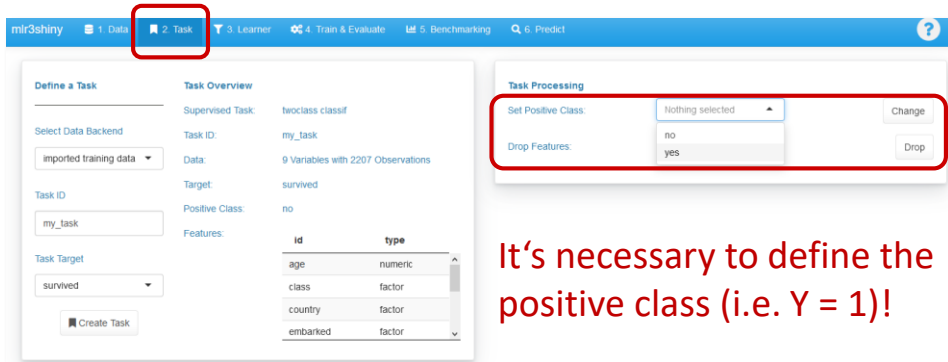
Test detects infection.

$$P(\hat{Y} = 1 | Y = 1)$$

$$P(\hat{Y} = 0 | Y = 0)$$



Confusion Matrix



It's necessary to define the positive class (i.e. $Y = 1$)!

| | $\hat{Y} = 1$ | $\hat{Y} = 0$ | |
|---------|---------------|---------------|--|
| $Y = 1$ | | | |
| $Y = 0$ | | | |
| | | | |

1. Accuracy

2. Sensitivity

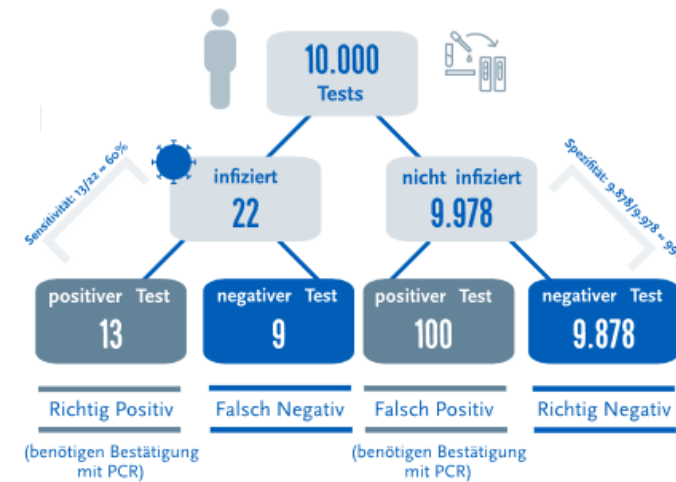
3. Specificity

4. Precision

Test detects infection.

$$P(\hat{Y} = 1 | Y = 1)$$

$$P(\hat{Y} = 0 | Y = 0)$$



Confusion Matrix

| | $\hat{Y} = 1$ | $\hat{Y} = 0$ | |
|---------|---------------|---------------|--|
| $Y = 1$ | | | |
| $Y = 0$ | | | |
| | | | |

1. Accuracy

2. Sensitivity

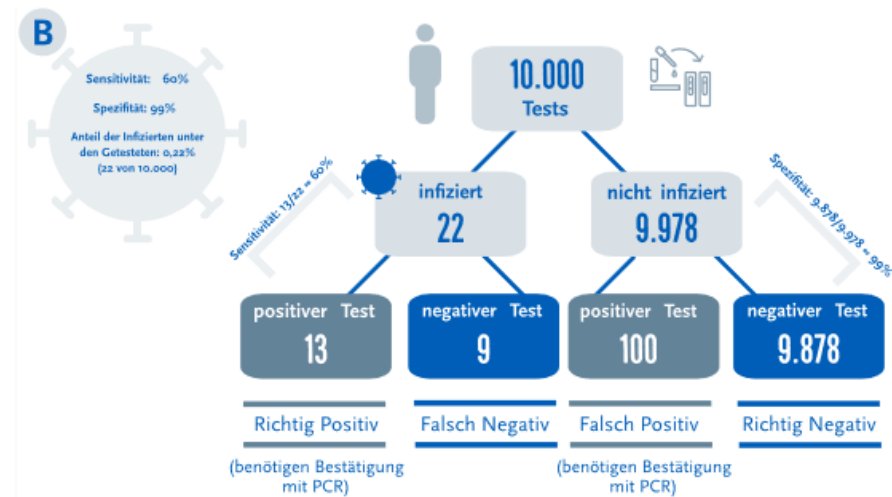
3. Specificity

4. Precision

Test detects infection.

$$P(\hat{Y} = 1|Y = 1)$$

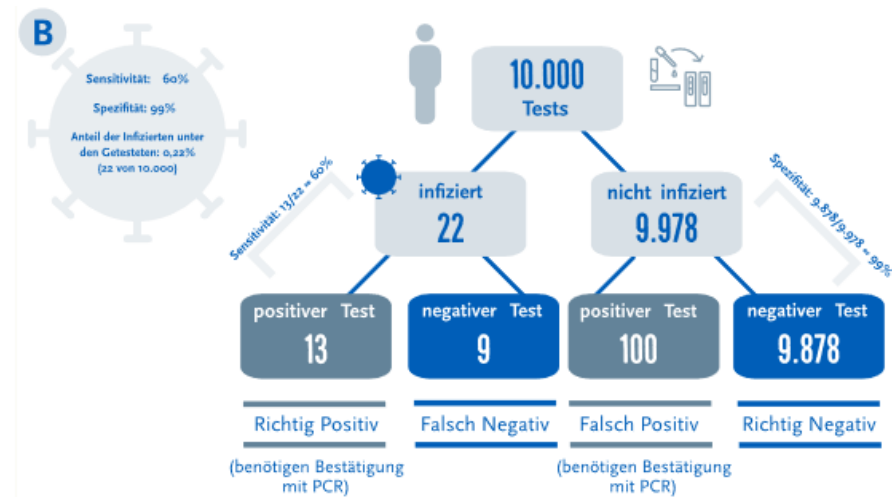
$$P(\hat{Y} = 0|Y = 0)$$



Confusion Matrix

| | $\hat{Y} = 1$ | $\hat{Y} = 0$ | |
|---------|---------------|---------------|--|
| $Y = 1$ | | | |
| $Y = 0$ | | | |
| | | | |

