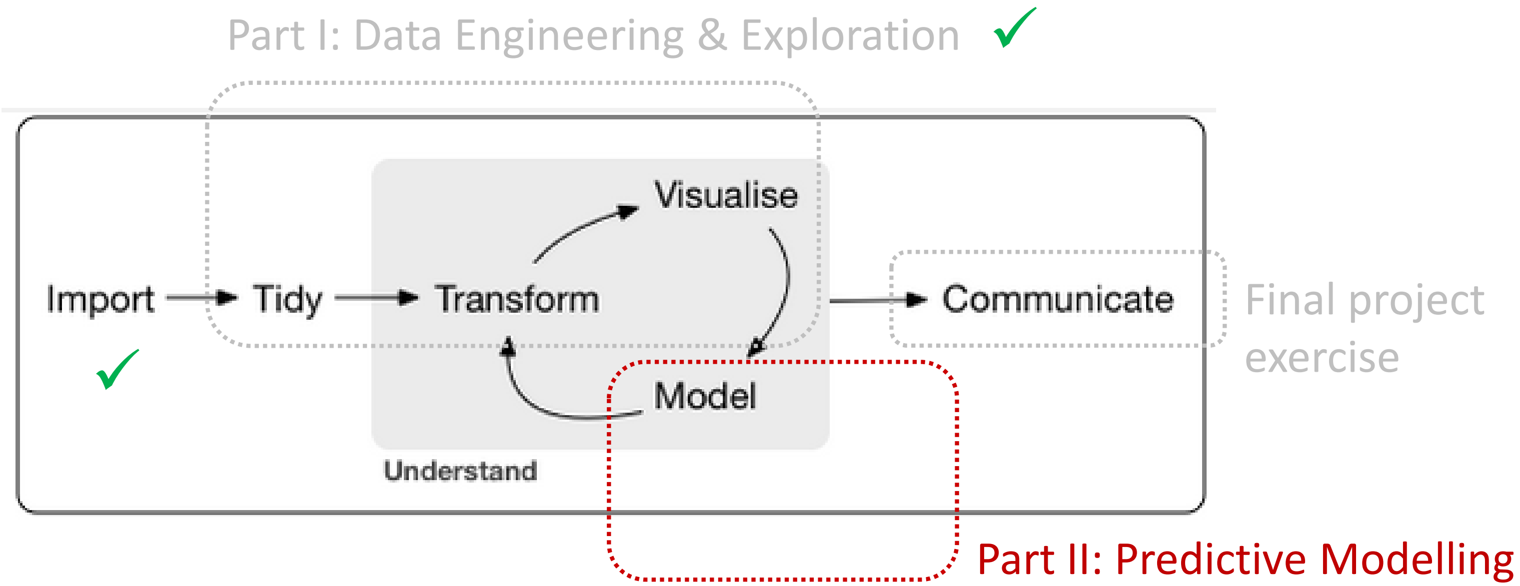
Data Science

Predictive Modelling

Gero Szepannek

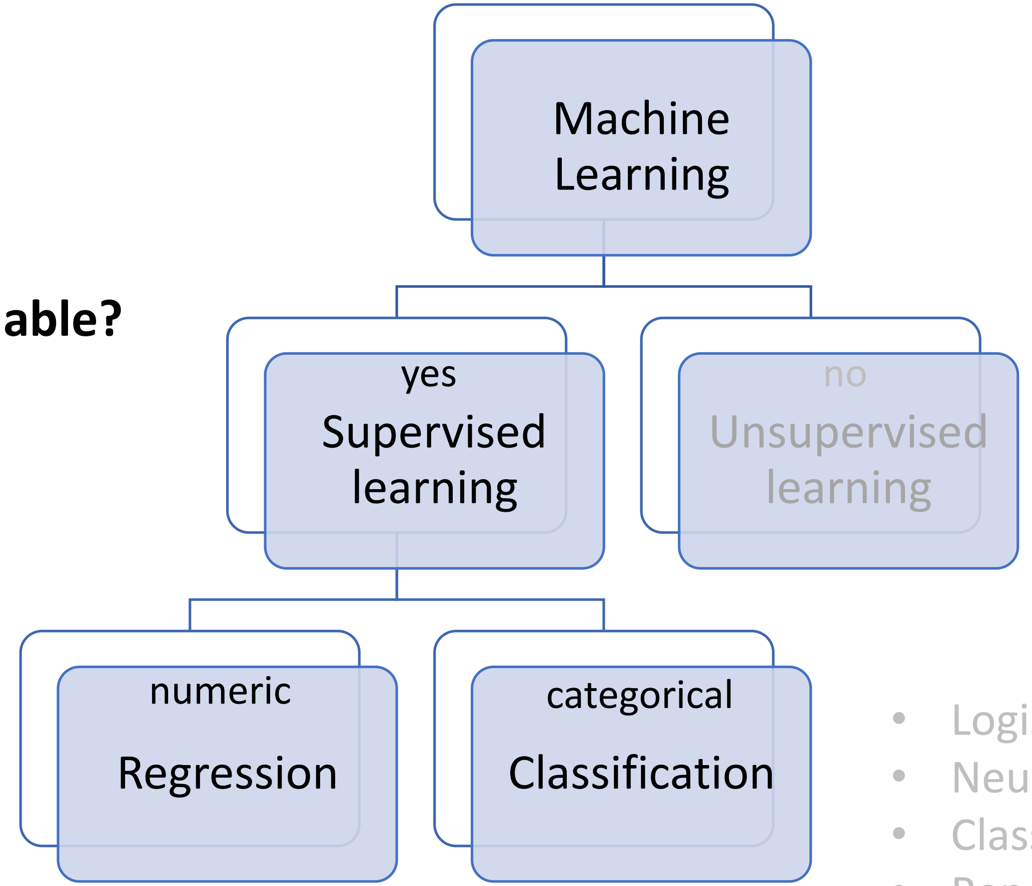
Data Science Process



Prediction Model

$$y = f(x) + \varepsilon \qquad \longrightarrow \qquad \hat{y}_i = \hat{f}(x_i)$$

Regression vs. Classification



Target variable available?

- Cluster Analysis, e.g. k means

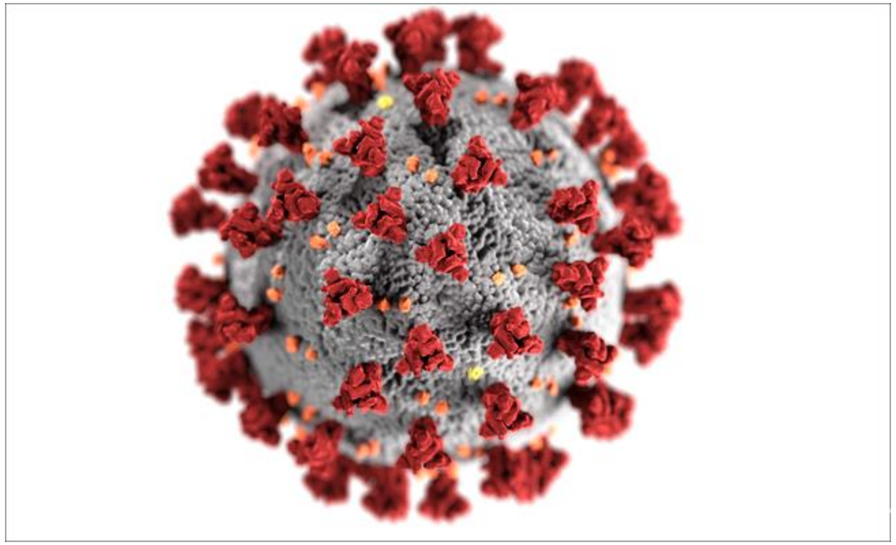
Type of target variable?

- Linear Regression ✓
- Neural Networks
- Regression Trees ✓
- Random Forests
- …

- Logistic Regression
- Neural Networks
- Classification Trees
- Random Forests
- …

