MEDNARODNA
PODIPLOMSKA ŠOLA
JOŽEFA STEFANA

Data and Text Mining

Petra Kralj Novak

October 23, 2019

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

Data and Text Mining

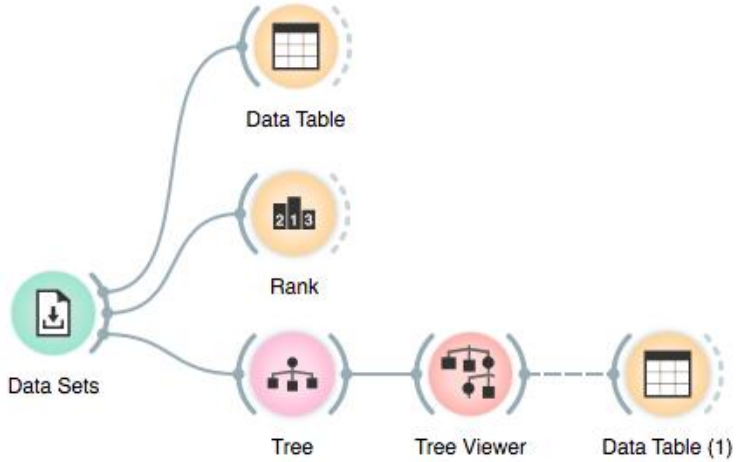
Course scope:

- Data preprocessing	Prof. dr. Bojan Cestnik
- Data mining	Prof. dr. Nada Lavrač
	Doc. dr. Petra Kralj Novak
- Text Mining	Prof. dr. Dunja Mladenić

Book: Max Bramer: Principles of data mining (2007)

- Skip Chapter 5
- Additional material on selected topics

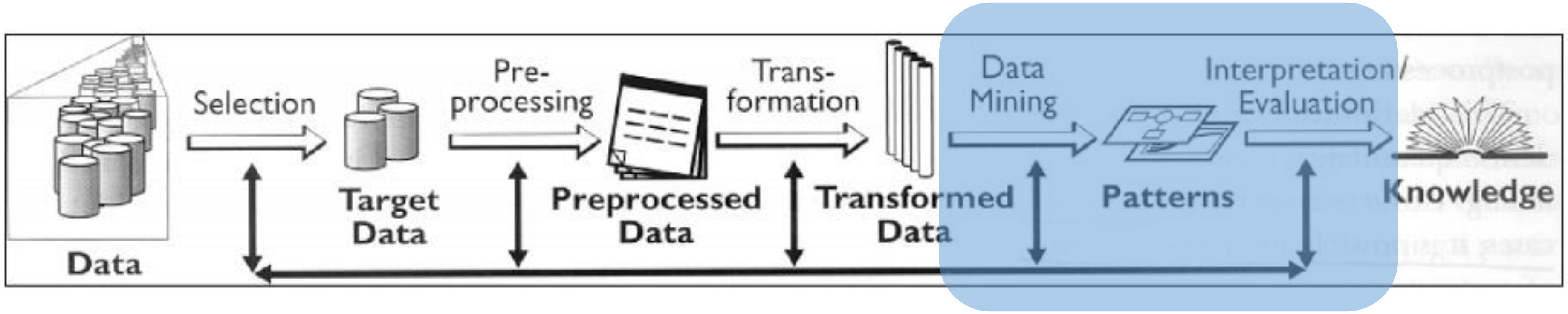
- Theory and exercises
- Hands-on orange
  - Open source machine learning and data visualization
  - Interactive data analysis workflows with a large toolbox
  - Visual programming
- Machine learning in Python with **scikit-learn**
  - The gold standard of Python machine learning
  - Simple and efficient tools for data mining and data analysis
  - Well documented



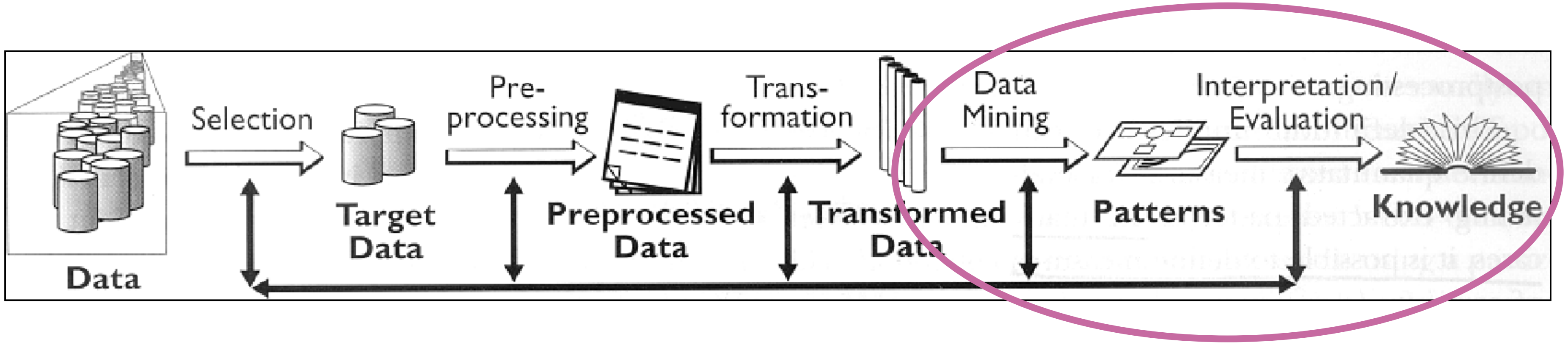

```
# ------------------------------------------------------------------------------------------------
print("Train and test classification models")
classifiers = [
    # ("Naive Bayes", naive_bayes.MultinomialNB()),
    ("Logistic regression", linear_model.LogisticRegression(C=1e5, solver='lbfgs', multi_class='multinomial', max_iter=600)),
    ("MultinomialNB", MultinomialNB()),
    ("SVC", svm.LinearSVC()),
    ("SVC-RBF", svm.SVC(gamma='scale', decision_function_shape='ovo'))]


for name, classifier in classifiers:
    classifier.fit(train_data, y_train)
    predictions = classifier.predict(test_data)
    classifier.confusion_matrix = metrics.confusion_matrix(predictions, y_test, labels=["negative", "neutral", "positive"])
    classifier.accuracy = metrics.accuracy_score(predictions, y_test)
    print(name, classifier.accuracy, "\n Confusion matrix: \n", classifier.confusion_matrix)
    pickle_clf(classifier, path="./models/"+name+".pkl")
```

KDD vs. ML/DM

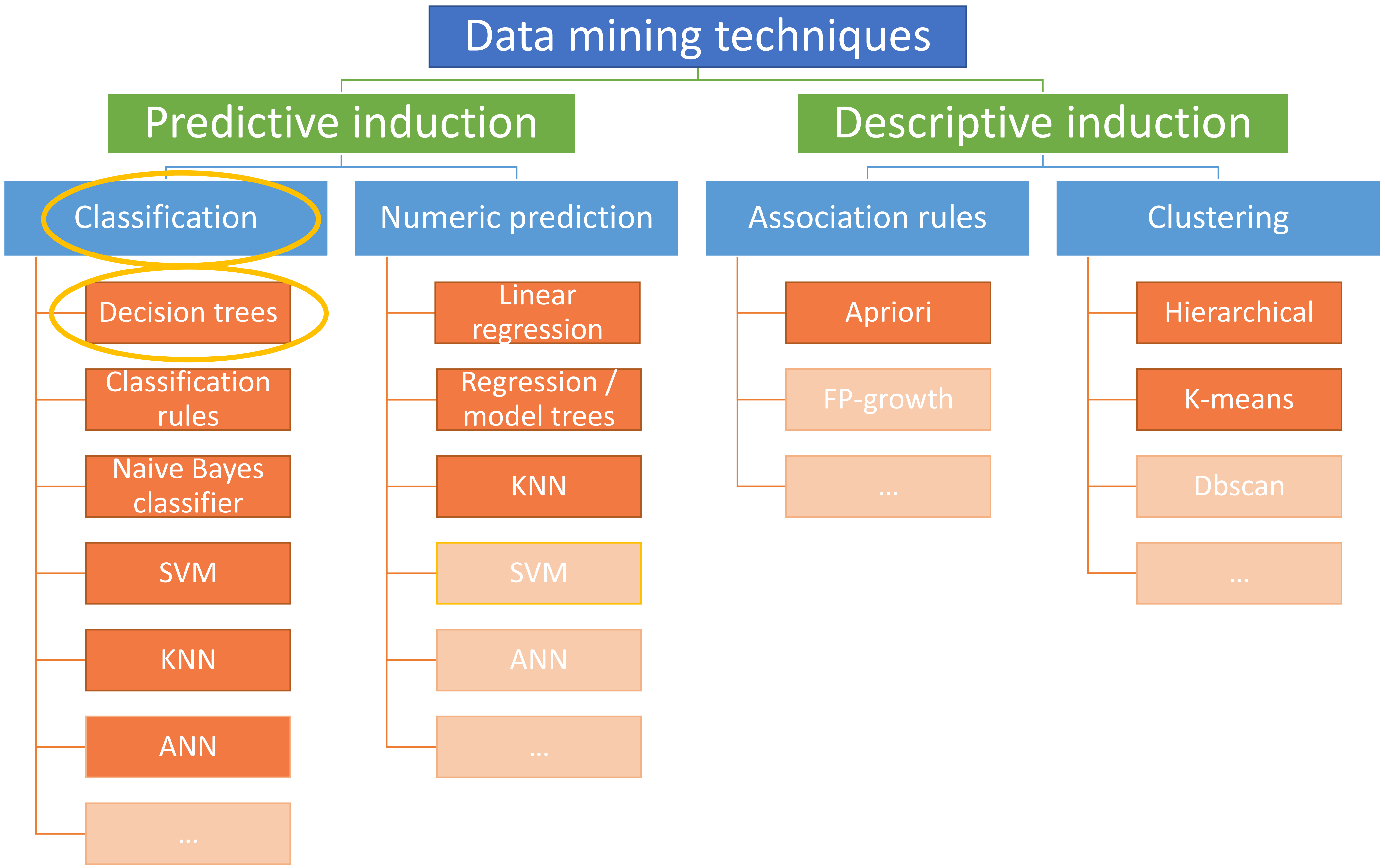
- Knowledge Discovery from Databases vs. Machine Learning/Data Mining



