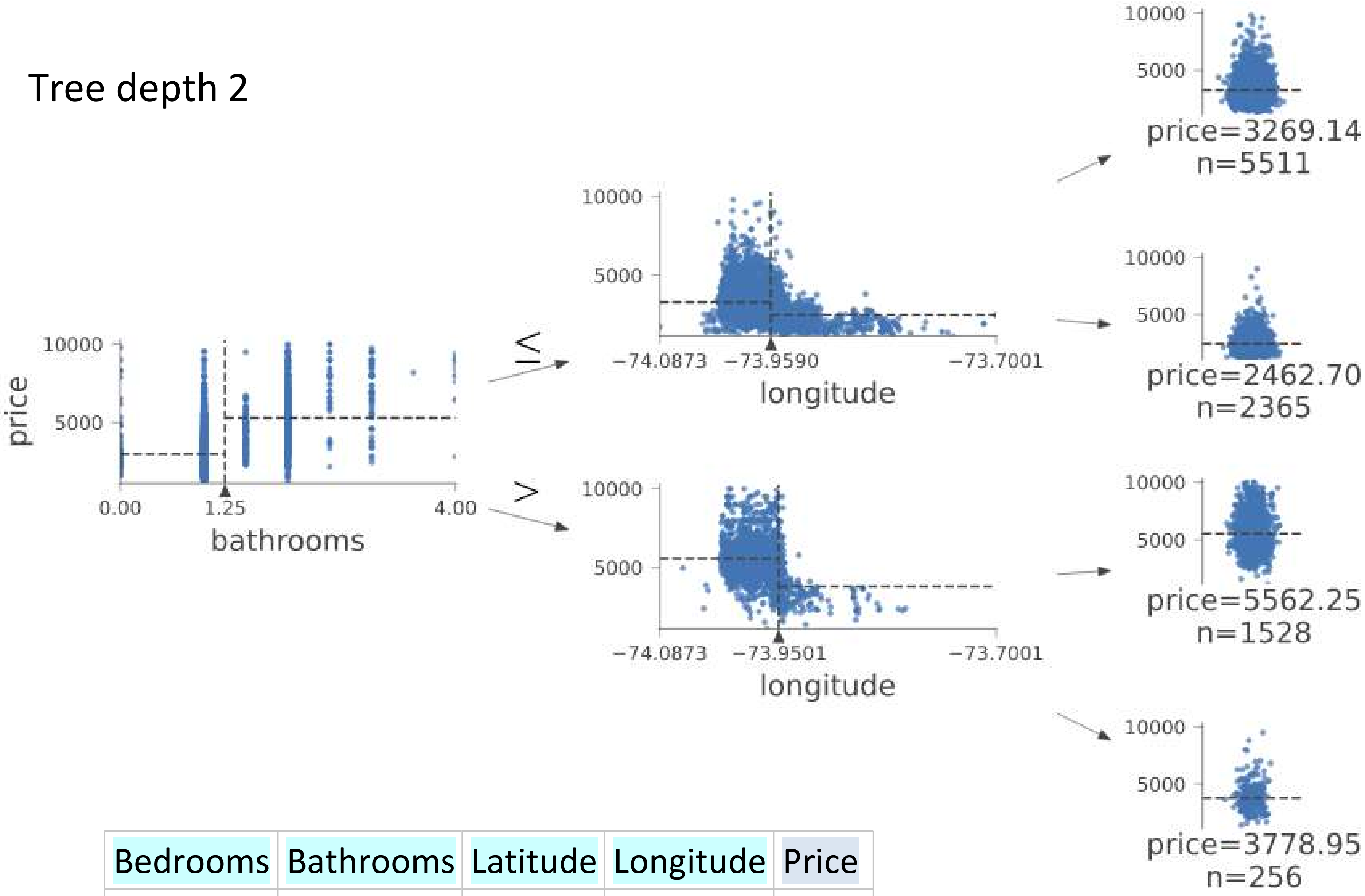
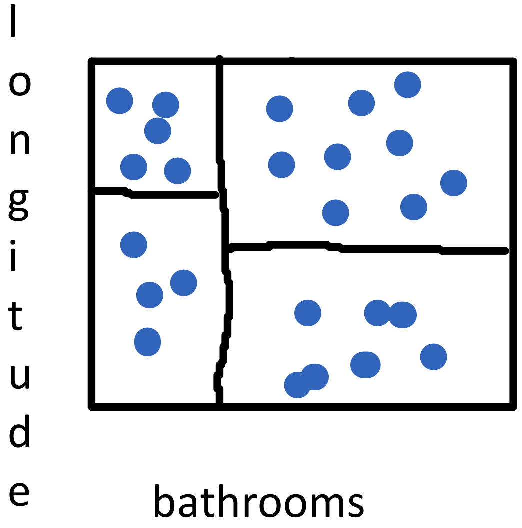
Regression - Rent prices of Appartments 1

Prof. Dr. Ingo Claßen

Donnerstag, 1. Juli 2021 13:58

Tree depth 2





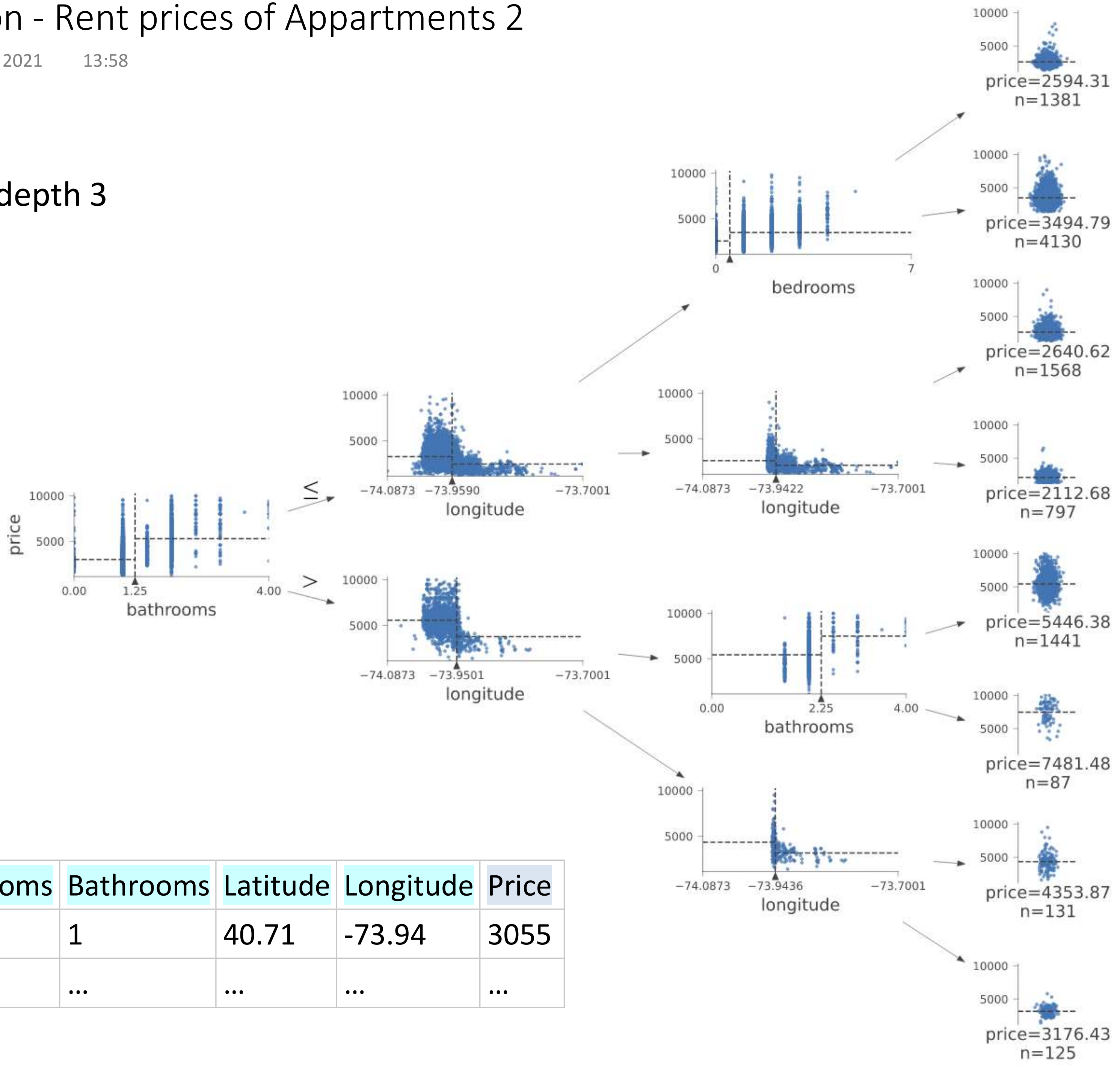
Bedrooms	Bathrooms	Latitude	Longitude	Price
1	1	40.71	-73.94	3055
…	…	…	…	…

Regression - Rent prices of Appartments 2

Donnerstag, 1. Juli 2021 13:58

Prof. Dr. Ingo Claßen

Tree depth 3



Bedrooms	Bathrooms	Latitude	Longitude	Price
1	1	40.71	-73.94	3055
…	…	…	…	…

Decomposition of Feature Space

Mittwoch, 30. Juni 2021 11:50

Feature space:

- 2 features: x1, x2: plane
- 3 features: x1, x2, x3: cube
- n features: x1, …, xn: n-dimensional cube

