



MEDNARODNA
PODIPLOMSKA ŠOLA
JOŽEFA ŠTEFANA

INFORMATION AND COMMUNICATION TECHNOLOGIES
Master study programme

Data and Text Mining

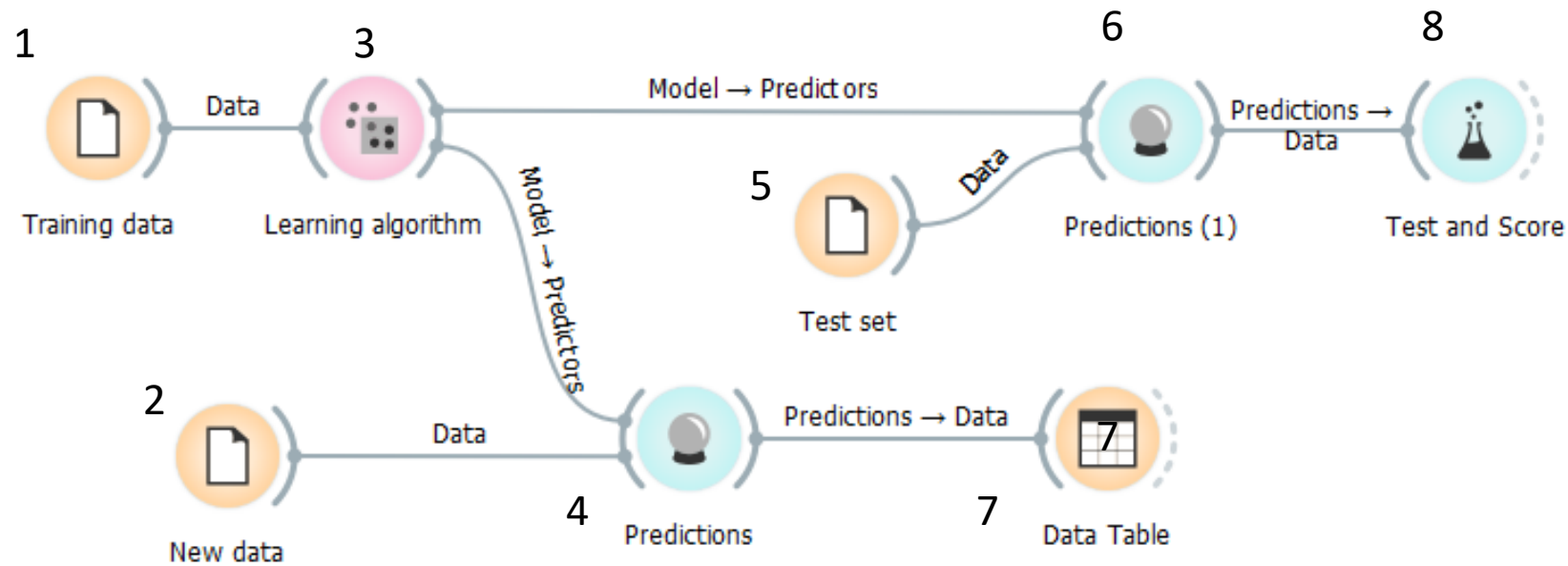
Petra Kralj Novak

November 6, 2019

http://kt.ijs.si/petra_kralj/dmkd.html

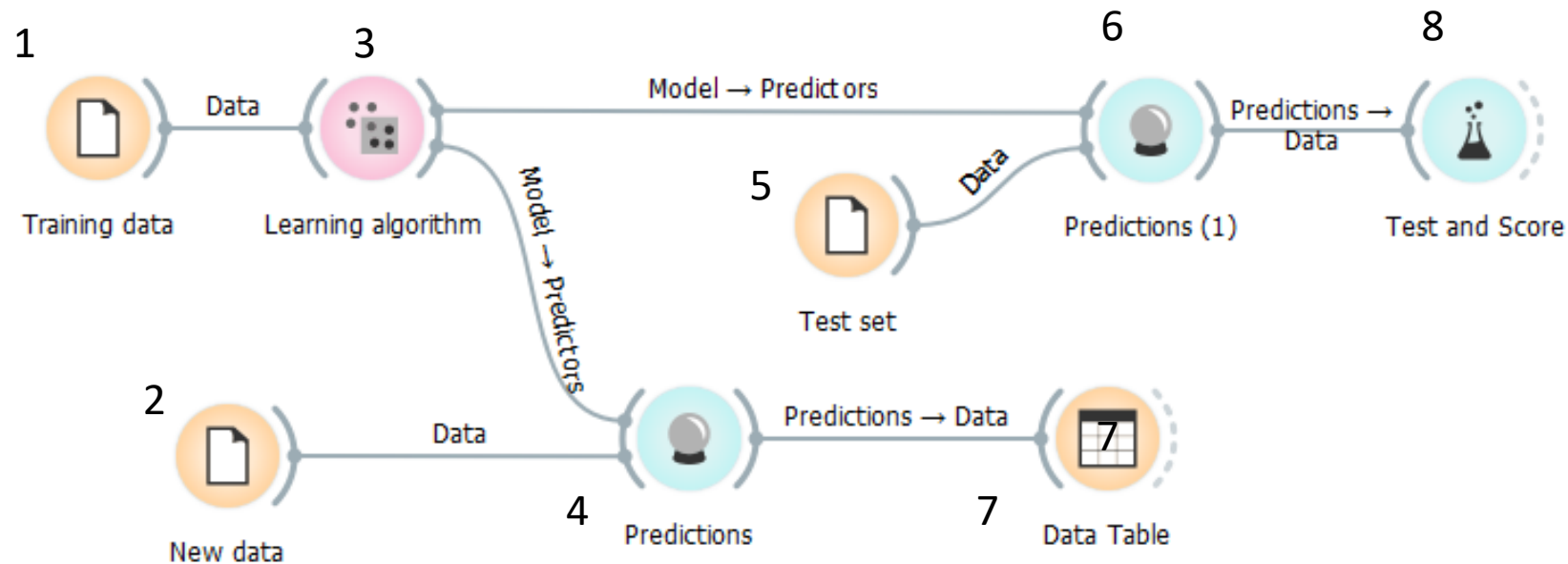
Classification

1. Train the model on train data
2. Test the model on test data
3. Classify new data with the model



Classification

1. Train the model on train data: 1,3
2. Test the model on test data: 5,6,8
3. Classify new data with the model: 2,4,7

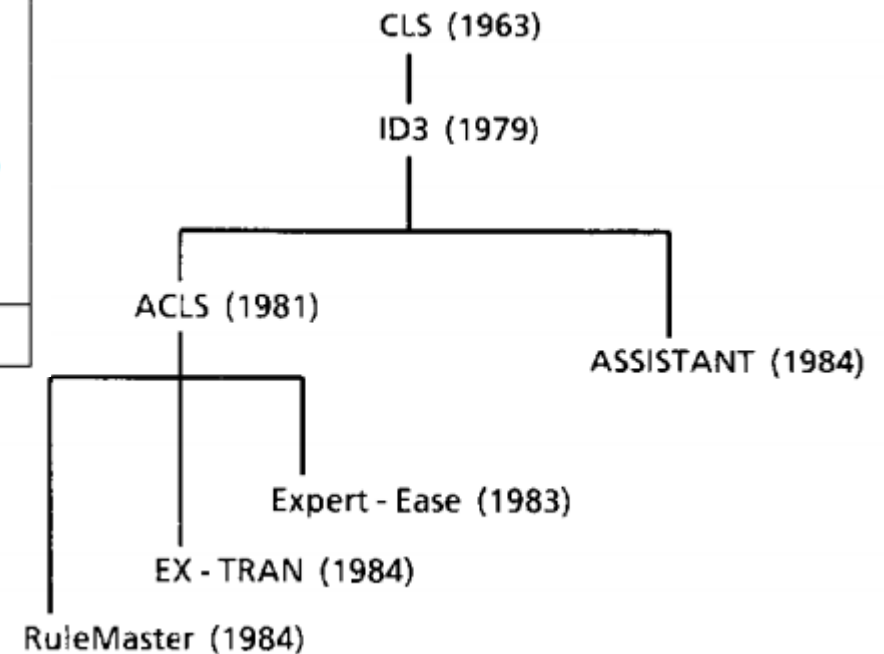


The TDIDT family of learning systems

TDIDT: BASIC ALGORITHM

IF all the instances in the training set belong to the same class
THEN return the value of the class
ELSE (a) Select an attribute A to split on⁺
(b) Sort the instances in the training set into subsets, one
for each value of attribute A
(c) Return a tree with one branch for each *non-empty* subset,
each branch having a descendant subtree or a class
value produced by applying the algorithm recursively

⁺ Never select an attribute twice in the same branch



Decision tree induction with ID3

Induce a decision tree on set S :

1. Compute the **entropy** $E(S)$ of the set S
2. **IF** $E(S) = 0$
3. The current set is “clean” and therefore a leaf in our tree
4. **IF** $E(S) > 0$
5. Compute the **information gain** of each attribute $\text{Gain}(S, A)$
6. The attribute A with the highest information gain becomes the root
7. Divide the set S into subsets S_i according to the values of A
8. Repeat steps 1-7 on each S_i


Exercise: Train and test a decision tree (ID3)

Person	Age	Prescription	Astigmatic	Tear_Rate	Lenses
P1	young	myope	no	normal	YES
P2	young	myope	no	reduced	NO
P3	young	hypermetrope	no	normal	YES
P4	young	hypermetrope	no	reduced	NO