Keywords



- Data
  - Attribute, example, attribute-value data, target variable, class, discretization, market basket data
- Algorithms
  - Decision tree induction, entropy, information gain, overfitting, Occam's razor, model pruning, naïve Bayes classifier, KNN, association rules, support, confidence, classification rules, Laplace estimate, numeric prediction, regression tree, model tree, hierarchical clustering, dendrogram, k-means clustering, centroid, Apriori, heuristics vs. exhaustive search, predictive vs. descriptive DM, language bias, artificial neural networks, deep learning, backpropagation,…
- Evaluation
  - Train set, test set, accuracy, confusion matrix, cross validation, true positives, false positives, ROC space, AUC, error, precision, recall, F1, MSE, RMSE, rRMSE, support, confidence


Data mining techniques
Predictive induction
Descriptive induction
Classification
Numeric prediction
Association rules
Clustering
Decision trees
Classification rules
Naive Bayes classifier
SVM
KNN
ANN
…
Linear regression
Regression / model trees
KNN
SVM
ANN
…
Apriori
FP-growth
…
Hierarchical
K-means
Dbscan
…

Data for Data Mining

Example: the "adult" dataset

Attributes

Examples

	y	age	sex	education-num	occupation	relationship	race	hours-per-week
1	<=50K	39.000	Male	13.000	Adm-clerical	Not-in-family	White	40.000
2	<=50K	50.000	Male	13.000	Exec-managerial	Husband	White	13.000
3	<=50K	38.000	Male	9.000	Handlers-clean...	Not-in-family	White	40.000
4	<=50K	53.000	Male	7.000	Handlers-clean...	Husband	Black	40.000
5	<=50K	28.000	Female	13.000	Prof-specialty	Wife	Black	40.000
6	<=50K	37.000	Female	14.000	Exec-managerial	Wife	White	40.000
7	<=50K	49.000	Female	5.000	Other-service	Not-in-family	Black	16.000
8	>50K	52.000	Male	9.000	Exec-managerial	Husband	White	45.000
9	>50K	31.000	Female	14.000	Prof-specialty	Not-in-family	White	50.000
10	>50K	42.000	Male	13.000	Exec-managerial	Husband	White	40.000
11	>50K	37.000	Male	10.000	Exec-managerial	Husband	Black	80.000
12	>50K	30.000	Male	13.000	Prof-specialty	Husband	Asian-Pac-Islan...	40.000
13	<=50K	23.000	Female	13.000	Adm-clerical	Own-child	White	30.000
14	<=50K	32.000	Male	12.000	Sales	Not-in-family	Black	50.000
15	>50K	40.000	Male	11.000	Craft-repair	Husband	Asian-Pac-Islan...	40.000
16	<=50K	34.000	Male	4.000	Transport-movi...	Husband	Amer-Indian-Es...	45.000
17	<=50K	25.000	Male	9.000	Farming-fishing	Own-child	White	35.000
18	<=50K	32.000	Male	9.000	Machine-op-in...	Unmarried	White	40.000

Types of attributes

- Categorical
  - Nominal (Colors: red, blue, green)
  - Binary (Gender: male, female)
  - Ordinal (Size: small, medium, large)
- Numerical
  - Integer (Number of car sits: 2, 5, …)
  - Real (Temperature in degrees: 21℃, 23.4℃,…)
  - Ordinal
  - Binary

Mining complex data types

- Time series analysis
  - Financial time series, heart-rate monitoring,…
- Text mining
  - News, comments, Wikipedia, books, … for content, sentiment, style, word meaning…
- Graph mining
  - Maps, molecules, citation networks, hyperlinks, …. for classification, patterns,…
- Social media mining (graphs + text)
  - Facebook, Twitter, …. Information spreading, hate speech…
- Images
  - Image classification

Lab exercise 1

Data for data mining in Orange

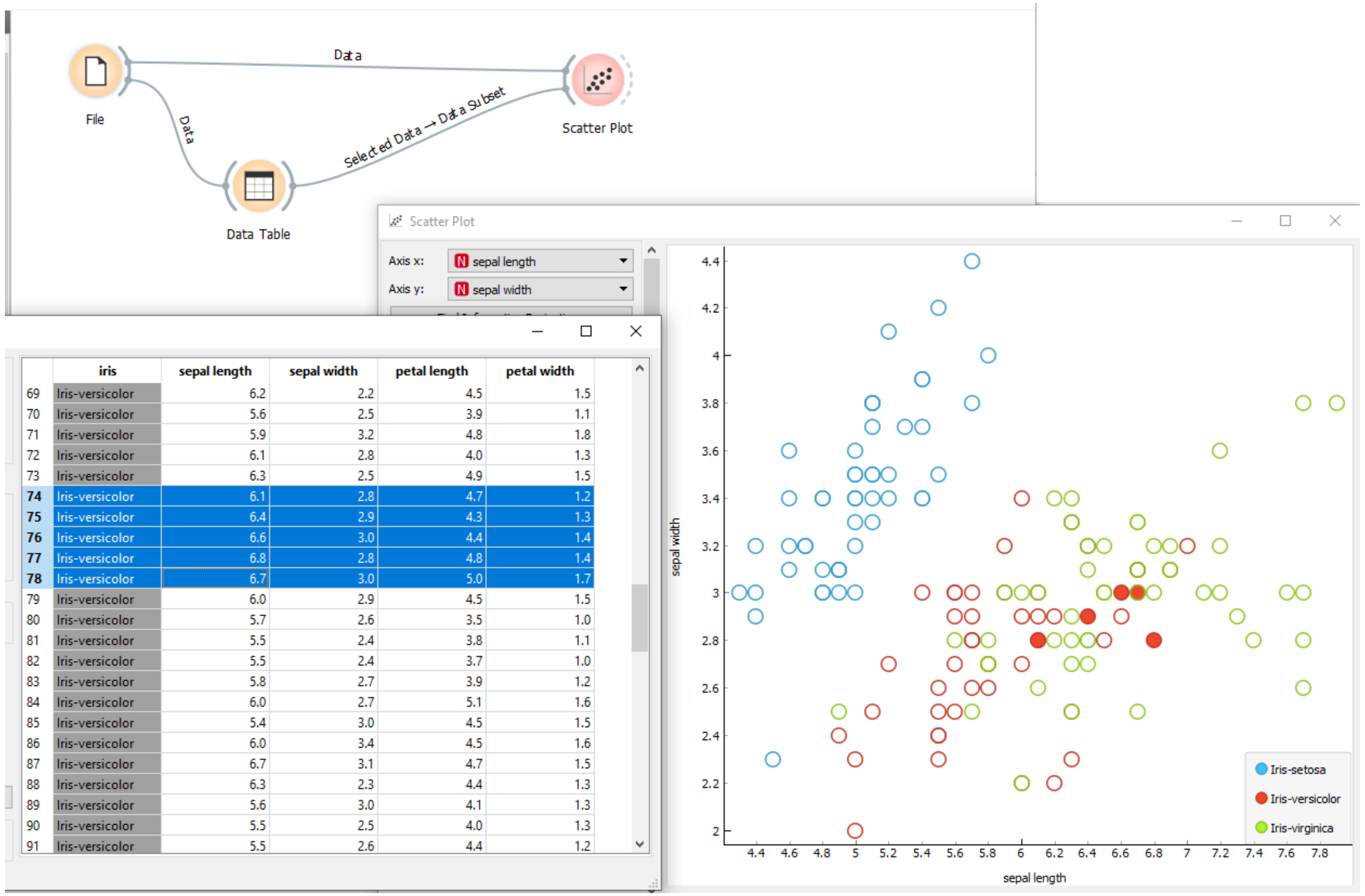
Exercise1: Use Orange to fill in the following table

	Number of examples	Number of attributes	Number of numeric attributes	Number of categorical attributes	Target variable	Number of ordinal attributes
Zoo						
Iris						
Auto-mpg						
Wine						
Titanic						

Exercise 2: Use a text editor to view (and understand) the .tab data format.