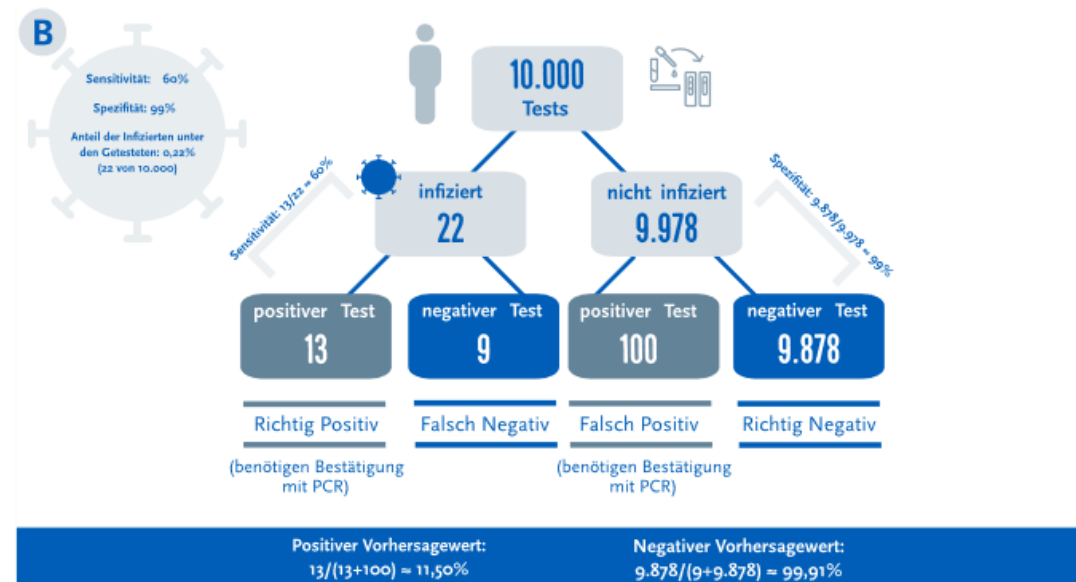
1. Accuracy
2. Sensitivity
3. Specificity
4. Precision



Confusion Matrix

| | $\hat{Y} = 1$ | $\hat{Y} = 0$ | |
|---------|---------------|---------------|--|
| $Y = 1$ | | | |
| $Y = 0$ | | | |
| | | | |

1. Accuracy
2. Sensitivity
3. Specificity
4. Precision / PPV

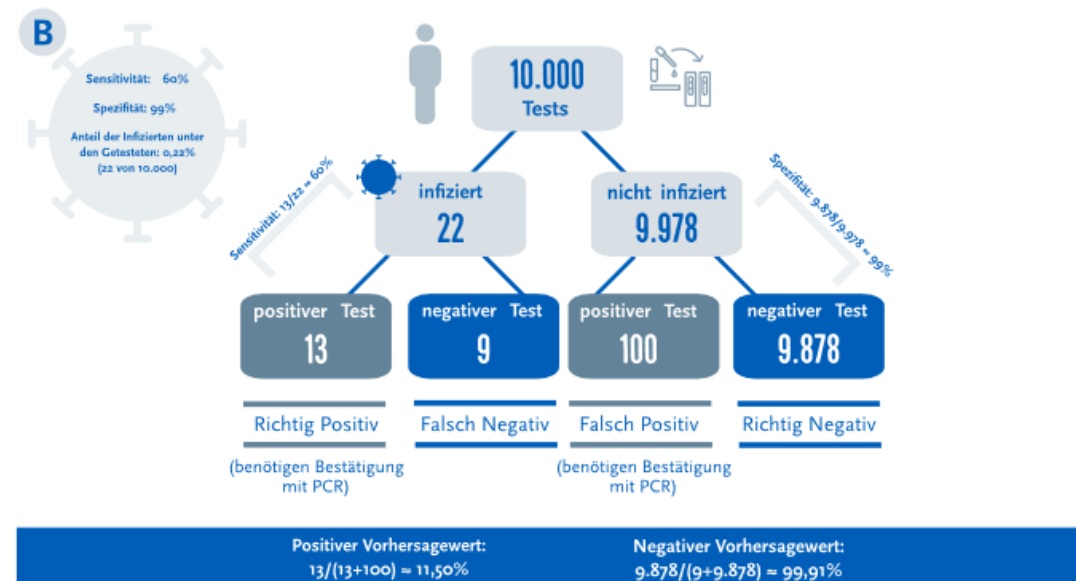


Confusion Matrix

| | $\hat{Y} = 1$ | $\hat{Y} = 0$ | |
|---------|---------------|---------------|--|
| $Y = 1$ | | | |
| $Y = 0$ | | | |
| | | | |

1. Imagine: How can we change the cut off in order to increase the sensitivity of the test?
2. ...What will typically happen with the precision (aka PPV) at the same time?

$$\hat{y}_i = \begin{cases} 1 & \dots \text{if } \hat{P}(Y_i = 1) \geq \text{cut off} \\ 0 & \dots \text{if } \hat{P}(Y_i = 1) < \text{cut off} \end{cases}$$



Takeaway...

Bei Veranstaltungen nicht auf Schnelltests verlassen!

Deshalb sei es gefährlich, sich bei Einlasskontrollen auf das Ergebnis eines Schnelltests zu verlassen - etwa beim Theater- oder Konzert-Besuch, an der Eingangstür eines Restaurants. Wenn ein Schnelltest eine Infektion "übersieht", wird diese Person herumlaufen in der Annahme, dass sie nicht ansteckend ist - und kann so mitunter andere infizieren. "Es ist nicht so simpel, wie es in der Politik dargestellt wird - nach dem Motto: Jetzt kann alles öffnen, weil wir ja die Schnelltests haben." Zwischen 40 Prozent und 60 Prozent der Infektionen werden bei Schnelltests übersehen, so Drosten.

