

Supervised Machine Learning principles

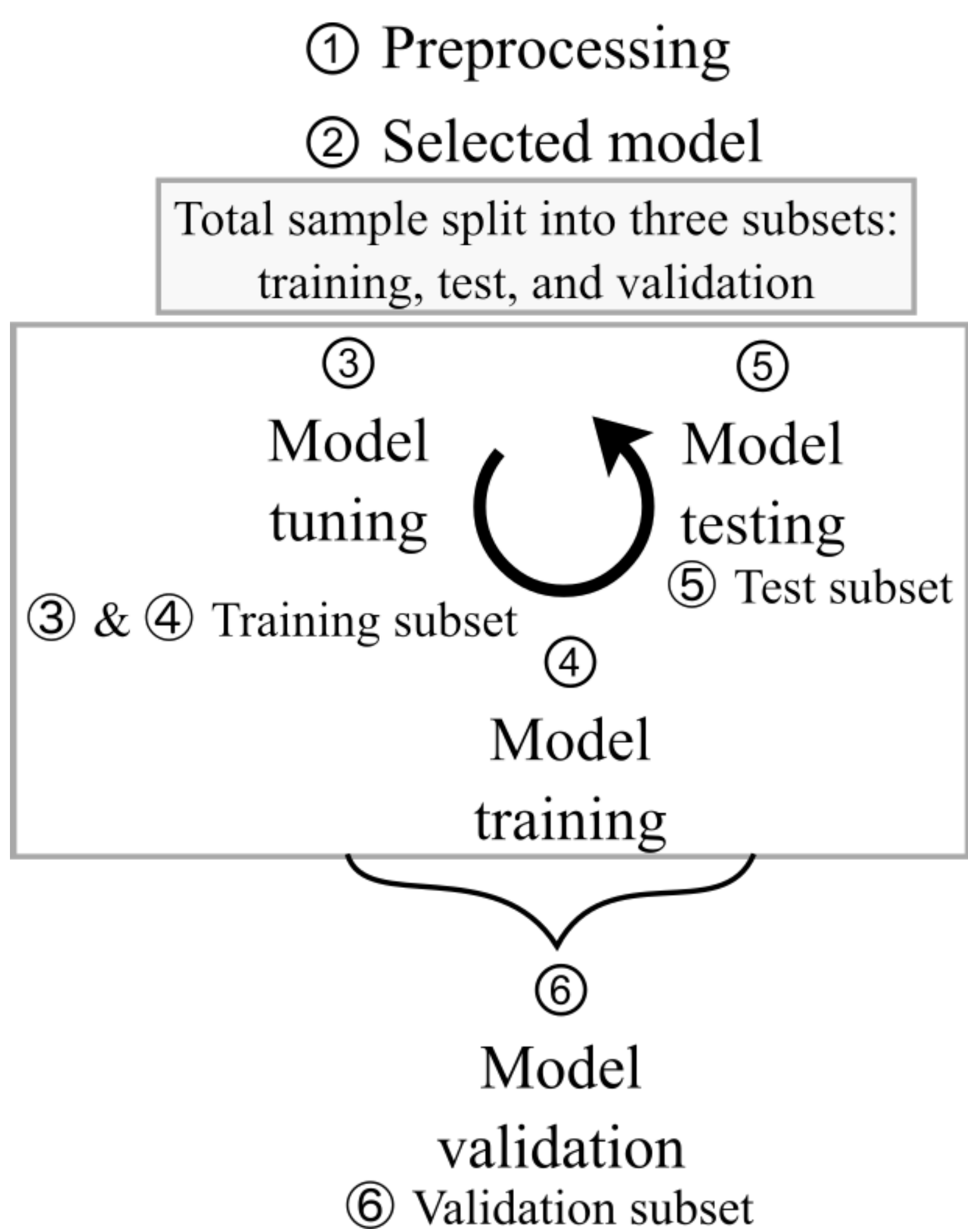
Marcel Miché (marcel.miche@unibas.ch), Thea Zander-Schellenberg, Karina Wahl & Roselind Lieb
Faculty of Psychology, Clinical Psychology and Epidemiology, University of Basel, Basel, Switzerland.



Introduction. Why Supervised Machine Learning (SML)?

S: A target variable **guides** the prediction model's **behavior**.
M: The prediction model runs on a computing **machine**.
L: The prediction model **learns** how the predictors and the target variable are associated, i.e., find optimal trade-off between minimizing prediction errors and maximizing generalizable prediction success.

How does a complete SML process look like?



Demonstration dataset (N = 5)

Predictor values: 4, 9, 10, 12, 15.
Outcome (see [SML](#), target variable) values: 25, 40, 55, 80, 100.

Step 1: Preprocessing

Assuming a nonlinear relationship between predictor and outcome, use the squared predictor values (16, 81, 100, 144, 255).