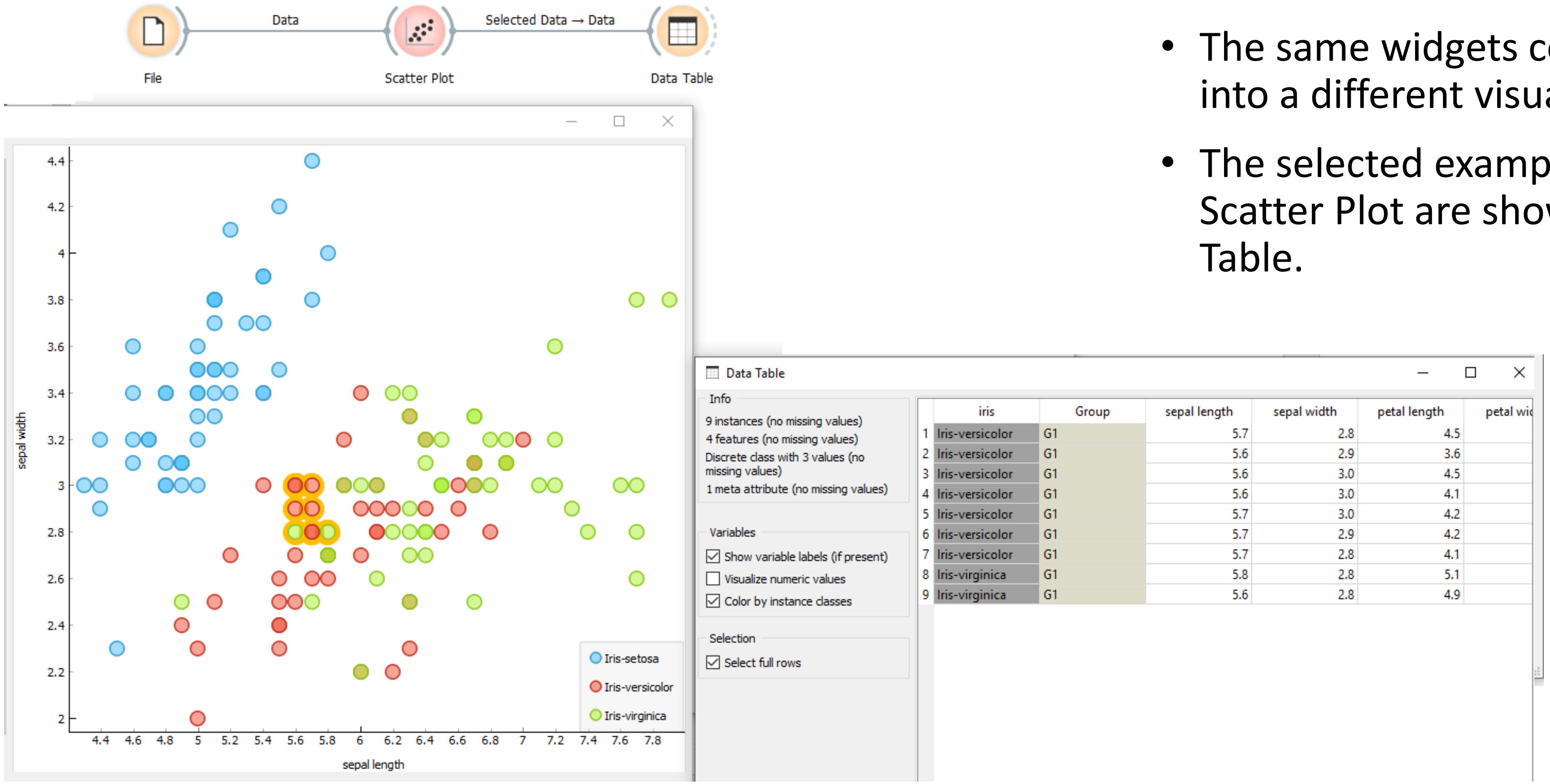
Exercise 3: Create two interesting data visualizations with Orange.

Interactive visualization in Orange



- The widgets File, Data Table and Scatter Plot are connected to form a visual program.
- The selected examples in the Data Table widget are displayed as full circles in the Scatterplot.
- Note: Scatter Plot has two inputs: Data and Data subset and they need to be connected correctly.

Interactive visualization in Orange



- The same widgets composed into a different visual program.
- The selected examples in Scatter Plot are shown in Data Table.

Classification

Classification problem

- Goal: Assign each example a category

- Examples
  - Magazine reader (or not)
  - Patients at risk for acquiring a certain illness
  - A patient needing antibiotics (or not)
  - Customers who are likely buyers
  - People who are likely to vote for a political party
  - Churn prediction
  - …

Classification problem

- Goal: Identifying to which one of a number of mutually exhaustive and exclusive categories (known as classes) an object belongs to.

  - Given a dataset of examples (described by attributes).
  - The target variable is a attribute that we are interested in predicting. In classification, the target is categorical.
  - The values of the target variable are called classes.
  - We train a model on the data that will predict the classes of new examples as accurately as possible.

Attribute-value data for classification

(nominal) target variable

attributes

Examples

or

instances

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**

classes

=

values of the (nominal) target variable

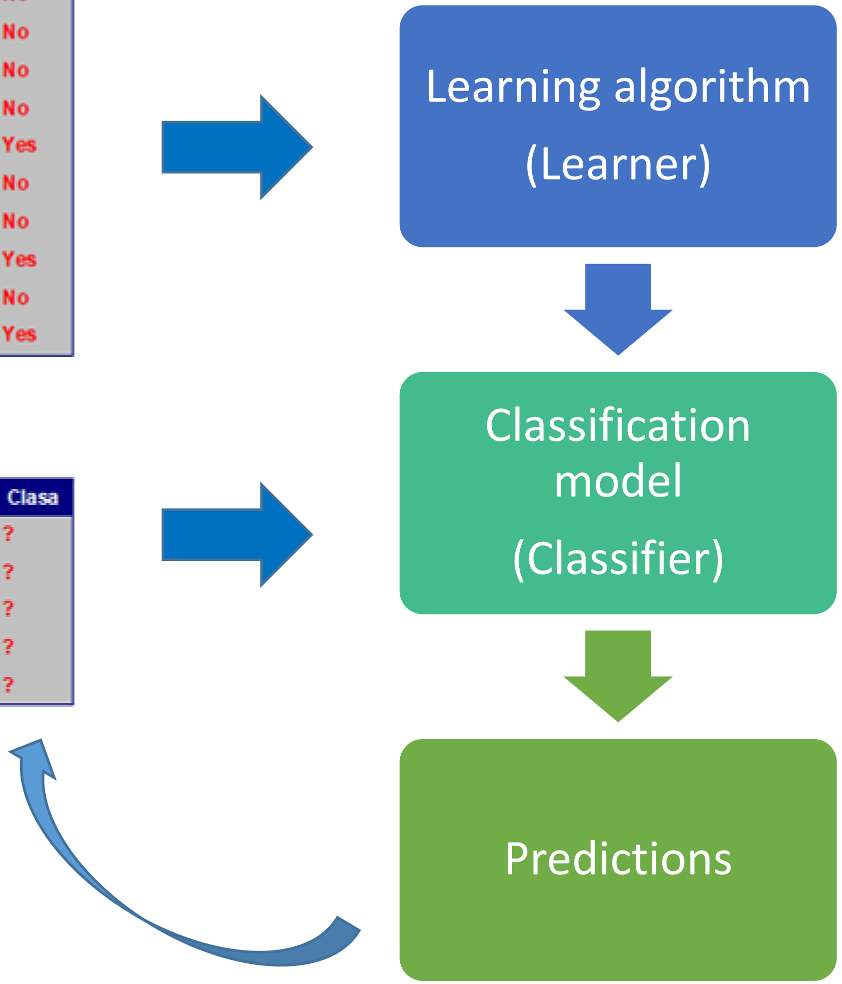
The basic classification schema

Šr	Atrib1	Atrib2	Atrib3	Clasa
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training set

Šr	Atrib1	Atrib2	Atrib3	Clasa
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

New data



- A classifier is a function that maps from the attributes to the classes
  - Classifier(attributes) = Classes
  - *f(X) = Y*
- In training, the attributes and the classes are known (training examples) and we are learning a mapping function *f* (the clasifier)
  - *?(X) = Y*
- When predicting, the attributes and the classifier are known and we are assigning the classes
  - *f(X) = ?*
- What about evaluation?