P5	young	myope	yes	normal	YES
P6	young	myope	yes	reduced	NO
P7	young	hypermetrope	yes	normal	YES
P8	young	hypermetrope	yes	reduced	NO
P9	pre-presbyopic	myope	no	normal	YES
P10	pre-presbyopic	myope	no	reduced	NO
P11	pre-presbyopic	hypermetrope	no	normal	YES
P12	pre-presbyopic	hypermetrope	no	reduced	NO
P13	pre-presbyopic	myope	yes	normal	YES
P14	pre-presbyopic	myope	yes	reduced	NO
P15	pre-presbyopic	hypermetrope	yes	normal	NO
P16	pre-presbyopic	hypermetrope	yes	reduced	NO
P17	presbyopic	myope	no	normal	NO
P18	presbyopic	myope	no	reduced	NO
P19	presbyopic	hypermetrope	no	normal	YES
P20	presbyopic	hypermetrope	no	reduced	NO
P21	presbyopic	myope	yes	normal	YES
P22	presbyopic	myope	yes	reduced	NO
P23	presbyopic	hypermetrope	yes	normal	NO
P24	presbyopic	hypermetrope	yes	reduced	NO

Split the dataset into a training and a test set

Person	Age	Prescription	Astigmatic	Tear_Rate	Lenses
P1	young	myope	no	normal	YES
P2	young	myope	no	reduced	NO
P3	young	hypermetrope	no	normal	YES
P4	young	hypermetrope	no	reduced	NO
P5	young	myope	yes	normal	YES
P6	young	myope	yes	reduced	NO
P7	young	hypermetrope	yes	normal	YES
P8	young	hypermetrope	yes	reduced	NO
P9	pre-presbyopic	myope	no	normal	YES
P10	pre-presbyopic	myope	no	reduced	NO
P11	pre-presbyopic	hypermetrope	no	normal	YES
P12	pre-presbyopic	hypermetrope	no	reduced	NO
P13	pre-presbyopic	myope	yes	normal	YES
P14	pre-presbyopic	myope	yes	reduced	NO
P15	pre-presbyopic	hypermetrope	yes	normal	NO
P16	pre-presbyopic	hypermetrope	yes	reduced	NO
P17	presbyopic	myope	no	normal	NO
P18	presbyopic	myope	no	reduced	NO
P19	presbyopic	hypermetrope	no	normal	YES
P20	presbyopic	hypermetrope	no	reduced	NO
P21	presbyopic	myope	yes	normal	YES
P22	presbyopic	myope	yes	reduced	NO
P23	presbyopic	hypermetrope	yes	normal	NO
P24	presbyopic	hypermetrope	yes	reduced	NO