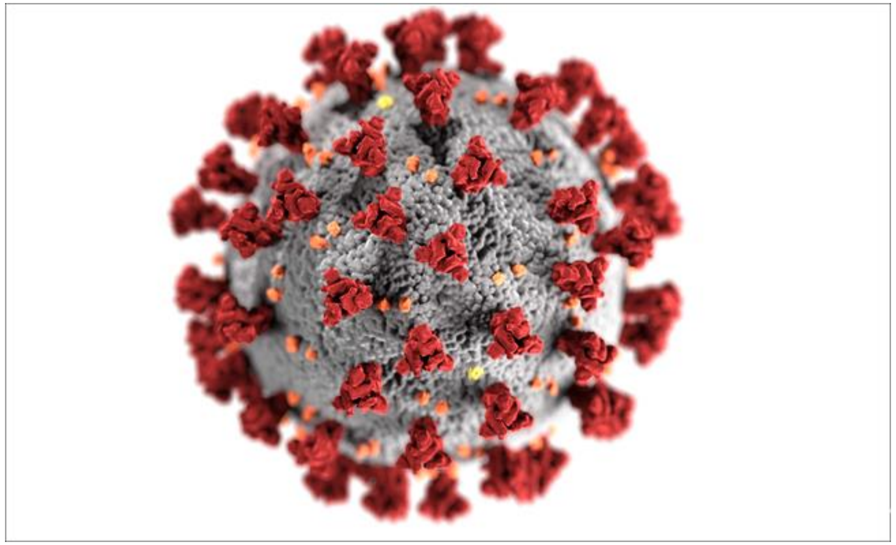

POOR
GOOD
CREDIT SCORE


PAYBACK
PAYBACK.de
dm
GALERIA KAUFHOF
real
REWE
und viele weitere Partner



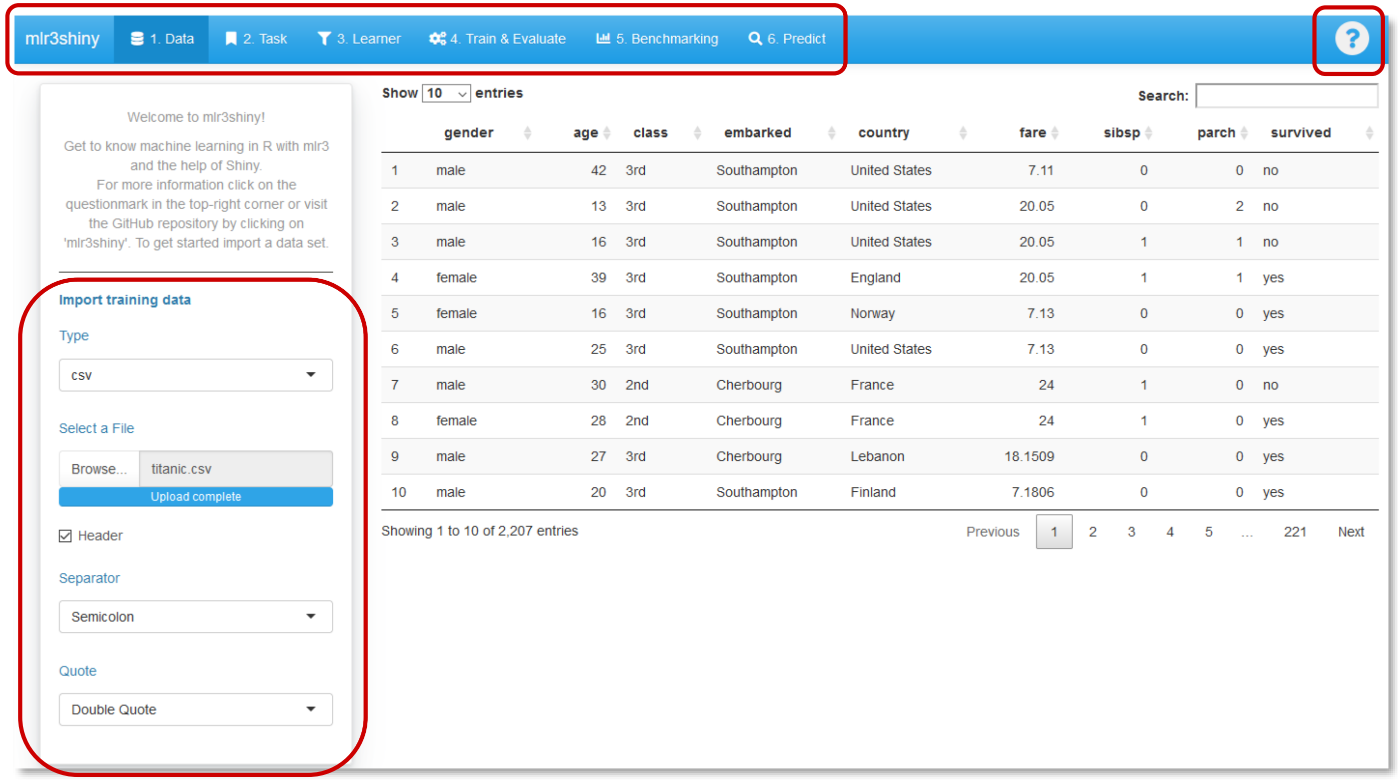

POOR
GOOD
CREDIT
SCORE


PAYBACK
PAYBACK
PAYBACK.de
dm
GALERIA KAUFHOF
real
REWE
und viele weitere Partner



A Point-and-click Framework developed @

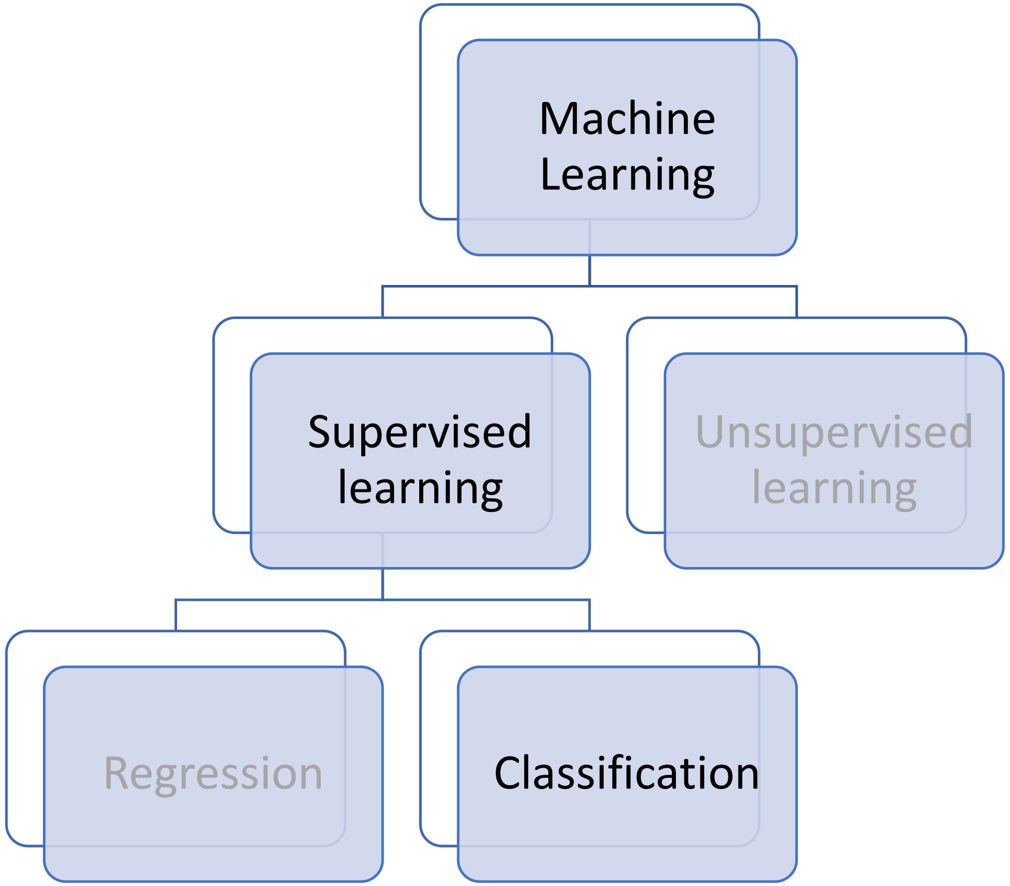
```r
# install
remotes::install_github("https://github.com/LamaTe/mlr3shiny.git")

# start
library(mlr3shiny)
launchMlr3Shiny()
```