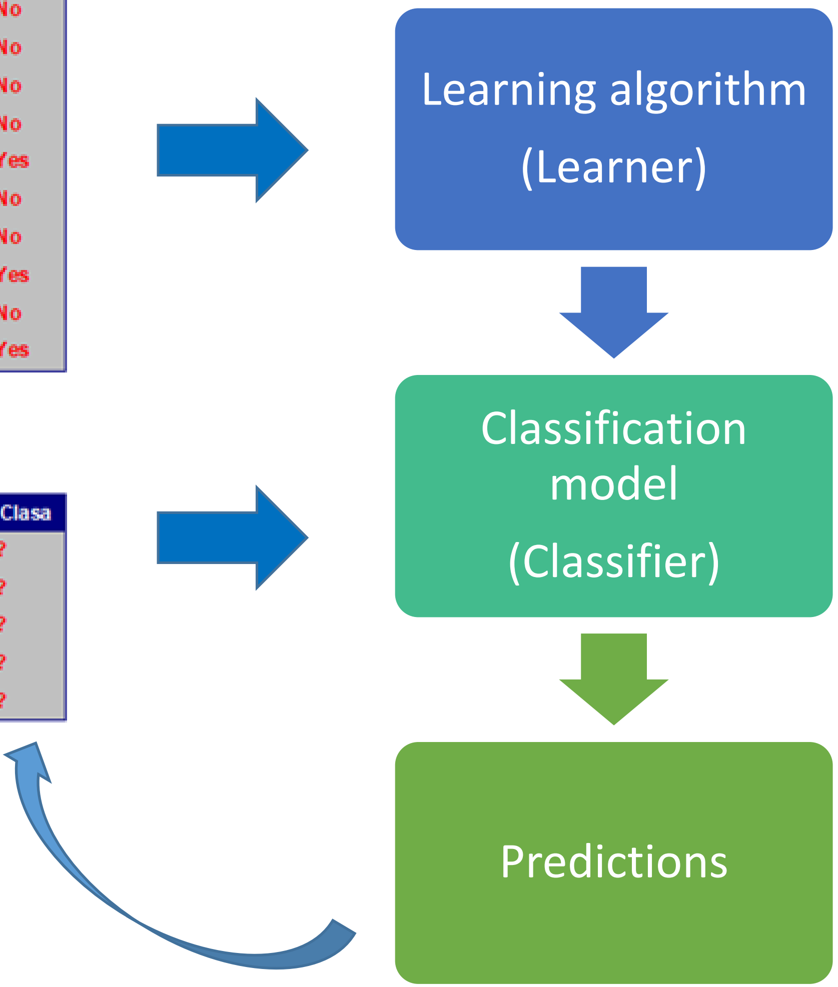
The basic classification schema

Št	Atrib1	Atrib2	Atrib3	Clasa
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training set

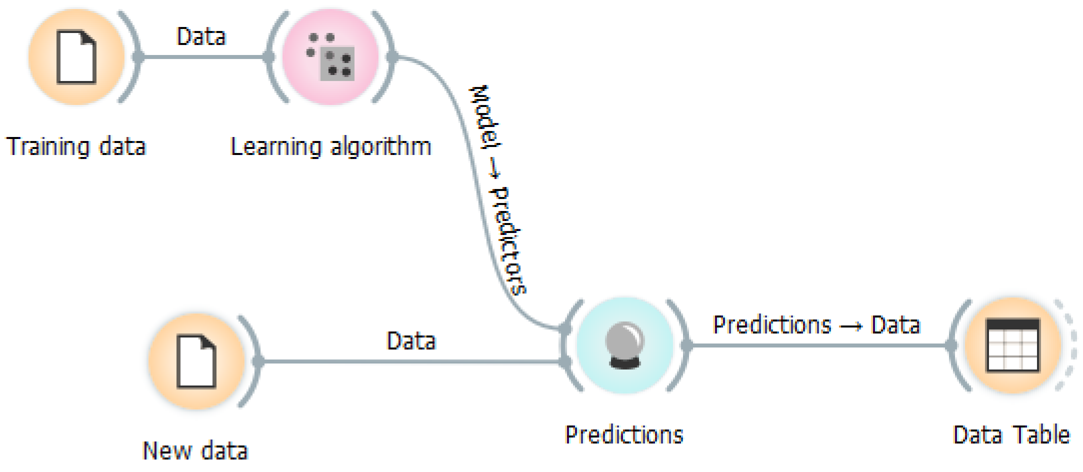
Št	Atrib1	Atrib2	Atrib3	Clasa
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

New data



- A classifier is a function that maps from the attributes to the classes
  - Classifier(attributes) = Classes
  - $f(X) = Y$
- In training, the attributes and the classes are known (training examples) and we are learning a mapping function *f* (the clasifier)
  - $?(X) = Y$
- When predicting, the attributes and the classifier are known and we are assigning the classes
  - $f(X) = ?$
- When evaluating, *f*, *X* and *Y* are known. We compute the predictions $Y_p = f(X)$ and evaluate the difference between *Y* and Y_p.

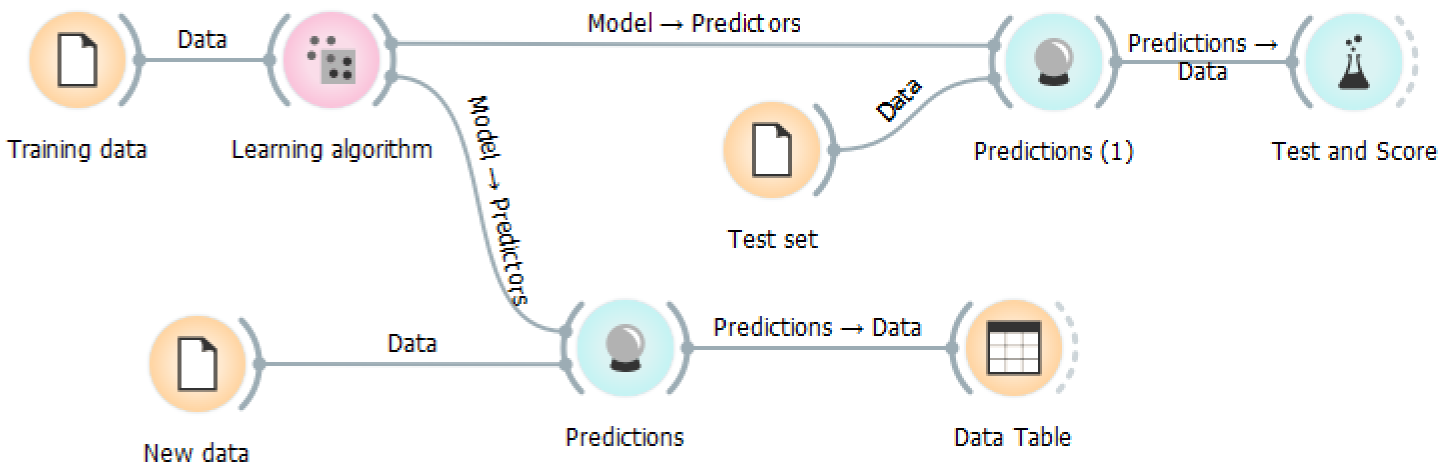
Basic classification schema in Orange



- We train the model on the train set
- We predict the target for the new instances
- There are several classification algorithms:
  - Decision trees
  - Naive Bayes classifier
  - K nearest neighbors (KNN)
  - Artificial neural networks (ANN)
  - ….

