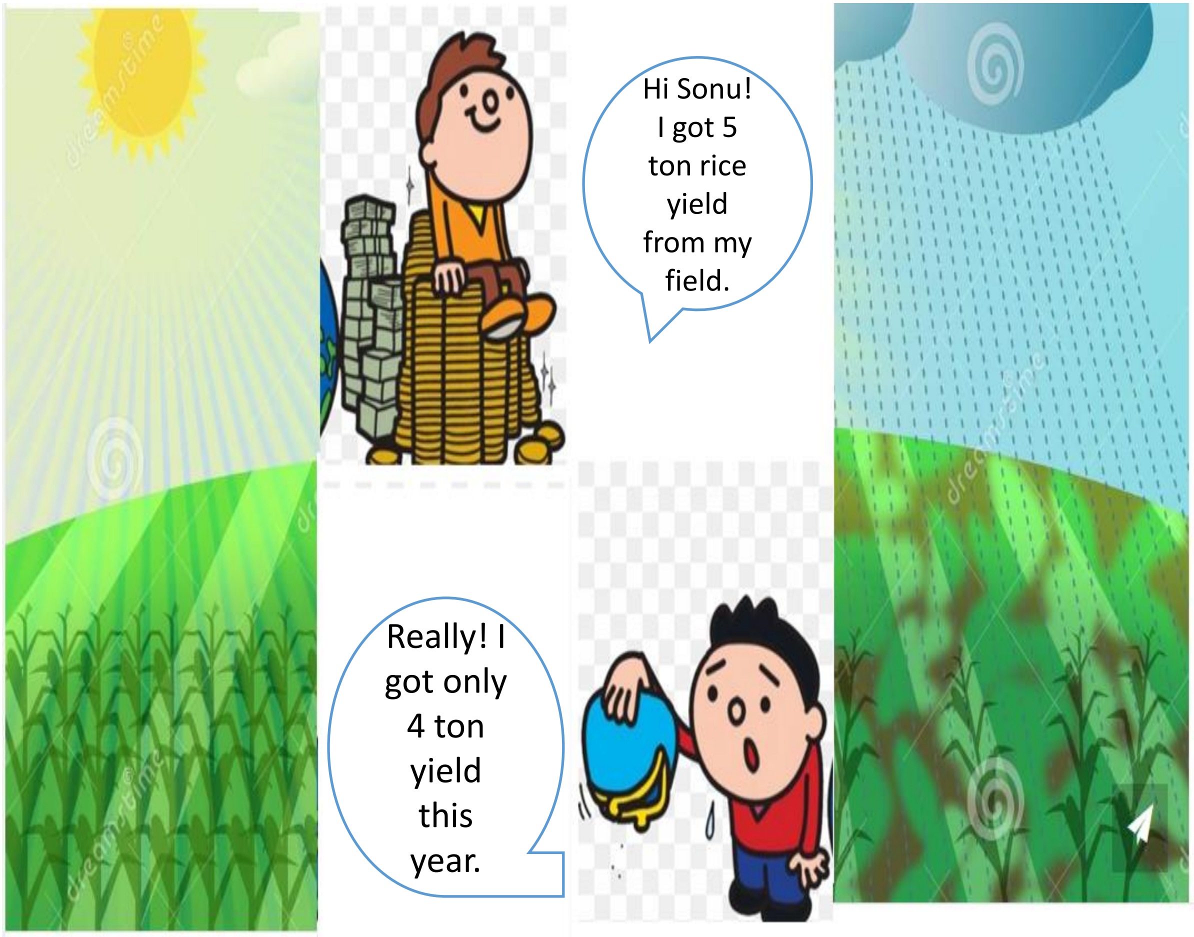
Machine learning Introduction and use in yield and yield gap decomposition



Hi Sonu! I got 5 ton rice yield from my field.
Really! I got only 4 ton yield this year.

Was sowing delayed ?
Less Fertilizer ?
No residue ?
More weed and pest ?
Missed any irrigation ?
More weed and pest ?
Was seed poor ?

Machine learning

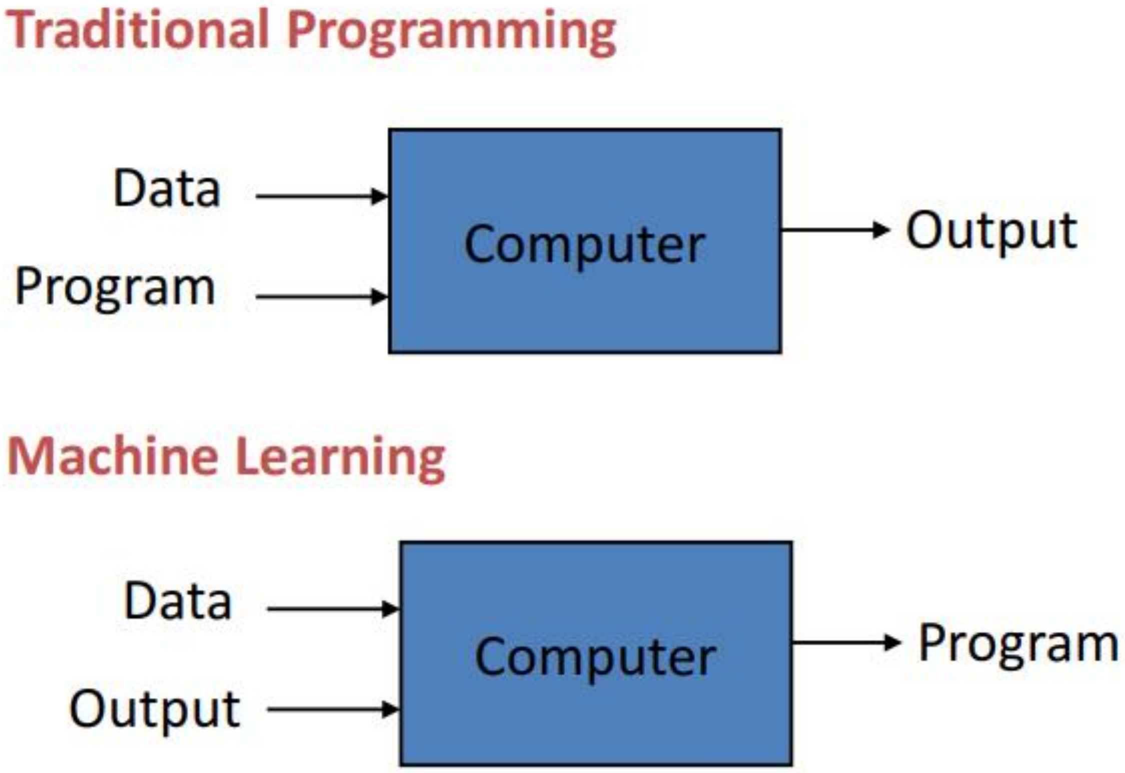
- Learning pattern from data

https://observablehq.com/@yizhe-ang/interactive-visualization-of-linear-regression

Say computer to fit a linear regression or frontier or boundary line based on some assumption about what should be the property of parameters, How residuals should be distributed ?

You provide your management and yield to the computer and ask computer how these set of management leads to this yield. How management interacts to affect yield ?

By default, the model are black box



Slide credit: Pedro Domingos

There exists some other technique for model explanation !!

Machine learning not new for you !

Amazon:

Recommended for you, bought together, recently viewed, etc. ML clusters the things based on search criteria

YouTube: If you regularly sees news in morning, When you open you tube in morning, mostly news comes.

If you search for flights in google, you may get an advertisement related to flight price while watching a video

(Video recommendation come based on recent searches ,clicks, likes and dislikes, watch time, and shares)

Movie recommendation from Netflix:

Browsing history and ratings issued, movie type and popularity, seasonal trends, and item-item similarity

Establishing relationship between two things based on data, maybe we don’t know earlier how they are related

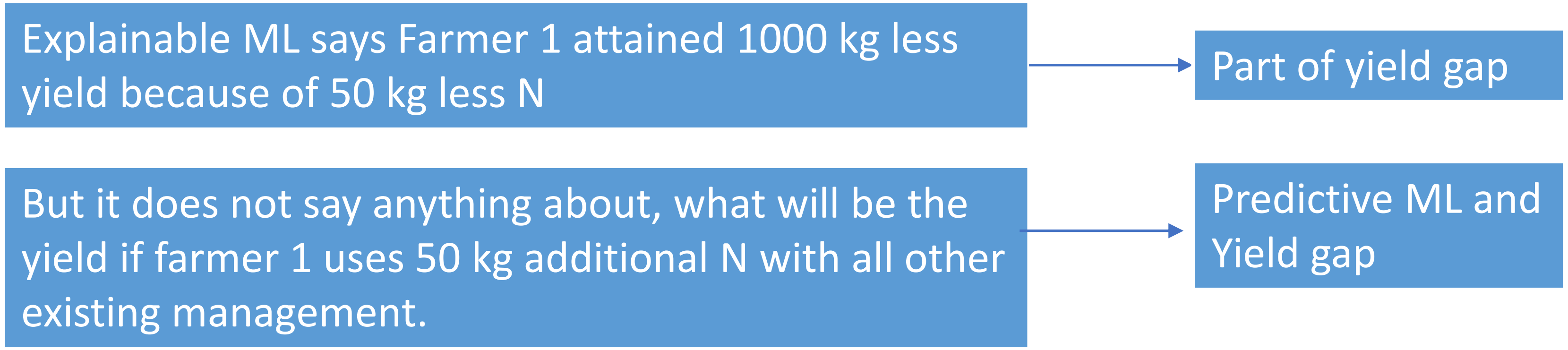
Yield decomposition vs Yield gap decomposition

Explainable Machine learning says how management directly and interacting with other variable affects yield ?

For example, ML says the farms with low irrigation and late sowing has less yield.

Does not it mean, management of these factors can improve yield and hence close yield gap ?

N	P	Irrigation	Sowing	Weed	Soil			Yield
100	50	5	05-Nov	Most	Heavy	Farmer 1		2500
150	55	5	06-Nov	Most	Heavy	Farmer 2		3500



Explainable ML says Farmer 1 attained 1000 kg less yield because of 50 kg less N

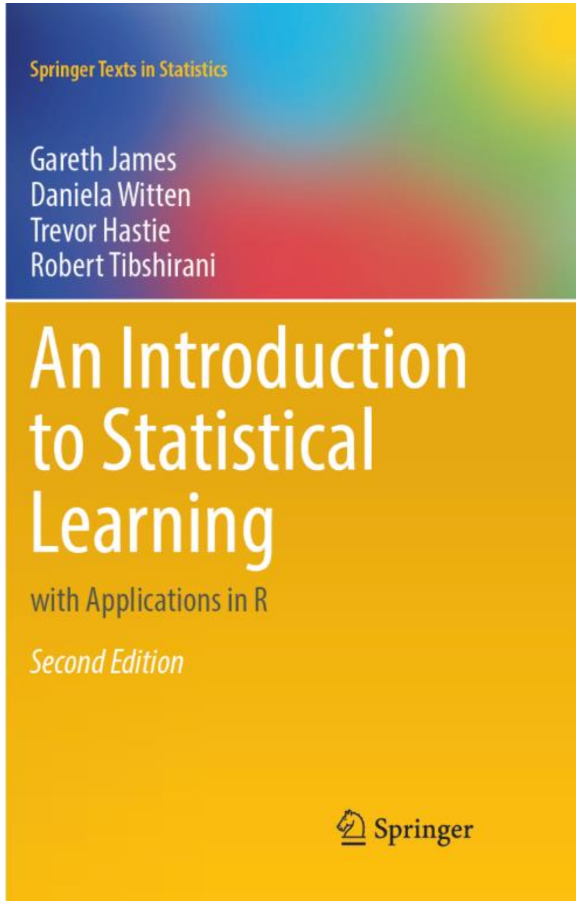
Part of yield gap

But it does not say anything about, what will be the yield if farmer 1 uses 50 kg additional N with all other existing management.