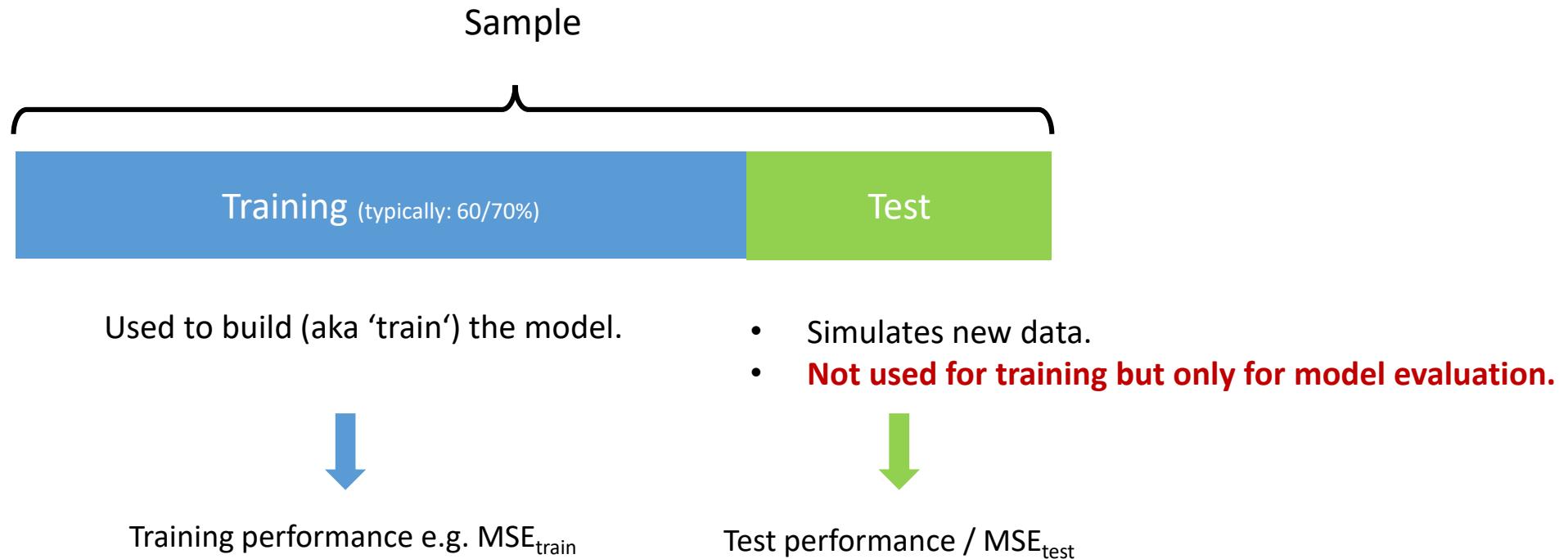
An Schulen machen Antigen-Schnelltests Sinn

Drosten stellt aber auch klar: In Schulen sei der Einsatz von Antigen-Schnelltests trotzdem gerechtfertigt - wenn die Schülerinnen und Schüler mindestens zweimal in der Woche getestet werden. "Selbst wenn bei einer Testung nicht alle Infektionen entdeckt werden, bei der nächsten Testung nach zwei oder drei Tagen werden die Infektionen dann nachgewiesen. In Clustern ist solch ein geringer zeitverzögerter Effekt kein Problem", meint der Virologe. Wichtig sei nur, Infektionen in einem Cluster aufzuspüren, um so die Kontrolle in den Schulen während der Pandemie zu behalten und dann entsprechend schnell mit Cluster-Quarantäne zu reagieren.

<https://www.ndr.de/nachrichten/info/Drosten-Schnelltests-sind-wohl-weniger-zuverlaessig-als-gedacht,coronavirusupdate178.html>

Test Data

- A model should have strong predictive power also for new data!
- In practice: no new data is available to test the model ☹️.
- Therefore use a trick: (Randomly) split the data in two parts:



Performance Evaluation

mlr3shiny 1. Data 2. Task 3. Learner 4. Train & Evaluate 5. Benchmarking 6. Predict

Basic Workflow

Use Resampling

Select a new learner:

☒ Learner1

Basic Workflow Overview

Task: my_task

Learner: Learner1 classif.rpart

State of Workflow: scored

Model: rpart

Performance Training Set: classif.acc: 0.817

Performance Test Set: classif.acc: 0.802

Train Model

Select percentage of training data for model training and evaluation. Set a seed to reproduce the random partitioning anytime. The default value to start sampling the training data from is 42.

42

0 70 100

Train model

Predict Target Variable

Use the trained model to predict the target values on the remaining test data of the training-test split as well as on the data partition the model was trained on.

Predict target

Measure Prediction Performance

accuracy

Score

Predicted Target on Test Data

| | row_ids | truth | response |
|----|---------|-------|----------|
| 1 | 5 | yes | no |
| 2 | 7 | no | no |
| 3 | 20 | no | no |
| 4 | 23 | yes | yes |
| 5 | 24 | no | yes |
| 6 | 28 | no | no |
| 7 | 34 | no | no |
| 8 | 38 | no | no |
| 9 | 40 | no | no |
| 10 | 41 | no | no |

Previous 1 2 3 4 5 ... 67 Next

Predicted Target on Training Data

| | row_ids | truth | response |
|---|---------|-------|----------|
| 1 | 1177 | no | no |
| 2 | 1098 | no | no |
| 3 | 1252 | yes | yes |

...Lab



On your own:

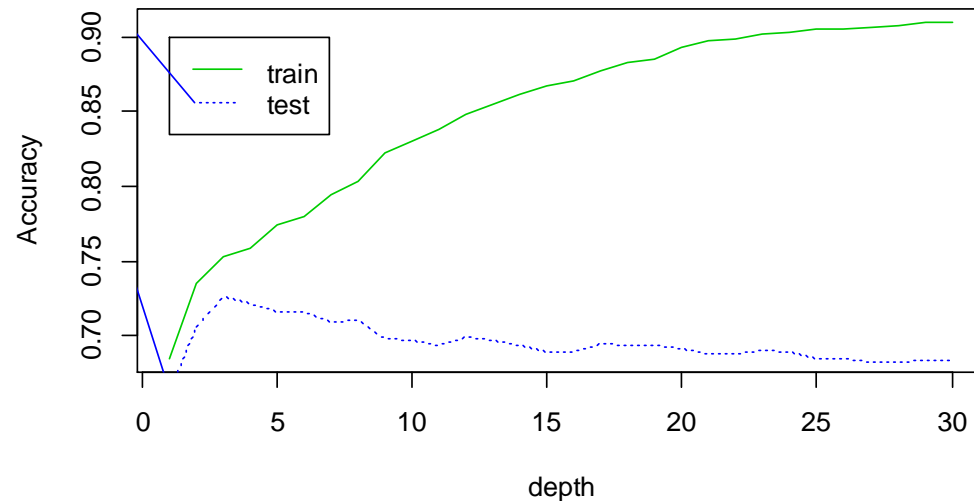
1. Modify the hyperparameters of the tree...
2. ...Try to find the best model, i.e. the model that maximizes accuracy!
3. What do you observe w.r.t. the accuracy on both:
 1. Training data...
 2. ...and Test data?

Learner Parameters

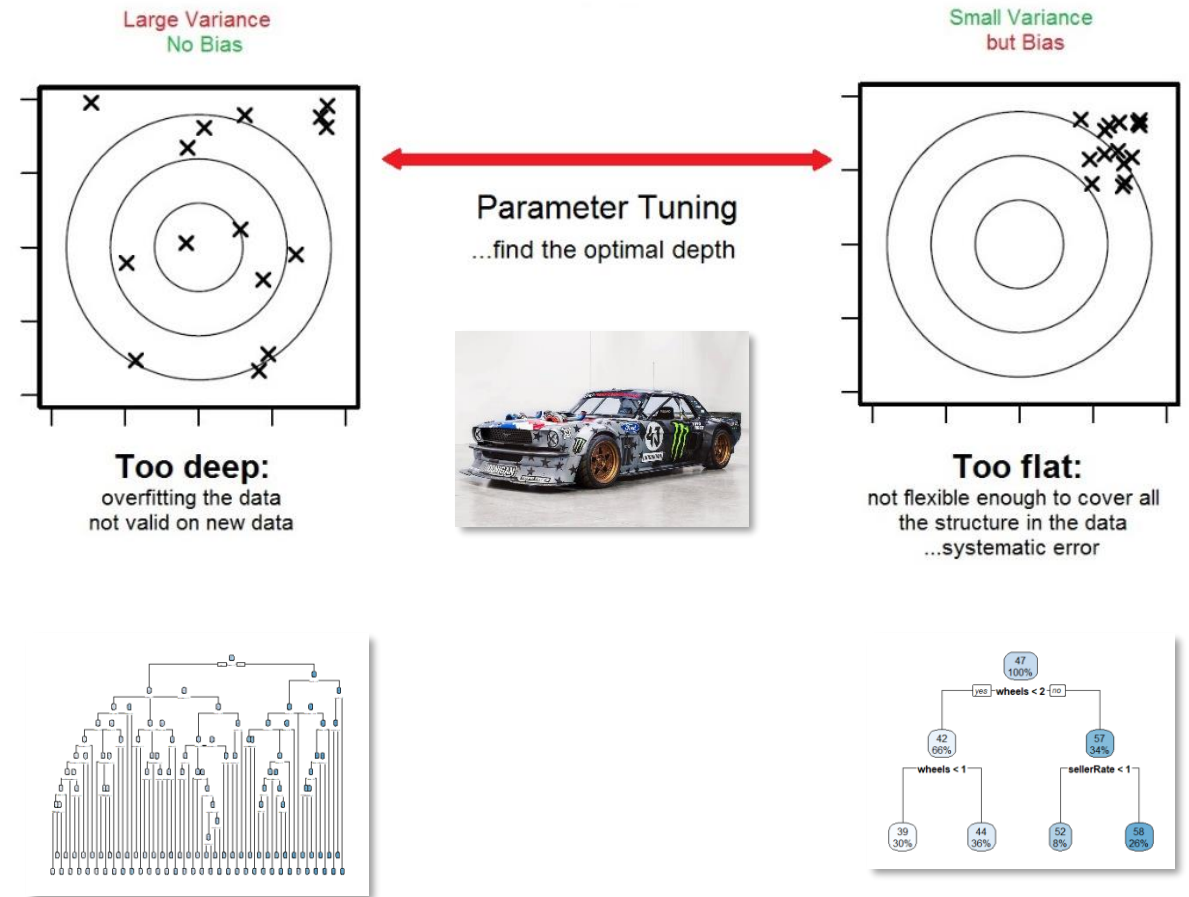
| | | | |
|----------|----------|------------|-----------------------------------|
| minsplit | Lower: 1 | Upper: Inf | <input type="text" value="20"/> |
| cp | Lower: 0 | Upper: 1 | <input type="text" value="0,01"/> |
| maxdepth | Lower: 1 | Upper: 30 | <input type="text" value="30"/> |

[Change Parameters](#)