Classification with evaluation

- We train the model on the train set
- We evaluate on the test set
- We classify the new instances



Example: “titanic” dataset

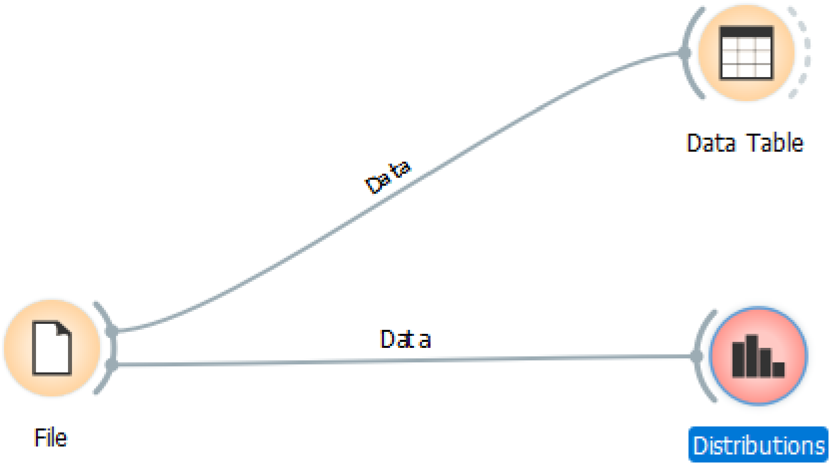
Target variable

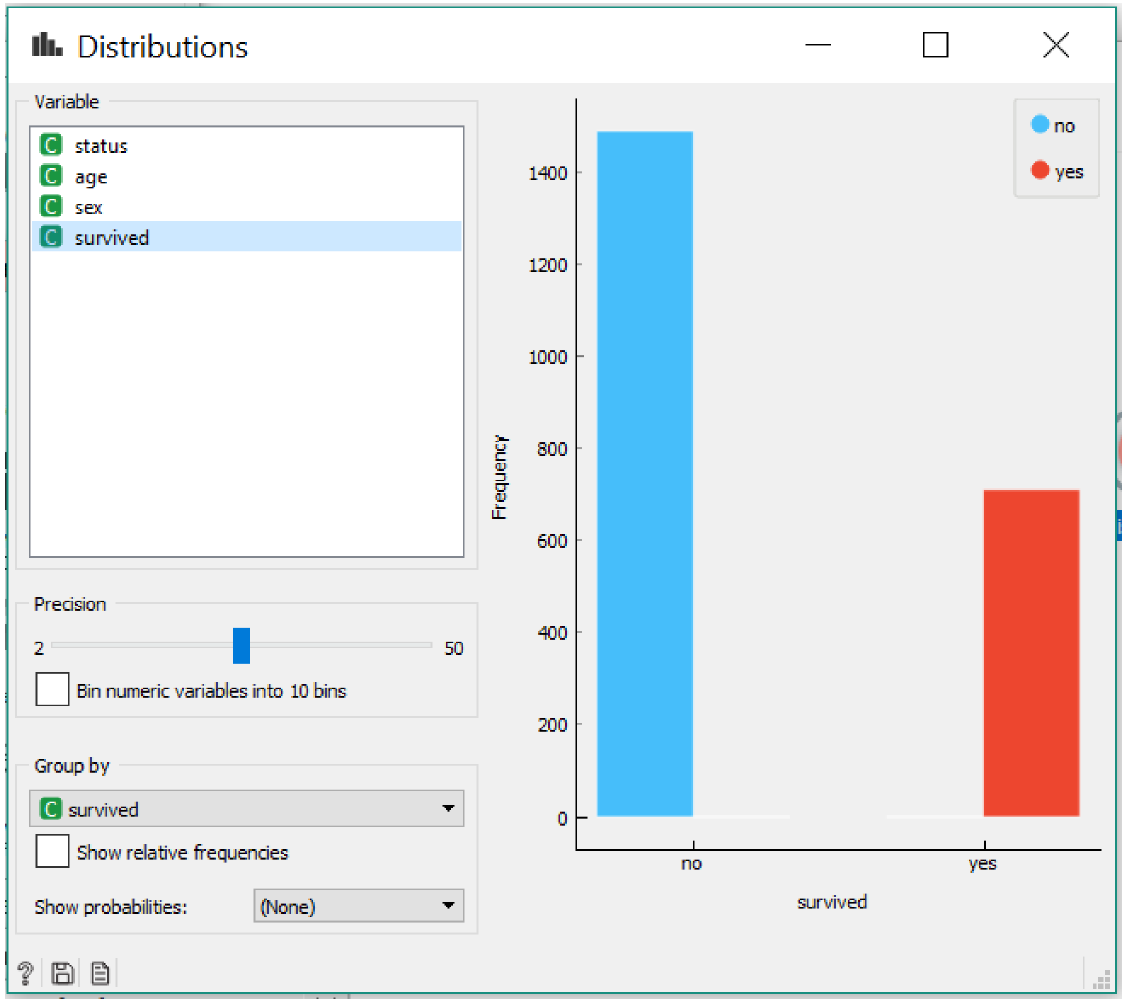
Attributes

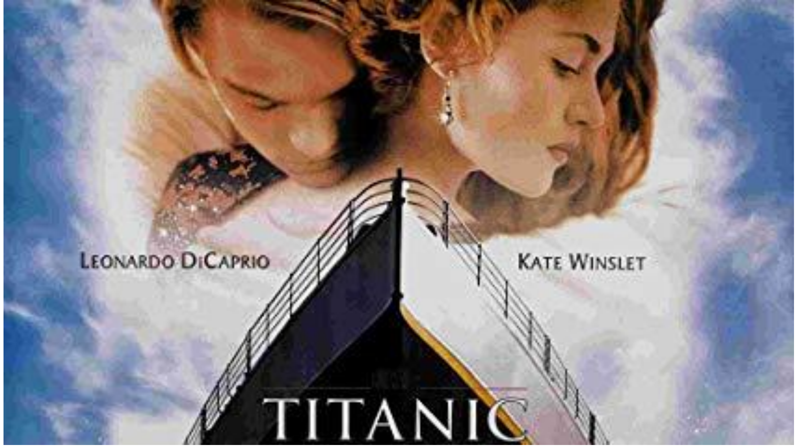
Examples

	survived	status	age	sex
1281	no	third	child	male
1282	no	third	child	male
1283	no	third	child	male
1284	no	third	child	male
1285	no	third	child	male
1286	yes	third	child	female
1287	yes	third	child	female
1288	yes	third	child	female
1289	yes	third	child	female
1290	yes	third	child	female
1291	yes	third	child	female
1292	yes	third	child	female
1293	yes	third	child	female
1294	yes	third	child	female
1295	yes	third	child	female
1296	yes	third	child	female
1297	yes	third	child	female
1298	yes	third	child	female
1299	yes	third	child	female
1300	no	third	child	female