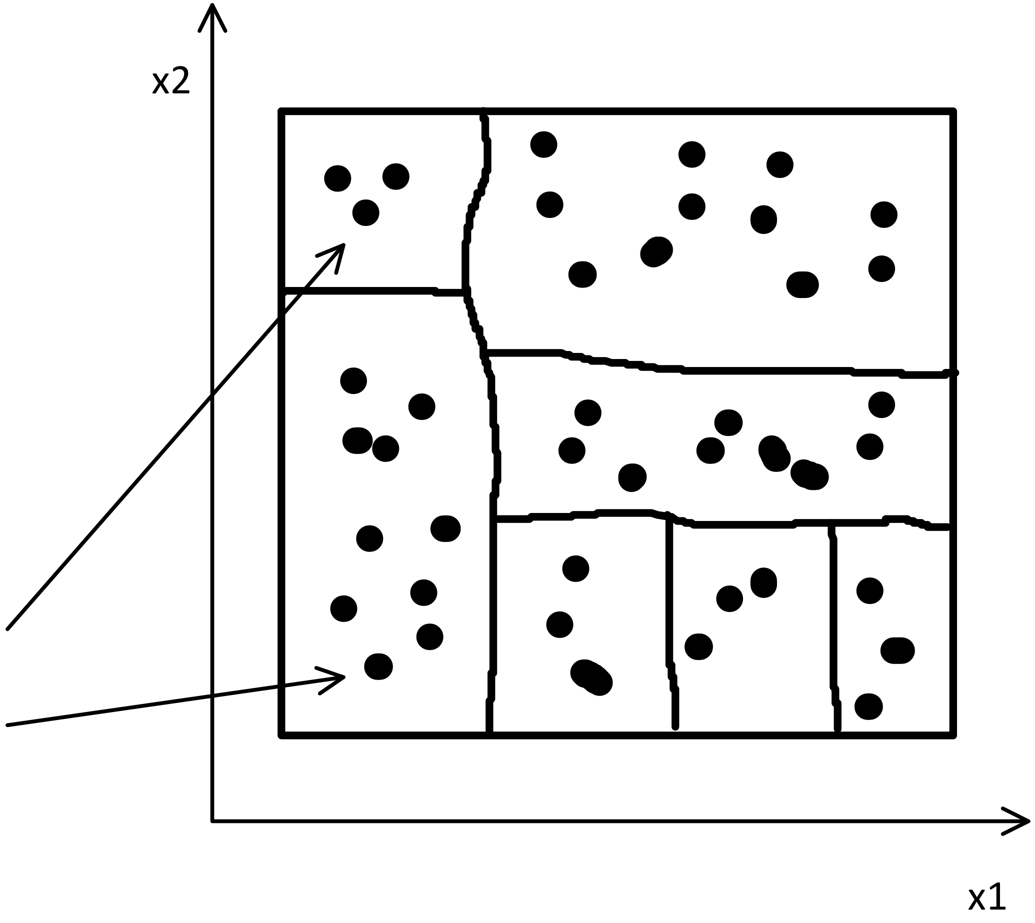
Simplifcation for presentation purposes

- Only two features: x1, x2
- Regions are rectangles

Find rectangular non-overlapping decomposition of space

- Leads to n-dimensional sub-cubes (regions)
- Each region should be as "uniform" as possible
- Average value of all records in region (regression)
- Majority vote in region (classification)

Instances



Finding an optimal decomposition automatically is infeasable (combinatorical explosion)

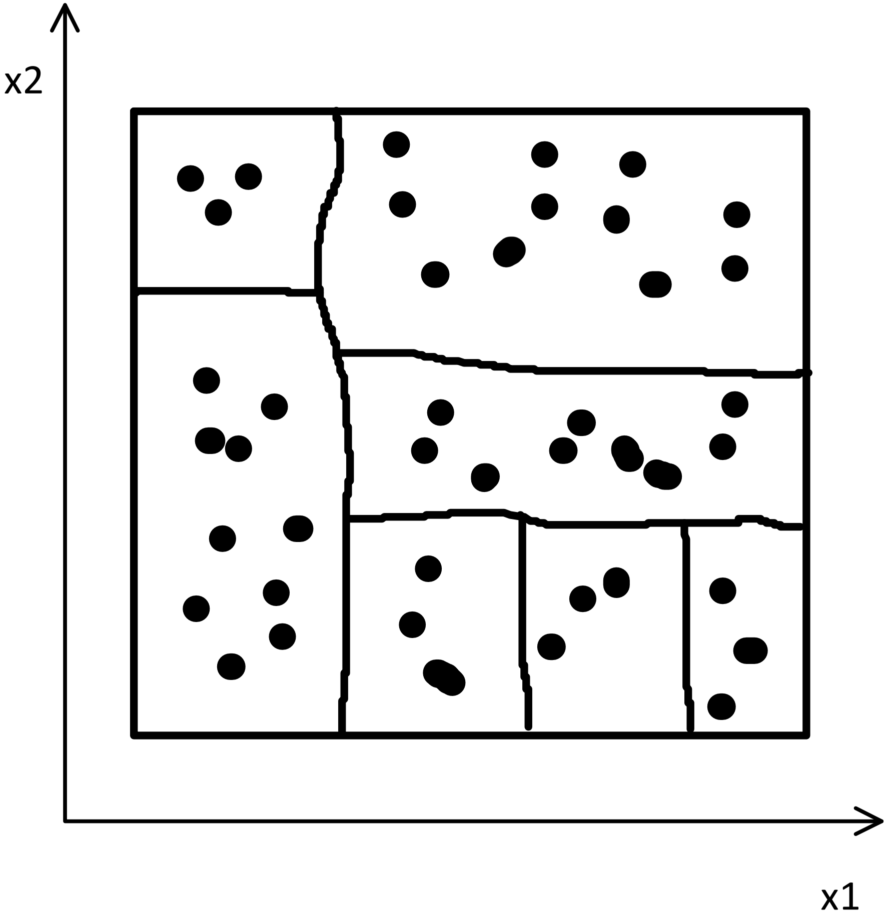
Concept of Decision Trees

Prof. Dr. Ingo Claßen

Mittwoch, 30. Juni 2021 11:50

General idea of decision trees:

- Gready decomposition of feature space by recursive binary splitting
- Stop splitting according to criterion, e.g.
  - minimal numbers of instances in region
  - max depth of tree
  - minimal performance gain



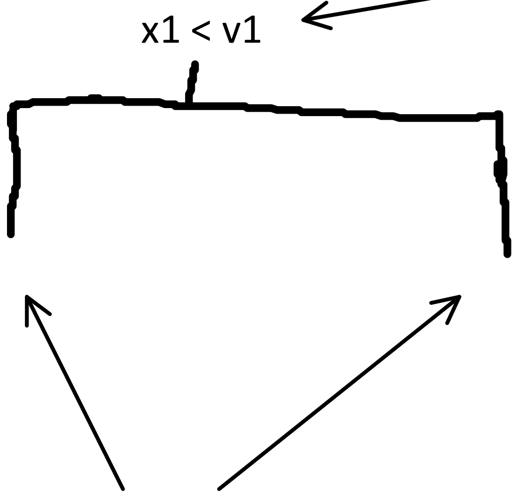
Finding splits

- Search through all features
- For each feature: consider all split values
- Take "best" (feature, splitval) combination

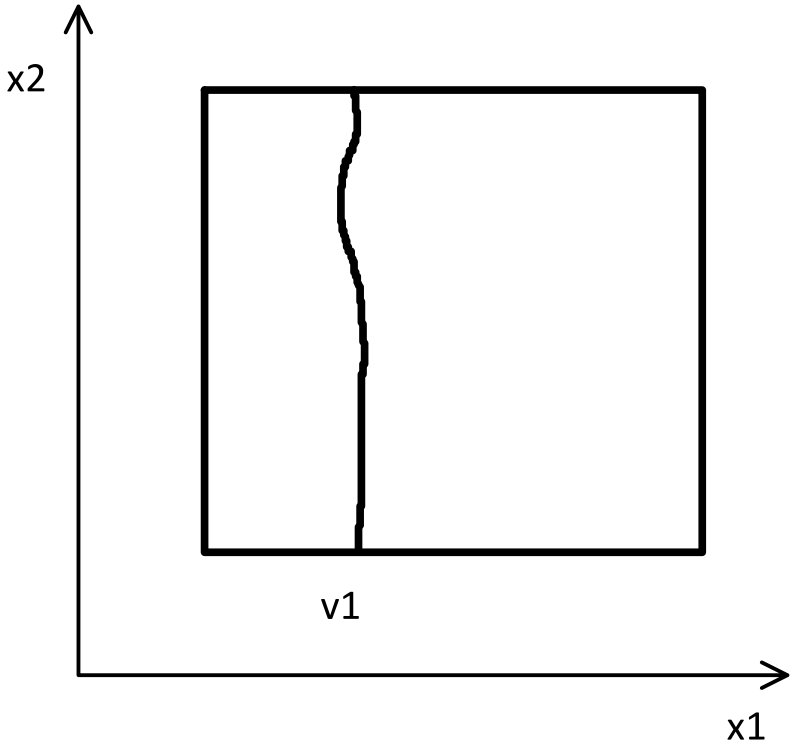
Recursive Binary Splitting 1

Mittwoch, 30. Juni 2021 14:39

split feature	split value
x1	v1



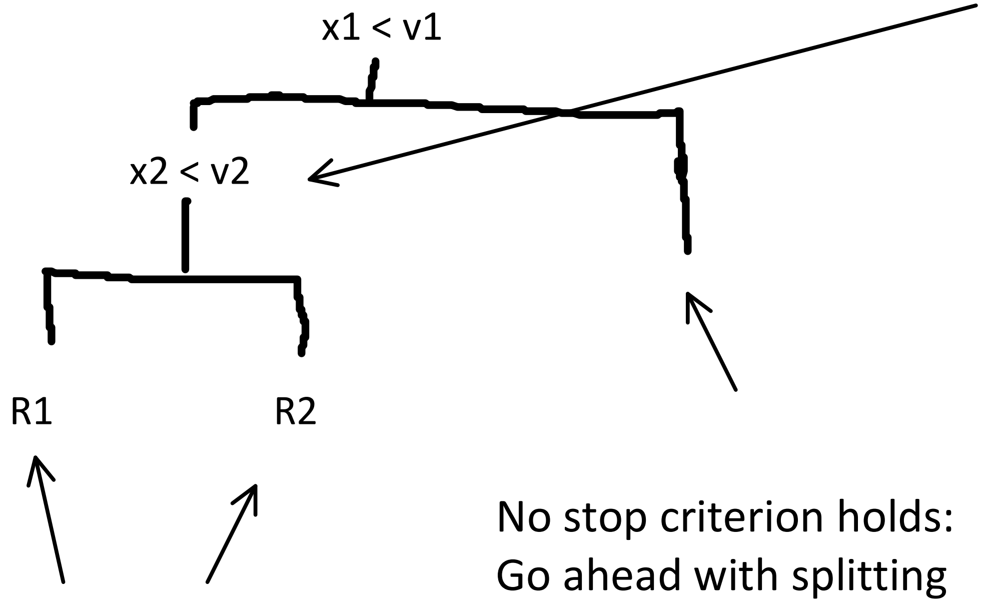
No stop criterion holds:
Go ahead with splitting