Overfitting and Underfitting

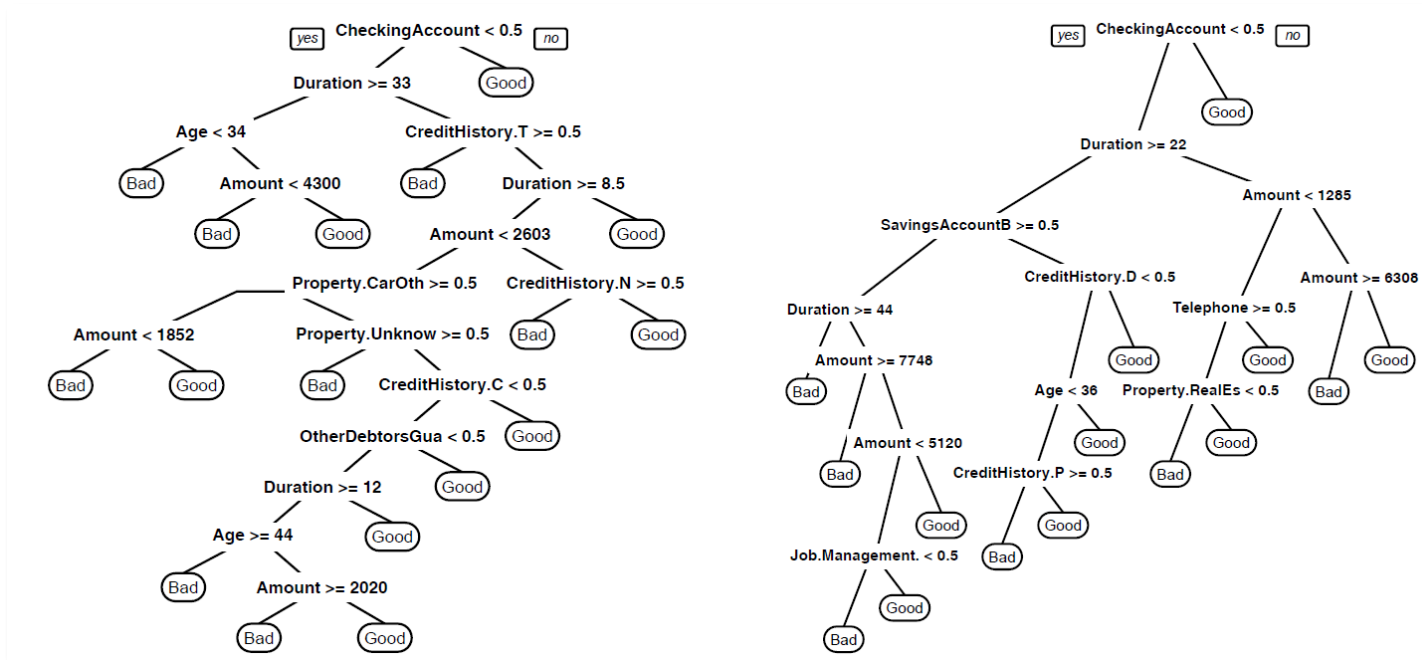


...e.g. set the parameters `minsplit = minbucket = 1` and `cp = 0` such that and then increase `maxdepth`.



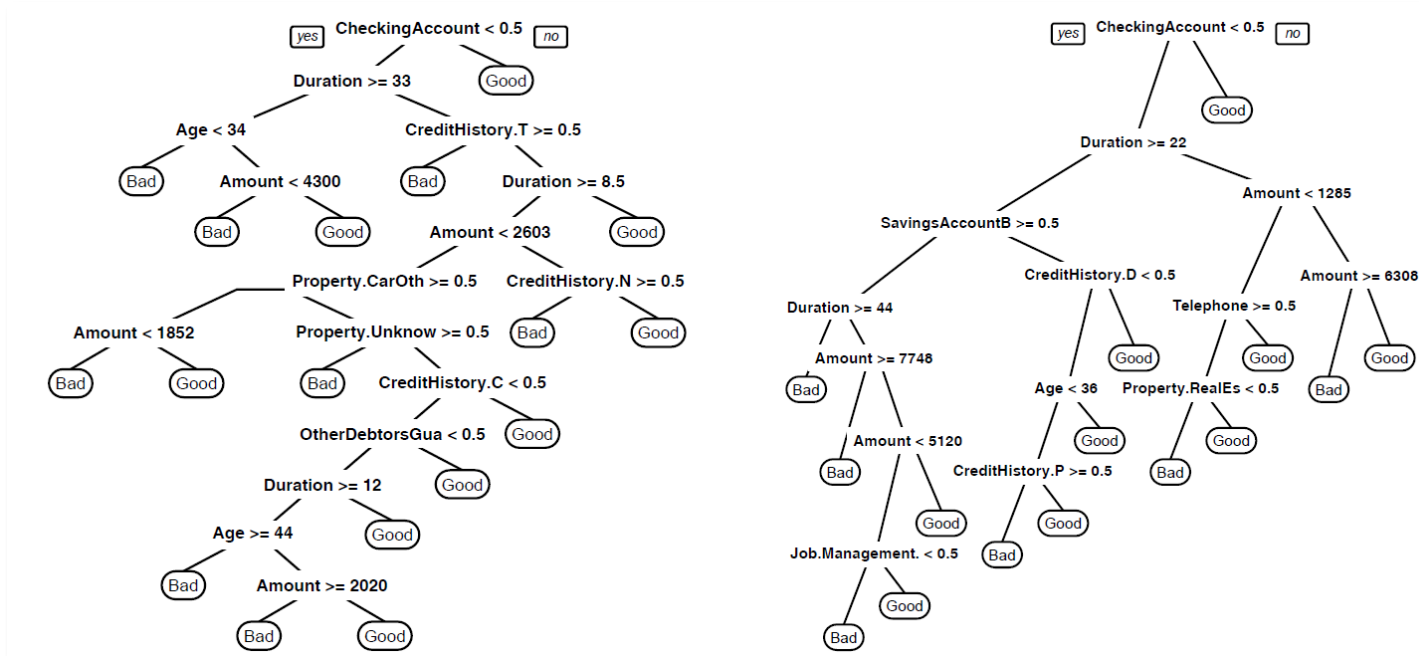
Variance of Models

- Deep trees will look pretty different on different samples...
- ...except for the first splits.
- **Try to explain, why?**



Variance of Models

- Deep trees will look pretty different on different samples...
- ...except for the first splits.
- **Try to explain, why?**



...from Statistics, we know:

$$Var(\bar{X}) = \frac{Var(X)}{n}$$

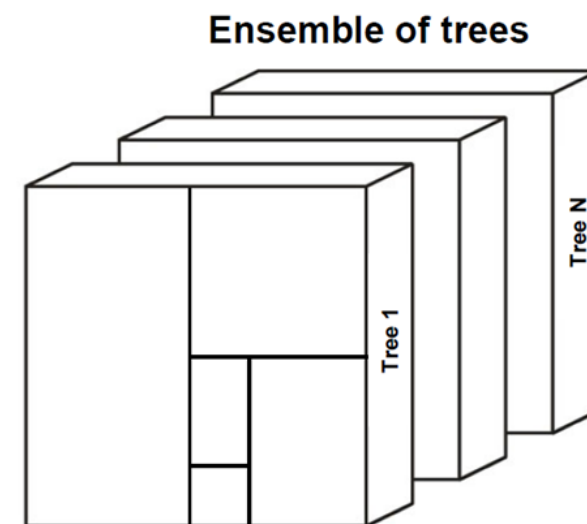
1. **What denotes \bar{X} in Statistics?**
2. ...How can we translate this to the variance of a single model?