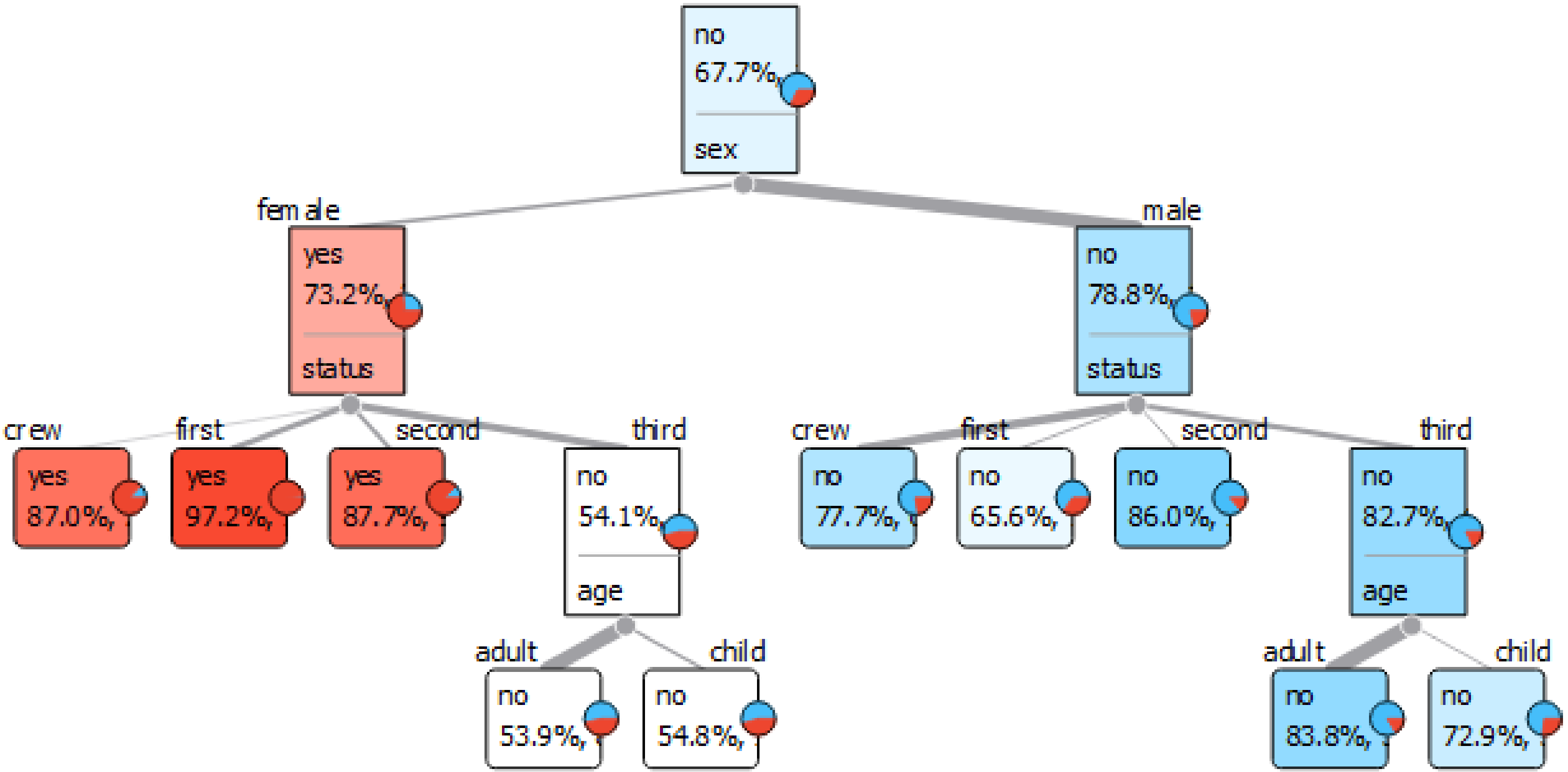
Classification: distribution of the target variable



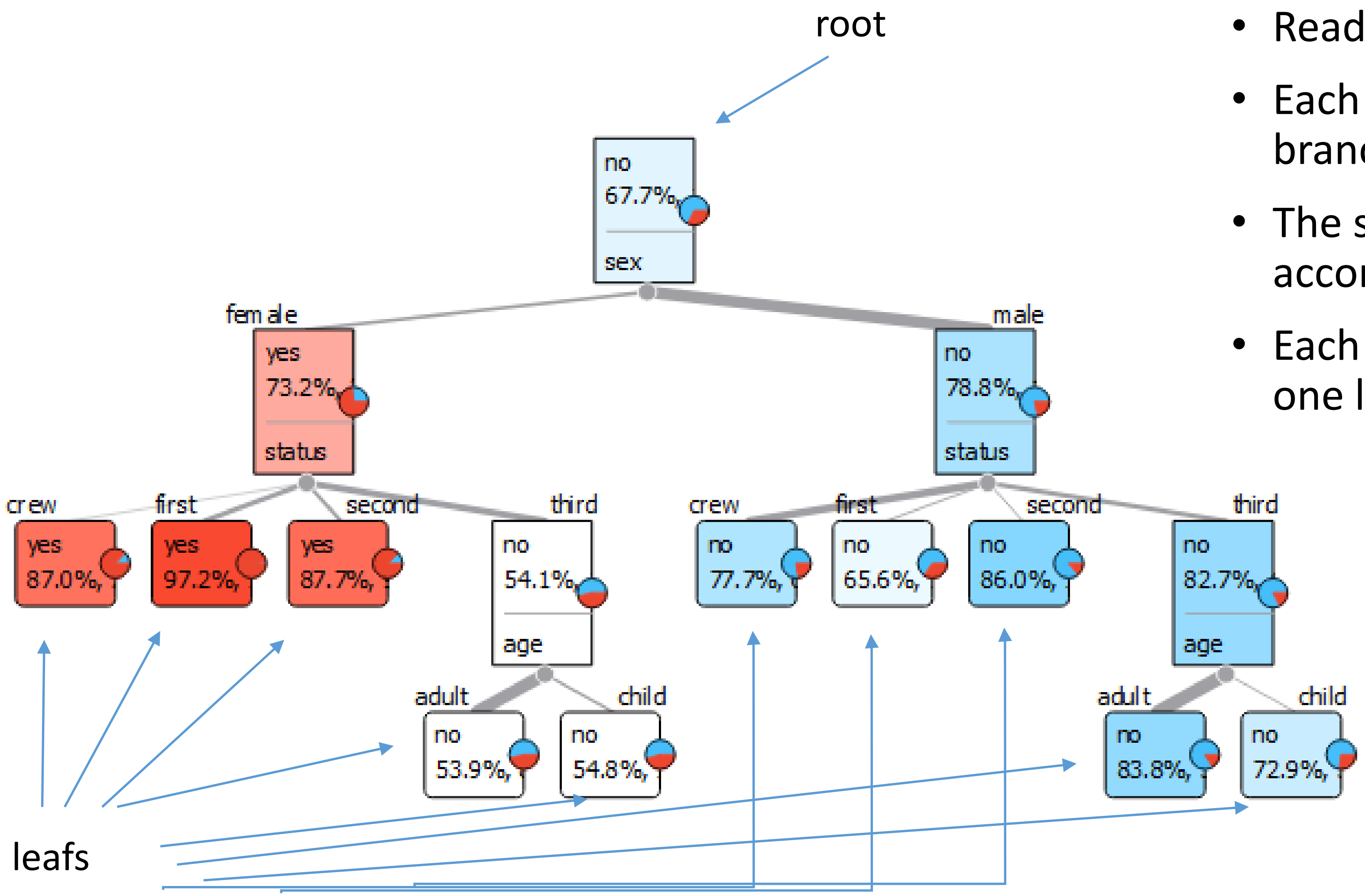




Who survived on the Titanic?