Predictive ML and Yield gap

Till end of the day,

1. Development of Model
2. Feature engineering and fine tuning
3. Model interpretation for yield explanation…



Springer Texts in Statistics

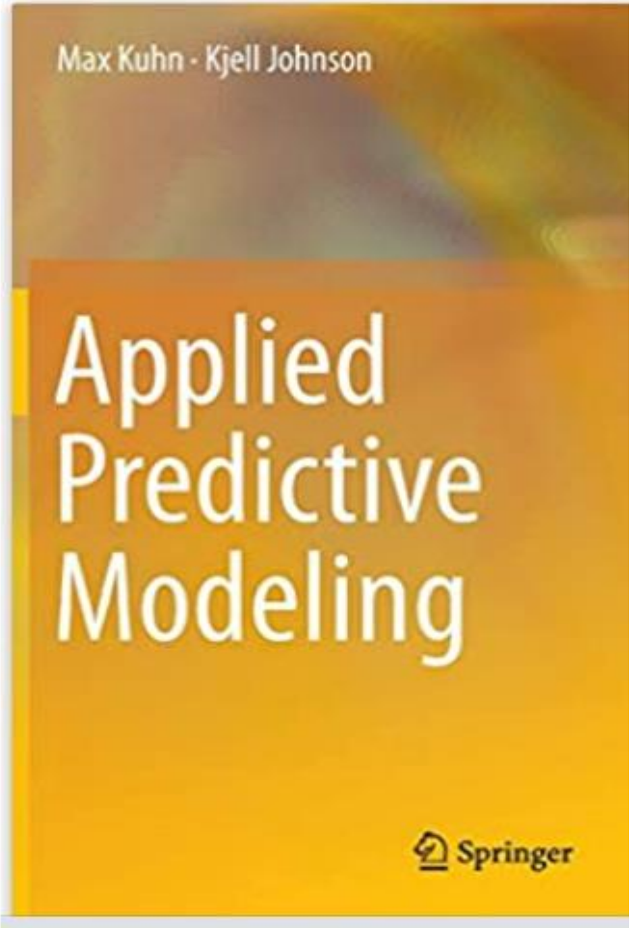
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Second Edition

Springer



Max Kuhn · Kjell Johnson

Applied Predictive Modeling

Springer

Step 1: Model comparison

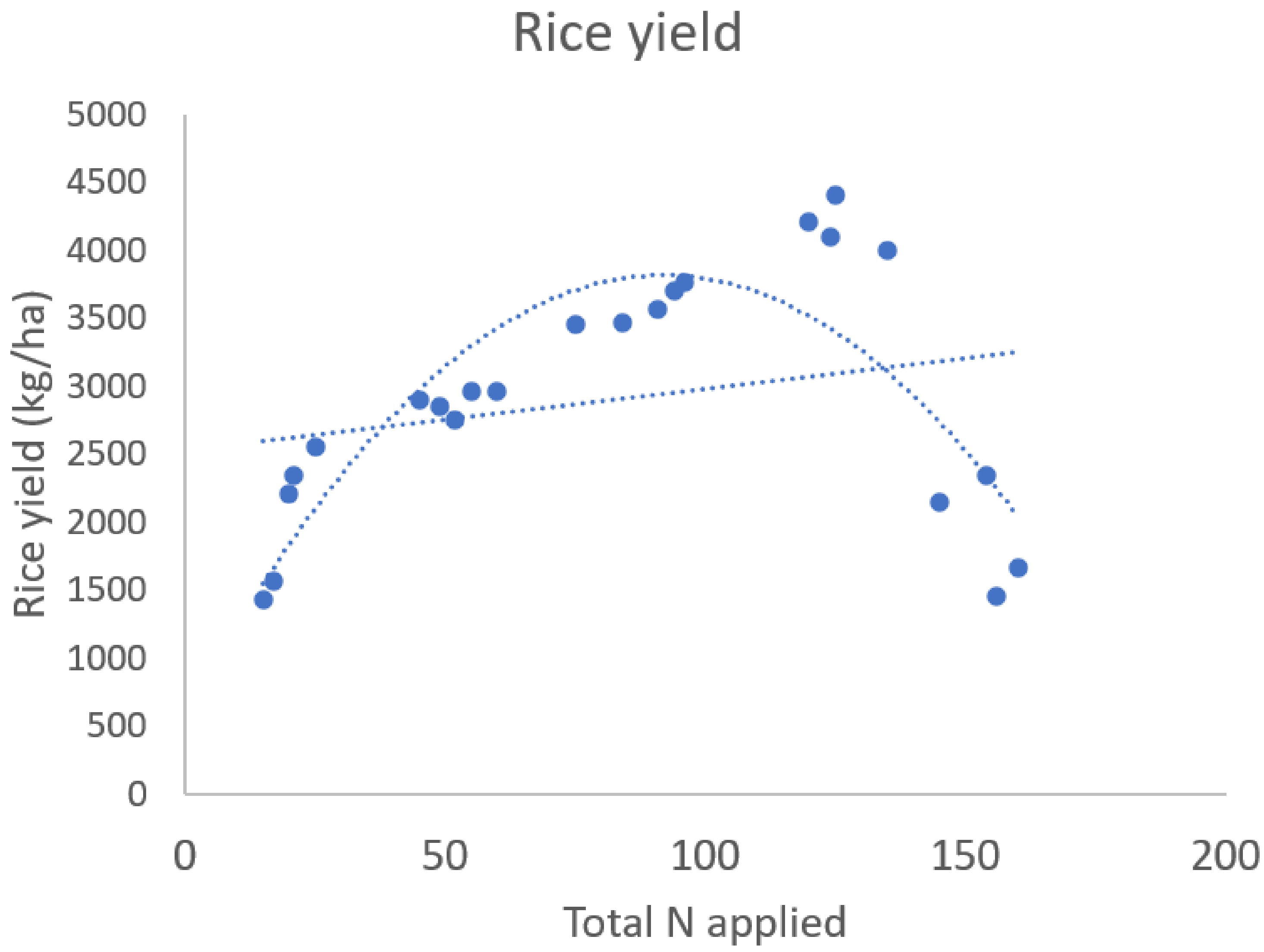
Why we need this ?

- Based on data structure various model may fit differently.
- If underlying structure is linear and small data set linear regression, Lasso or Ridge regression may perform better.
- If too much complex, nonlinear interactions tree-based models works best.
- if too-too much complex ensemble model may be required (mixture of two model one overcoming the others limitation).

Useful in many agronomic data not limited to yield gap only…. (Clustering of farms, Resource use efficiency, homogenous of weed occurrence, …)