Random Forests and Bootstrapping

- A large number of random samples of size n with replacement is drawn.
- For each sample a (deep) tree is built.
- The final forest model counts the number of predictions of each class for an observation.
- The class with the most predictions is assigned.
- This process of aggregating predictions is called „voting“.

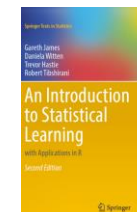
| Stichprobe | 1 | 2 | 3 | 4 | 5 | 6 |
|--------------------|---|---|---|---|---|---|
| Bootstrap-Sample 1 | 6 | 6 | 2 | 5 | 4 | 4 |
| Bootstrap-Sample 2 | 5 | 1 | 4 | 5 | 3 | 5 |
| Bootstrap-Sample 3 | 6 | 2 | 3 | 6 | 6 | 1 |
| Bootstrap-Sample 4 | 3 | 4 | 6 | 1 | 6 | 6 |





...From Bootstrapping to Random Forests¹

- In addition to Bootstrapping random forests make use of a 2nd trick:
- For each tree at each split **only a subset of variables is randomly offered** for split search.
- **Idea:**
 - Variance reduction formula is for independent observations.
 - ...If all variables are available to all splits of all trees the trees will look pretty much the same ('be correlated').



Random Forests: chp. 8.2

¹ Breiman, L. (2001): Random Forests, Machine Learning 45(1), 5 – 32..

Random Forests in Practice

mlr3shiny 1. Data 2. Task 3. Learner 4. Train & Evaluate 5. Benchmarking 6. Predict ?

Select a machine-learning algorithm

Learner 1 decision tree Go

Learner 2 random forest Go

- decision tree
- random forest
- support vector machine
- logistic regression

Learner 1 Learner 2

Learner Information

| | | | |
|--------------------------|----------------|--------------------------|---|
| Algorithm: | Random Forest | Properties: | importance, multiclass, oob_error, twoclass, weights |
| Current Predict Type: | response | Supported Feature Types: | logical, integer, numeric, character, factor, ordered |
| Supported Predict Types: | response, prob | | |

Learner Parameters

| | | | |
|---------------|----------|------------|-----|
| num.trees | Lower: 1 | Upper: Inf | 500 |
| mtry | Lower: 1 | Upper: 8 | |
| min.node.size | Lower: 1 | Upper: Inf | 1 |