	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

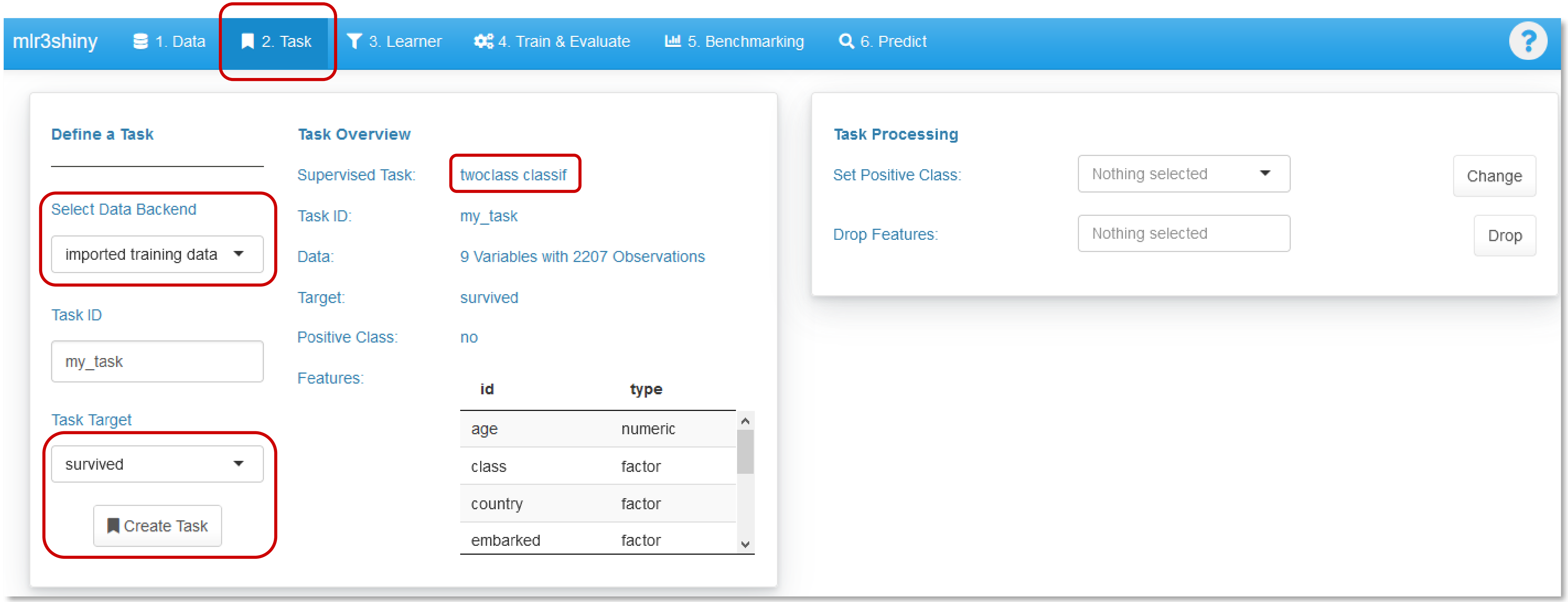
$\hat{y}_i$



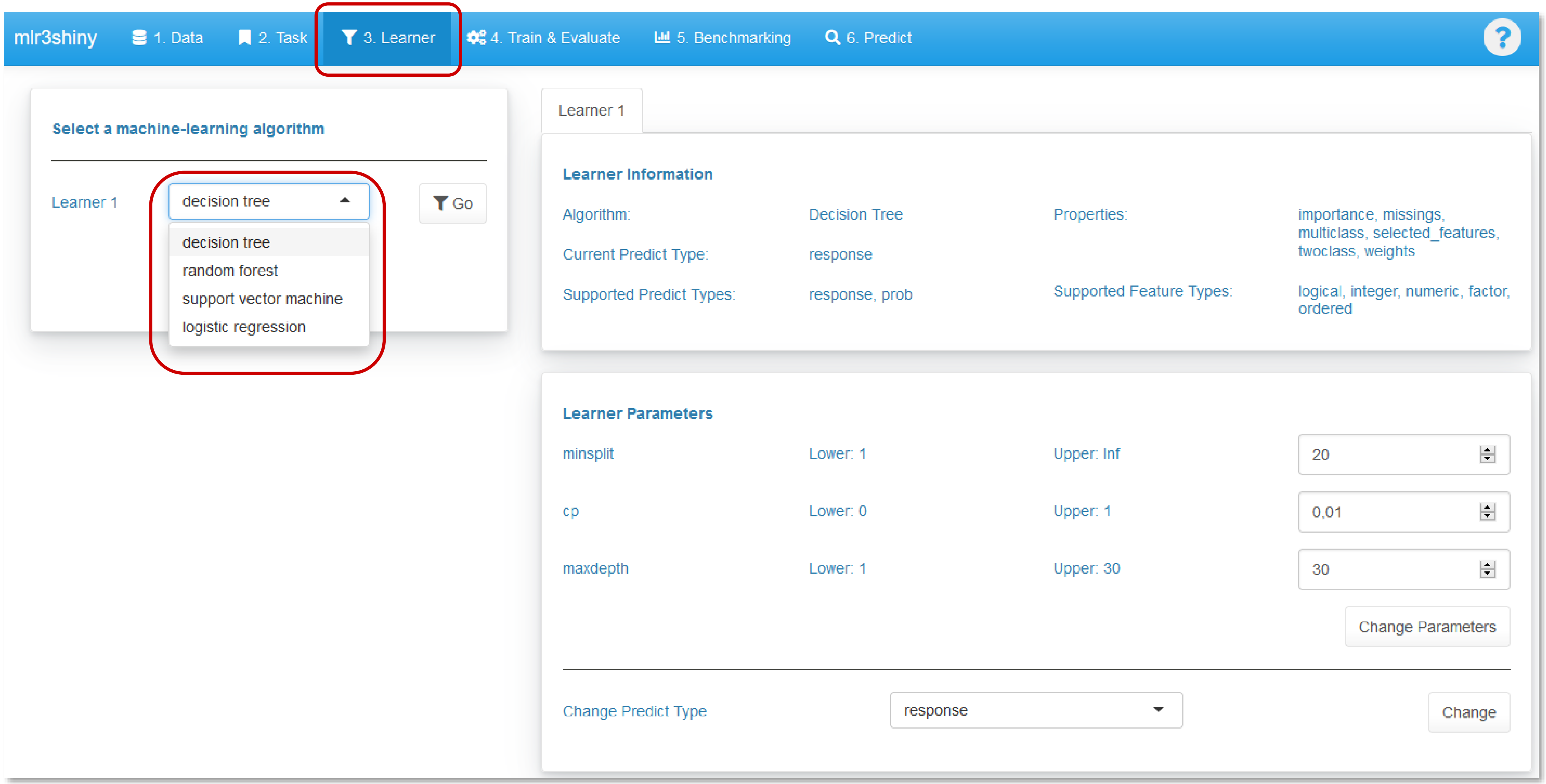
$$\hat{y}_i = \hat{f}(x_i)$$

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

Classification



Classification Algorithms



Flavours of Classification Models

$\hat{y}_i$

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

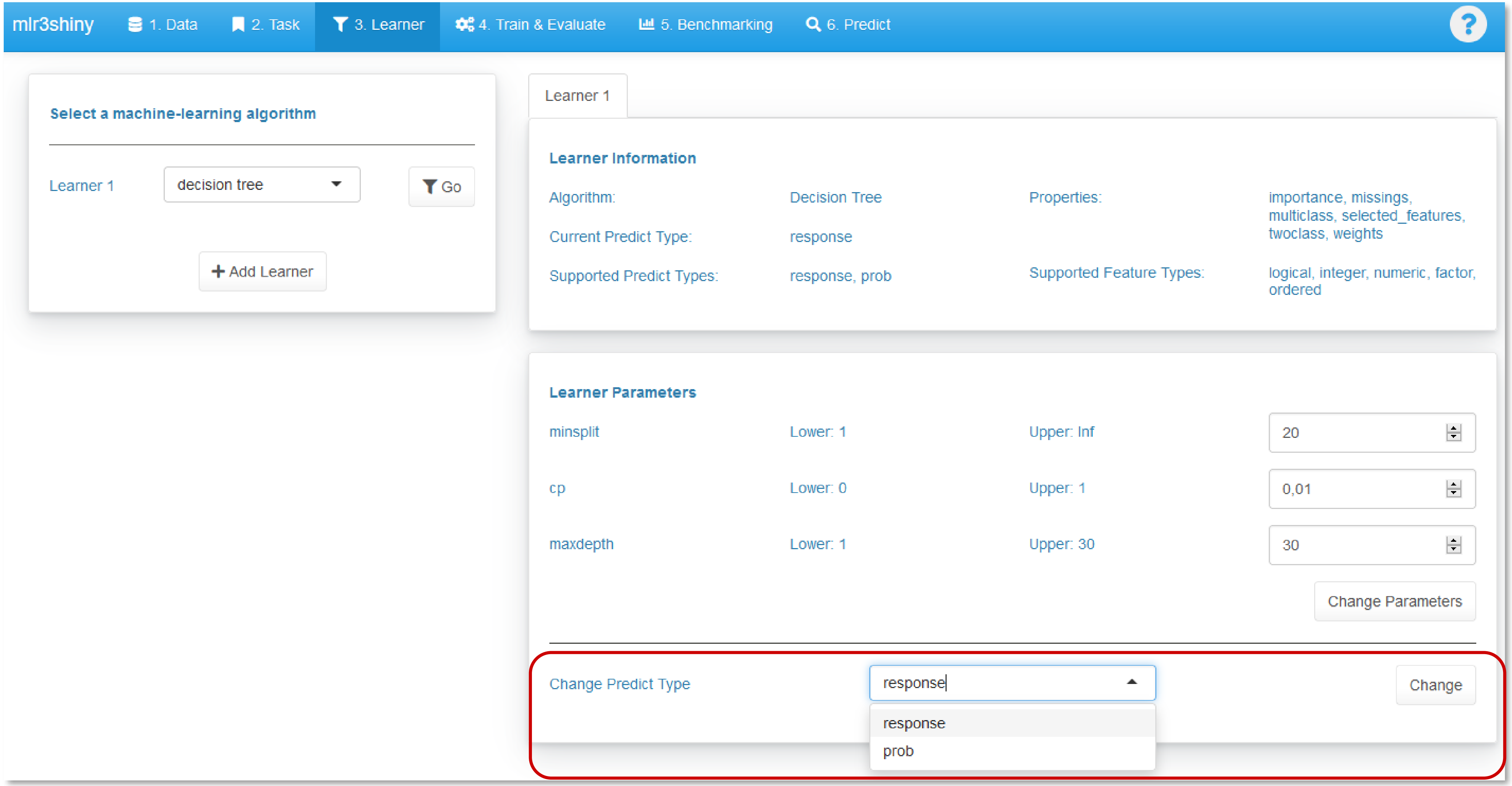
1. In classification y_i is categorical.
2. …most relevant practical case: binary classification → $y_i \in \{0,1\}$
3. Predicting class labels: $\hat{y}_i \in \{0,1\}$
4. Predicting posterior probabilities: $\hat{P}(Y_i = 1) \in [0,1]$

?

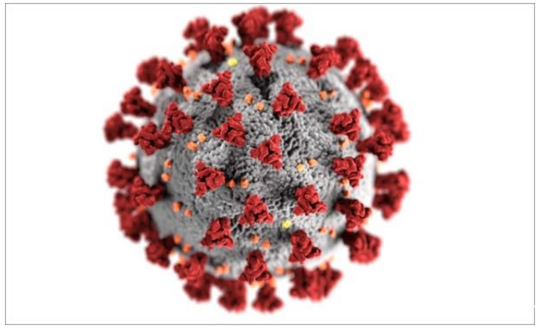
What is a natural rule for prediction of class labels in order to minimize the probability of making an error?

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

Classification Algorithms



Performance Measures for Classification



Epidemiologisches Bulletin | 8|2021 | 25. Februar 2021 (online vorab)

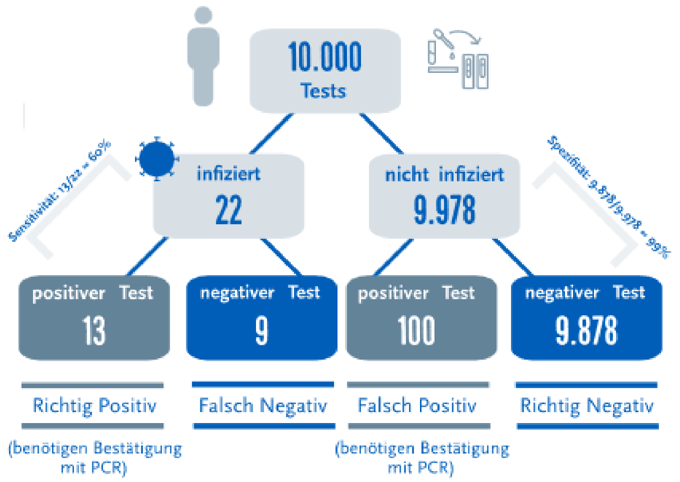
Was ist bei Antigentests zur Eigenanwendung (Selbsttests) zum Nachweis von SARS-CoV-2 zu beachten?

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

Test detects infection.