30% of examples are
(randomly)
selected for testing



Information gain

number of examples in the subset S_v

set S attribute A

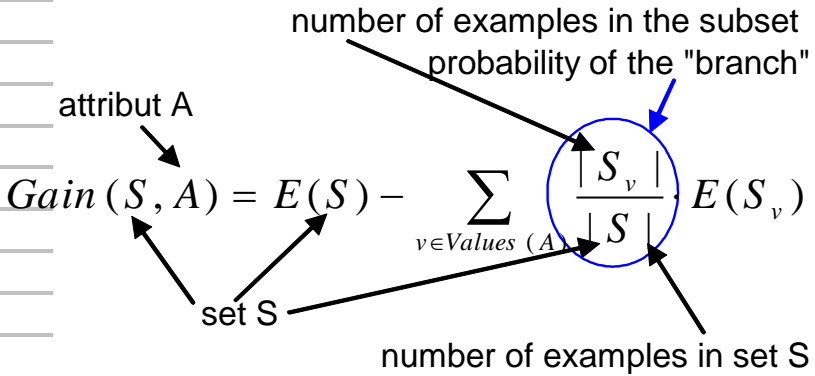
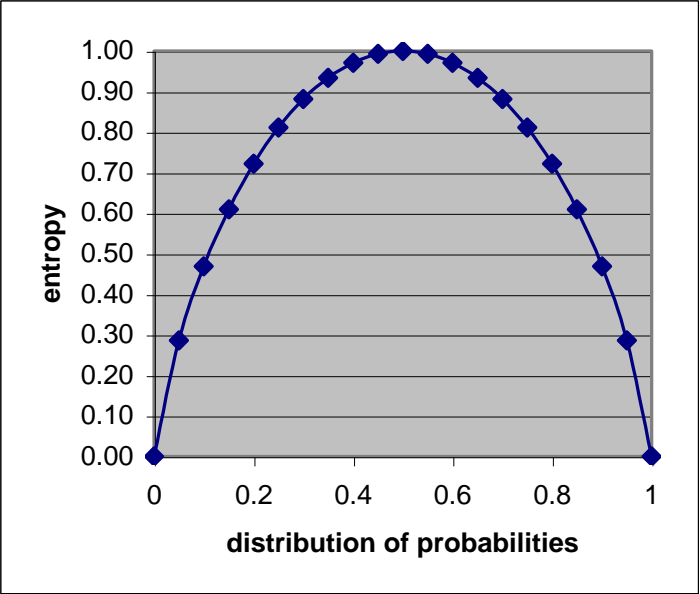
$$Gain(S, A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot E(S_v)$$

number of examples in set S

Weight = probability of a branch

Entropy and information gain

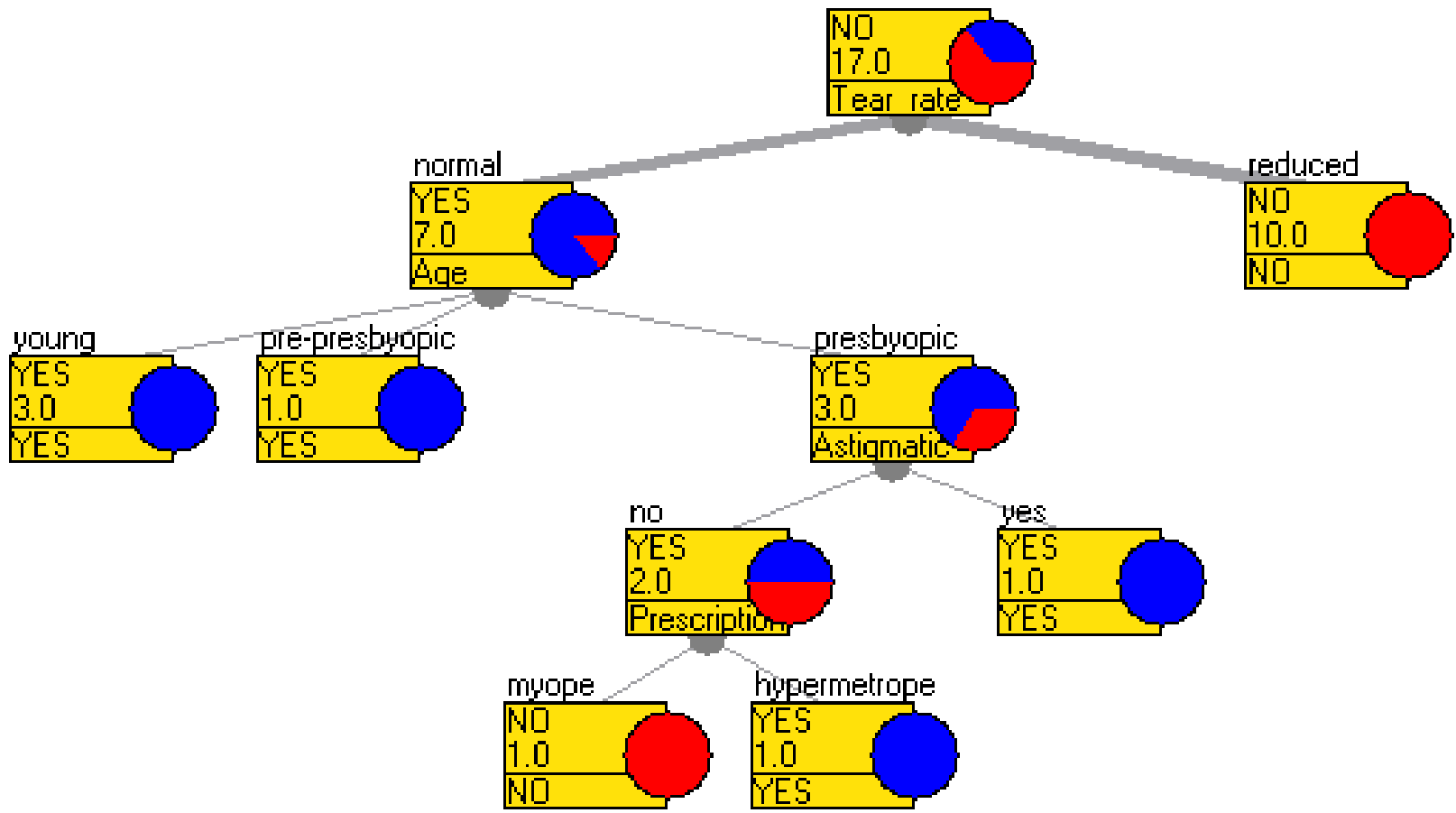
probability of class 1	probability of class 2	entropy $E(p_1, p_2) = -p_1 \cdot \log_2(p_1) - p_2 \cdot \log_2(p_2)$
p_1	$p_2 = 1 - p_1$	
0	1	0.00
0.05	0.95	0.29
0.10	0.90	0.47
0.15	0.85	0.61
0.20	0.80	0.72
0.25	0.75	0.81
0.30	0.70	0.88
0.35	0.65	0.93
0.40	0.60	0.97
0.45	0.55	0.99
0.50	0.50	1.00
0.55	0.45	0.99
0.60	0.40	0.97
0.65	0.35	0.93
0.70	0.30	0.88
0.75	0.25	0.81
0.80	0.20	0.72
0.85	0.15	0.61
0.90	0.10	0.47
0.95	0.05	0.29
1	0	0.00