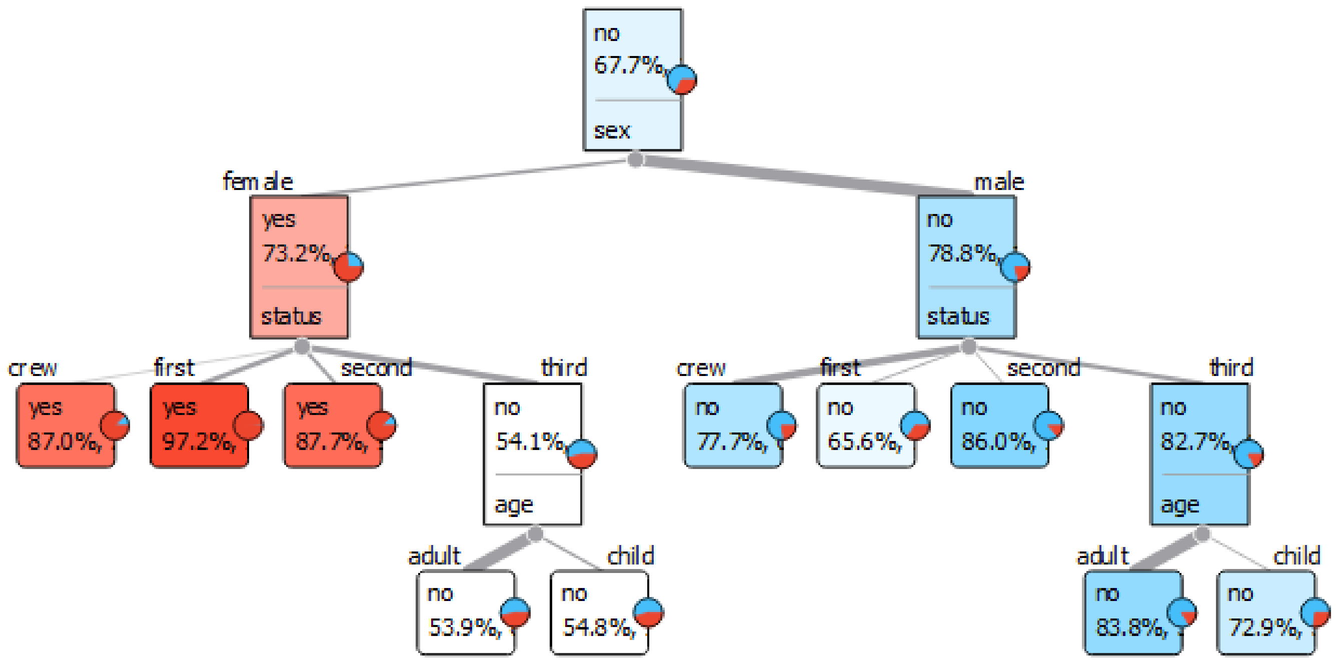


Decision tree



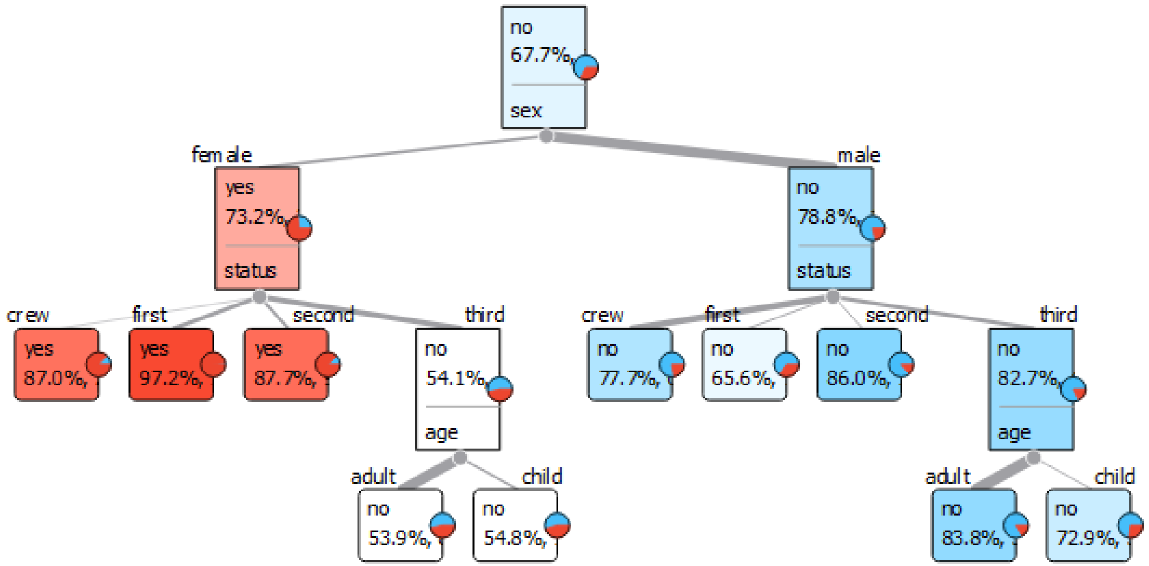
- Read top-down
- Each node is an attribute which branches according to its values
- The set of examples splits according to attribute values
- Each example end up in exactly one leaf

Exercise: Classify the data instances



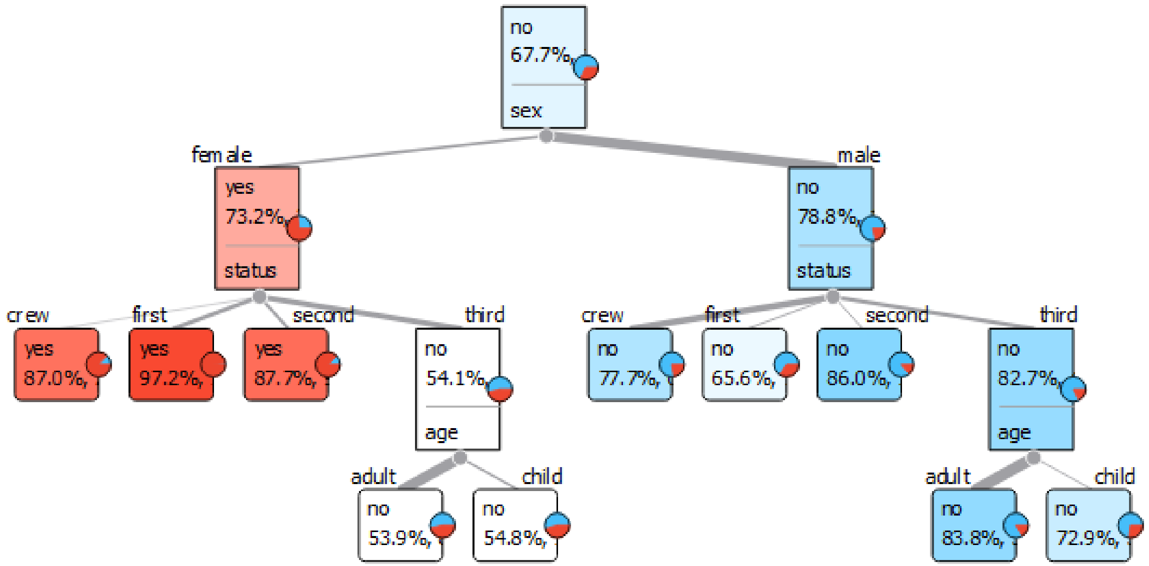
	status	age	sex	survived?
1	third	child	male	
2	third	child	female	
3	crew	adult	male	
4	first	adult	male	
5	second	adult	male	
6	third	adult	male	
7	first	adult	female	
8	second	adult	female	
9	third	adult	female	
10	third	child	male	

We can rewrite the tree as a set of rules



- One rule for each leaf

We can rewrite the tree as a set of rules

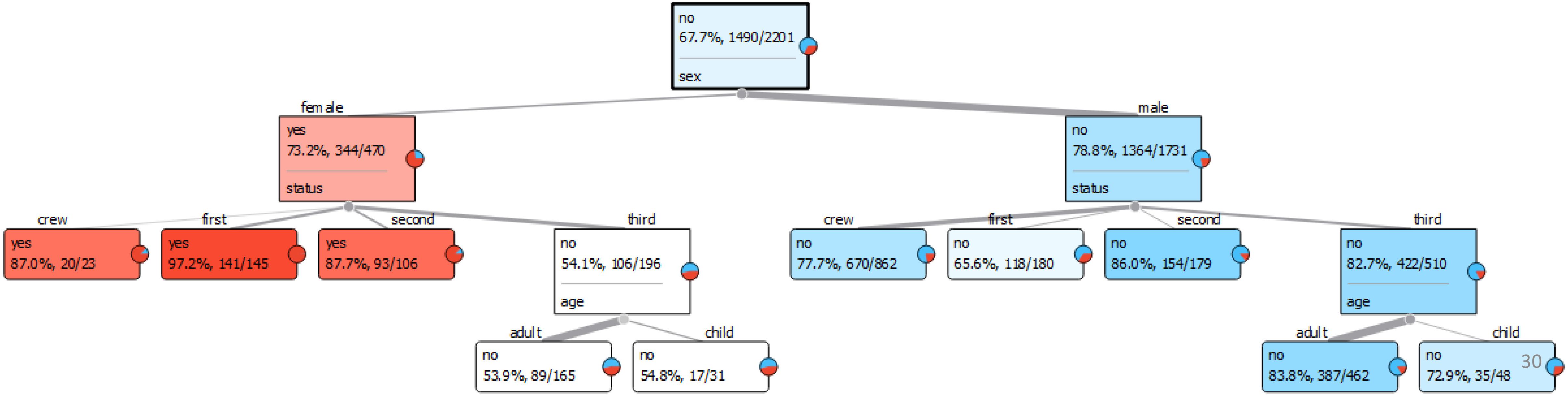


- sex = female & status = crew → survived = yes
- sex = female & status = first → survived = yes
- sex = female & status = second → survived = yes
- sex = female & status = third & age = adult → survived = no
- sex = female & status = third & age = child → survived = no
- sex = male & status = crew → survived = no
- sex = male & status = first → survived = no
- sex = male & status = second → survived = no
- sex = male & status = third & age = adult → survived = no
- sex = male & status = third & age = child → survived = no

- Rule: a path from root leaf
- Each example *fires* exactly one rule

We can interpret decision trees

- Which is the most informative attribute?
- Visualization in orange:
  - The number of examples in each node
  - Percentage of examples belonging to the majority class
  - Colour intensity = certainty of the prediction
  - Thickness of the branch proportional to the number of examples



TDIDT
Top Down Induction of Decision Trees

TDIDT – Top Down Induction of Decision Trees

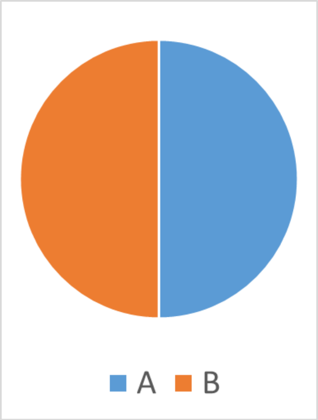
- We induce decision trees top-down
- There is many possible decision trees for a given dataset
- It is very important which attribute we choose as the root
- Heuristic: we choose the attribute which **best separates** the classes