Model 1

Linear and quadratic regression

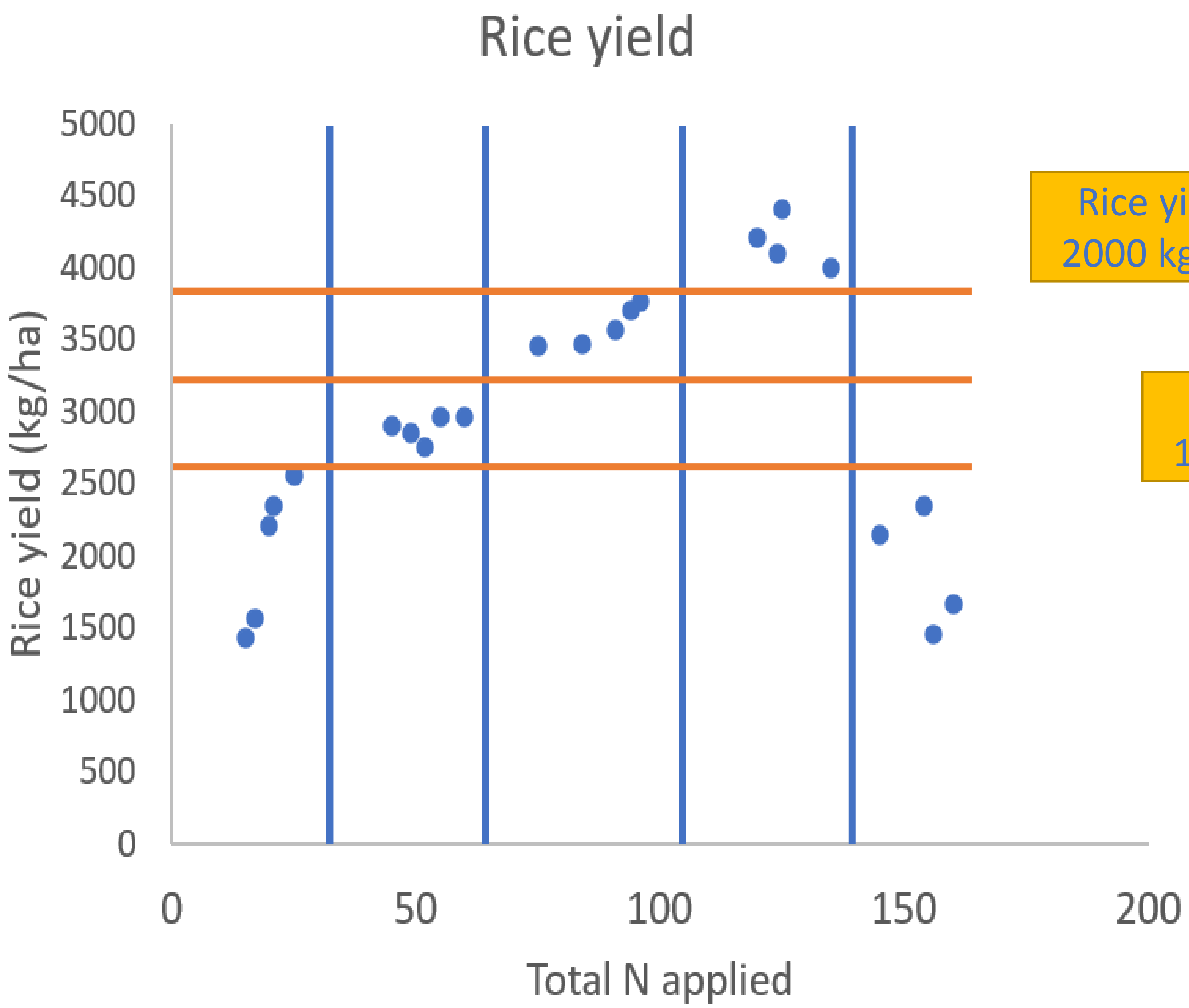


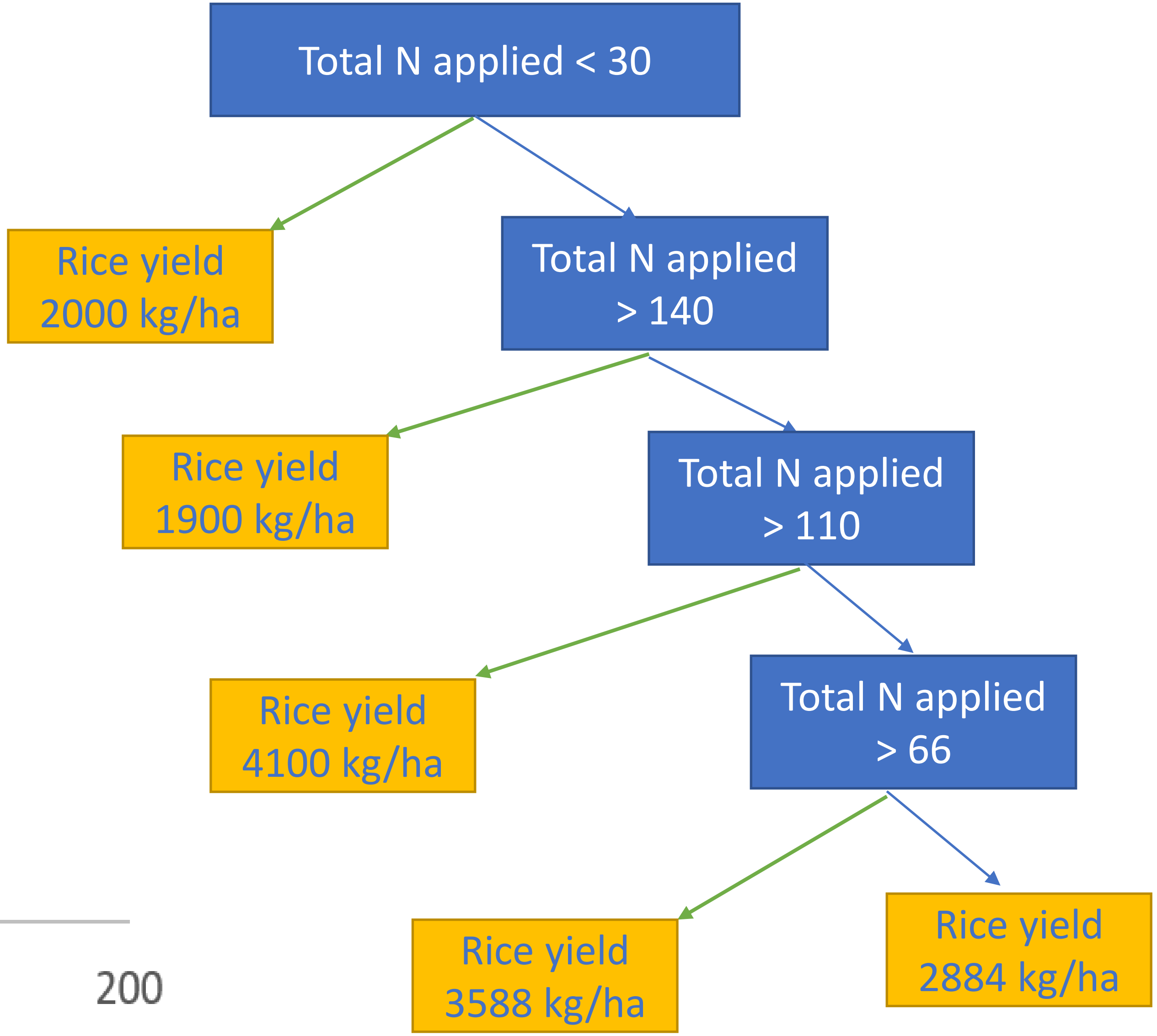
1. We get coefficient which can be directly interpreted
2. No iteration required
3. Easy to calculate and interpret