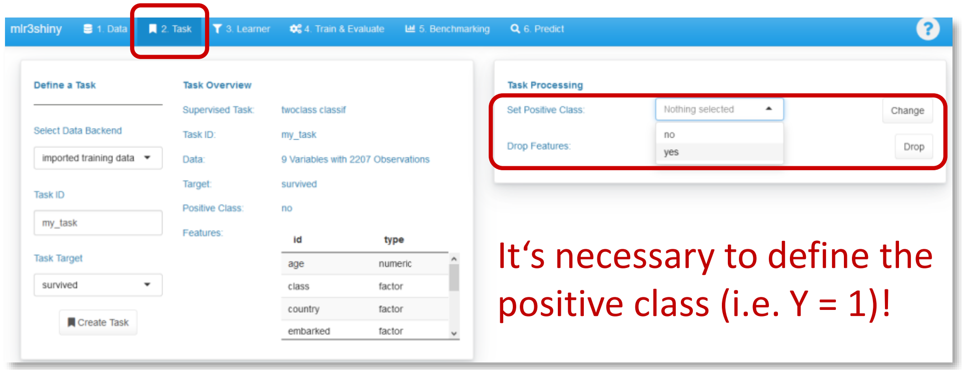
$$P(\hat{Y} = 1 | Y = 1)$$

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



https://edoc.rki.de/bitstream/handle/176904/7826/EB-8-21-Heimtest-Beitrag_22-02-21-zur%20Ver%C3%B6ffentlichung.pdf?sequence=1&isAllowed=y

Confusion Matrix



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

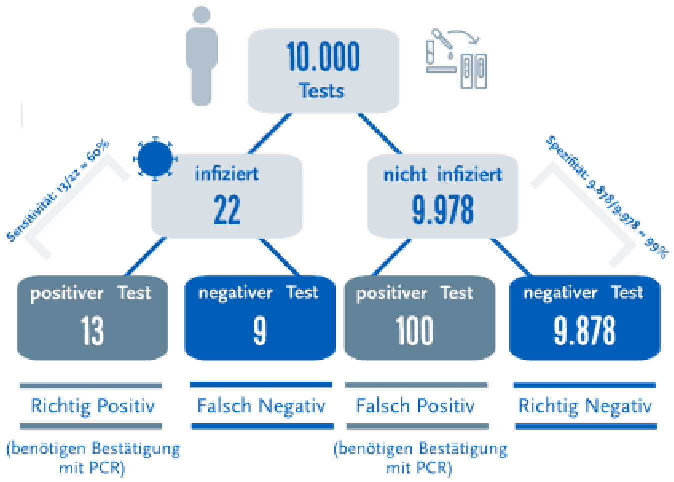
	$\hat{Y} = 1$	$\hat{Y} = 0$	
$Y = 1$			
$Y = 0$			

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

Test detects infection.

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

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



Confusion Matrix

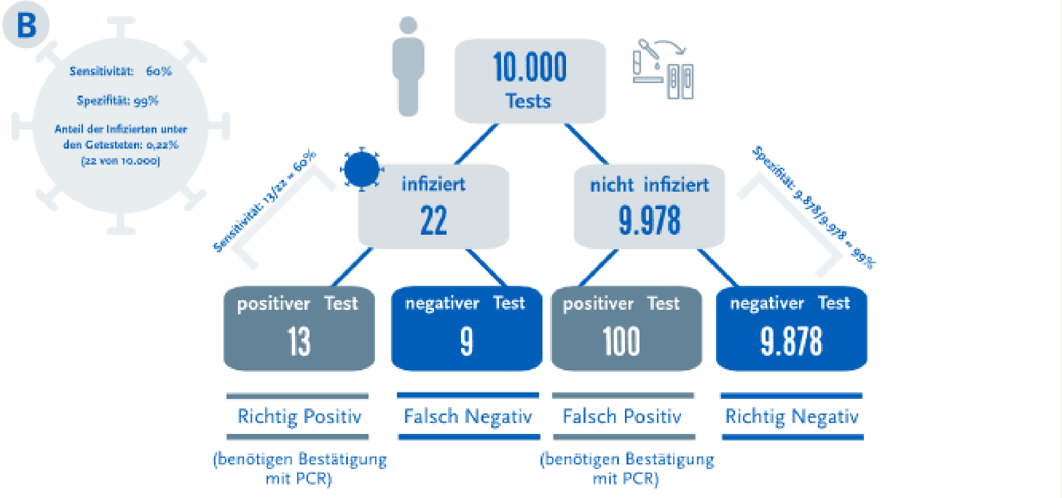
	$\hat{Y}=1$	$\hat{Y}=0$	
$Y=1$			
$Y=0$			

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

Test detects infection.

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

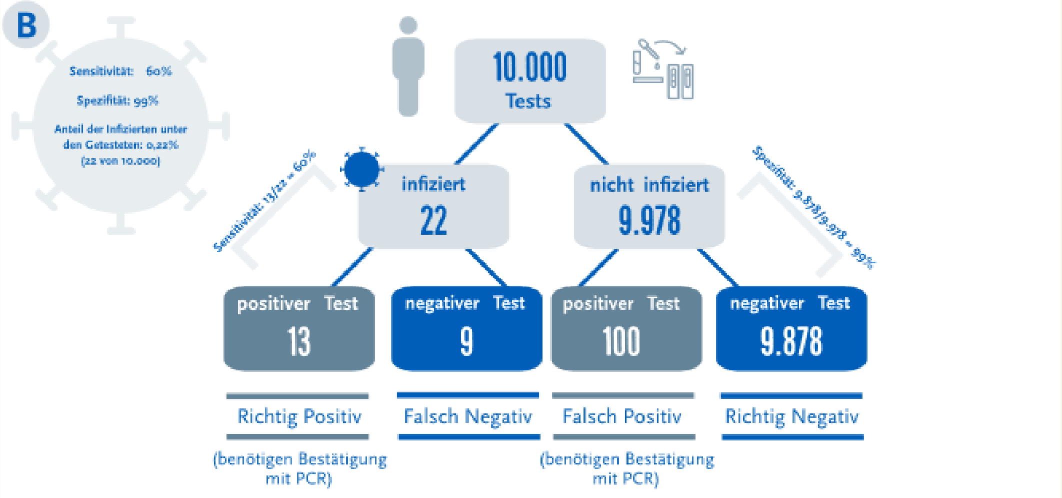
$$P(\hat{Y}=0|Y=0)$$



Confusion Matrix

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

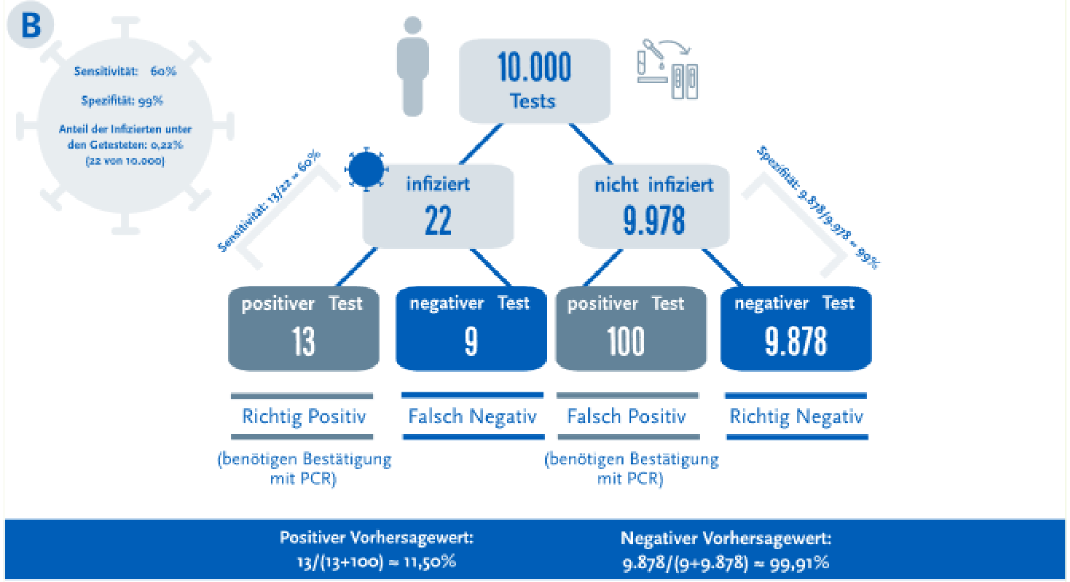
1. Accuracy
2. Sensitivity
3. Specifity
4. Precision



Confusion Matrix

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

1. Accuracy
2. Sensitivity
3. Specifity
4. **Precision / PPV**

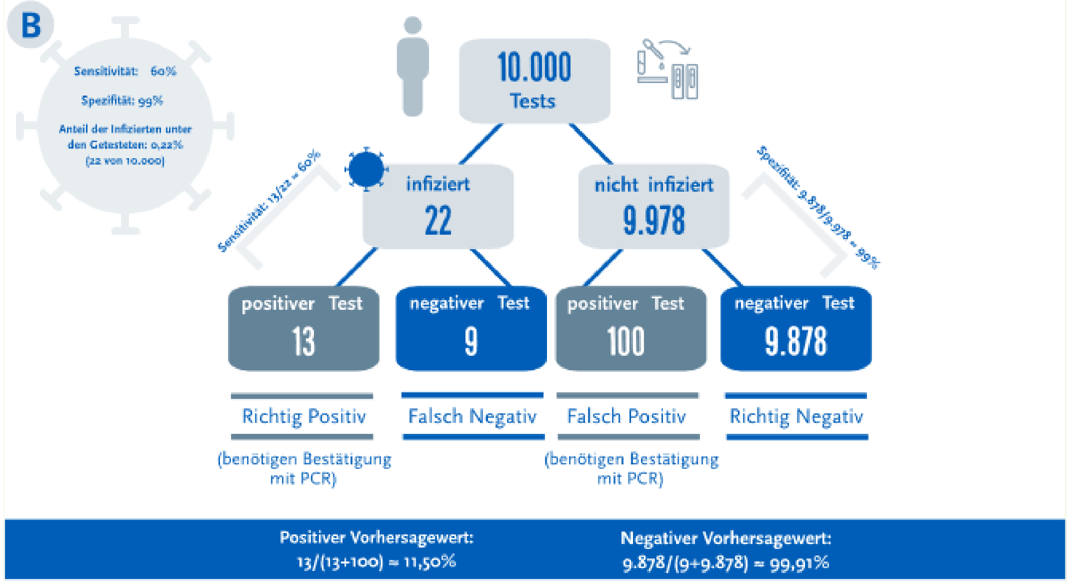


Confusion Matrix

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

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?**

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



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.