Recursive Binary Splitting 2

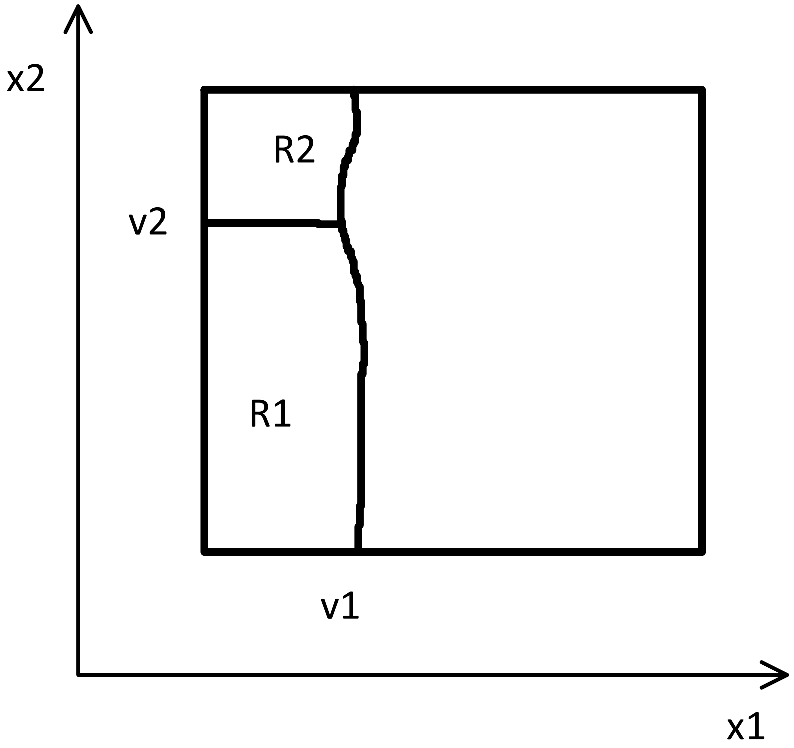
Mittwoch, 30. Juni 2021 14:39

split feature	split value
x2	v2



No stop criterion holds:
Go ahead with splitting

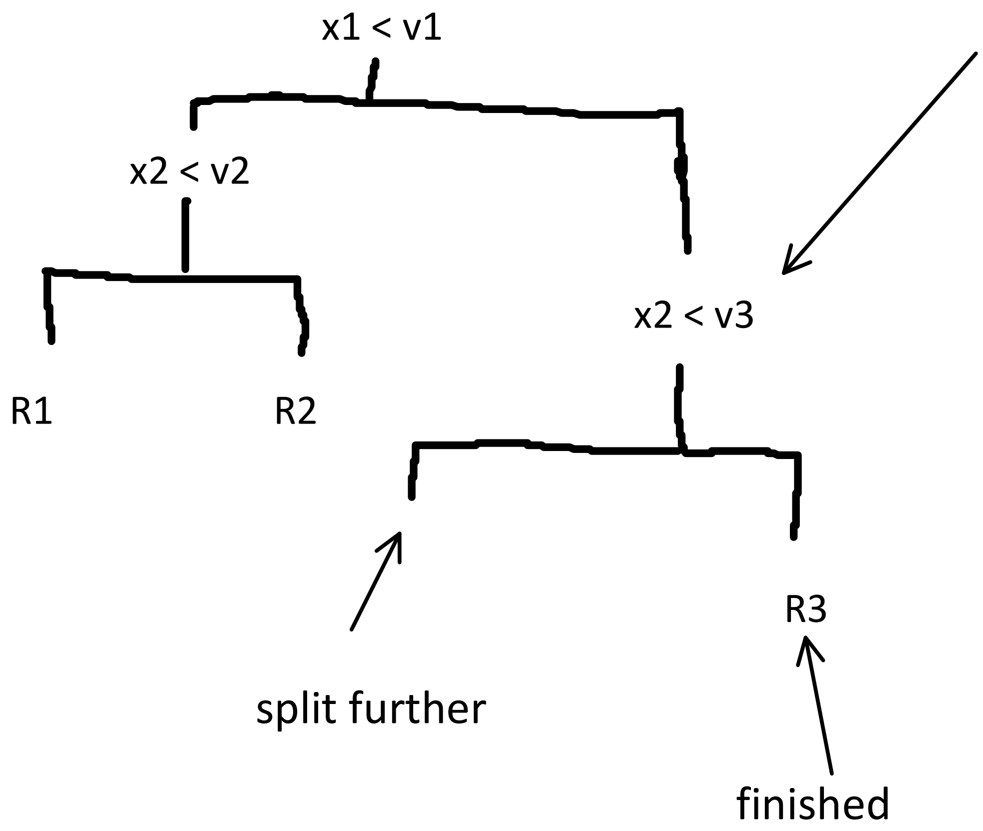
Stop criterion holds:
do not further split