Change Parameters

Change Predict Type response Change

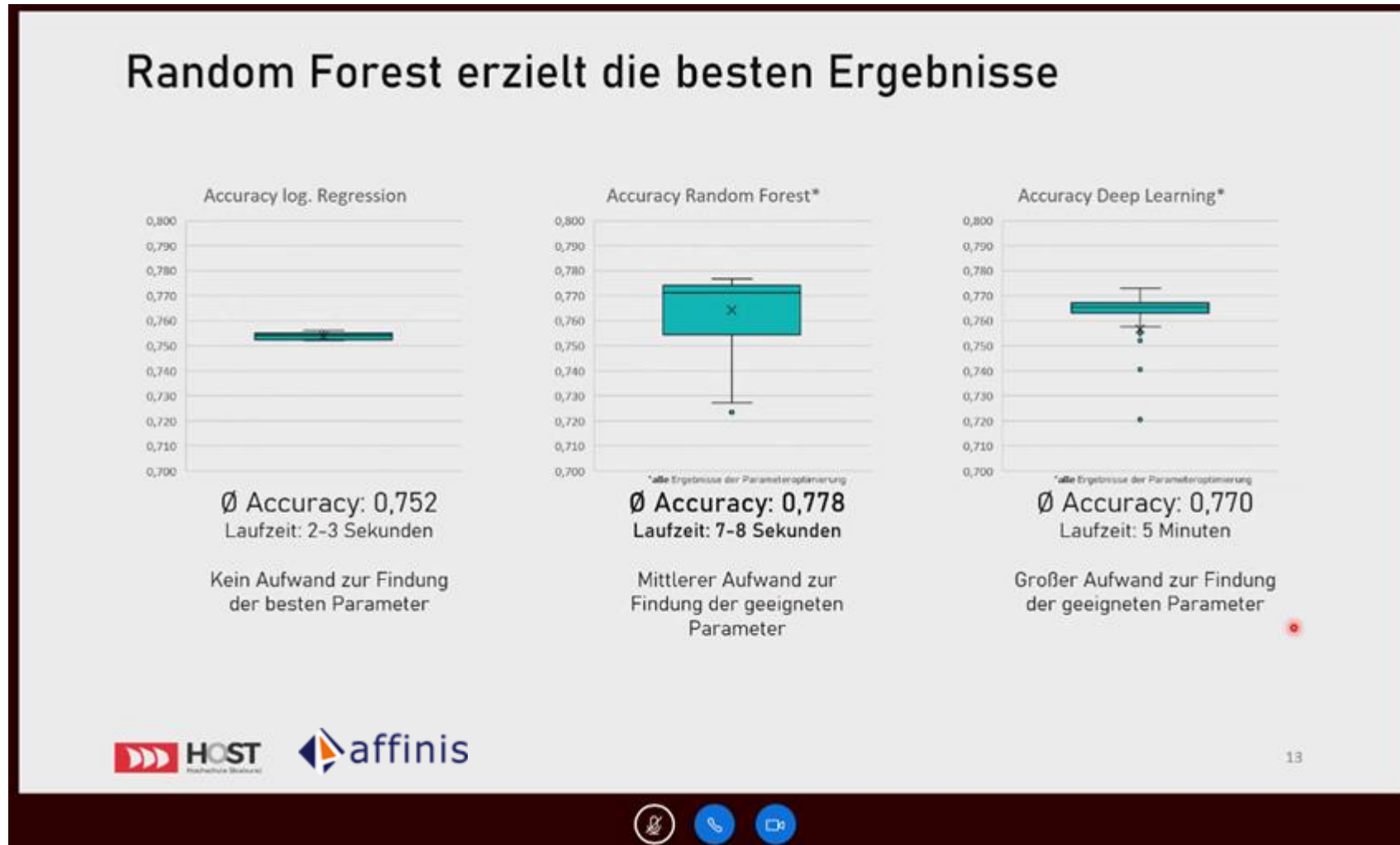
Random Forest Tuning Parameters

| Parameter | Description | Name in | Default |
|---------------------|--|----------|--------------------------------|
| # Trees | Number of trees to grow: <ul style="list-style-type: none">- Generally: the larger the better, but increases computation time- ...thumb rule: not smaller than 100. | ntree | 500 |
| Min. nodesize | Tree parameter: Minimal number of observations in each terminal node: <ul style="list-style-type: none">- Idea of forests: deep trees (overfitting will be averaged out).- Strong impact on computation time => can be increased for large data. | nodesize | 5 |
| # Variables Split | Number of randomly sampled variables @ each split: <ul style="list-style-type: none">- Important parameter, optimal choice depends on data:- ...too large: similar trees- ...too small: trees/splits with only unproductive values may occur. | mtry | $\sqrt{\#Variables\ in\ data}$ |

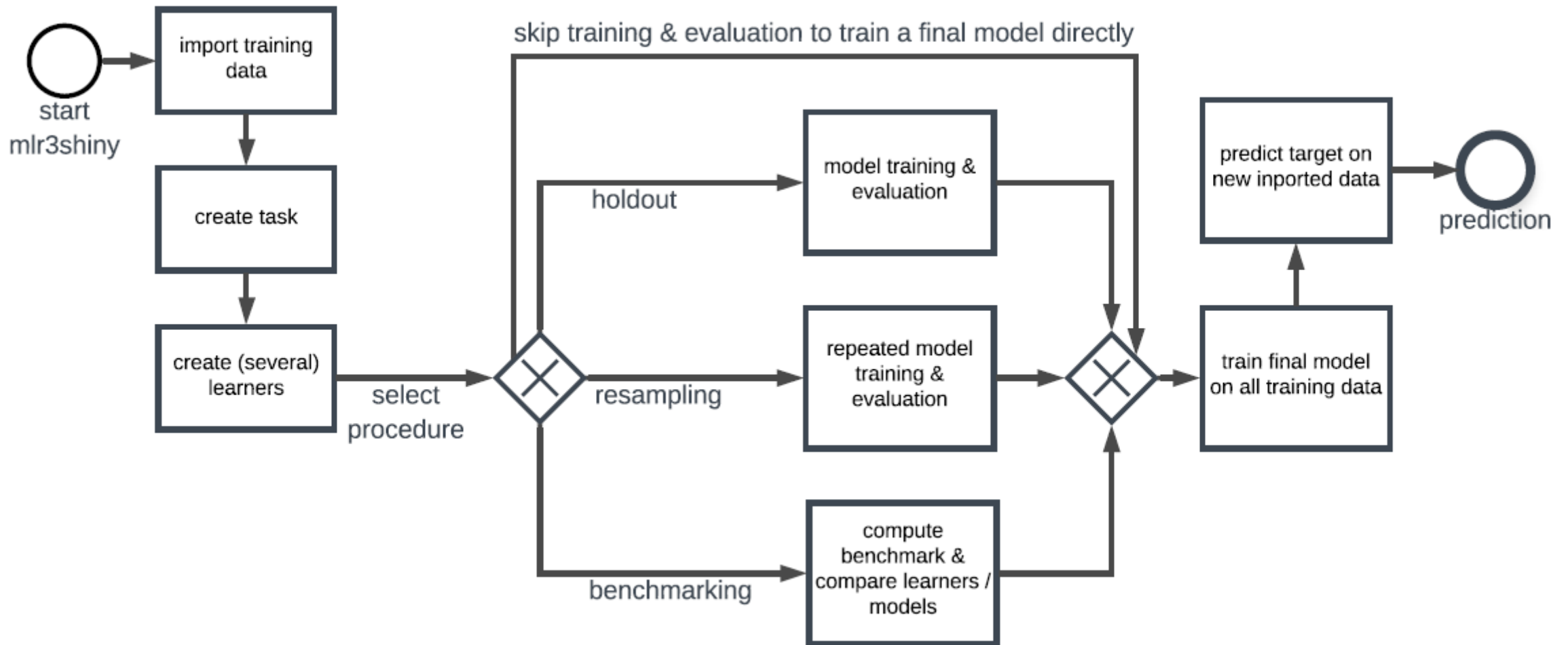
For further information, see:

Szepannek, G. (2017): On the Practical Relevance of Modern Machine Learning Algorithms for Credit Scoring Applications, In: Mucha, H.: *Big Data Clustering: Data Preprocessing, Variable Selection and Dimension Reduction*, WIAS Report 29, S. 88-96, DOI: [10.20347/WIAS.REPORT.29](https://doi.org/10.20347/WIAS.REPORT.29).

Recent Example



The General Machine Learning Workflow



Finding the Best Learner: Benchmarking

mlr3shiny 1. Data 2. Task 3. Learner 4. Train & Evaluate 5. Benchmarking 6. Predict ?

Create a Benchmark Design

Select learners to include in benchmarking:

☒ Learner1
☒ Learner2

Select a resampling strategy:

☒ holdout
☐ cross-validation
☐ bootstrap

Restart benchmarking

Benchmarking Overview

Task: iris

Learners: Learner1: classif.rpart, Learner2: classif.ranger

Resampling Strategy: holdout

Performed Iterations: iterations in 0 resamplings

Best Learner: Learner1 classif.rpart; classif.acc: 0.96

Export Benchmark Result

Resampling Parameter Settings

Fraction in (0, 1) of the data used for training the model

0,666666666666667

Benchmark

Benchmarking large datasets may take a while.