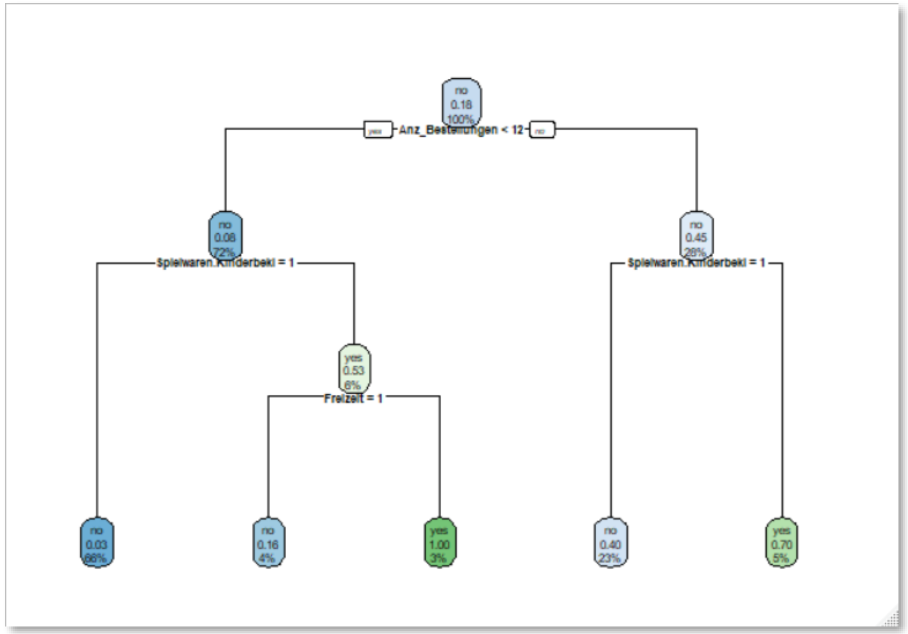
An Schulen machen Antigen-Schnelltests Sinn

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.

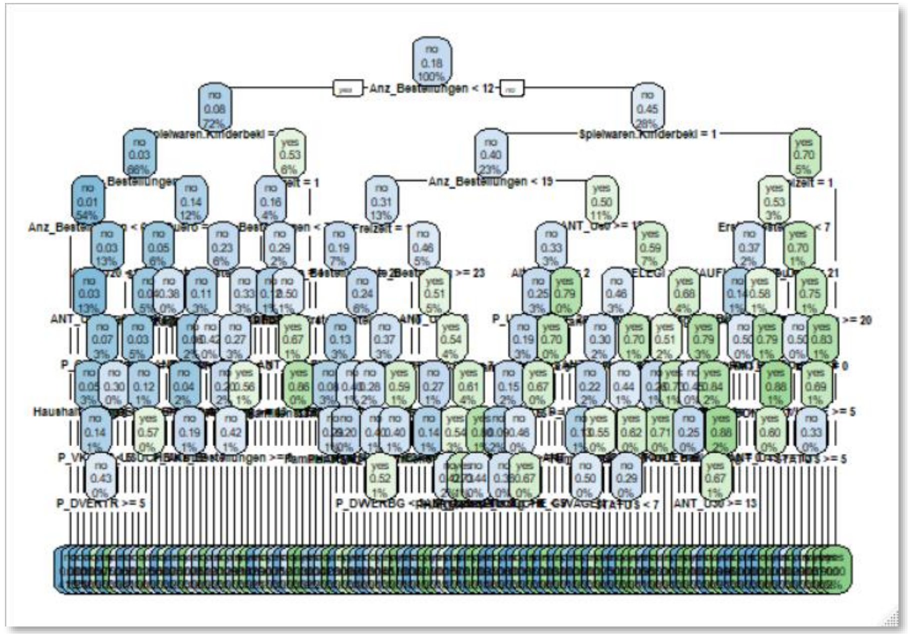
https://www.ndr.de/nachrichten/info/Drosten-Schnelltests-sind-wohl-weniger-zuverlaessig-als-gedacht,coronavirusupdate178.html

Optimizing a Decision Tree

"Flat tree"



"Deep tree"



1. What might be the problem with a flat tree (w.r.t. accuracy)?
2. …What might be the problem with a deep tree?

Test Data

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

Sample

Training (typically: 60/70%)

Test

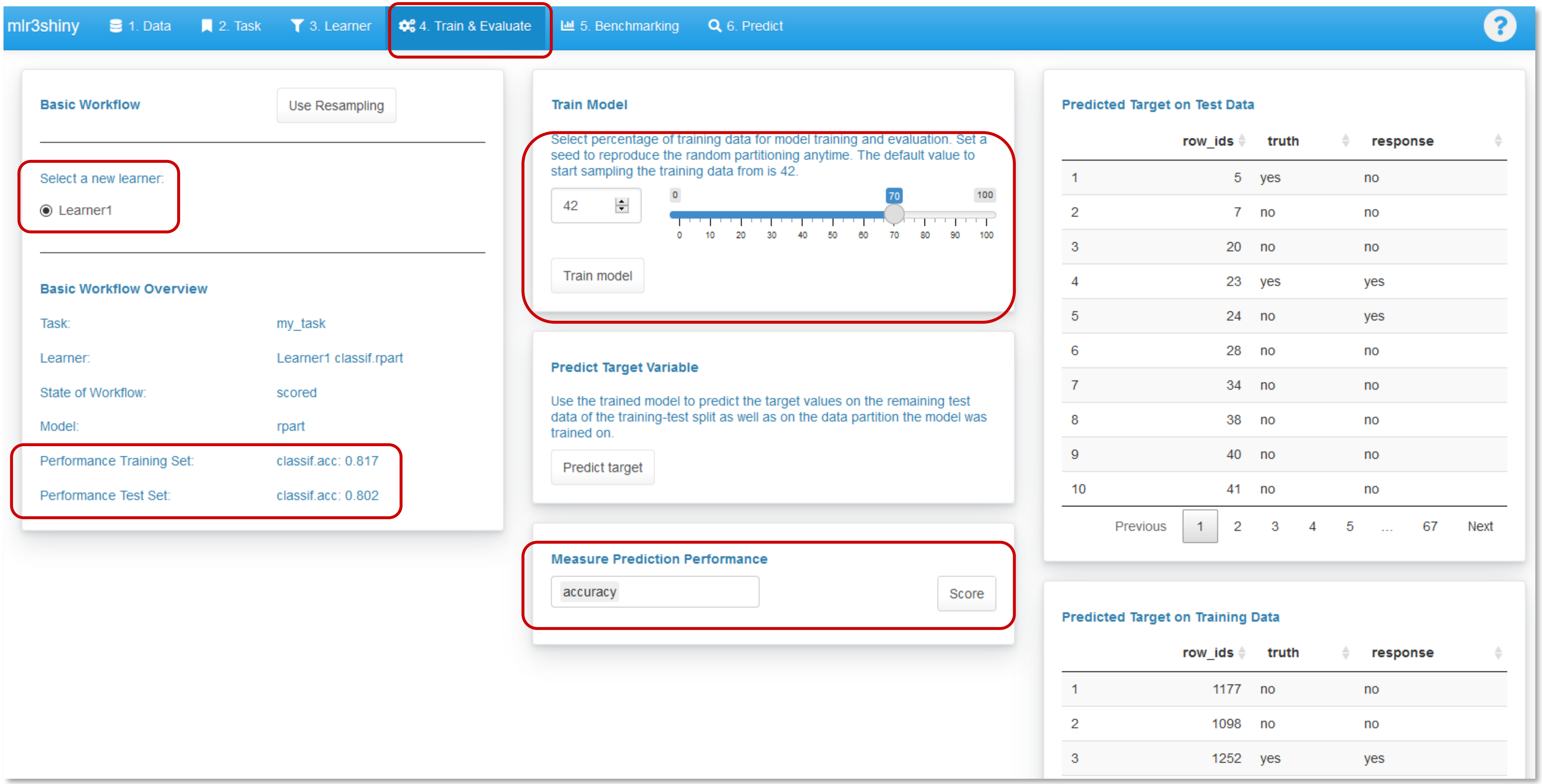
Used to build (aka ‘train‘) the model.