1. Not good for nonlinear and complex interactions

Model 2

Regression tree intuition





Total N applied < 30

Rice yield 2000 kg/ha

Total N applied > 140

Rice yield 1900 kg/ha

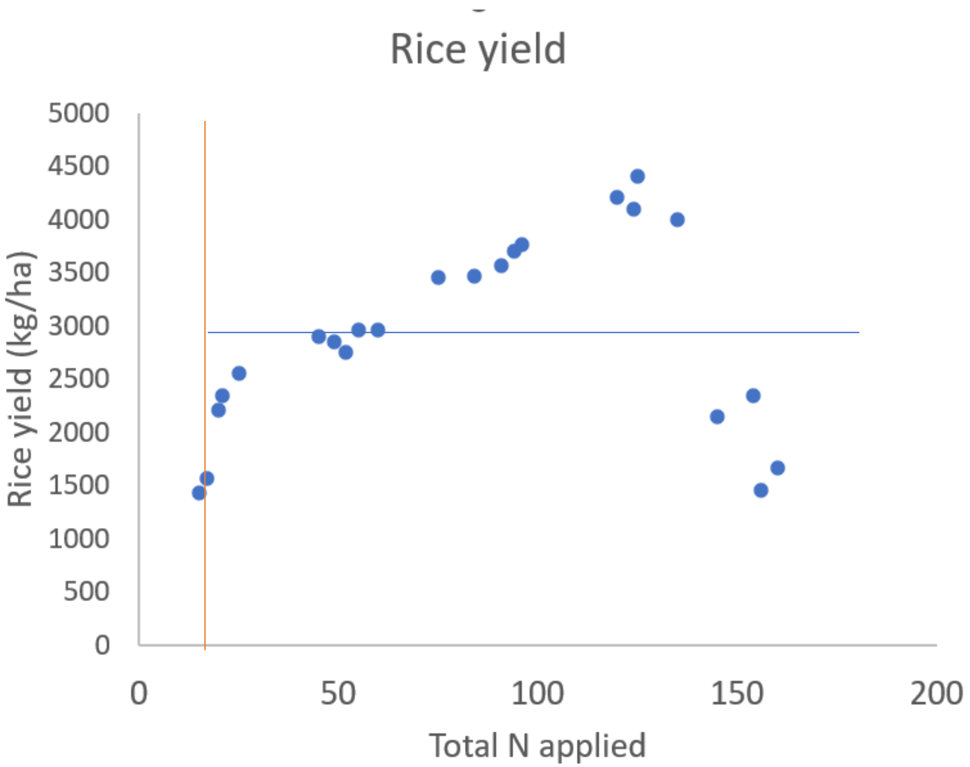
Total N applied > 110

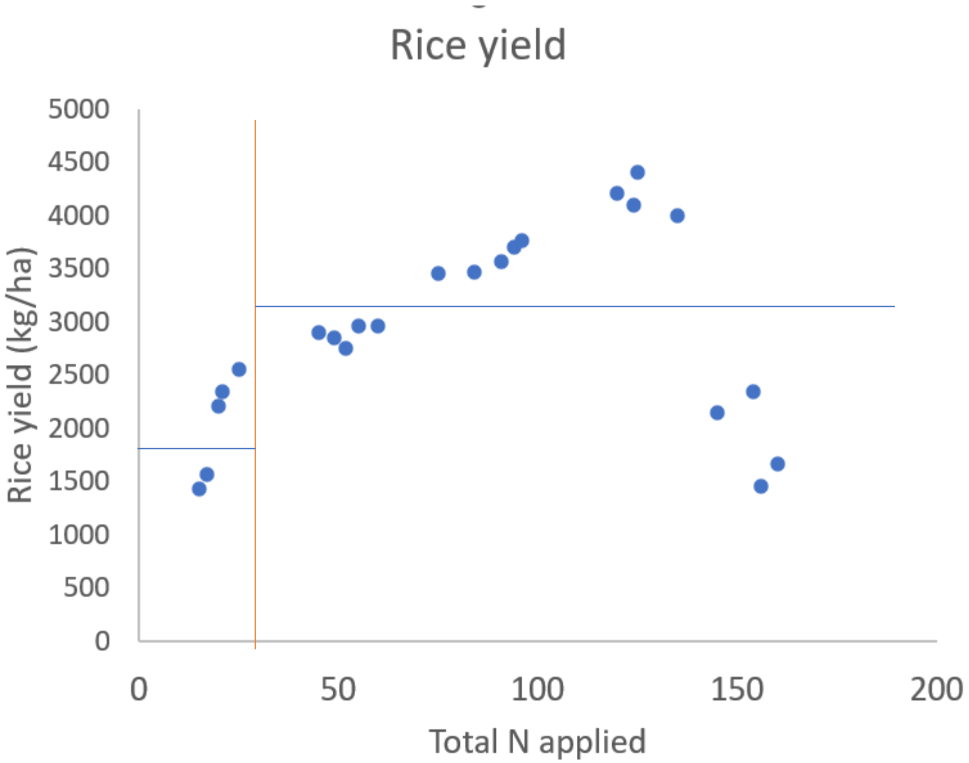
Rice yield 4100 kg/ha

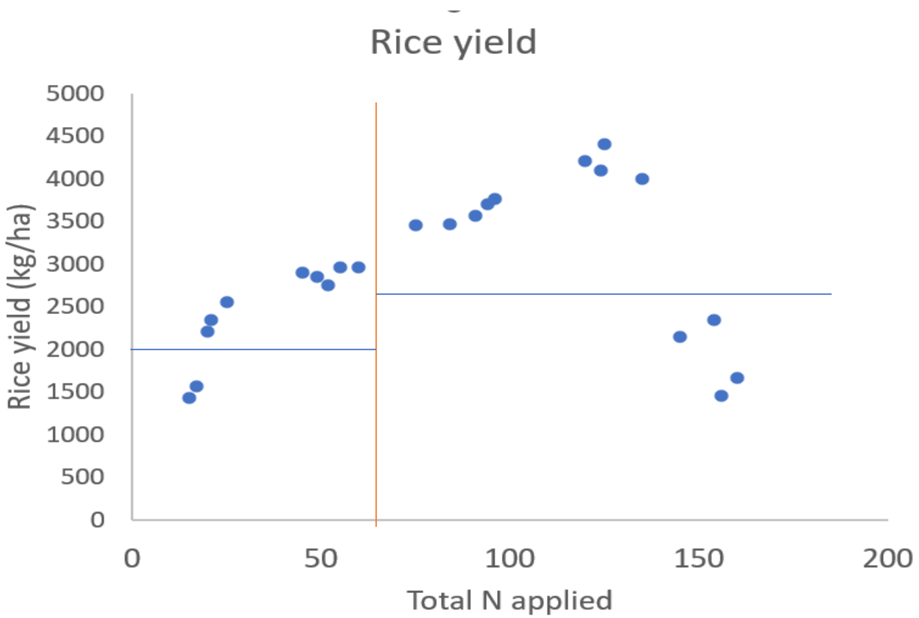
Total N applied > 66

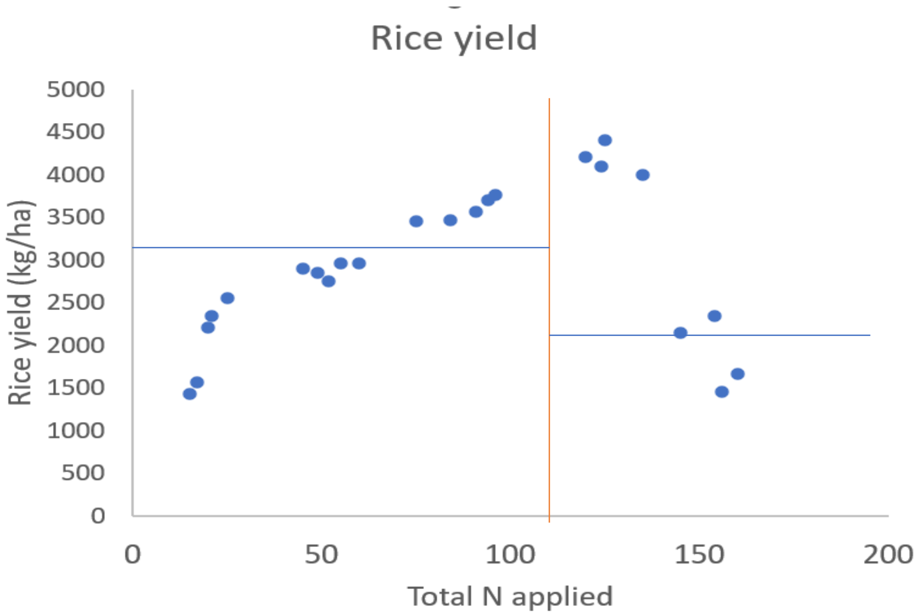
Rice yield 3588 kg/ha

Rice yield 2884 kg/ha

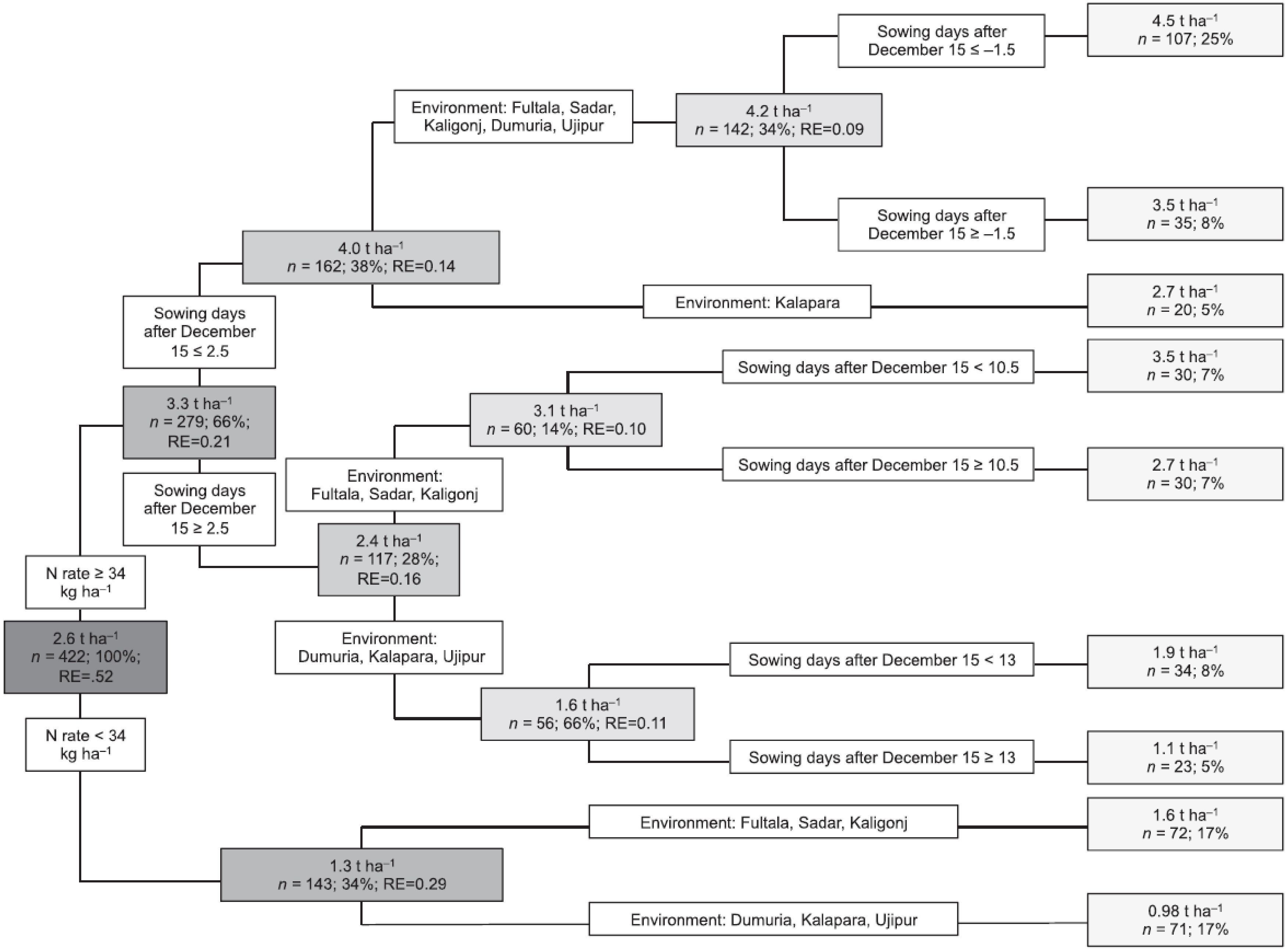








How regression tree choose variable and threshold



2.6 t ha^{-1}
n = 422; 100%; RE=.52

N rate ≥ 34 kg ha^{-1}

3.3 t ha^{-1}
n = 279; 66%; RE=0.21

Sowing days after December 15 ≤ 2.5

4.0 t ha^{-1}
n = 162; 38%; RE=0.14

Environment: Fultala, Sadar, Kaligonj, Dumuria, Ujipur

4.2 t ha^{-1}
n = 142; 34%; RE=0.09

Sowing days after December 15 ≤ –1.5

4.5 t ha^{-1}
n = 107; 25%

Sowing days after December 15 ≥ –1.5

3.5 t ha^{-1}
n = 35; 8%

Environment: Kalapara

2.7 t ha^{-1}
n = 20; 5%

Sowing days after December 15 ≥ 2.5

2.4 t ha^{-1}
n = 117; 28%; RE=0.16

Environment: Fultala, Sadar, Kaligonj

3.1 t ha^{-1}
n = 60; 14%; RE=0.10

Sowing days after December 15 < 10.5

3.5 t ha^{-1}
n = 30; 7%

Sowing days after December 15 ≥ 10.5

2.7 t ha^{-1}
n = 30; 7%

Environment: Dumuria, Kalapara, Ujipur

1.6 t ha^{-1}
n = 56; 66%; RE=0.11

Sowing days after December 15 < 13

1.9 t ha^{-1}
n = 34; 8%

Sowing days after December 15 ≥ 13

1.1 t ha^{-1}
n = 23; 5%

N rate < 34 kg ha^{-1}

1.3 t ha^{-1}
n = 143; 34%; RE=0.29

Environment: Fultala, Sadar, Kaligonj

1.6 t ha^{-1}
n = 72; 17%

Environment: Dumuria, Kalapara, Ujipur

0.98 t ha^{-1}
n = 71; 17%

(Krupnik et al., 2015)

Features of Classification and regression tree

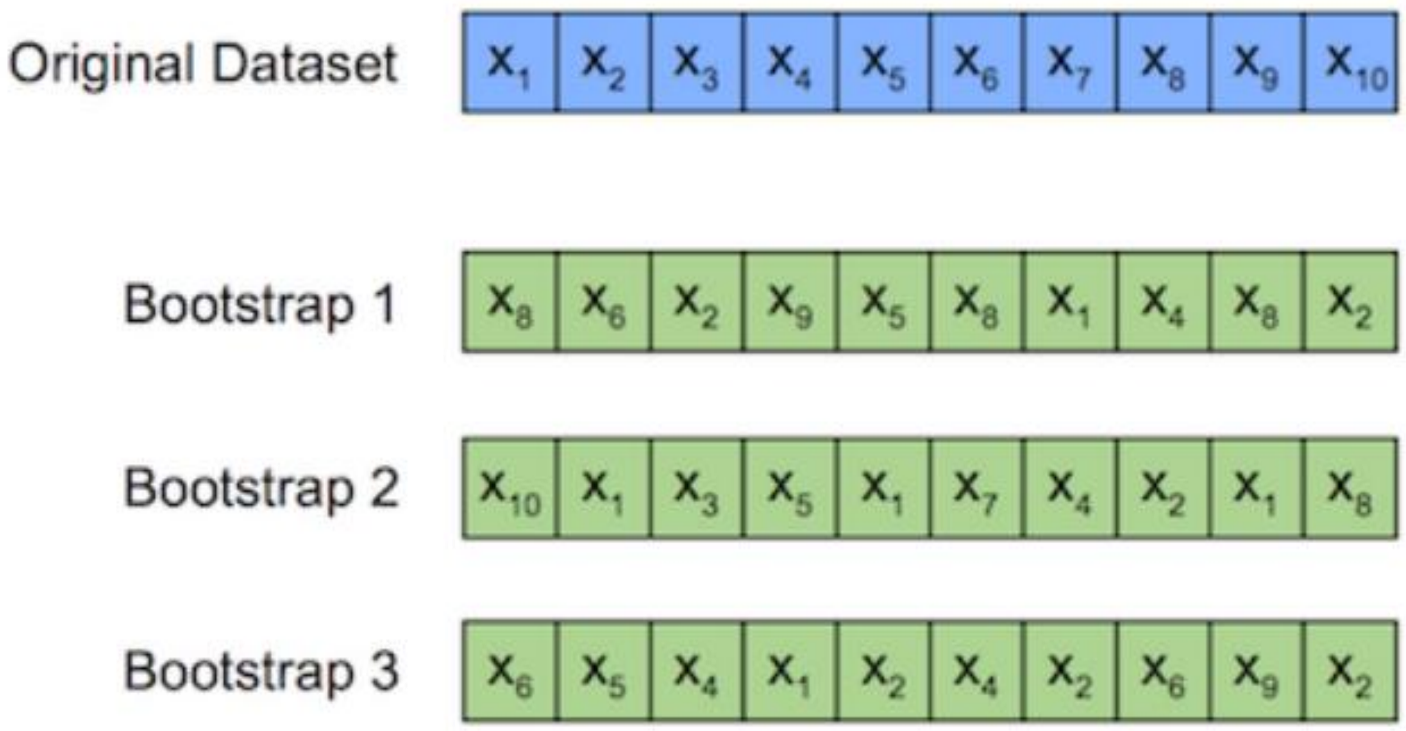
1. Interpretation is easy based on some rule (Structure of tree for model explanation)
2. Variable selection is based on some criteria (minimization of sum of square of error)
3. We can grow tree up to any complexity depth (Hyperparameter; cp)
4. Sensitive to data, variable, and outlier
5. Solution require Iterative running
6. Also, we can take classification problem based on Ginni’s Purity index

7. rpart, rpart.plot

Model 3

Random forest; bagging method

Step 1: Create a “bootstrapped” dataset

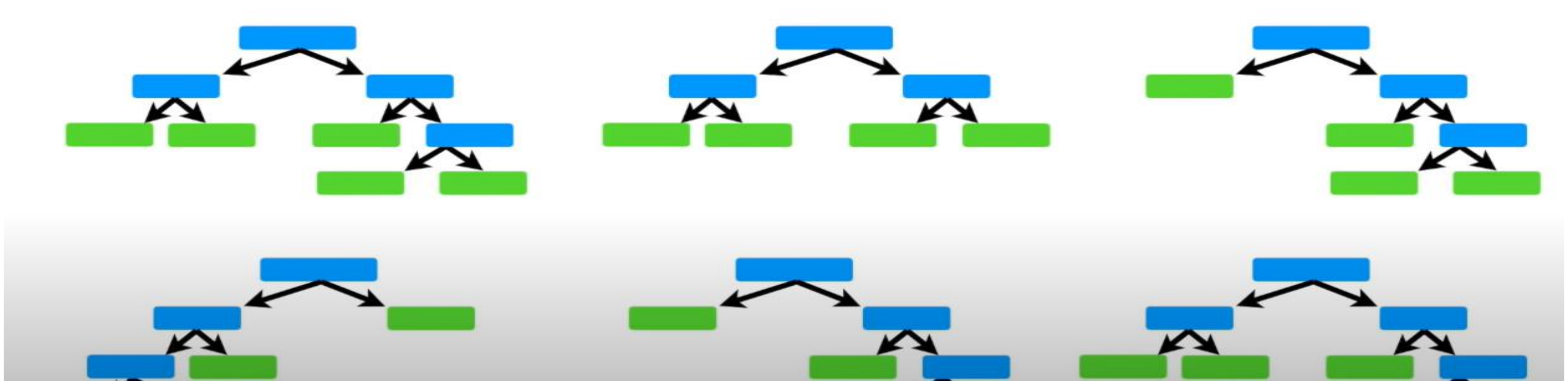


Original Dataset	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Bootstrap 1	x_8	x_6	x_2	x_9	x_5	x_8	x_1	x_4	x_8	x_2
Bootstrap 2	x_{10}	x_1	x_3	x_5	x_1	x_7	x_4	x_2	x_1	x_8
Bootstrap 3	x_6	x_5	x_4	x_1	x_2	x_4	x_2	x_6	x_9	x_2