Training set

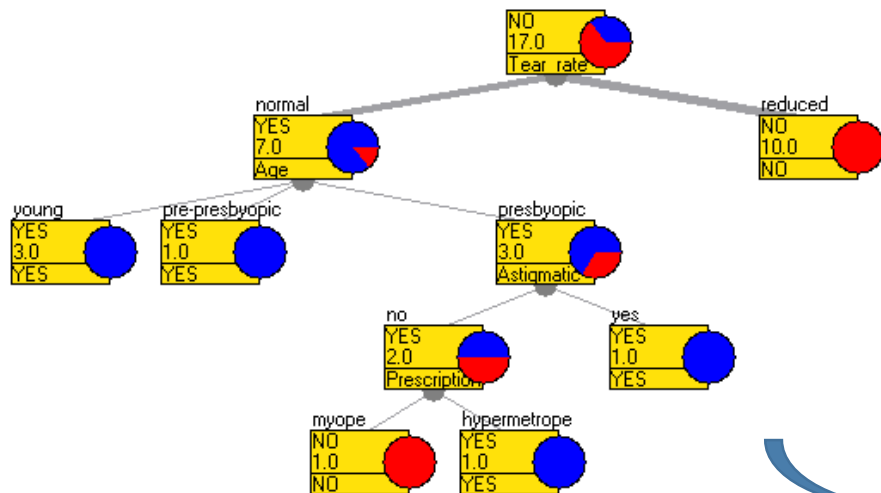
Person	Age	Prescription	Astigmatic	Tear_Rate	Lenses
P1	young	myope	no	normal	YES
P2	young	myope	no	reduced	NO
P4	young	hypermetrope	no	reduced	NO
P5	young	myope	yes	normal	YES
P6	young	myope	yes	reduced	NO
P7	young	hypermetrope	yes	normal	YES
P8	young	hypermetrope	yes	reduced	NO
P10	pre-presbyopic	myope	no	reduced	NO
P11	pre-presbyopic	hypermetrope	no	normal	YES
P14	pre-presbyopic	myope	yes	reduced	NO
P17	presbyopic	myope	no	normal	NO
P18	presbyopic	myope	no	reduced	NO
P19	presbyopic	hypermetrope	no	normal	YES
P20	presbyopic	hypermetrope	no	reduced	NO
P21	presbyopic	myope	yes	normal	YES
P22	presbyopic	myope	yes	reduced	NO
P24	presbyopic	hypermetrope	yes	reduced	NO

The induced decision tree



Classification with the tree

Person	Age	Prescription	Astigmatic	Tear_Rate	Lenses
P3	young	hypermetrope	no	normal	YES
P9	pre-presbyopic	myope	no	normal	YES
P12	pre-presbyopic	hypermetrope	no	reduced	NO
P13	pre-presbyopic	myope	yes	normal	YES
P15	pre-presbyopic	hypermetrope	yes	normal	NO
P16	pre-presbyopic	hypermetrope	yes	reduced	NO
P23	presbyopic	hypermetrope	yes	normal	NO