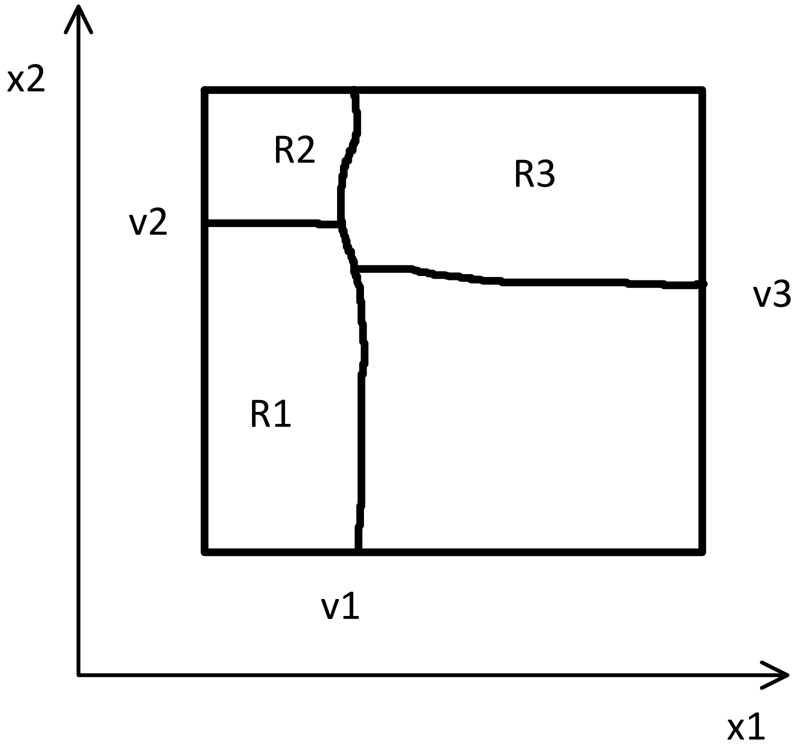


Recursive Binary Splitting 3

Mittwoch, 30. Juni 2021 14:39

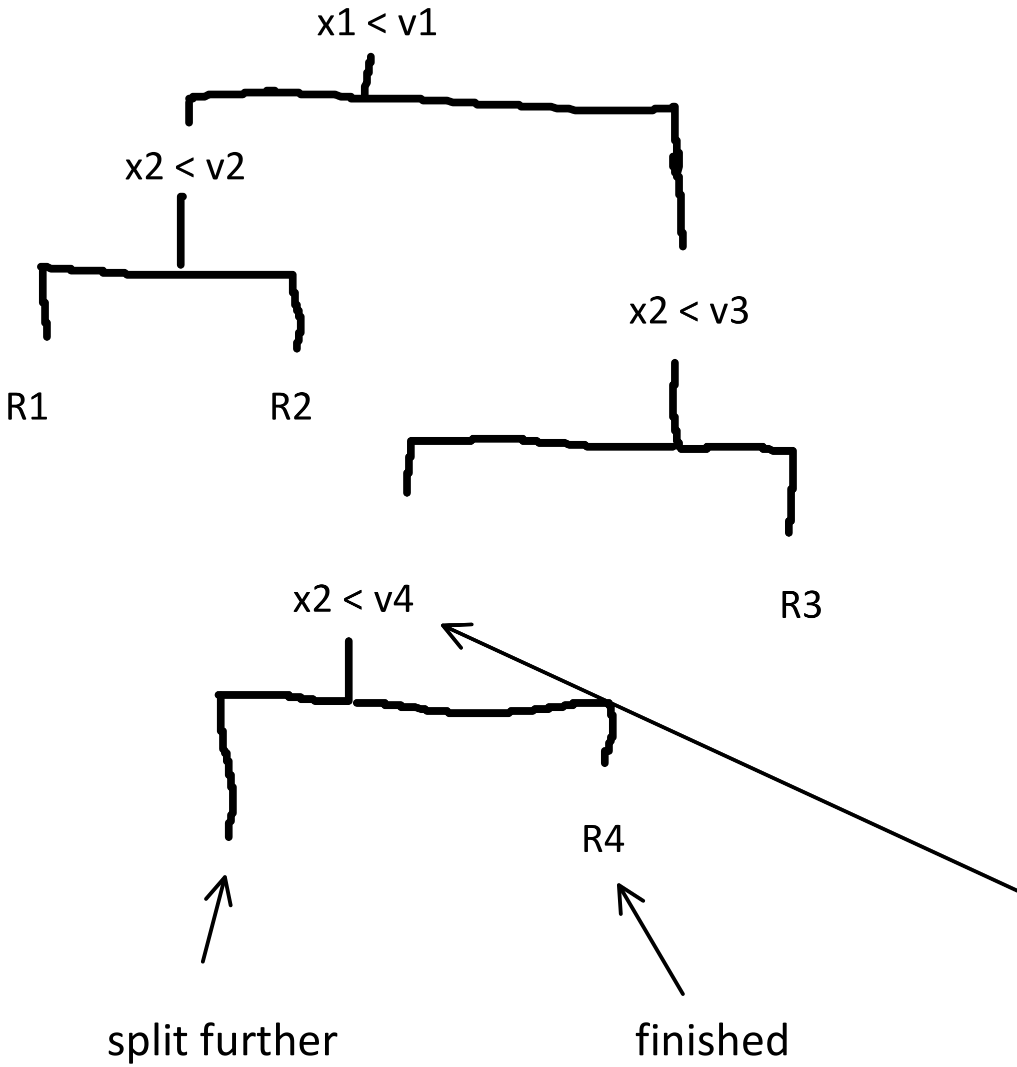
split feature	split value
x2	v3

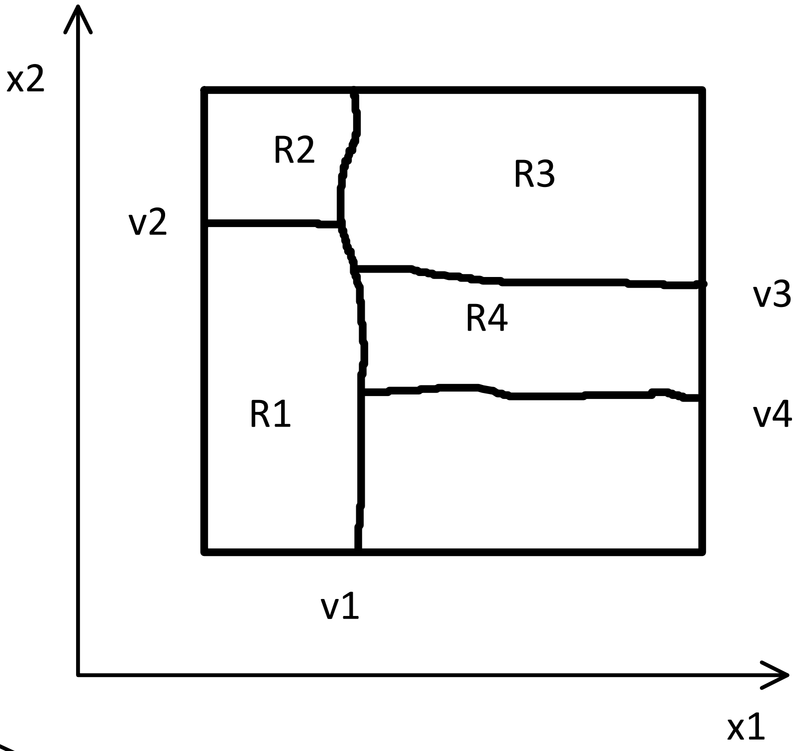




Recursive Binary Splitting 4

Mittwoch, 30. Juni 2021 14:39



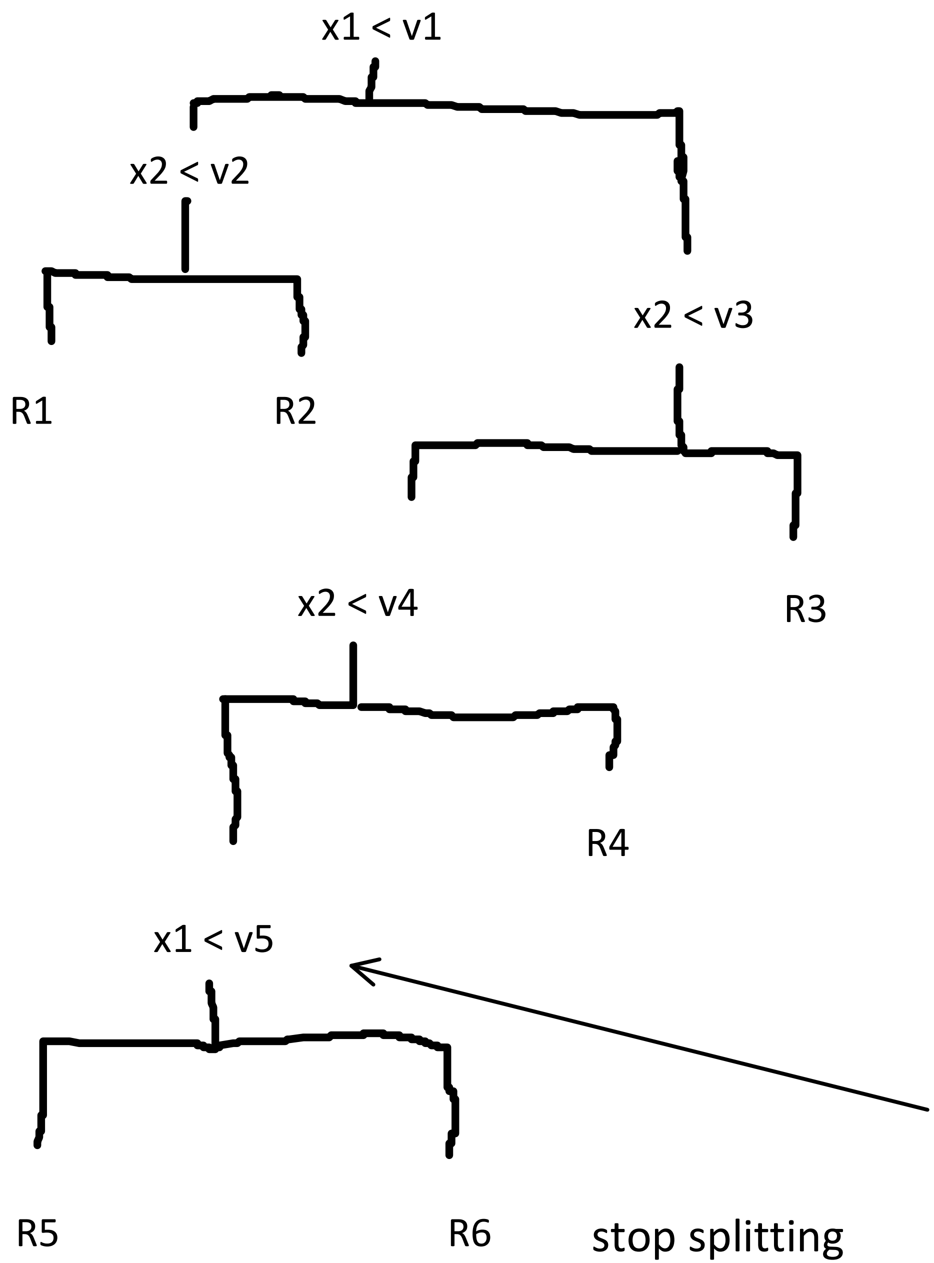


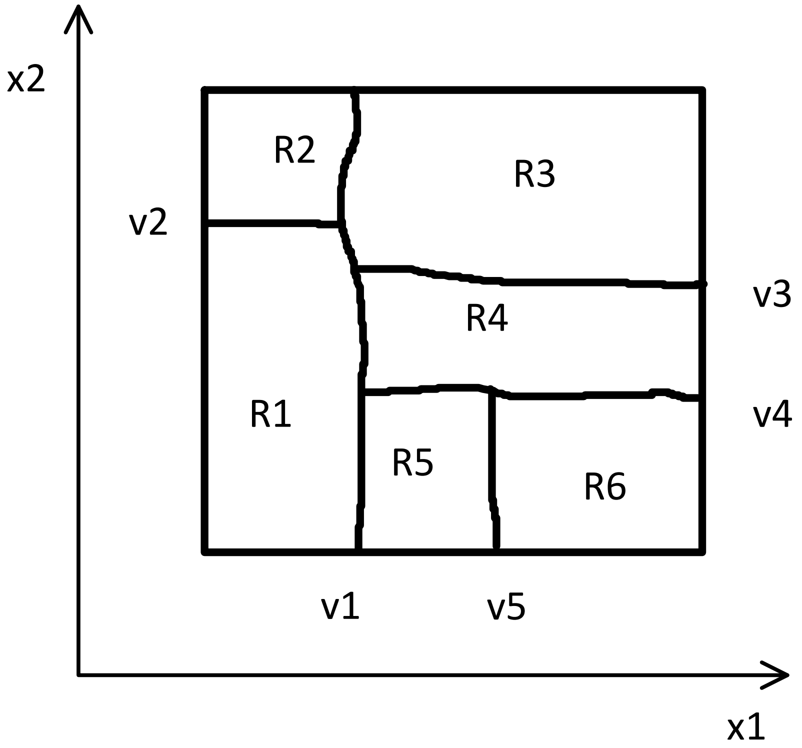
split feature	split value
x2	v4

Recursive Binary Splitting 5

Mittwoch, 30. Juni 2021 14:39

Prof. Dr. Ingo Claßen





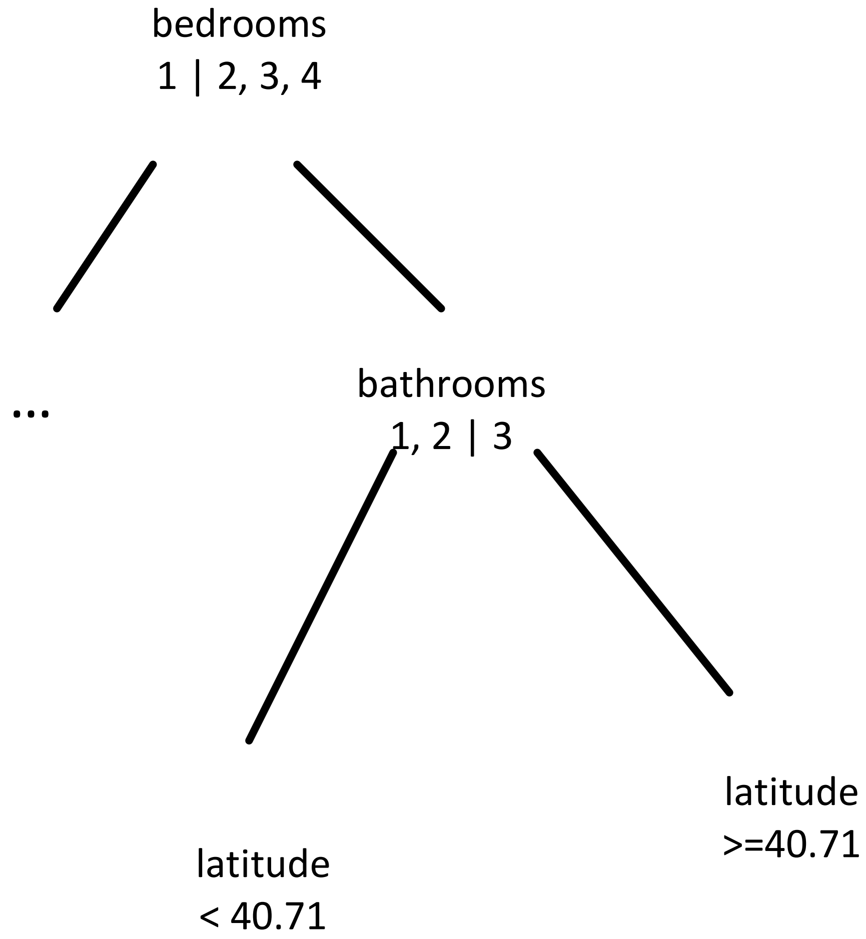
split feature	**split value**
x1	v5

Example for appartment rents

Prof. Dr. Ingo Claßen

Freitag, 18. Juni 2021 10:32

Split on best combination of feature and split point



bedrooms
1 | 2, 3, 4

...

bathrooms
1, 2 | 3

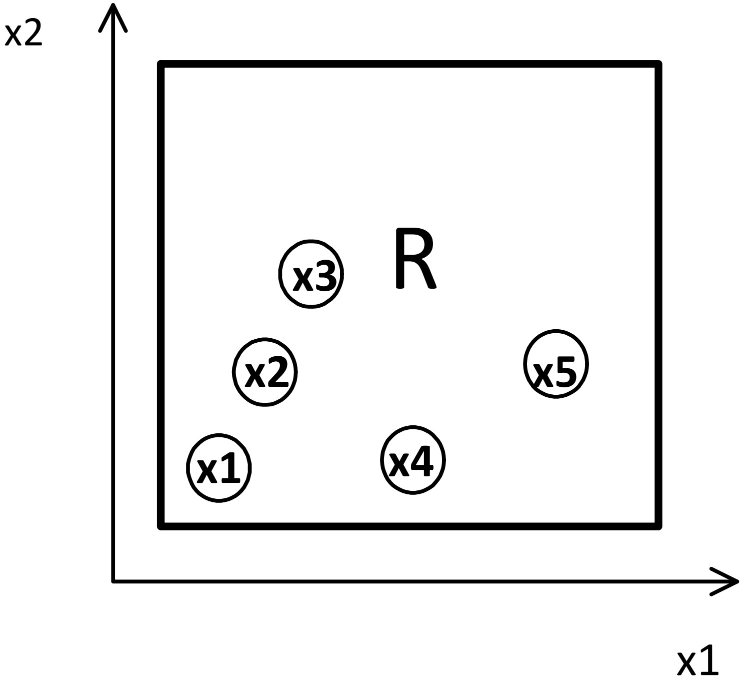