Step 2: Select model

Select the simple linear regression model.
Split total sample into training, test, and validation subsets:

Obs / Run	1	2	3	4	5
1	V				
2		V			
3			V		
4				V	
5					V

2	1.1	1.2	1.3	1.4
3	TE	TE		
4			TE	
5				TE

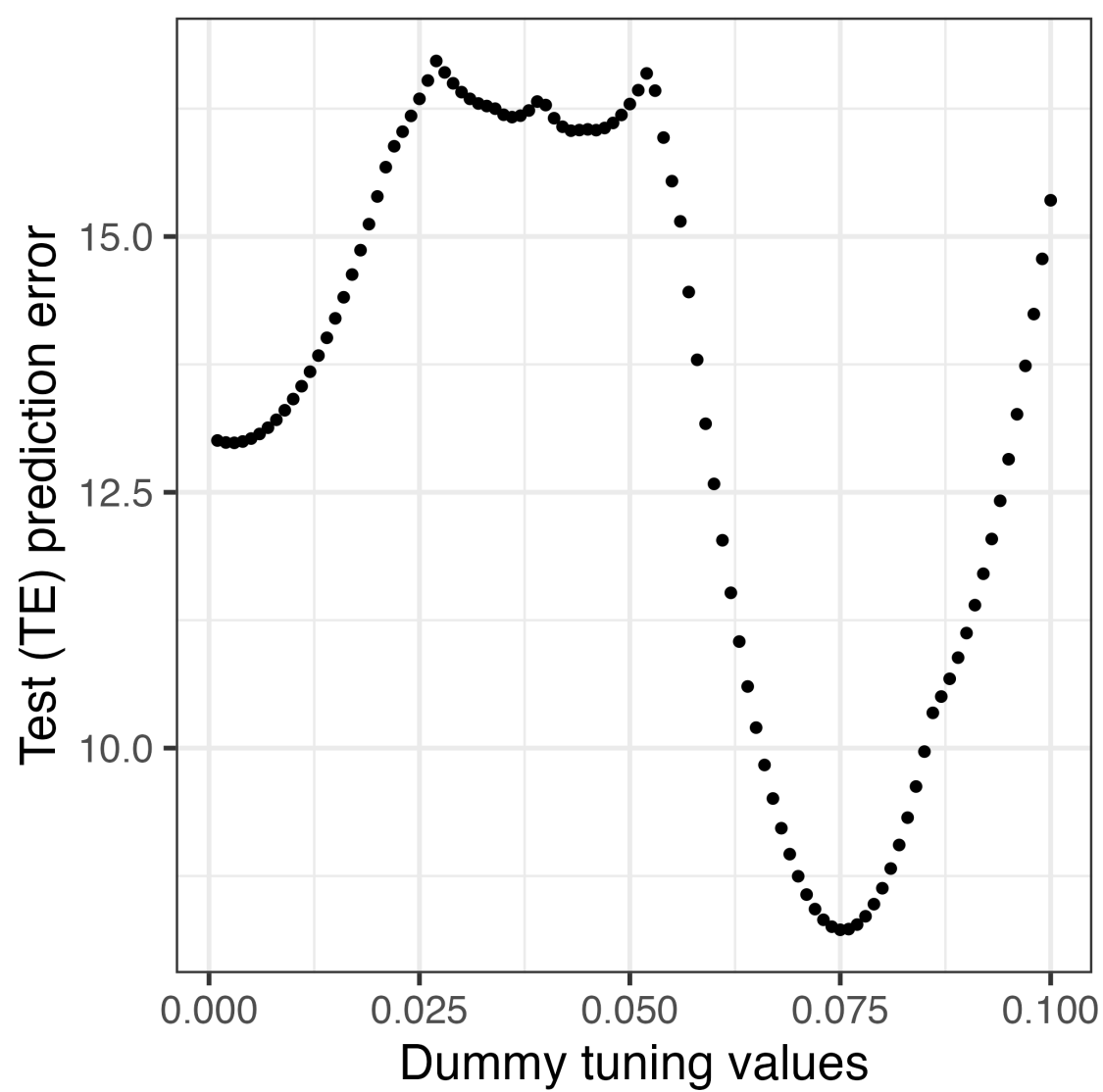
1	2.1	2.2	2.3	2.4
3	TE			
4		TE		
5			TE	

1	5.1	5.2	5.3	5.4
2	TE			
3		TE		
4			TE	

Obs, observation; V, validation; TE, test. Lower tables' empty cells, training.

Step 3: Model tuning

Experimentally influence the training model's regression weight (see [SML](#), find optimal trade-off between bias and variance).



100 different values between 0 and 0.1 were selected for the tuning.

Step 4: Model training. Use training subset (N = 3)

Per training session, add one tuning value to the regression weight.

Step 5: Model testing. Use test subset (N = 1)

Apply each tuned/trained model to the held-out test subset and record the squared prediction error. The lowest prediction error was obtained when adding 0.075 to the regression weight (see step 3).

Step 6: Model validation. Use validation subset (N = 1)

Apply the best test model (**TE**) to the held-out validation subset (see step 2, upper table) and record the squared prediction error.

Results. Best prediction = lowest squared prediction error.

T-	TE	V	T+	TE	V
Run 1	4.4	41.5	Run 1	0.6	37.0
Run 2	8.9	122.0	Run 2	2.8	205.9
Run 3	41.5	4.4	Run 3	0.5	23.0
Run 4	7.3	132.0	Run 4	23.0	0.5
Run 5	3.1	107.0	Run 5	14.2	74.1
Mn	13.0	81.4	Mn	8.2	68.1
Md	8.1	94.2	Md	5.5	52.5

T-, not tuned; TE test; V, validation; T+, tuned; Mn, mean; Md, median.

Tuned better than non-tuned model across the validation subsets.

Conclusion

The six SML principles are meant as a blueprint to eventually master SML introductory texts and scientific reports that apply SML.

Curious? Scan QR code for more SML information ...