Classification accuracy = $(3+2) / (3+2+2+0) = 71\%$

	Predicted „YES“	Predicted „NO“
Actual „YES“	TP=3	FN=0
ACTUAL „NO“	FP=2	TN=2

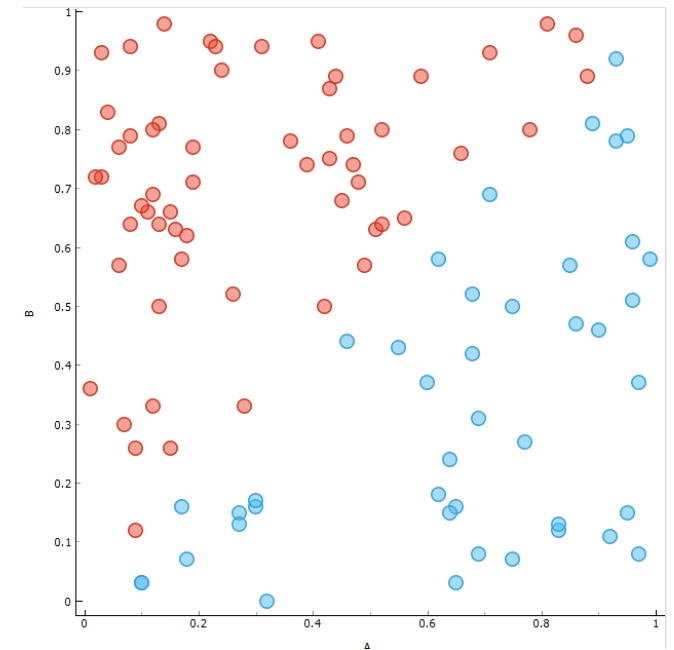


Questions

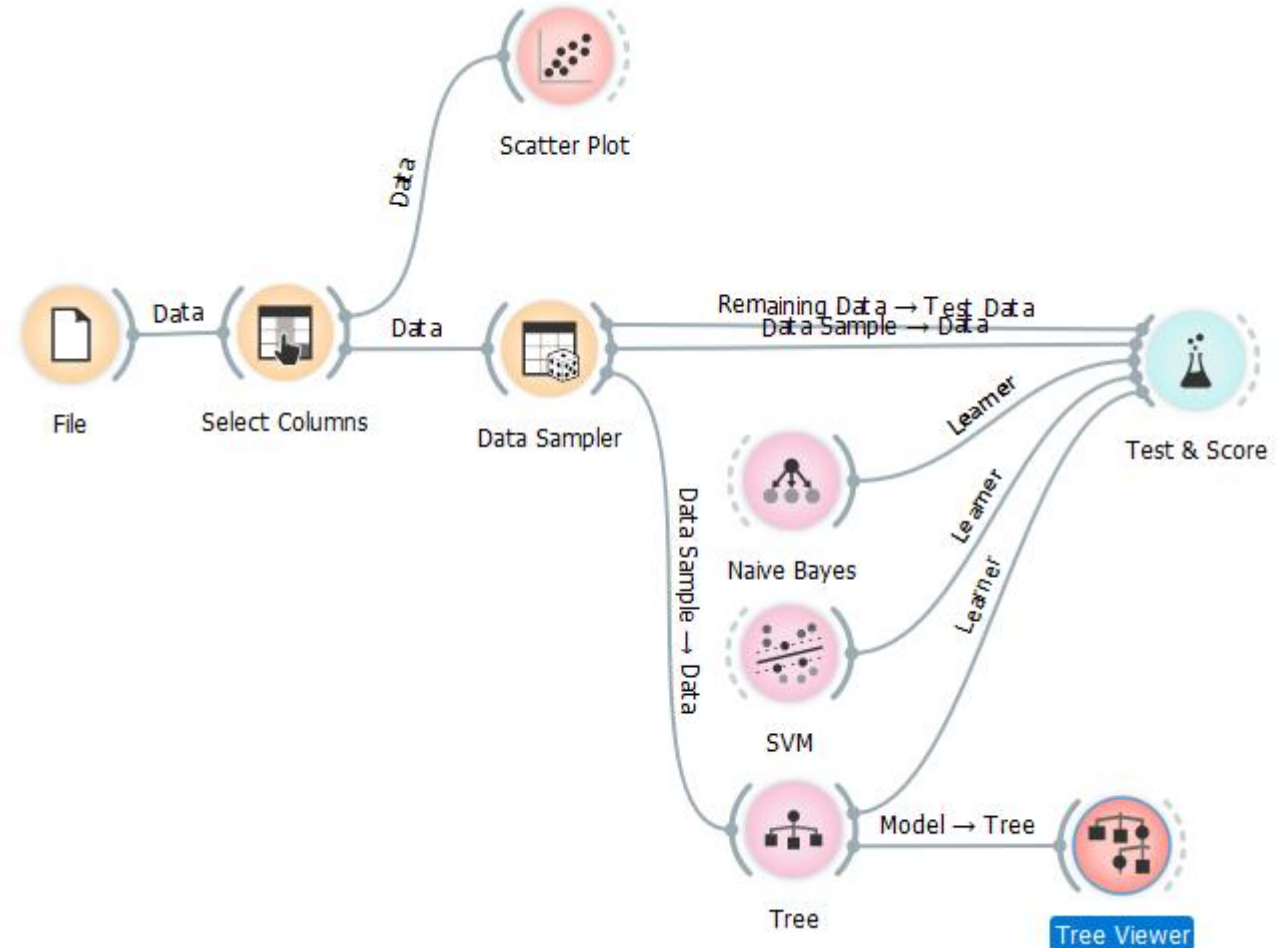
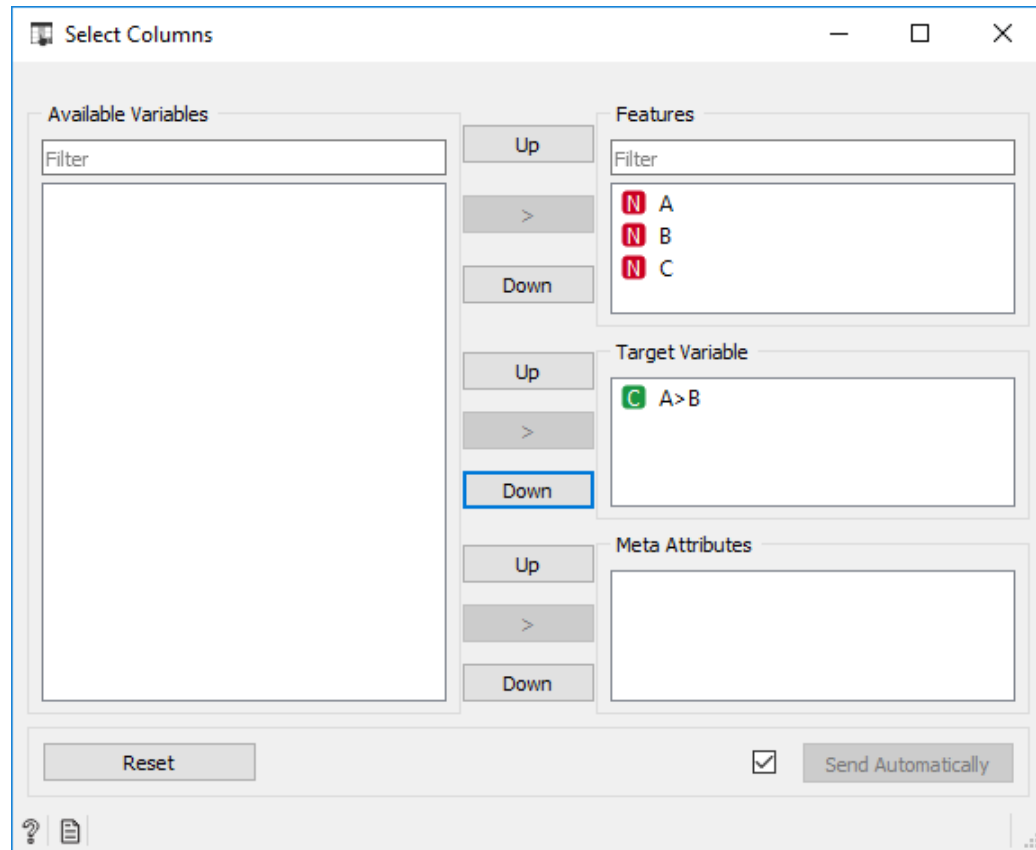
- Construct an attribute with Information gain =1.
- Construct an attribute with Information gain =0.
- Compute the Information gain of the attribute “Person”.
- How would you compute the information gain of a numeric attribute.
- What would be the classification accuracy of the decision tree (on the previous slide) if we pruned it at the node „Astigmatic“?