latitude
< 40.71

latitude
>=40.71

MSE Value of Region

Mittwoch, 30. Juni 2021 15:16

$$\text{MSE}(R) = \frac{1}{\text{count}(y_i)} \sum_{\mathbf{x}_i \in R} (y_i - \text{avg}(y_i))^2$$



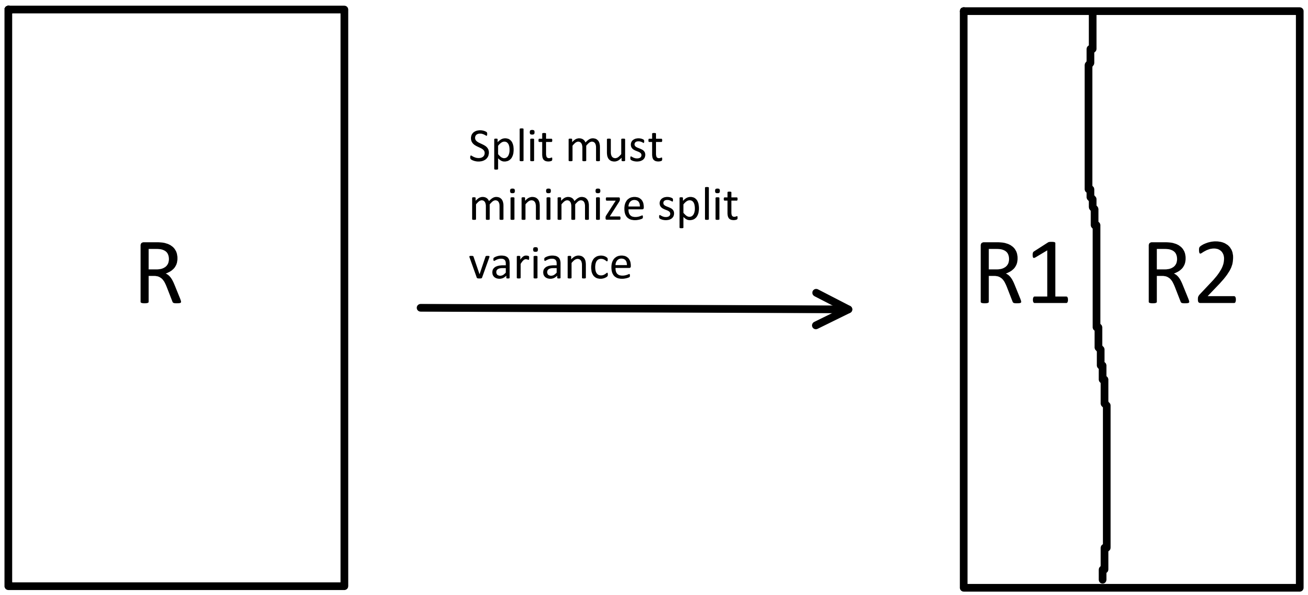
	x1	x2	y
x1	1	1	20
x2	1	2	30
x3	2	3	20
x4	3	1	10
x5	4	2	20

$$\text{avg}(y_i) = 20$$

$$\begin{aligned}\text{MSE}(R) &= \frac{1}{5}\left((20-20)^2 + (30-20)^2 + (20-20)^2 + (10-20)^2 + (20-20)^2\right) \\ &= \frac{1}{5}\left(0 + 100 + 0 + 100 + 0\right) \\ &= \frac{1}{5} * 200 \\ &= 40\end{aligned}$$

Split Criterion Regression

Mittwoch, 30. Juni 2021 18:11



Split must minimize split variance

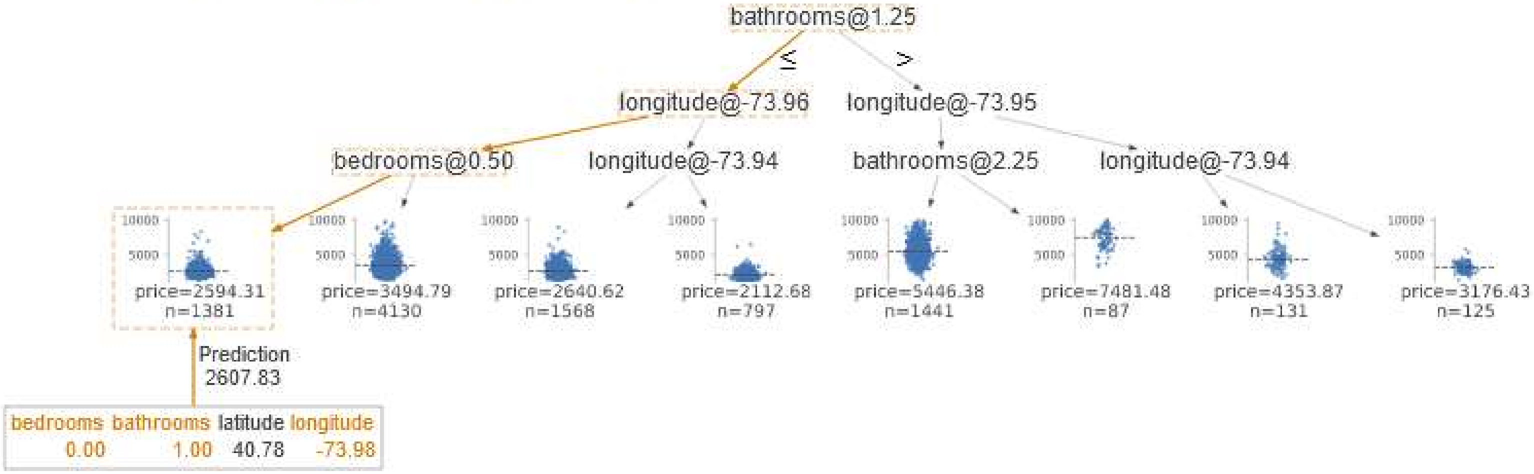
$$\text{SplitVariance} = \frac{1}{\text{count}(R_1)}\text{MSE}(R_1) + \frac{1}{\text{count}(R_2)}\text{MSE}(R_2)$$