- Simulates new data.
- **Not used for training but only for model evaluation.**

Training performance e.g. MSE_{train}

Test performance / MSE_{test}

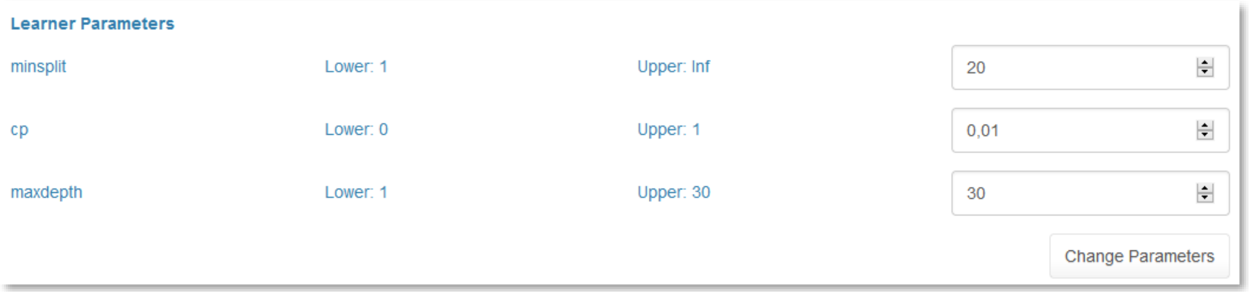
Performance Evaluation



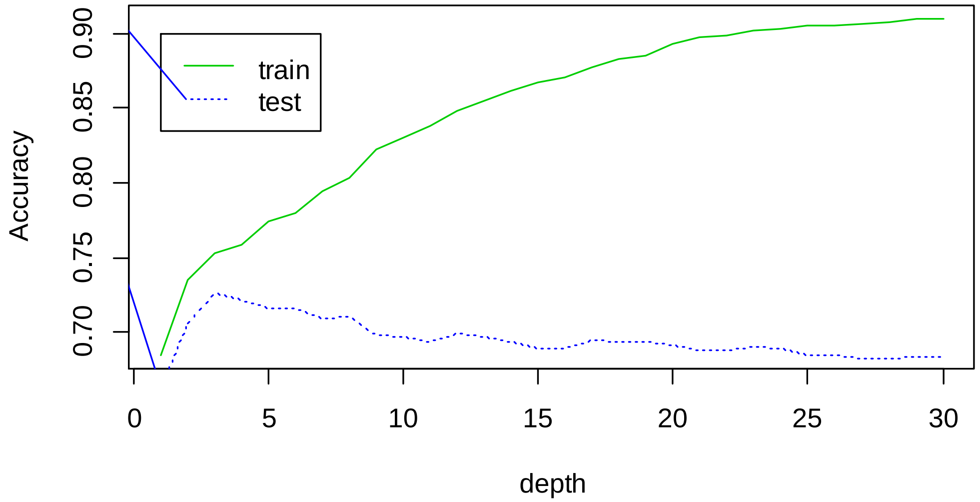
...Lab

On your own:

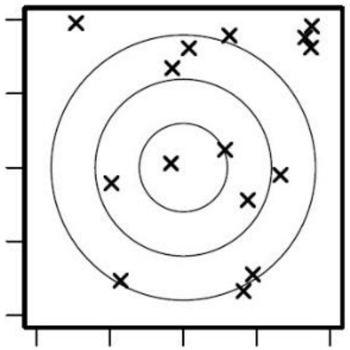
1. Modify the hyperparameters of the tree...
2. ...Try to find the best model, i.e. the model that maximizes accuracy!
3. What do you observe w.r.t. the accuracy on both:
    1. Training data...
    2. ...and Test data?



Overfitting and Underfitting



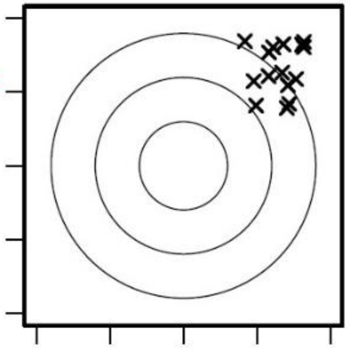
Large Variance
No Bias



Too deep:
overfitting the data
not valid on new data

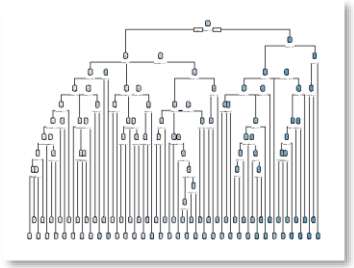
Parameter Tuning
...find the optimal depth

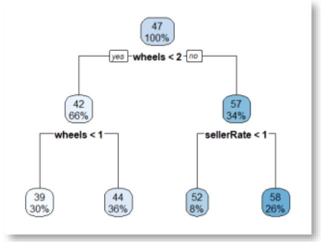
Small Variance
but Bias



Too flat:
not flexible enough to cover all
the structure in the data
...systematic error

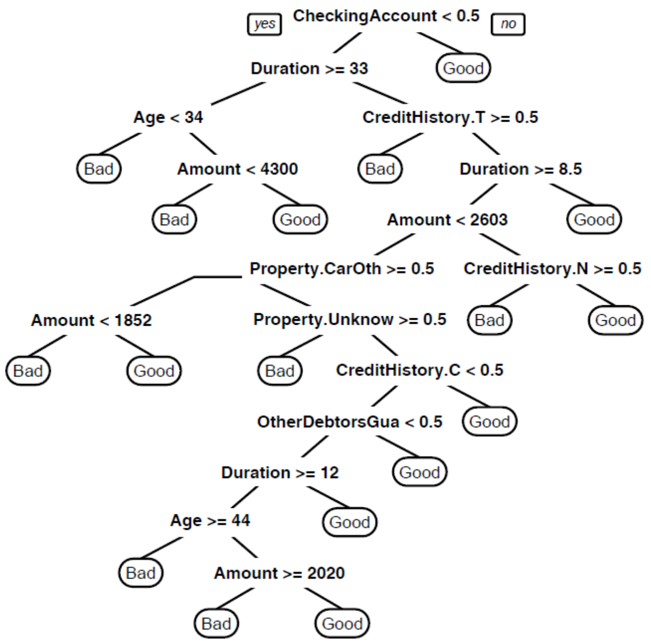
…e.g. set the parameters minsplit = minbucket = 1 and cp = 0 such that and then increase maxdepth.

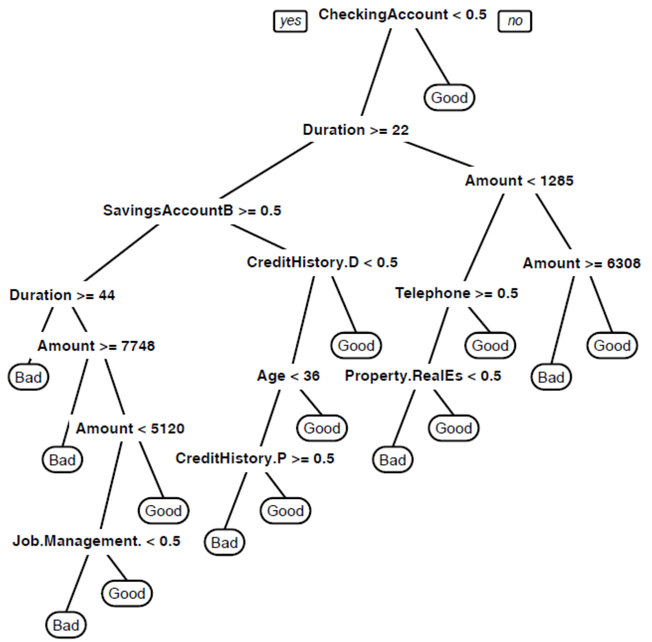




Variance of Models

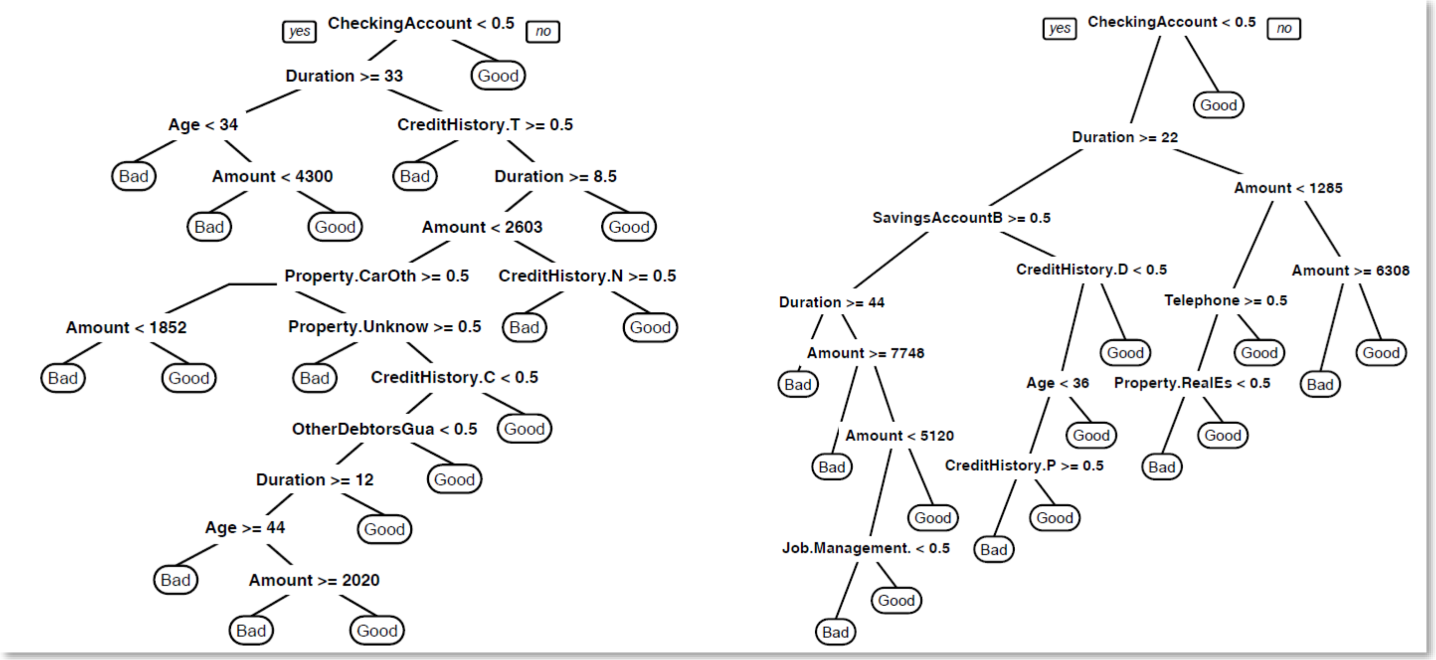
- Deep trees will look pretty different on different samples…
- …except for the first splits.
- **Try to explain, why?**





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:

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

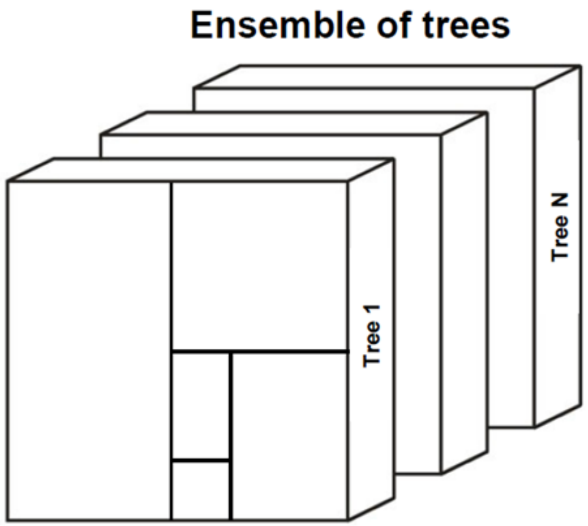
1. **What denotes $\bar{X}$ in Statistics?**
2. …How can we translate this to the variance of a single model?