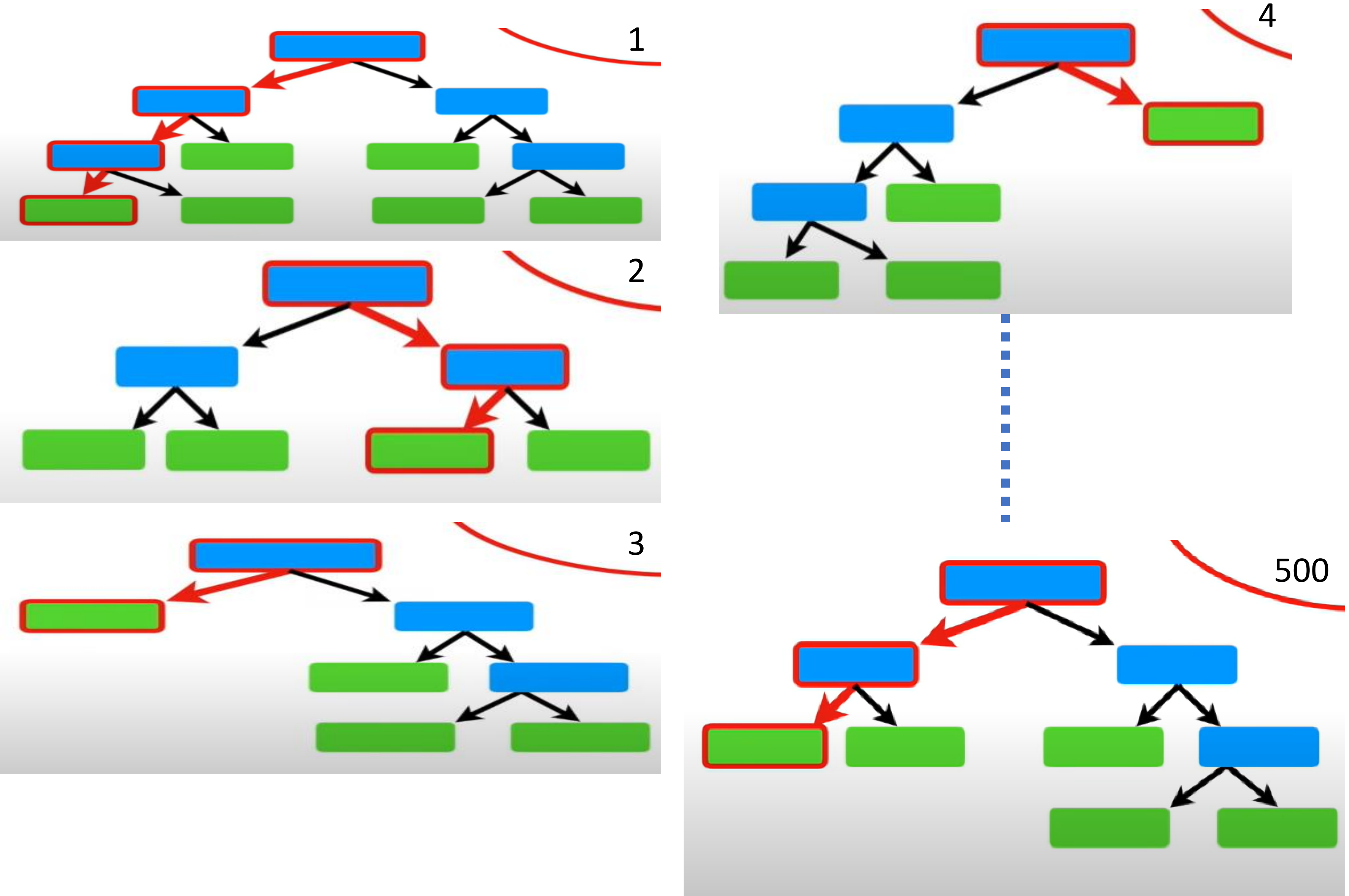
Step 2: Create a decision tree using the bootstrapped dataset, but only use a random subset of variables (or columns) at each step.

1) Using a bootstrapped dataset (Now less sensitive to outlier)
2) Only considering a random subset of variable at each step (mtry or max.nodesize)

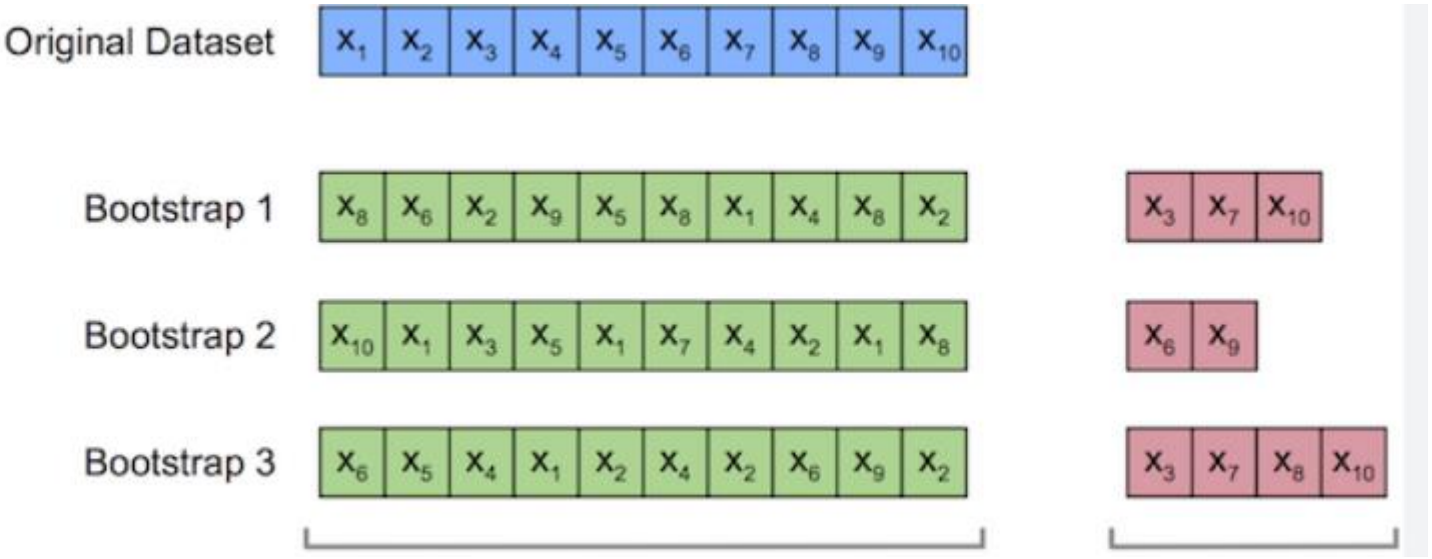
   (Not over rely on a variable)
3) Repeat ideally for 100, 500, or 1000 times (num.tree)
4) We get wide variety of tree (More effective than individual decision tree)



Bootstrap the data + Aggregating the decision = Bagging



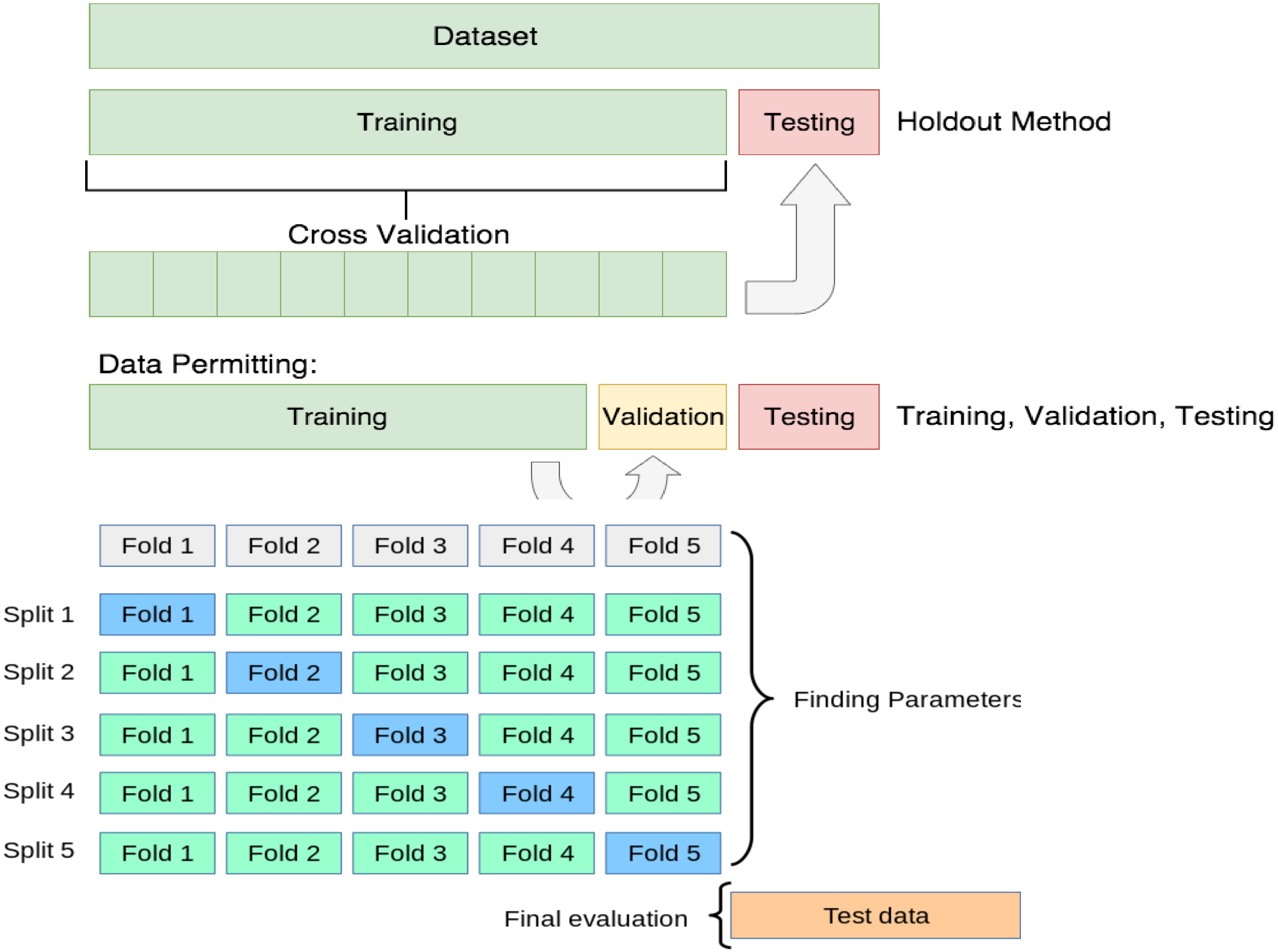
Internal check on performance on Out-of-bag or Out-of-boot data set



Original Dataset	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}					
Bootstrap 1	x_8	x_6	x_2	x_9	x_5	x_8	x_1	x_4	x_8	x_2		x_3	x_7	x_{10}	
Bootstrap 2	x_{10}	x_1	x_3	x_5	x_1	x_7	x_4	x_2	x_1	x_8		x_6	x_9		
Bootstrap 3	x_6	x_5	x_4	x_1	x_2	x_4	x_2	x_6	x_9	x_2		x_3	x_7	x_8	x_{10}

- Performance of random forest is checked based on correct prediction of out-of-bag sample.
- **Out-of-bag error**

Cross-validation: Techniques to check performance on unseen data set while building model



Dataset

Training | Testing — Holdout Method

Cross Validation

Data Permitting:

Training | Validation | Testing — Training, Validation, Testing

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

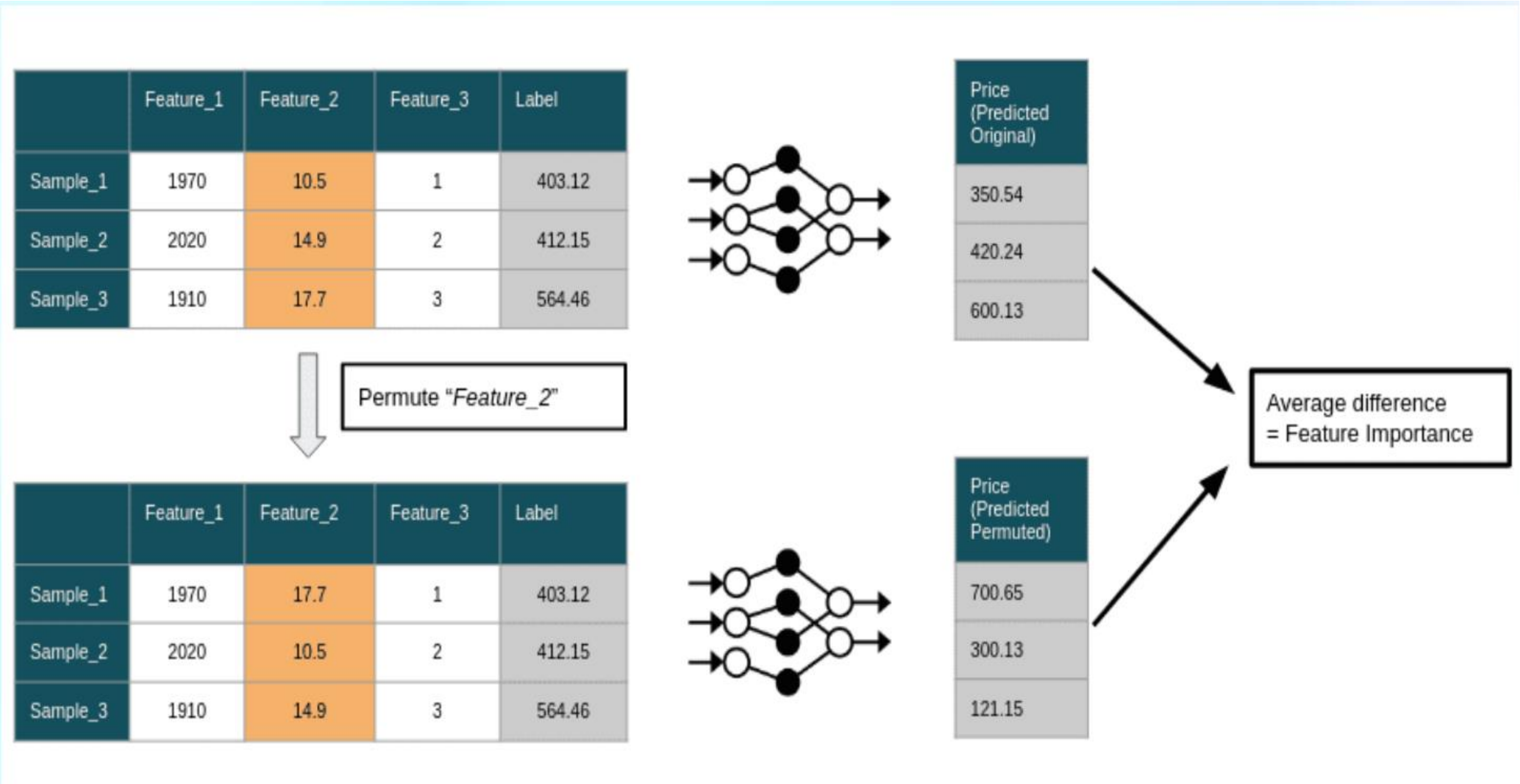
Finding Parameters

Final evaluation — Test data

We can use whole data set for model building

On the expectation that the model performance on the cross validated data set is an unbiased estimator on the model performance on the test set

Permutation based feature importance



	Feature_1	Feature_2	Feature_3	Label
Sample_1	1970	10.5	1	403.12
Sample_2	2020	14.9	2	412.15
Sample_3	1910	17.7	3	564.46

Permute "*Feature_2*"

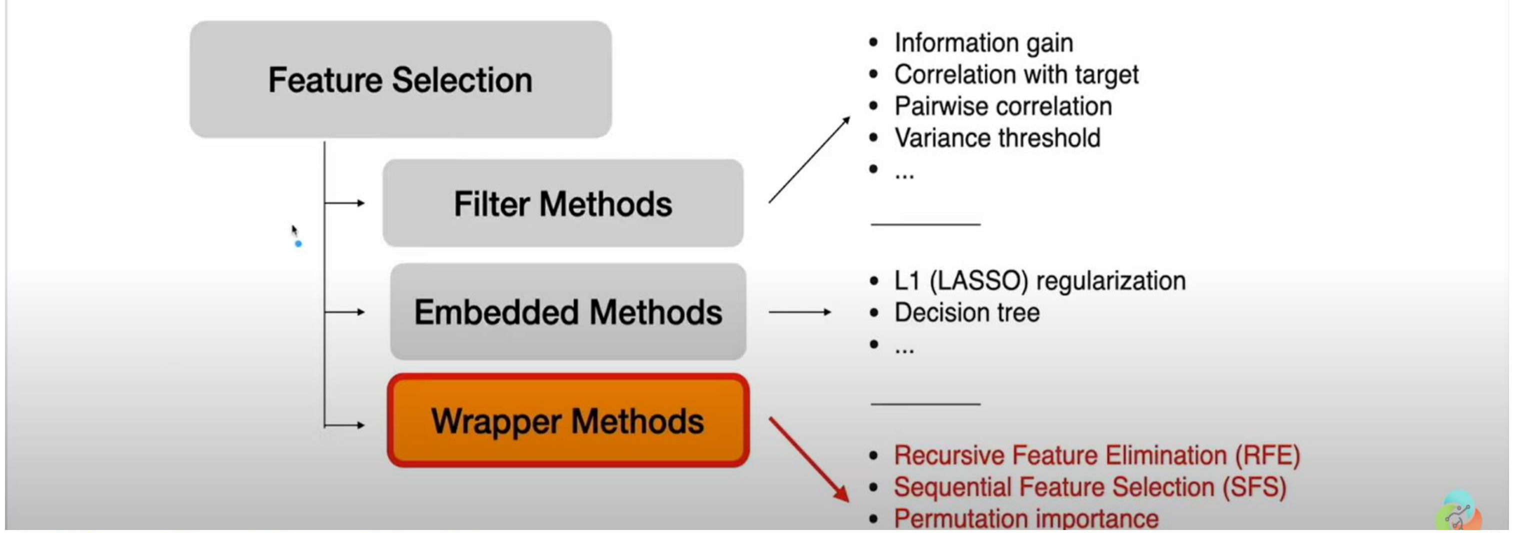
	Feature_1	Feature_2	Feature_3	Label
Sample_1	1970	17.7	1	403.12
Sample_2	2020	10.5	2	412.15
Sample_3	1910	14.9	3	564.46

Price (Predicted Original)
350.54
420.24
600.13

Price (Predicted Permuted)
700.65
300.13
121.15

Average difference
= Feature Importance

Feature selection methods: Recursive feature elimination



Feature Selection

Filter Methods

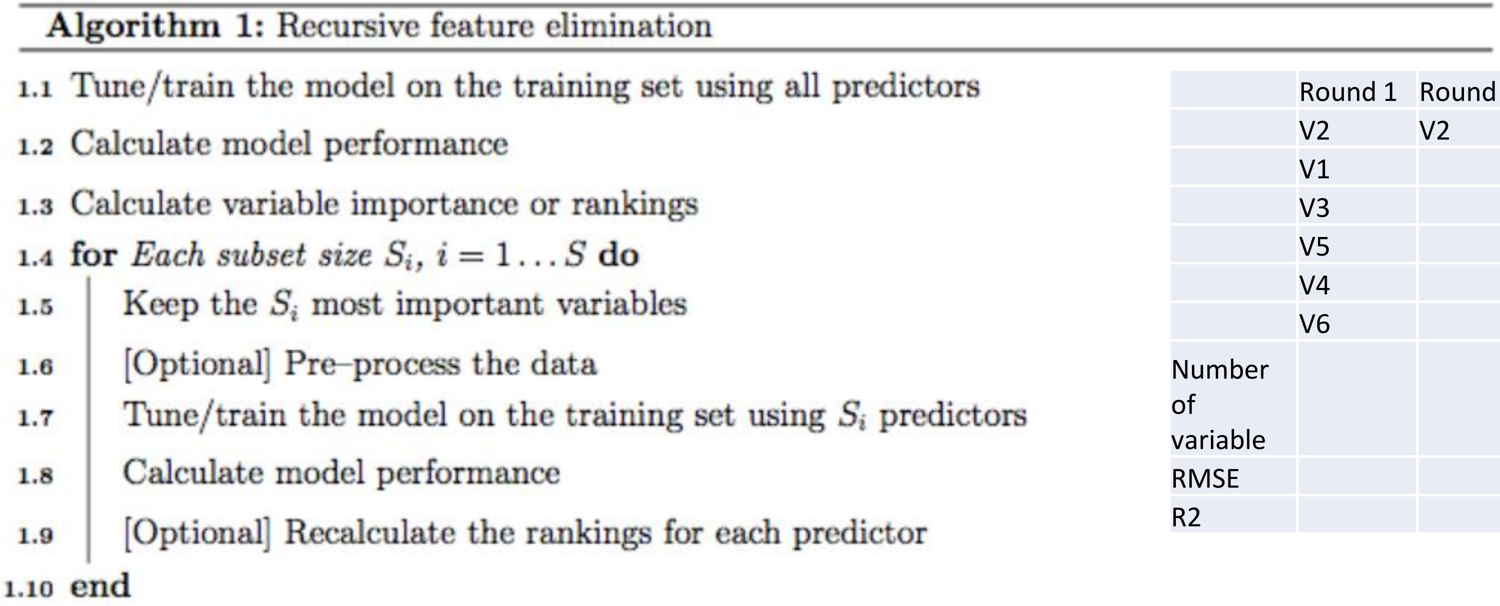
- Information gain
- Correlation with target
- Pairwise correlation
- Variance threshold
- ...

Embedded Methods

- L1 (LASSO) regularization
- Decision tree
- ...

Wrapper Methods

- Recursive Feature Elimination (RFE)
- Sequential Feature Selection (SFS)
- Permutation importance



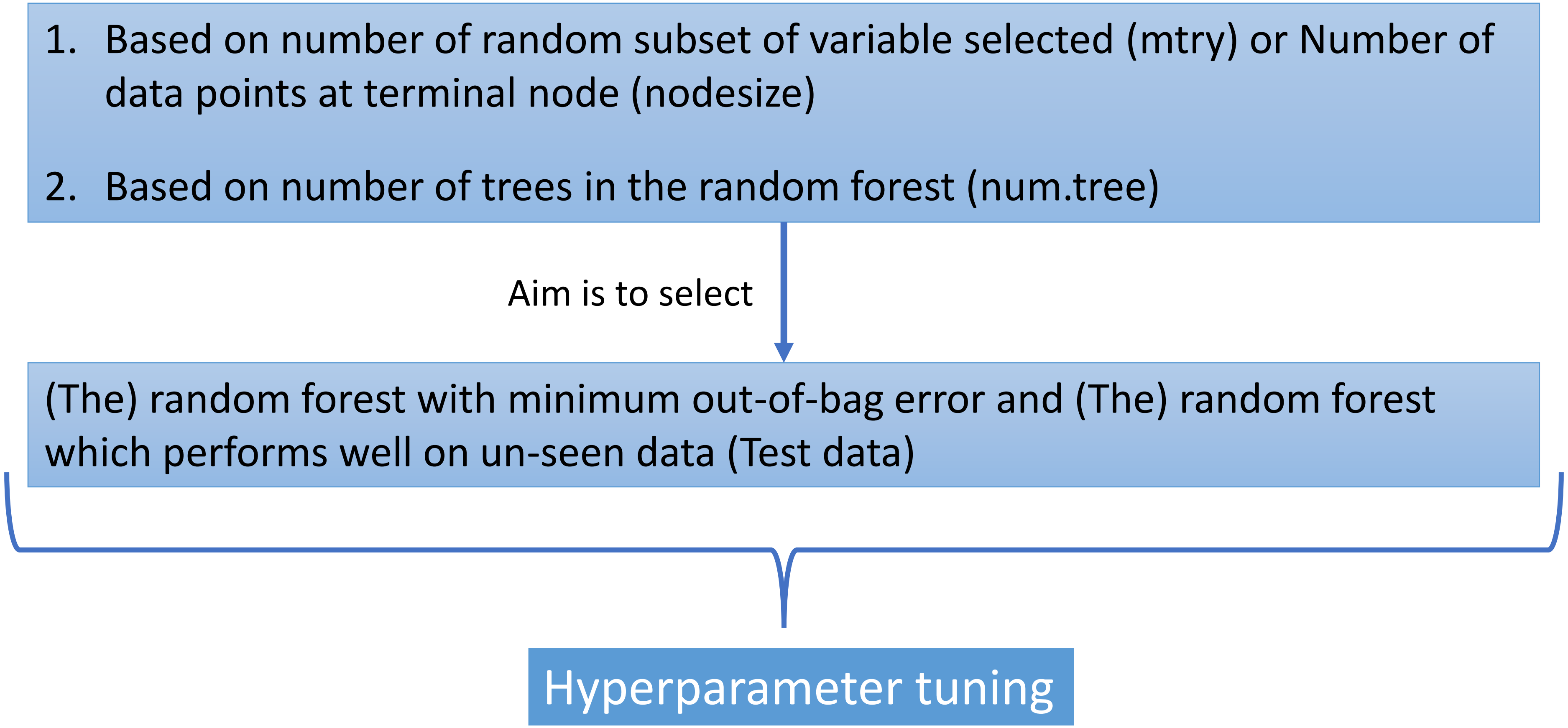
Algorithm 1: Recursive feature elimination

```
Tune/train the model on the training set using all predictors
Calculate model performance
Calculate variable importance or rankings
for Each subset size $S_i$, $i = 1 \ldots S$ do
    Keep the $S_i$ most important variables
    [Optional] Pre-process the data
    Tune/train the model on the training set using $S_i$ predictors
    Calculate model performance
    [Optional] Recalculate the rankings for each predictor
end
```

	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
	V2	V2	V2	V2	V2	V2	V2
	V1		V1	V1	V1	V1	V1
	V3			V3	V3	V3	V3
	V5				V5	V5	V5
	V4					V4	V4
	V6						V6
Number of variable		1	2	3	4	5	6
RMSE							
R2							

(https://www.youtube.com/watch?v=SljoN0cO95Q)

There can be many random forest..