Analysis of Single Instance 1

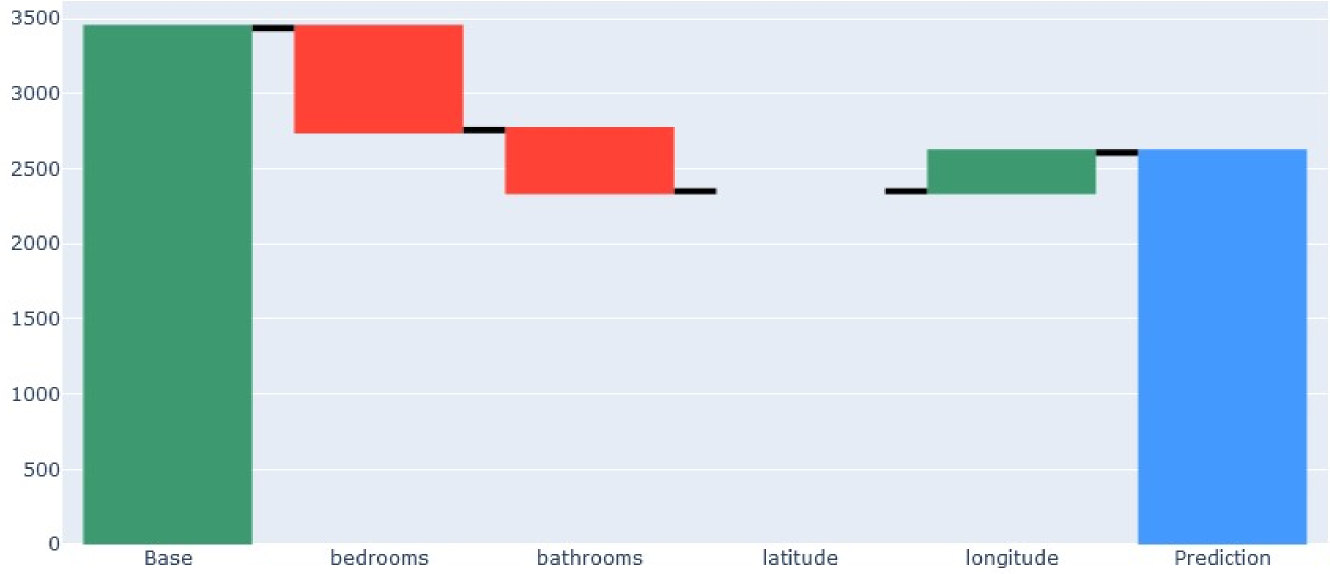
Donnerstag, 1. Juli 2021 15:33

```
       bedrooms  bathrooms  latitude  longitude  price
19791         0        1.0   40.7757    -73.981   2190
```

```
            Contributions
Base          3435.888690
bedrooms      -677.945519
bathrooms     -407.762466
latitude         0.000000
longitude      257.648895
Prediction    2607.829600
```

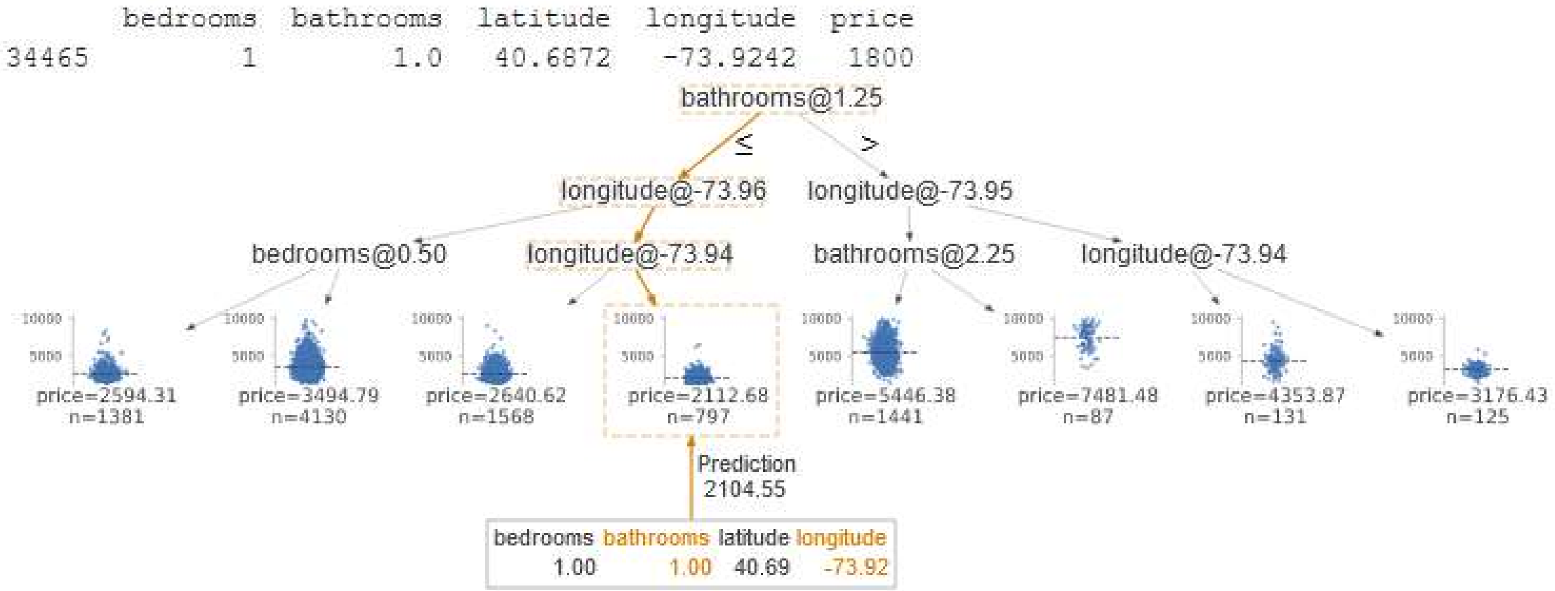


Prediction : 2607.8296004300305

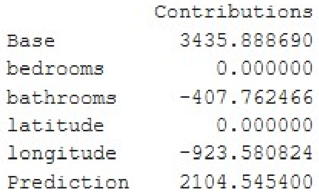


Analysis of Single Instance 2

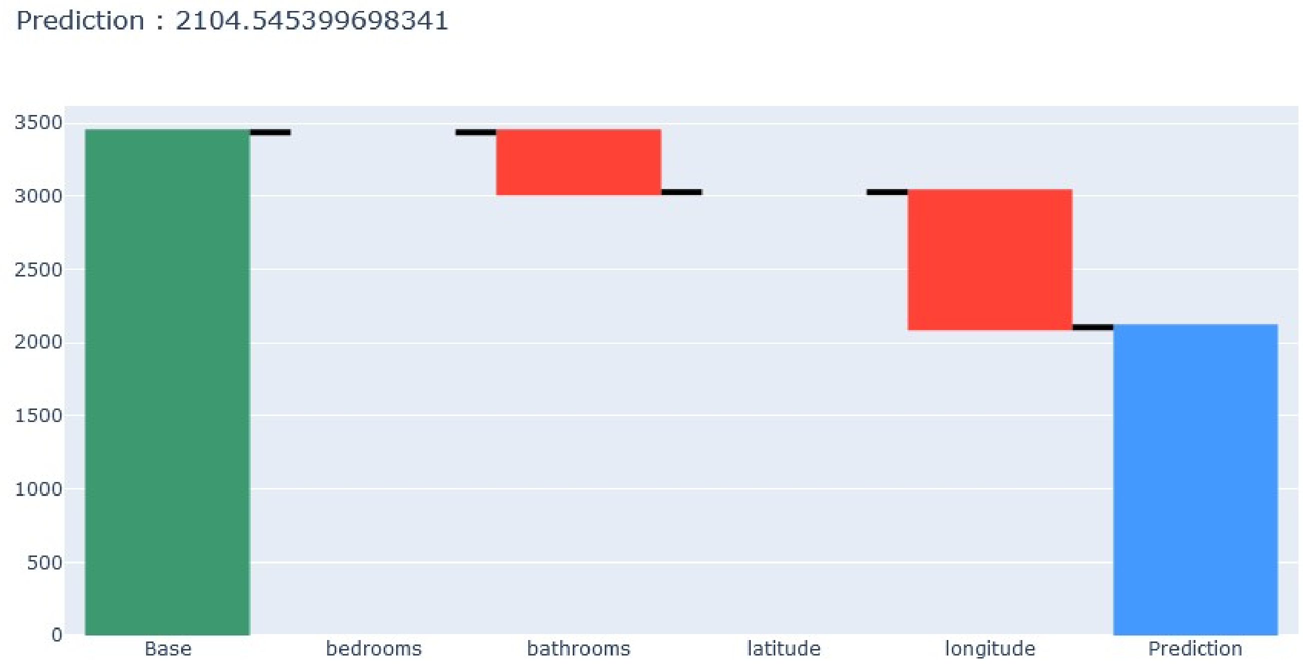
Donnerstag, 1. Juli 2021 17:45



```
       bedrooms  bathrooms  latitude  longitude  price
34465         1        1.0   40.6872   -73.9242   1800
```



```
            Contributions
Base          3435.888690
bedrooms         0.000000
bathrooms     -407.762466
latitude         0.000000
longitude     -923.580824
Prediction    2104.545400
```