Measure Aggregated Performance

accuracy

Score

Aggregated Benchmark Result

| | learner_id | resampling_id | classif.acc |
|---|----------------|---------------|-------------|
| 1 | classif.rpart | holdout | 0.96 |
| 2 | classif.ranger | holdout | 0.96 |

Final Retraining on All Data

mlr3shiny

1. Data

2. Task

3. Learner

4. Train & Evaluate

5. Benchmarking

6. Predict

Apply best learner on new data

Select the best learner to train on the entire training data:

Learner1

Learner2

Learner: Learner1 classif.rpart

Predict Type: response

Target: survived

Status: trained

Train Learner

Export learner

Import a new dataset

Type: csv

Select a File

Browse... titanic_pred.csv

Upload complete

☒ Header

Separator: Semicolon

Quote: Double Quote

Predict target

New dataset

| | gender | age | class | embarked | country | fare | sibsp | parch |
|---|--------|-----|-------|-------------|---------------|-------|-------|-------|
| 1 | male | 42 | 3rd | Southampton | United States | 7.11 | 0 | 0 |
| 2 | male | 13 | 3rd | Southampton | United States | 20.05 | 0 | 2 |
| 3 | male | 16 | 3rd | Southampton | United States | 20.05 | 1 | 1 |
| 4 | female | 39 | 3rd | Southampton | England | 20.05 | 1 | 1 |

Previous12345...221Next

Prediction

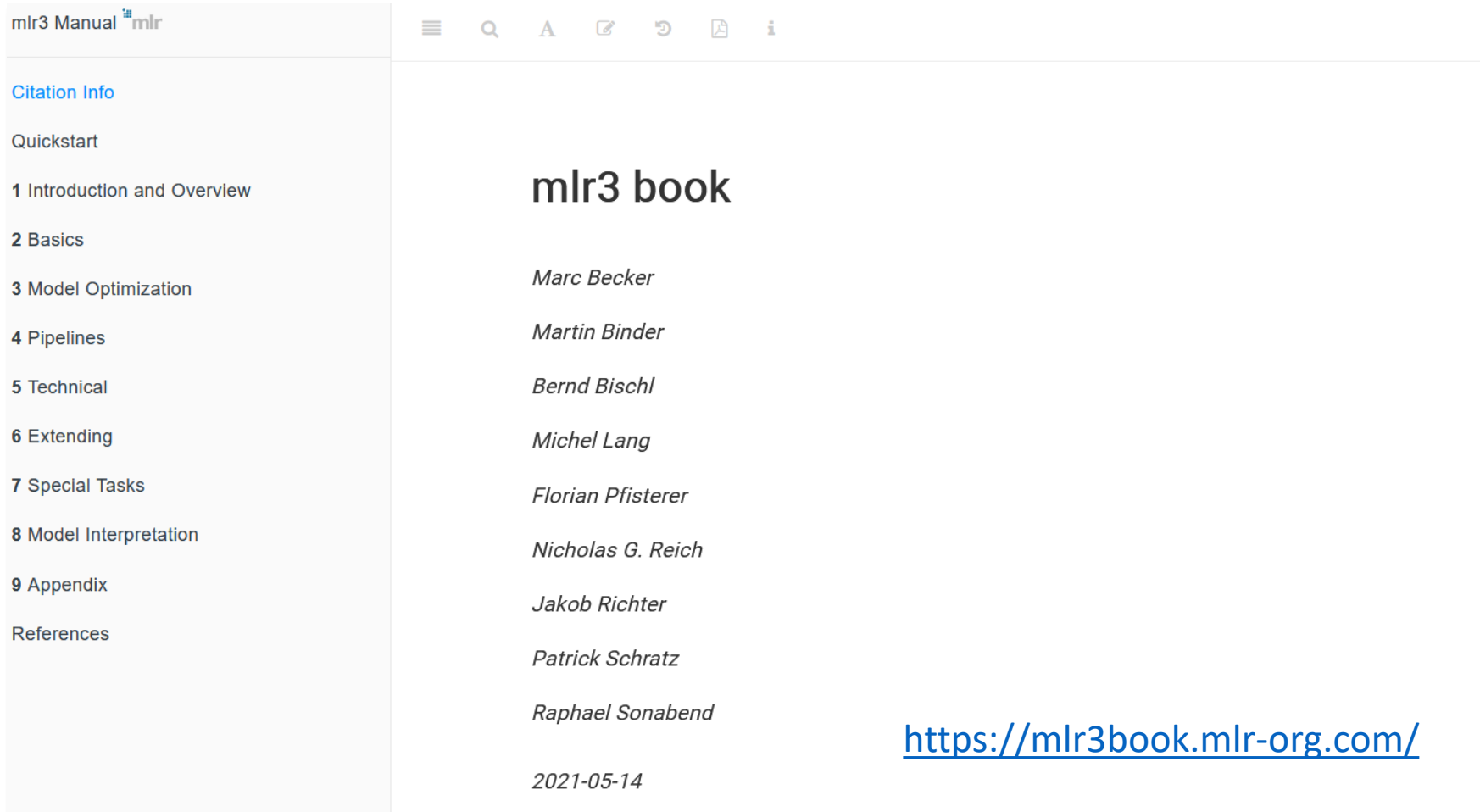
| | row_ids | truth | response |
|---|---------|-------|----------|
| 1 | 1 | | no |
| 2 | 2 | | no |
| 3 | 3 | | no |
| 4 | 4 | | yes |

Previous12345...221Next


Export as .csv

Export as .rds

Further Reading



The screenshot shows the 'mlr3 Manual' website. On the left is a sidebar with a table of contents. The main content area displays the title 'mlr3 book' and a list of authors. At the bottom right, there is a URL and a date.

mlr3 Manual 

[Citation Info](#)

Quickstart

1 Introduction and Overview

2 Basics

3 Model Optimization

4 Pipelines

5 Technical

6 Extending

7 Special Tasks

8 Model Interpretation

9 Appendix

References

mlr3 book

Marc Becker

Martin Binder

Bernd Bischl

Michel Lang

Florian Pfisterer

Nicholas G. Reich

Jakob Richter

Patrick Schratz

Raphael Sonabend

<https://mlr3book.mlr-org.com/>

2021-05-14