Lab exercise: Decision trees & Language bias

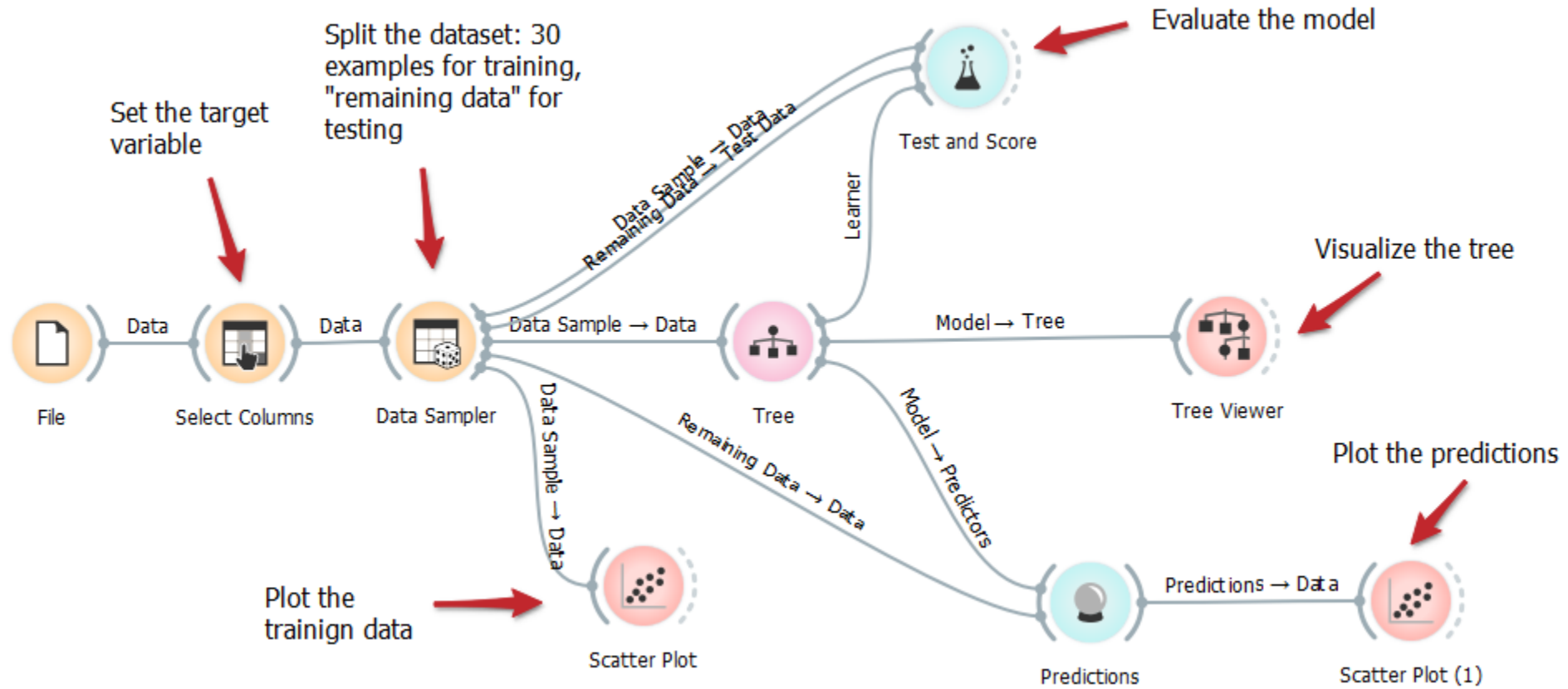
- Use a spreadsheet program (e.g. MS Excel) to generate 1000 examples:
 - Attributes A, B and C should have random values
 - Target variable „A>B“, should have value „true“ if A>B else “false”
 - Save the file
- Use Orange trees to predict „A>B“ from the attributes A, B in C
 - Set the target variable
 - Use separate test set for validation
 - Plot the training and classified data in “Scatter Plot”
- How good is your model?
- How does the training set size influence the model performance?
- MS Excel hints:
 - = RAND()
 - = IF(A2>B2, “true”, “false”)