Random Forests and Bootstrapping

- A large number of random samples of size n with replacement is drawn.

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

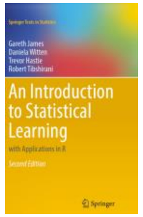
- For each sample a (deep) tree is built.
- The final forest model counts the number of predictions of each class for an observation.
- The class with the most predcitions is assigned.
- This process of aggregating predictions is called „voting“.



…From Bootstrapping to Random Forests[1]

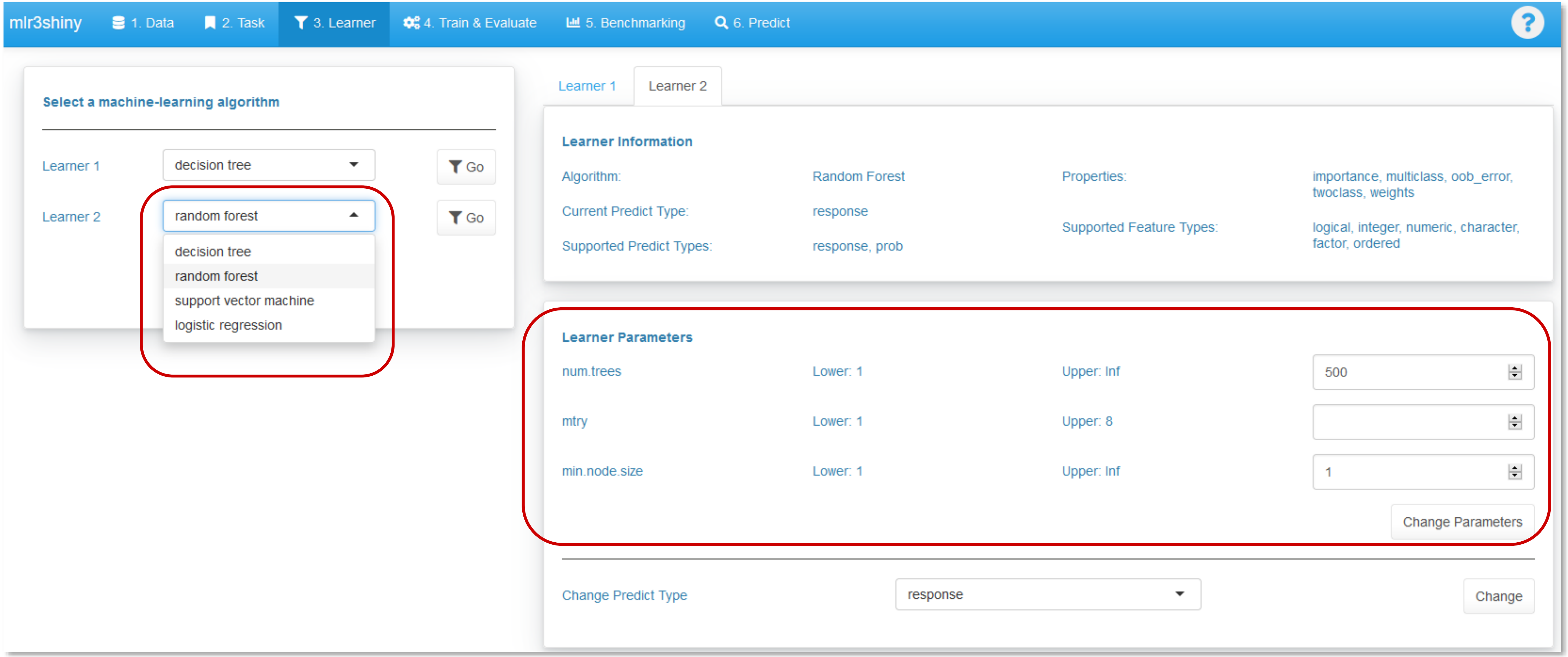
- In addition to Bootstrapping random forests make use of a 2nd trick:
- For each tree at each split **only a subset of variables is randomly offered** for split search**.**
- **Idea**:
  - Variance reduction formula is for independent observations.
  - …If all variables are available to all splits of 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



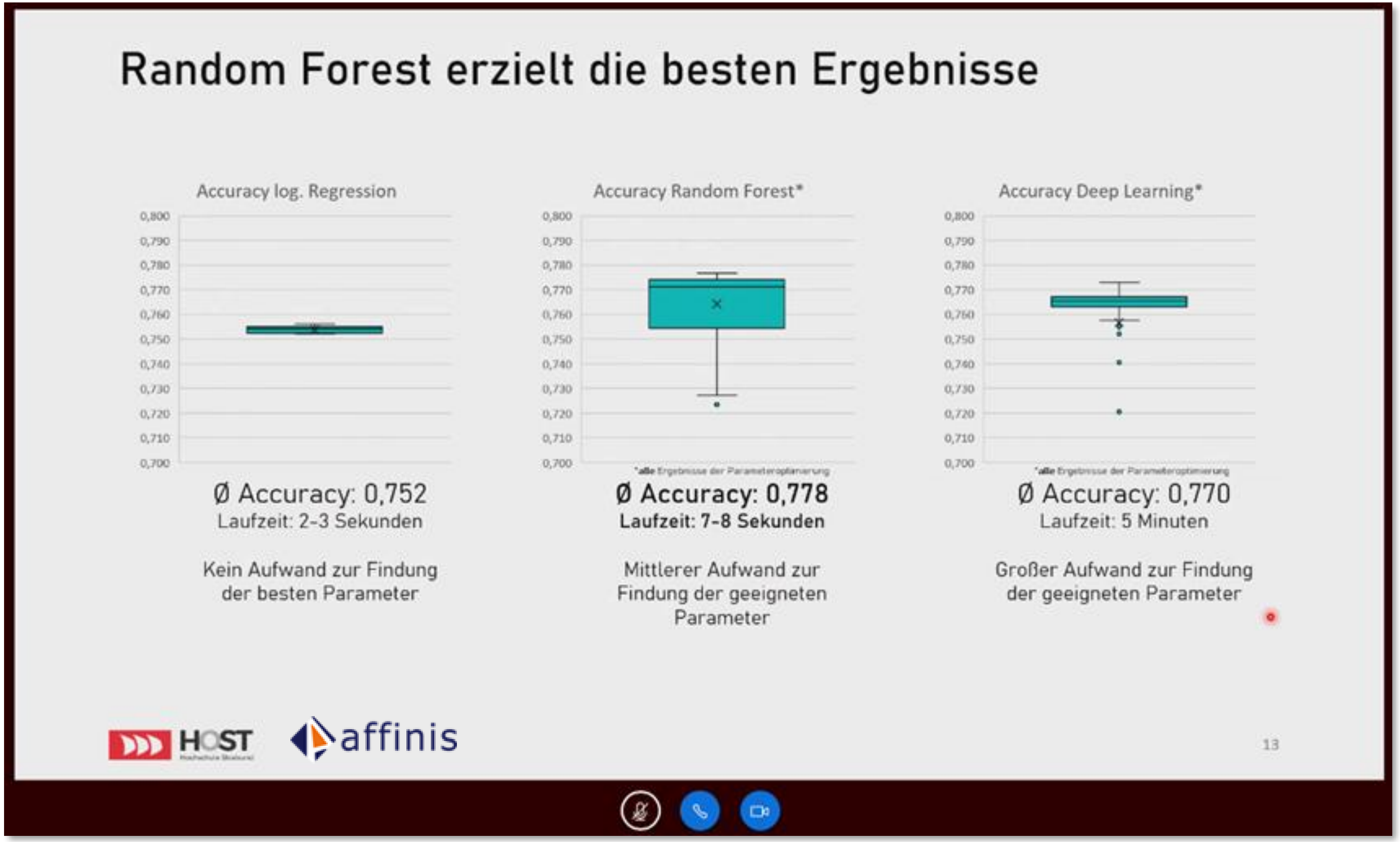
Random Forest Tuning Parameters

Parameter	Description	Name in	Default
# Trees	**Number of trees to grow:** - 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: - 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:** - Important parameter, optimal choice depends on data: - …too large: similar trees - …too small: trees/splits with only unpredictive values may occur.	`mtry`	$\sqrt{\#Variables\ in\ data}$