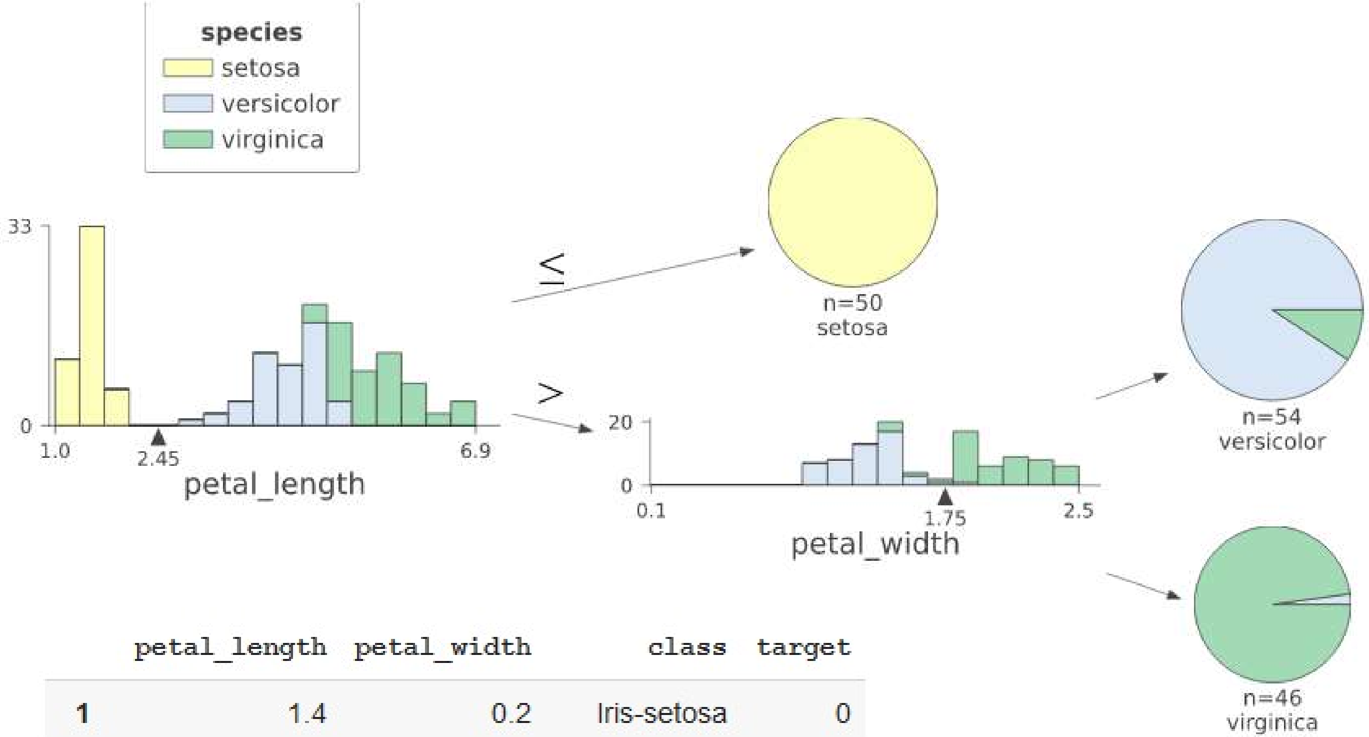


Prediction : 2104.545399698341

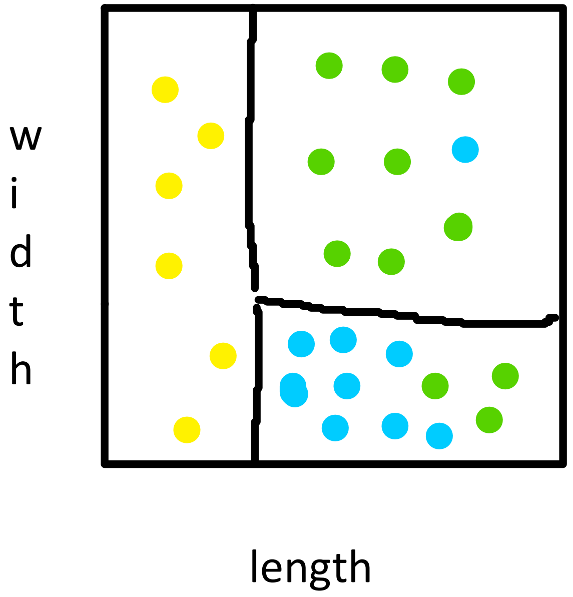
Classification - Iris Flowers 1

Donnerstag, 1. Juli 2021 14:16

Tree depth 2



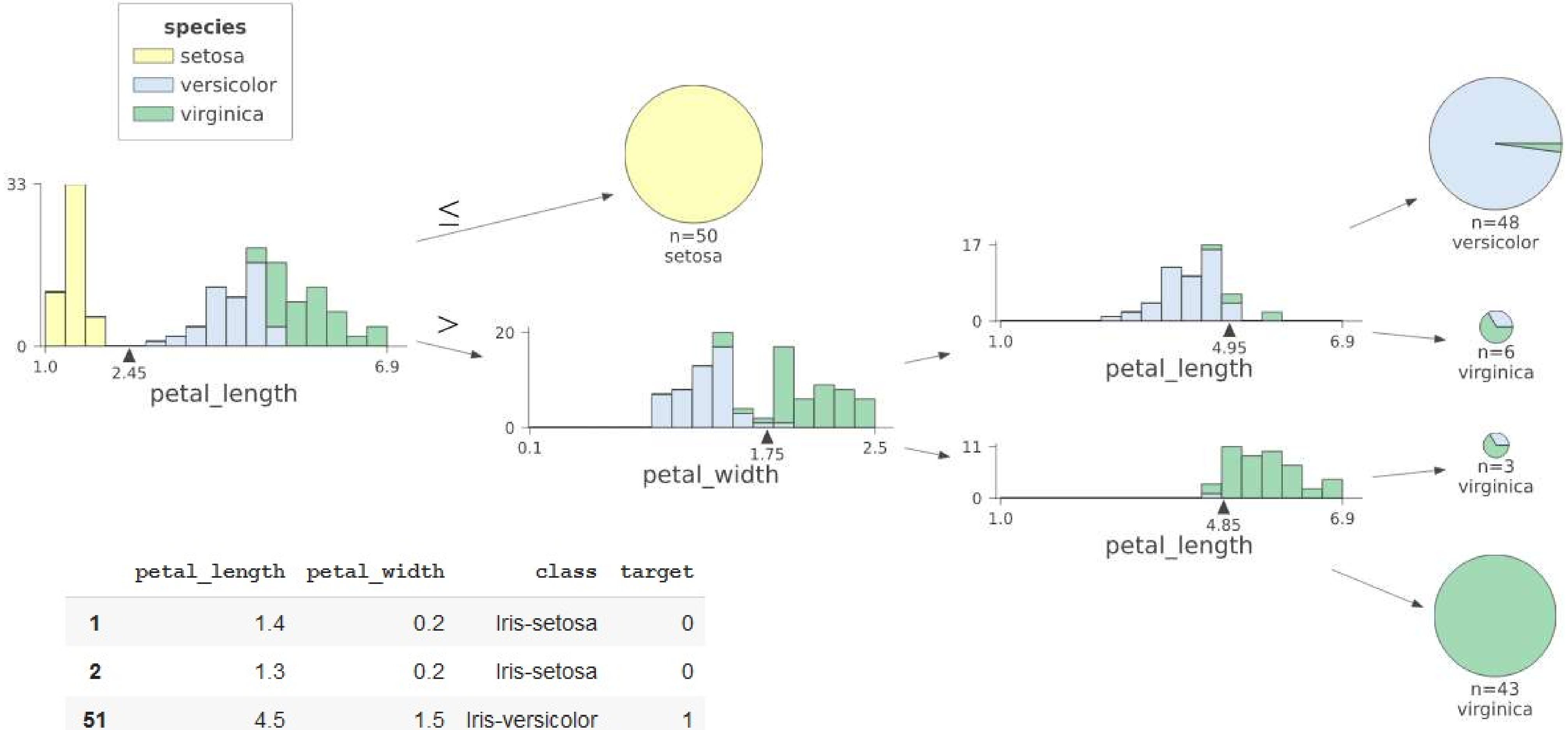
	petal_length	petal_width	class	target
1	1.4	0.2	Iris-setosa	0
2	1.3	0.2	Iris-setosa	0
51	4.5	1.5	Iris-versicolor	1
52	4.9	1.5	Iris-versicolor	1
101	5.1	1.9	Iris-virginica	2
102	5.9	2.1	Iris-virginica	2



Classification - Iris Flowers 2

Donnerstag, 1. Juli 2021 14:16

Tree depth 3



	petal_length	petal_width	class	target
1	1.4	0.2	Iris-setosa	0
2	1.3	0.2	Iris-setosa	0
51	4.5	1.5	Iris-versicolor	1
52	4.9	1.5	Iris-versicolor	1
101	5.1	1.9	Iris-virginica	2
102	5.9	2.1	Iris-virginica	2

Gini Value of Region

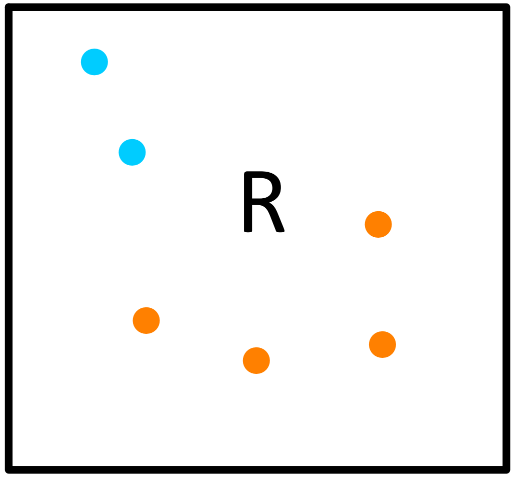
Mittwoch, 30. Juni 2021 15:17

$$\mathrm{Gini}(R) = 1 - \sum_{c \in C} \mathrm{Prob}(c, R)^2$$

R	Region
Prob	Probability
Gini	Gini Value
C	Set of all classes
c	Single class

Class tags:

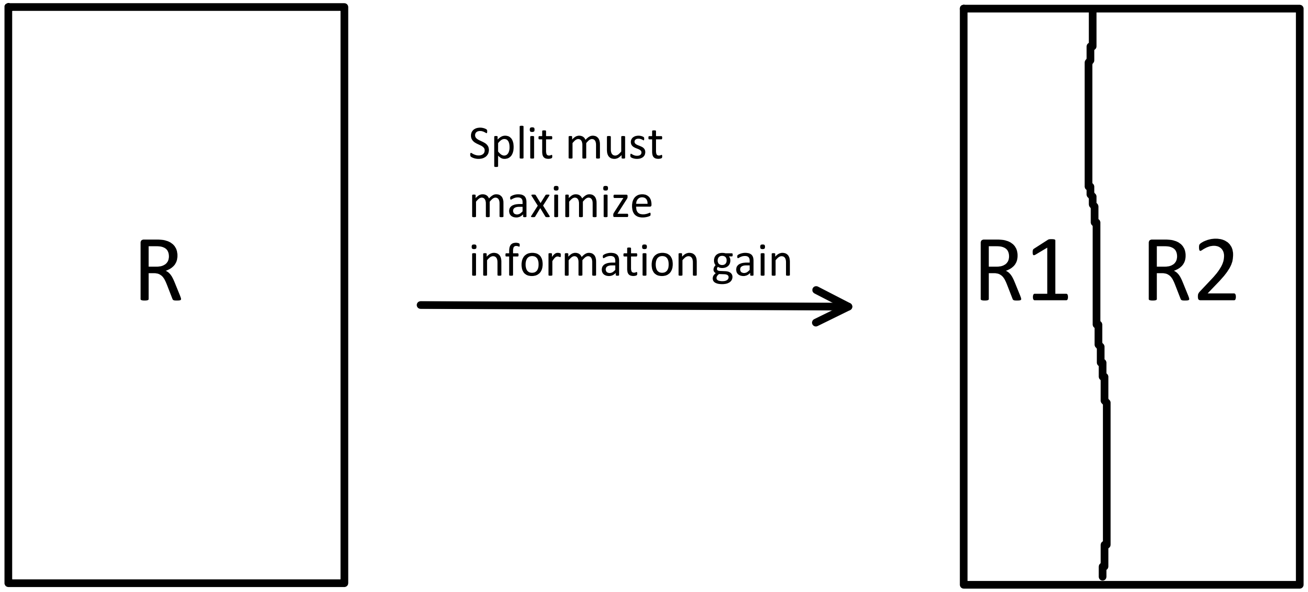
- Class1: orange
- Class2: blue



$$\begin{aligned}\mathrm{Gini}(R) &= 1 - \left(\left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2\right) \\ &= 1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right) \\ &= 1 - \left(\frac{1}{9} + \frac{4}{9}\right) \\ &= 1 - \frac{5}{9} \\ &= \frac{4}{9}\end{aligned}$$

Split Criterion Classification

Mittwoch, 30. Juni 2021 18:11



R

Split must maximize information gain

R1

R2

$$\text{InformationGain} = \text{Gini}(R) - \left(\frac{1}{\text{count}(R_1)}\text{Gini}(R_1) + \frac{1}{\text{count}(R_2)}\text{Gini}(R_2)\right)$$

Example Information Gain

Mittwoch, 30. Juni 2021 18:30

See:
https://towardsdatascience.com/decision-tree-an-algorithm-that-works-like-the-human-brain-8bc0652f1fc6

Tree Ensembles - Random Forest

Prof. Dr. Ingo Claßen

Freitag, 18. Juni 2021 15:22

Problems of trees

- Too specific
- Overfit on trainings data

Assume leafs only contain one record

- Trainings error would get zero

Train many tree models

- Randomly select subset of training data (with replacement, bootstrapping)
- Randomly select subset of features

Prediction - Regression

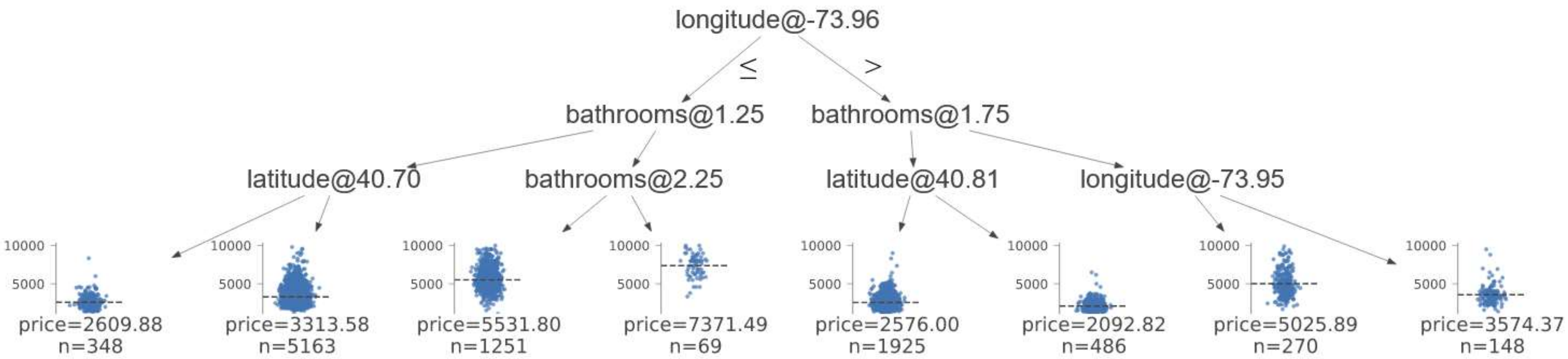
- Ask all trees for prediction
- Average results
- Two levels of averaging
  - Leaf level of each singular tree
  - Average of all trees

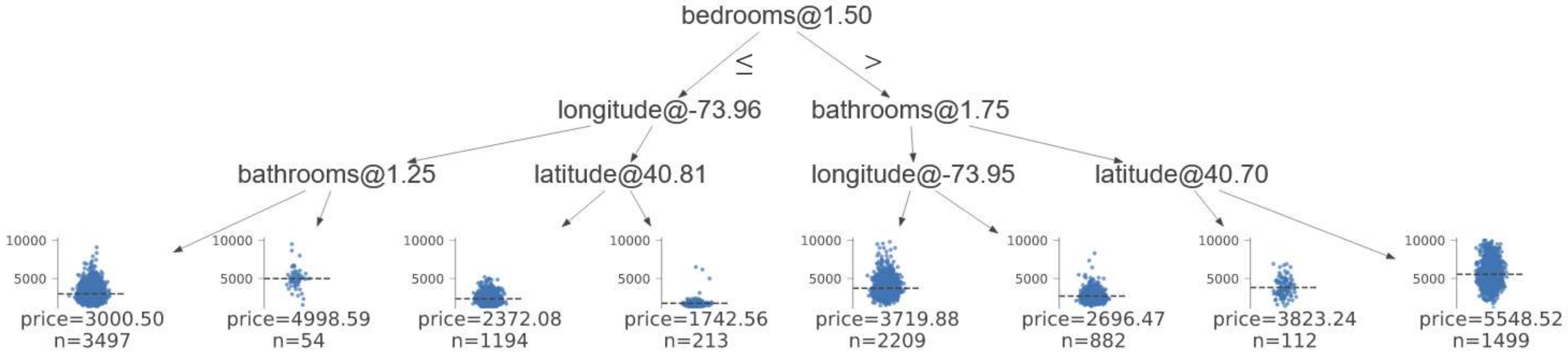
Prediction - Classification

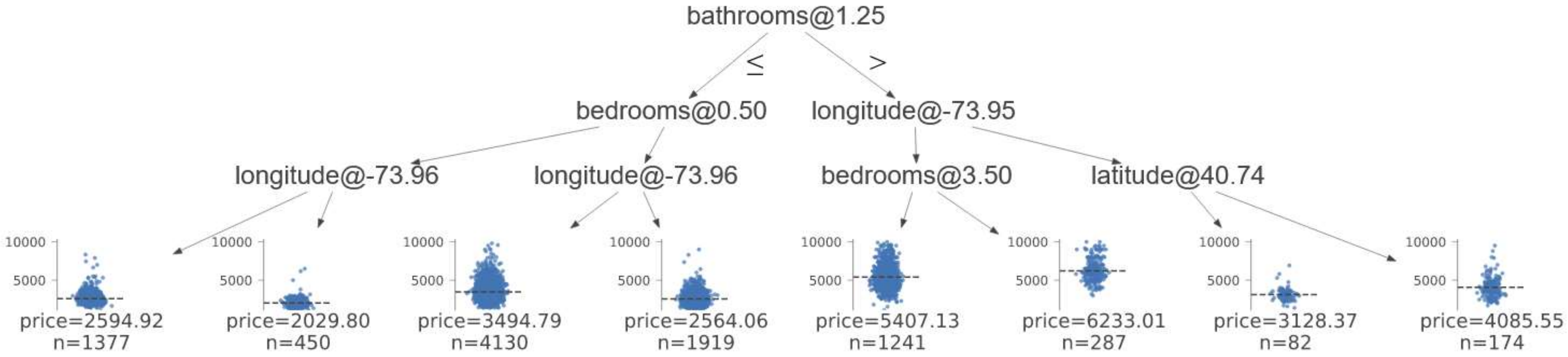
- Ask all trees for prediction
- Majority vote
- Two levels of voting
  - Leaf level of each singular tree
  - Combining votes of all trees

Selected Trees from the Forest

Donnerstag, 1. Juli 2021 15:31





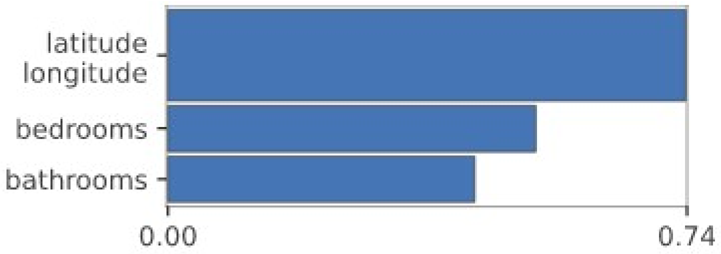


Feature Importance

Prof. Dr. Ingo Claßen

Donnerstag, 1. Juli 2021 14:34

Permutation Importance


latitude
longitude
bedrooms
bathrooms