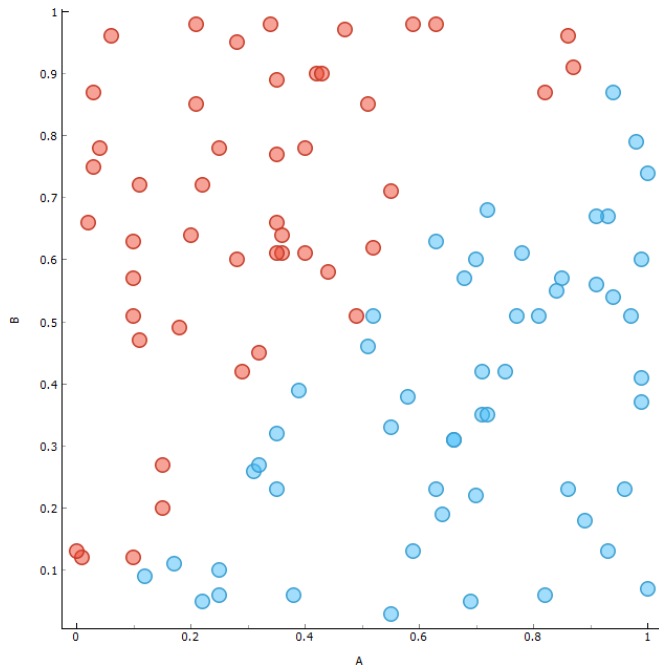
Lab exercise: Decision trees & Language bias



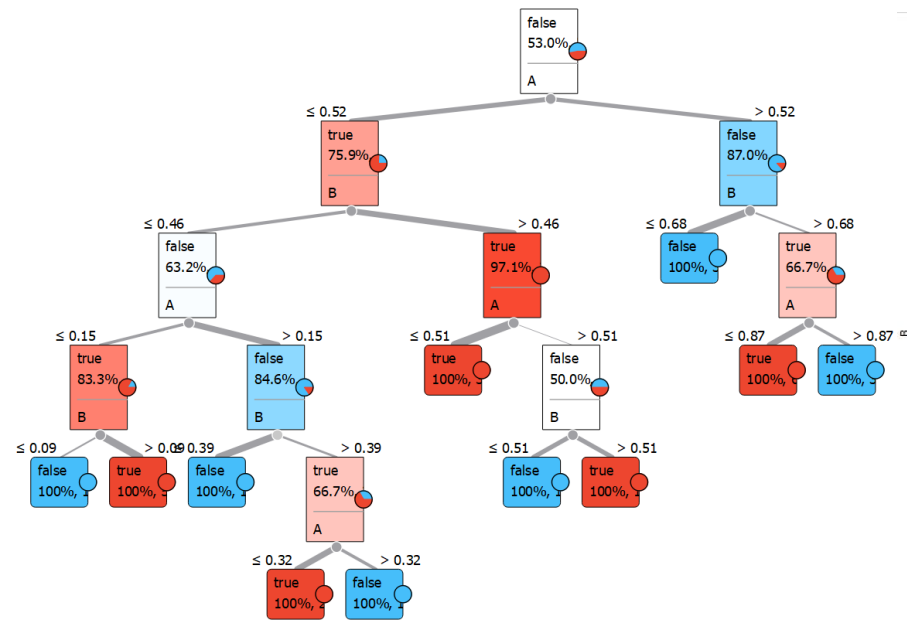
Complete workflow



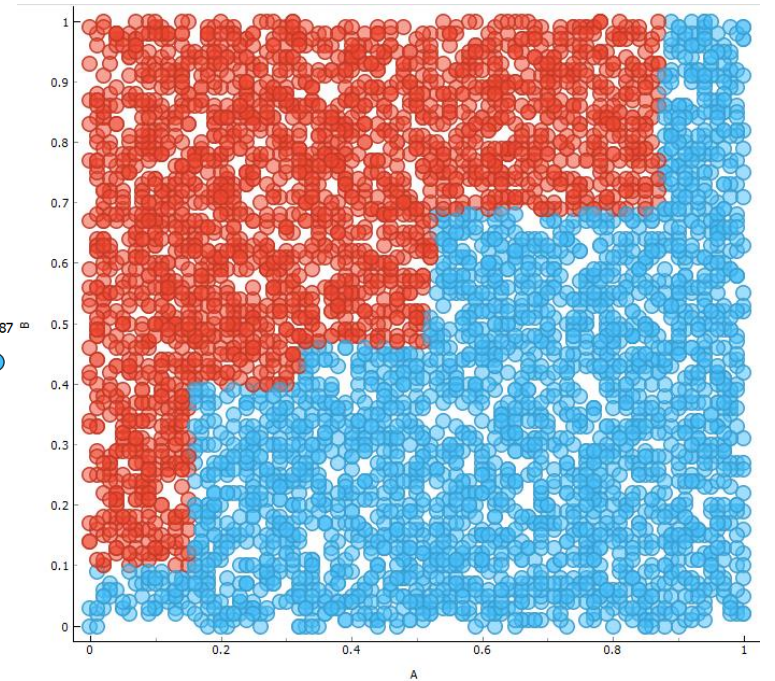
Lab exercise: Decision trees & Language bias