1. Based on number of random subset of variable selected (mtry) or Number of data points at terminal node (nodesize)

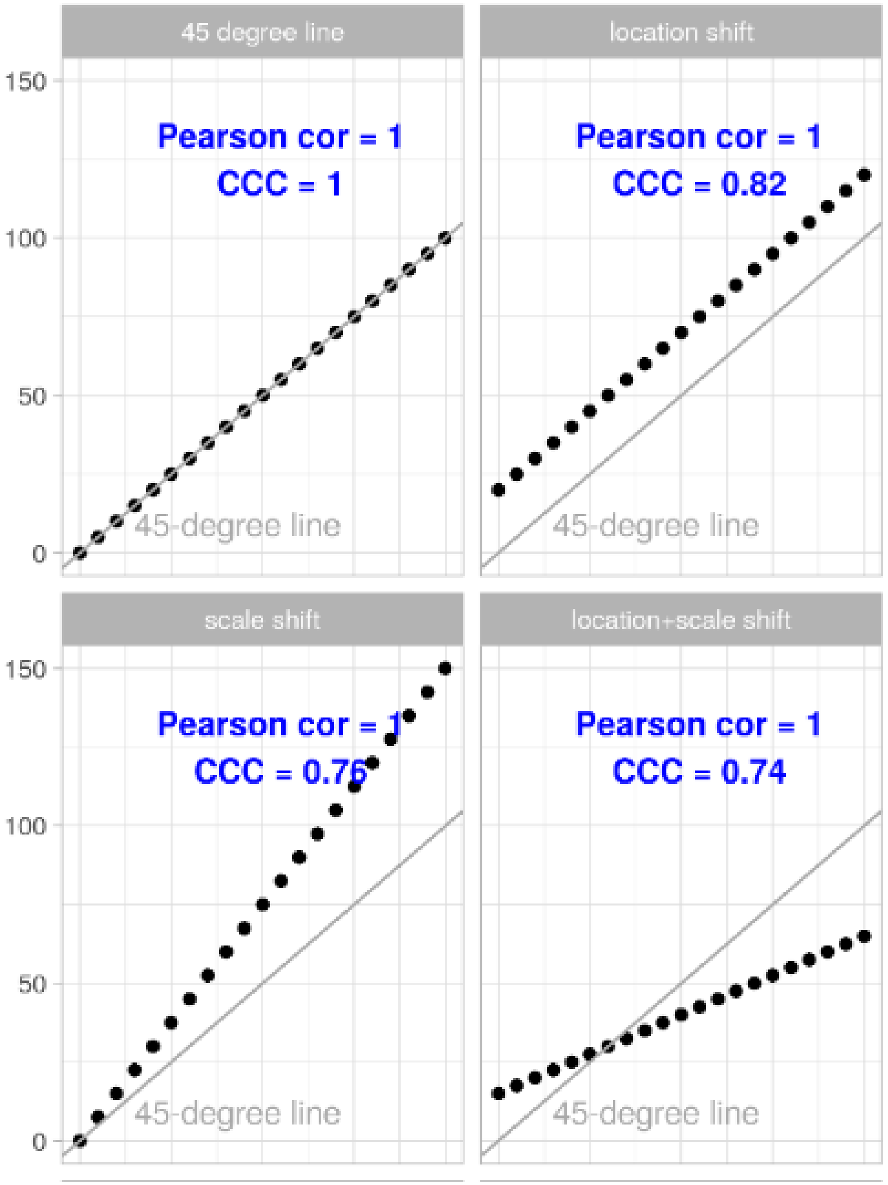
2. Based on number of trees in the random forest (num.tree)

Aim is to select

(The) random forest with minimum out-of-bag error and (The) random forest which performs well on un-seen data (Test data)

Hyperparameter tuning

Model evaluation statistics



Internal Only using the calibration/training information and model diagnostics

External Using external information, not used in model calibration/training

Cross-validation Simulating external assessment with the same dataset used for model building

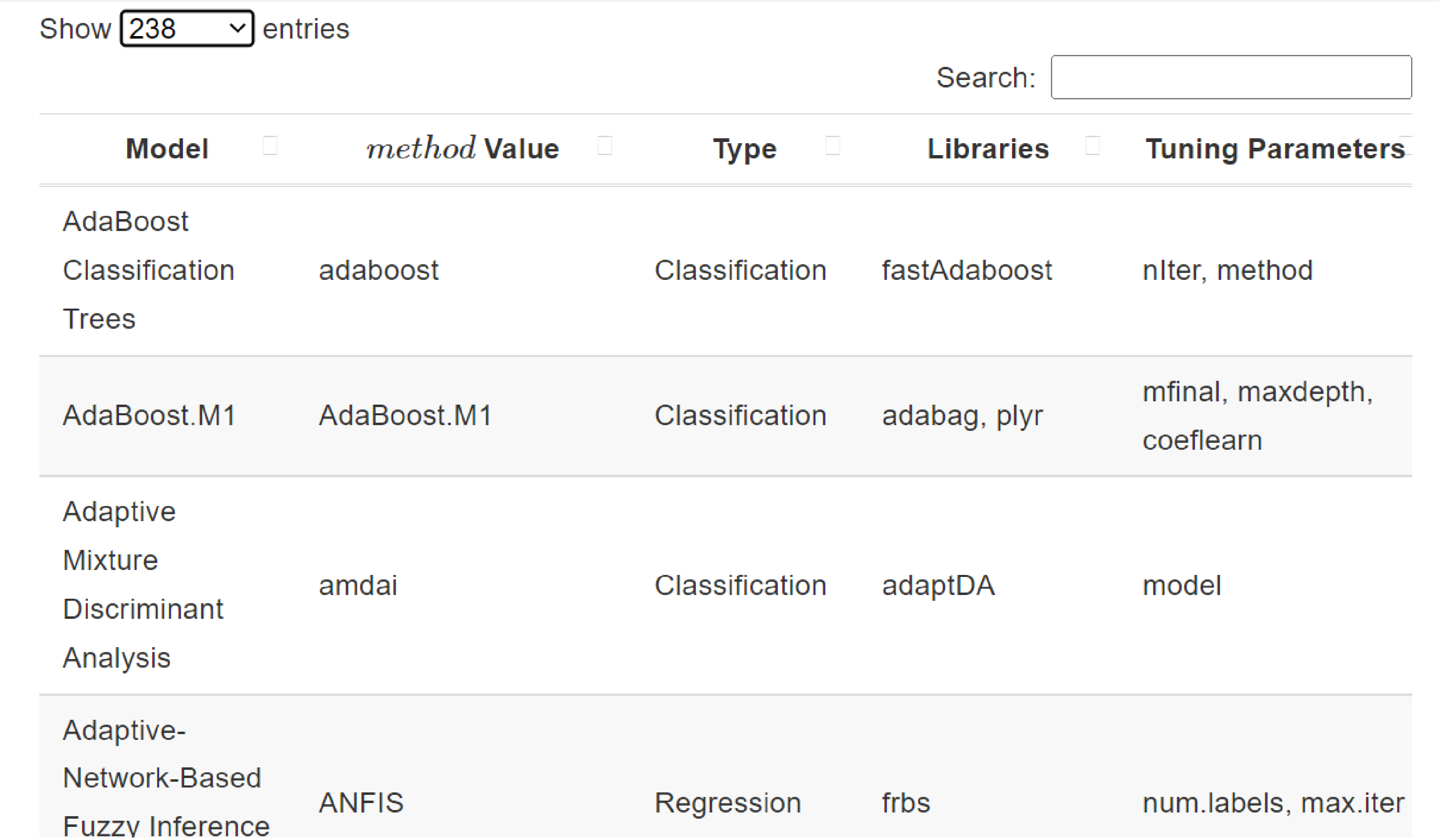
$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i) \qquad \mathrm{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2\right]^{1/2}$$

The CCC includes all sources of deviation from a perfect model:

- location shift (bias) $(\mu_1 - \mu_2)/\sqrt{\sigma_1\sigma_2}$
- scale shift (slope not 1) σ_1/σ_2
- lack of correlation (spread) $1 - \rho_{1,2}$

(David Rosetire)

Numerous model (and variants) exists in literature and are increasing day by day too…



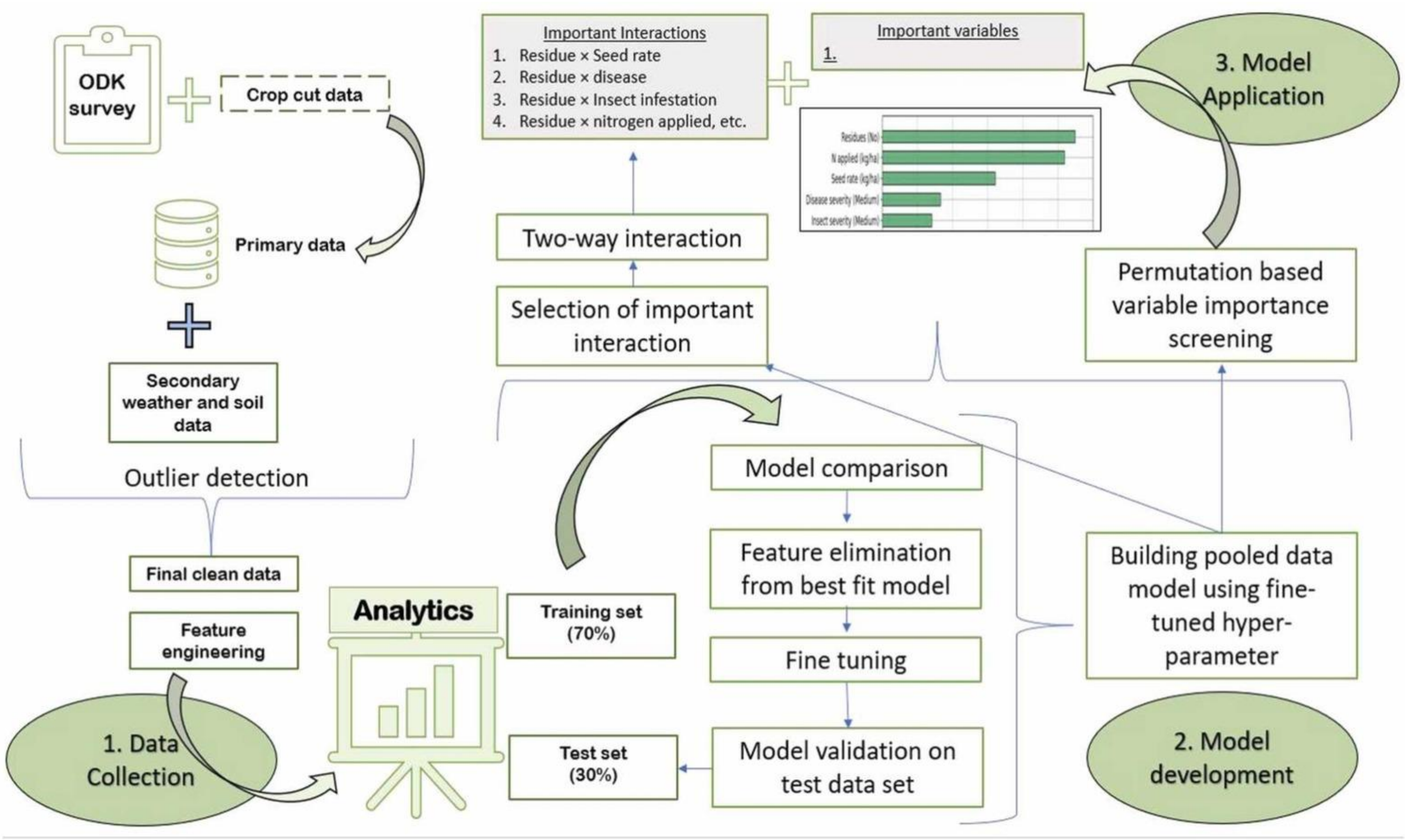
Show 238 entries

Search:

Model	*method* Value	Type	Libraries	Tuning Parameters
AdaBoost Classification Trees	adaboost	Classification	fastAdaboost	nIter, method
AdaBoost.M1	AdaBoost.M1	Classification	adabag, plyr	mfinal, maxdepth, coeflearn
Adaptive Mixture Discriminant Analysis	amdai	Classification	adaptDA	model
Adaptive-Network-Based Fuzzy Inference	ANFIS	Regression	frbs	num.labels, max.iter

Snapshot from caret github page on 13/05/2023

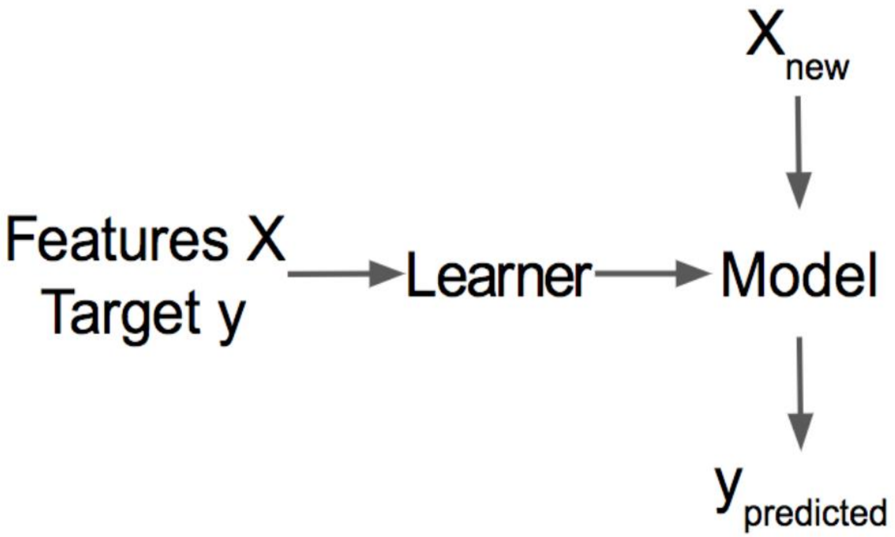
What a typical ML workflows should be ?

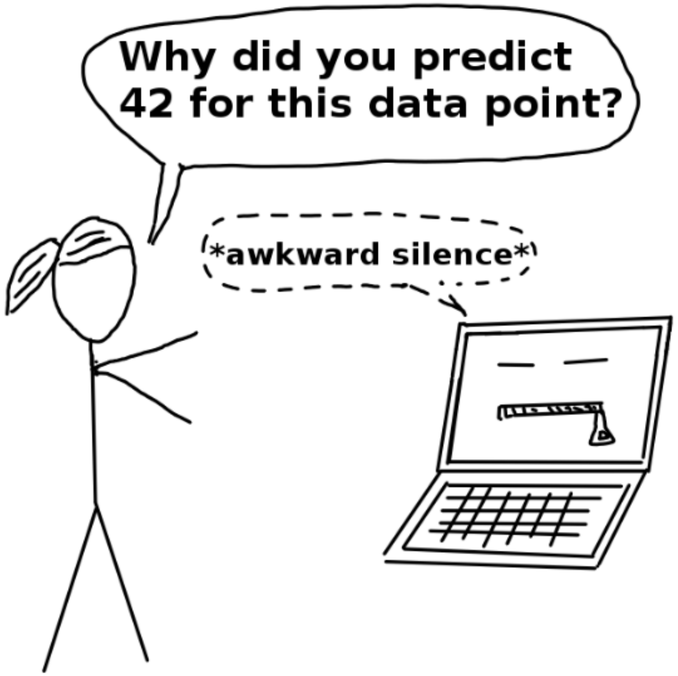


(Nayak, H.S., Silva J V., Parihar C.M., et al., 2022)

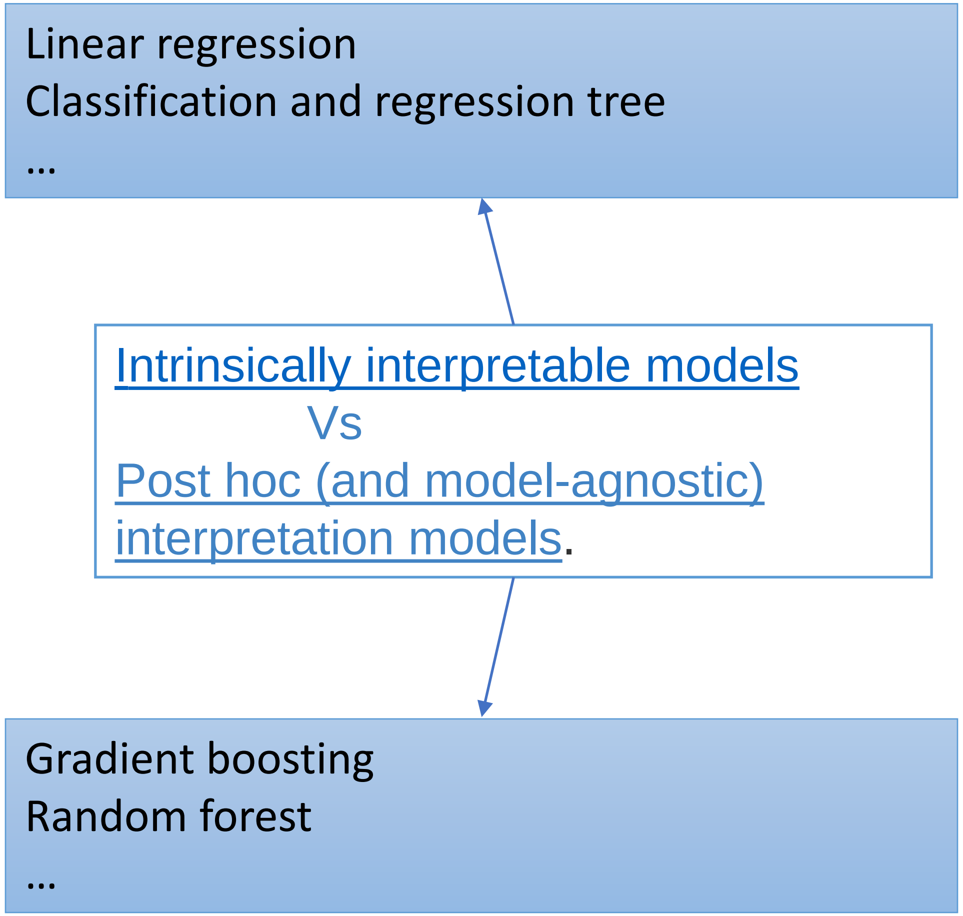
RF model developed! How to interpret ??

1. We don’t have any coefficient alike in linear regression for interpreting or publishing
2. We don’t have a regression tree like structure (We have 500s of trees)
3. How to say which variable are important ? And their effect ?





Model interpretation



Linear regression
Classification and regression tree
…

Intrinsically interpretable models
Vs
Post hoc (and model-agnostic) interpretation models.

Gradient boosting
Random forest
…

Local or global?

Which variable is more important for describing yield at population level ?

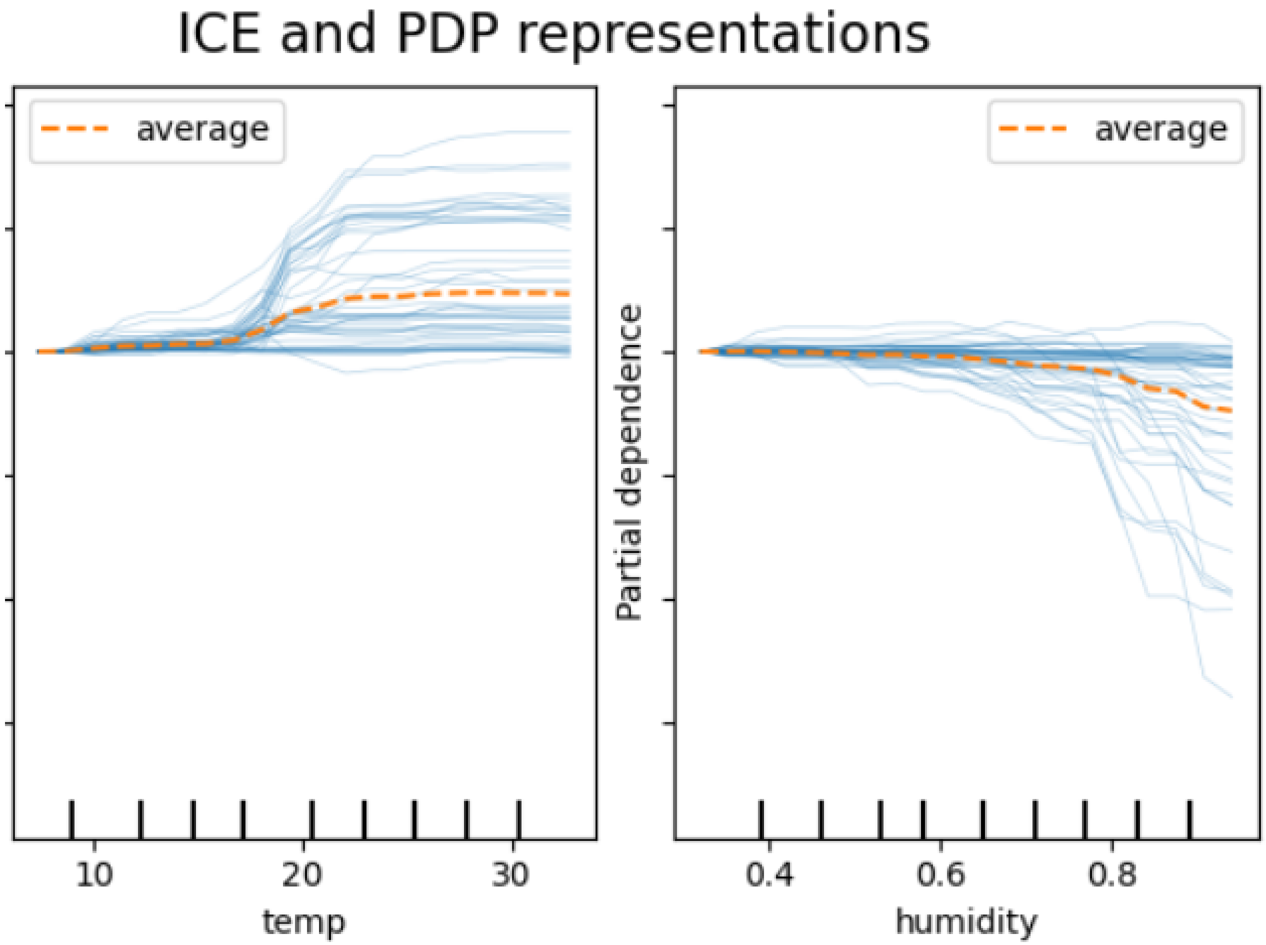
What is the effect of the variable (say N) on yield ?

With which factor N is interacting to affect yield ? (Important two-way interactions)

At farm level what is the effect of any production practices on rice yield ?

Effect of production practices on yield or (any outcome of interest)

Partial dependence plots (Single variable or two-way interactions)



$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}\left(x_S, X_C\right)\right]$$

By marginalizing the machine learning model output over the distribution of the features in set C, PDP shows the relationship between the features in set S we are interested in and the predicted outcome

Feature interaction

We will look at the exercise..

Thank You