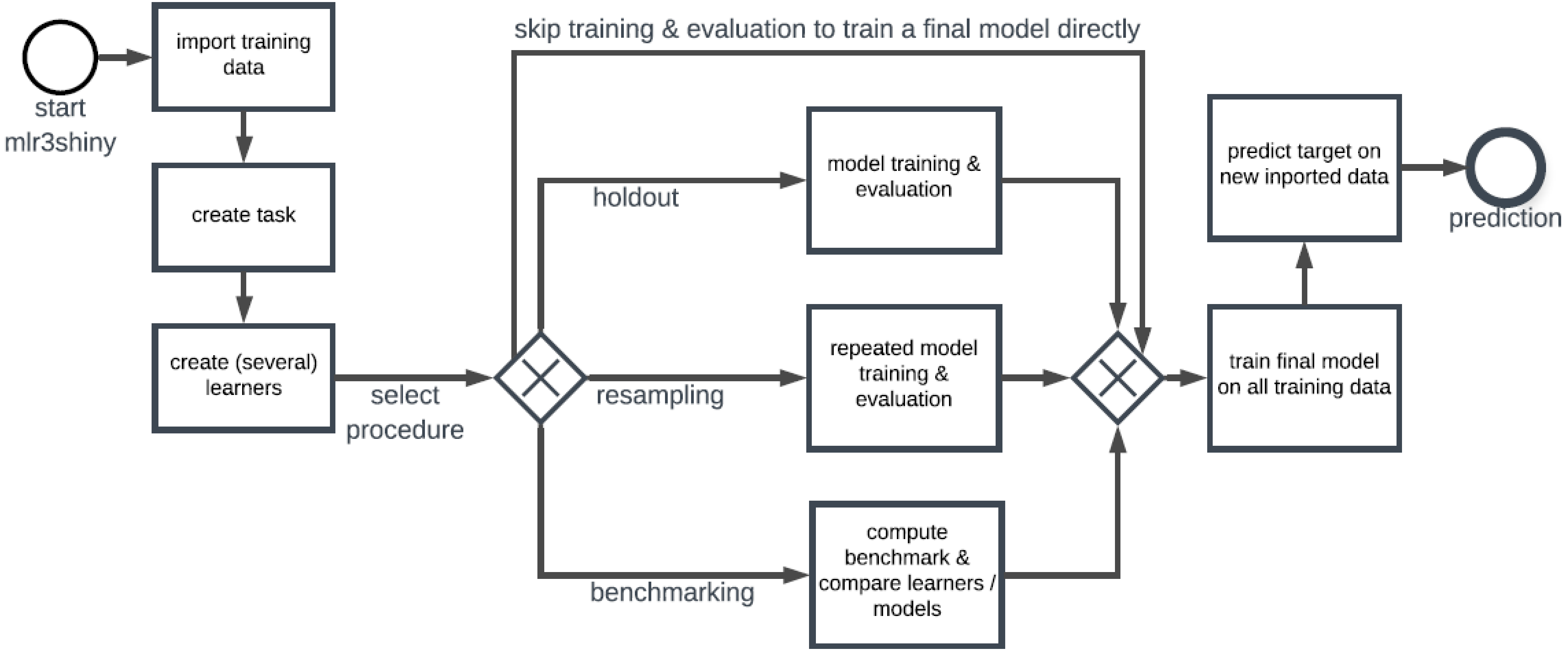
For further information, see:

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.

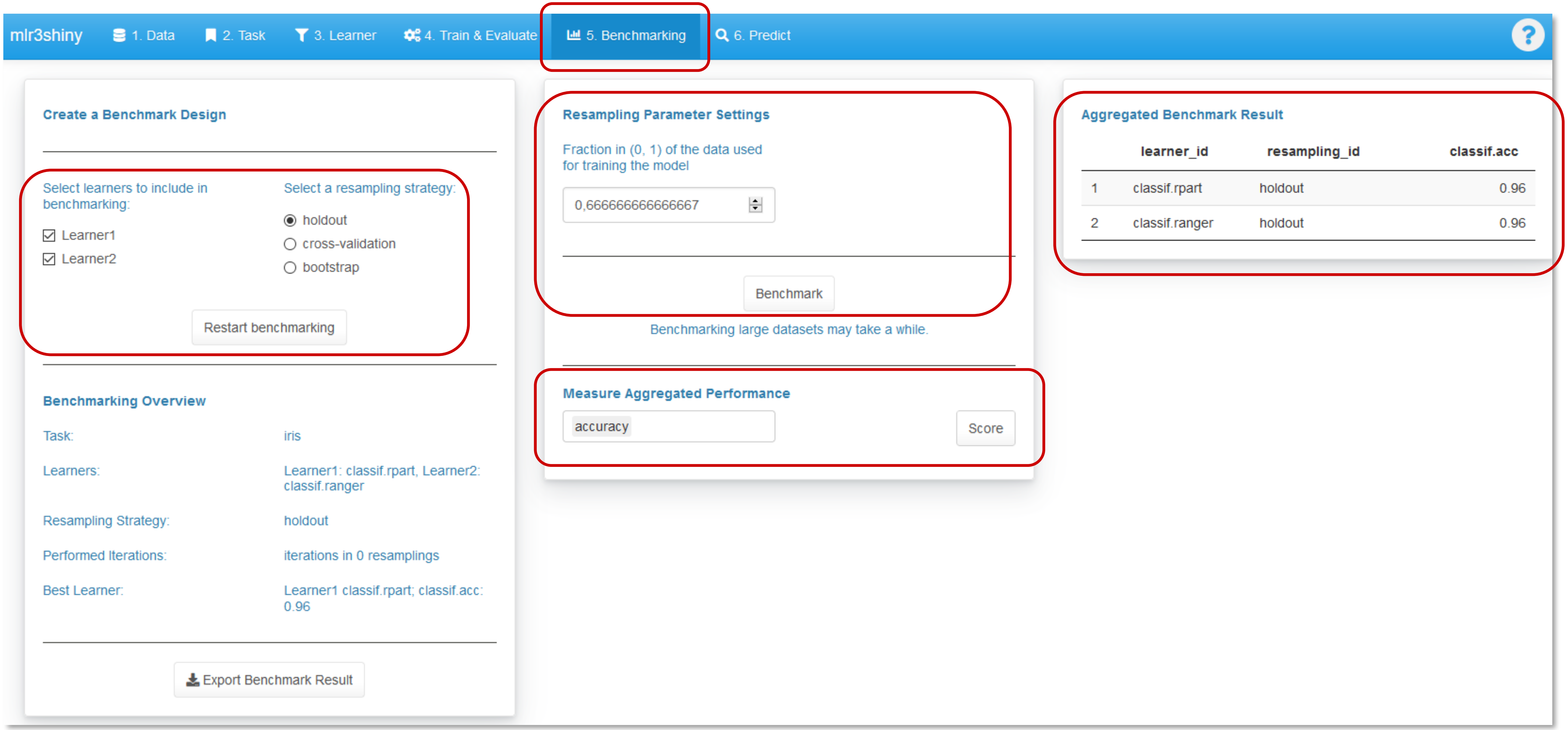
Recent Example



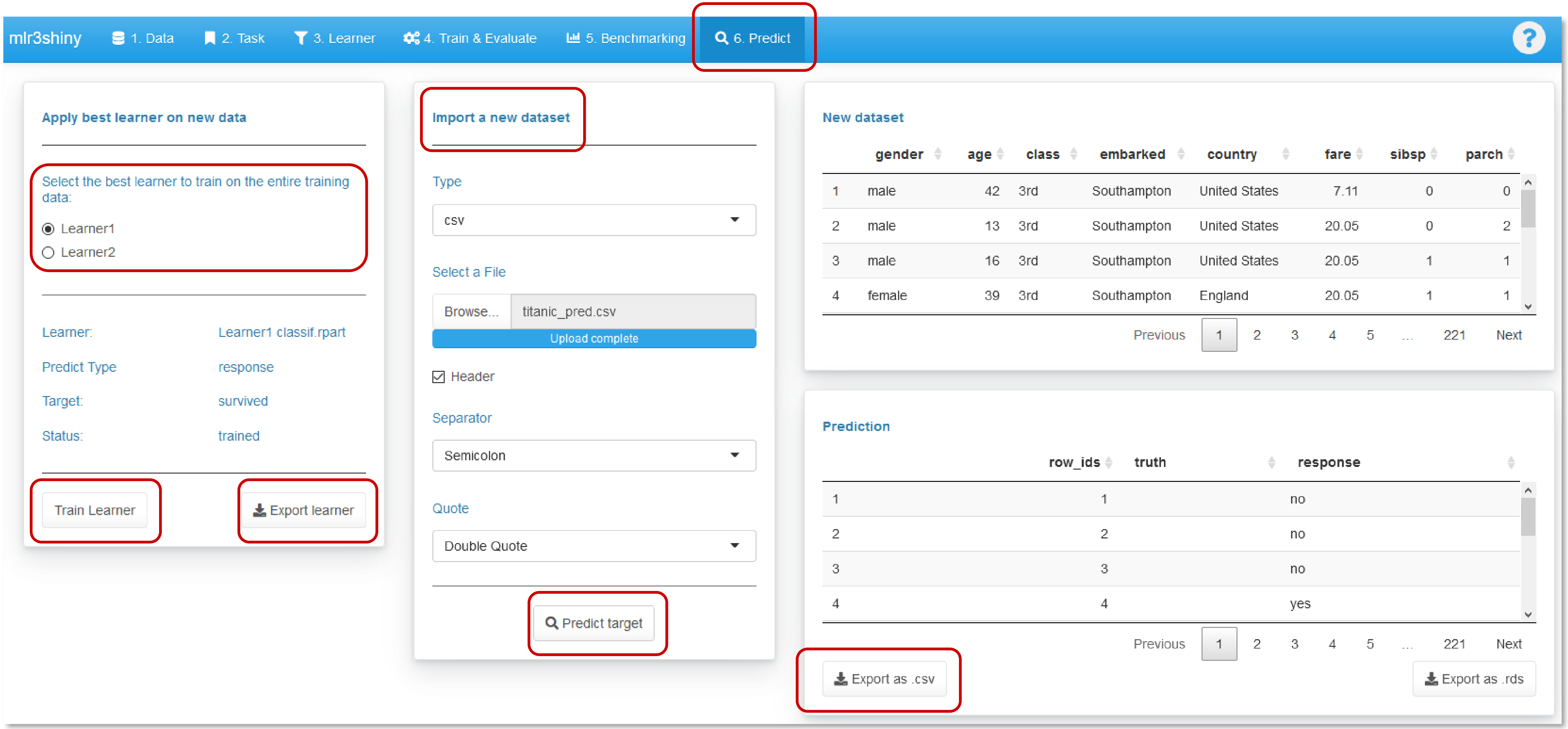
The General Machine Learning Workflow



Finding the Best Learner: Benchmarking



Final Retraining on All Data



Further Reading

mlr3 Manual mlr

mlr3 book

Marc Becker

Martin Binder

Bernd Bischl

Michel Lang

Florian Pfisterer

Nicholas G. Reich

Jakob Richter

Patrick Schratz

Raphael Sonabend

2021-05-14

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