Training set

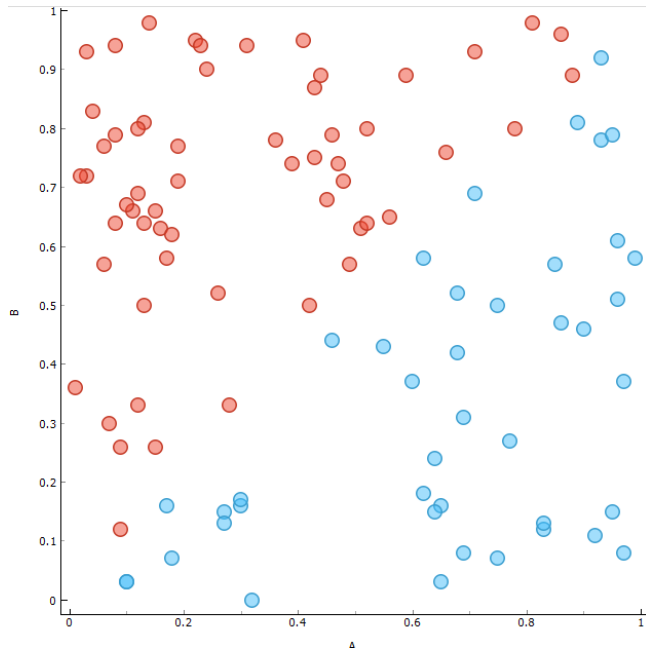


Decision tree

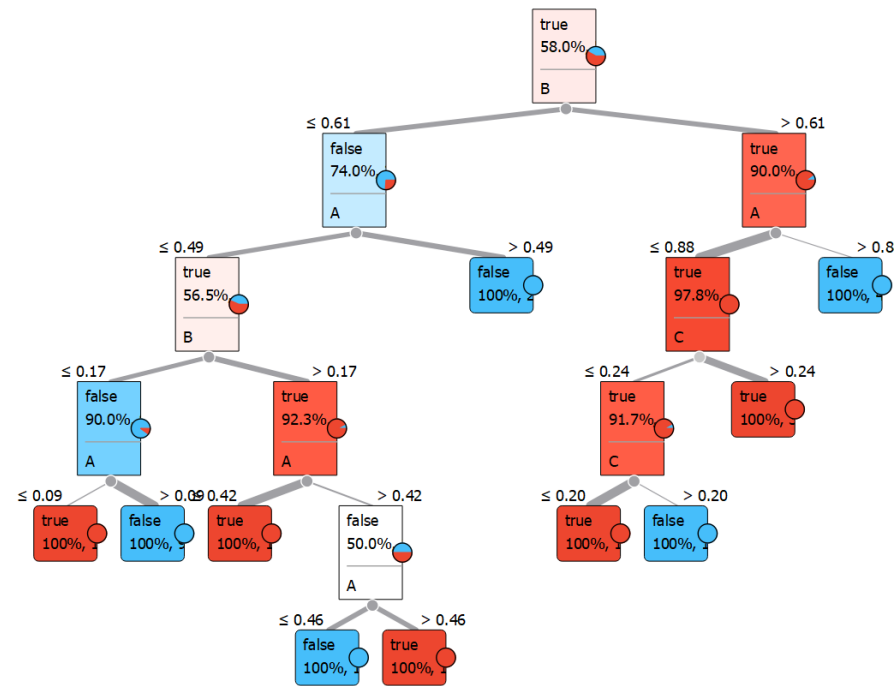


Test set

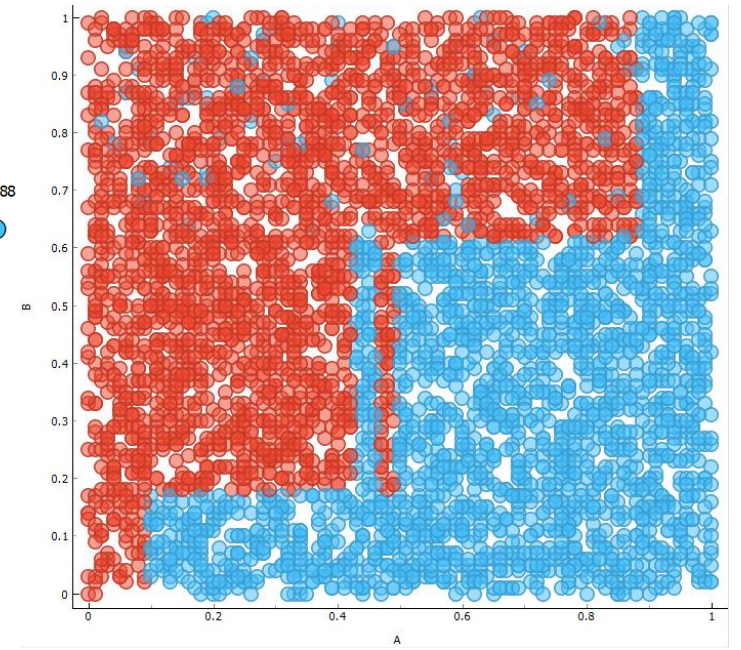
Same program, different random seed for training set selection



Training set



Decision tree



Test set

How to overcome this

- Feature engineering

- Create a new feature A>B

- Examples

- We have a person's height and body mass
→ Create a new attribute BMI (bod mass index)
- We have income and outcome data
→ Create a new attribute "profit"

$$BMI = \frac{Weight (kg)}{[Height(m)]^2}$$

- Ensemble

- We build more models that vote for the final classification
- Random forest: Several trees built on different subsets of the training set
- On this example, decision trees achieve CA 88,2% while random forest 90,8%
- As a general rule, classifier ensembles always outperform single classifiers

Homework

- Extend the workflow from the Lab exercise to use other ML algorithms:
 - Random forest
 - SVM with linear kernel
- Experiment with different random seeds (sample data with `data.sampler` several times) and observe the stability of results of different algorithms in different runs.
- Learn about random forests:
 - Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." *R news* 2.3 (2002): 18-22.

Literature

- Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." *R news* 2.3 (2002): 18-22.