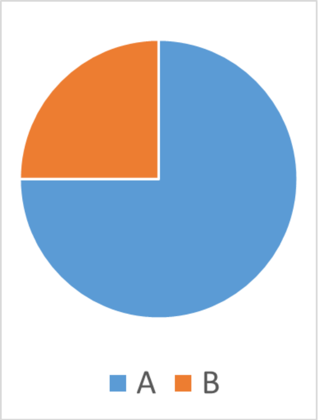
Information gain

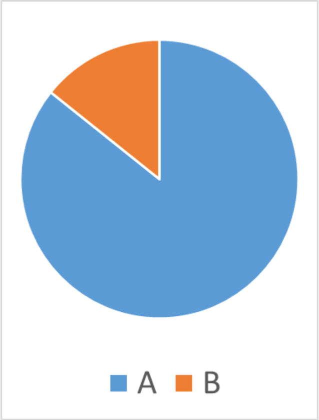
Entropy

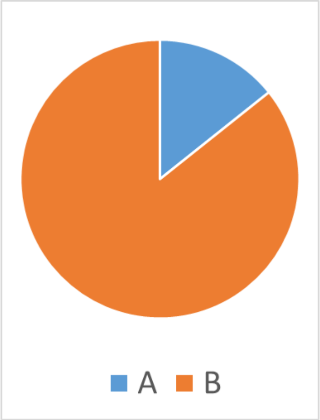
Entropy

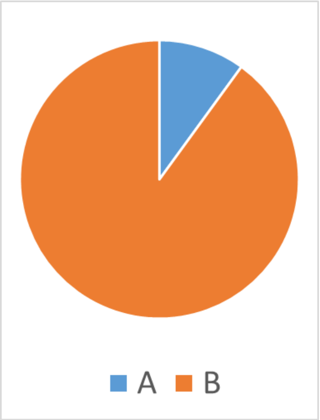
- Entropy (information theory) is a measure of uncertainty.

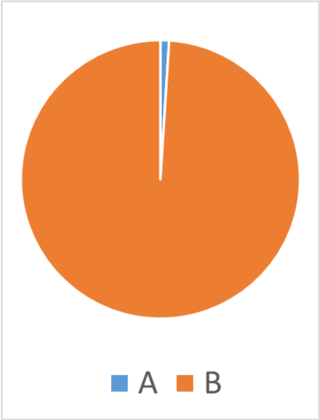












½ ¼ 1/7 6/7 9/10 99/100

Entropy

$$E(S) = -\sum_{c=1}^{N} p_c \cdot \log_2 p_c$$

- Calculate:

  E (0 , 1) =

  E (1/2 , 1/2) =

  E (1/4 , 3/4) =

  E (1/7 , 6/7) =

  E (6/7 , 1/7) =

  E (0.1 , 0.9) =

  E (0.001 , 0.999) =

Entropy

$$E(S) = -\sum_{c=1}^{N} p_c \cdot \log_2 p_c$$

- Calculate:

  E (0 , 1) = 0

  E (1/2 , 1/2) = 1

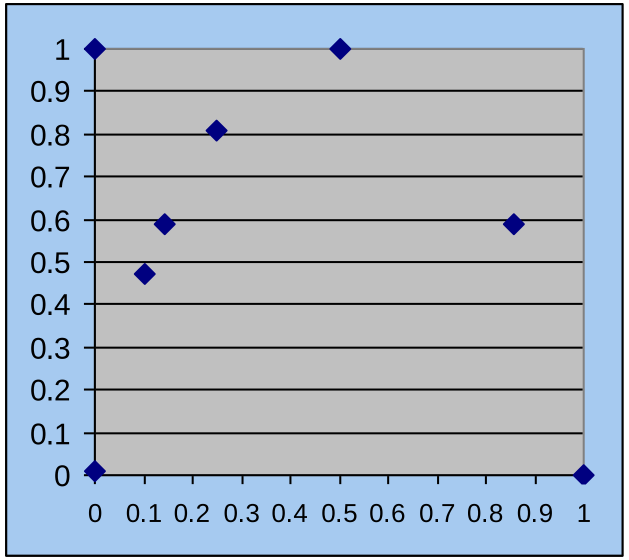
  E (1/4 , 3/4) = 0.81

  E (1/7 , 6/7) = 0.59

  E (6/7 , 1/7) = 0.59

  E (0.1 , 0.9) = 0.47

  E (0.001 , 0.999) = 0.01



Entropy

$$E(S) = -\sum_{c=1}^{N} p_c \cdot \log_2 p_c$$

- Calculate:

  E (0 , 1) = 0

  E (1/2 , 1/2) = 1

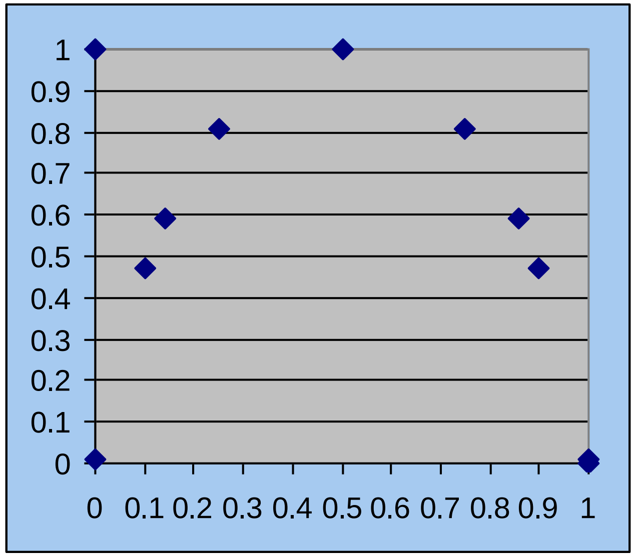
  E (1/4 , 3/4) = 0.81

  E (1/7 , 6/7) = 0.59

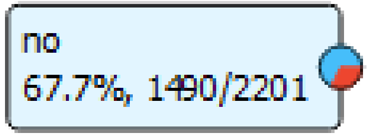
  E (6/7 , 1/7) = 0.59

  E (0.1 , 0.9) = 0.47

  E (0.001 , 0.999) = 0.01



Example: entropy of a dataset



Titanic survivers

- All passengers: 2201
- Survivers: 721

$$E(S) = -\sum_{c=1}^{N} p_c \cdot \log_2 \, p_c$$

- The entire dataset 2201 instances
- 1490 classifierd NO
- 721 classified YES

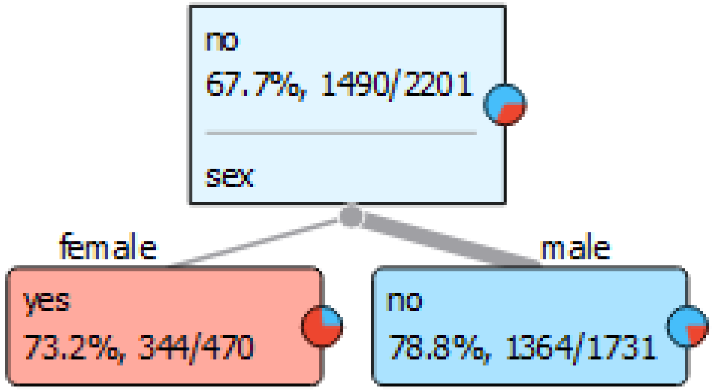
We compute the entropy

	NO	YES	total
	1490	721	2211
class probability	0.674	0.326	
pi * log (pi, 2)	-0.384	-0.527	
entropy	-0.911		

Information gain (of an attribute)

Information gain (IG) measures how much "information" a feature gives us about the class.

= How much the entropy is reduced by splitting the data according to the attribute



Information Gain

number of examples in the subset S_v

(probability of the branch)

set S

attribute A

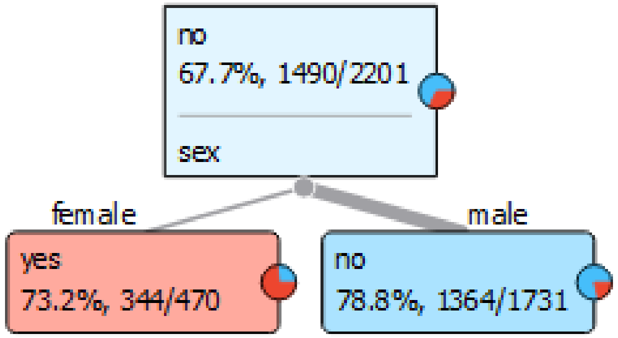
$$Gain(S, A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot E(S_v)$$

Entropy of the set S

Entropy of the subset S_v

number of examples in set S

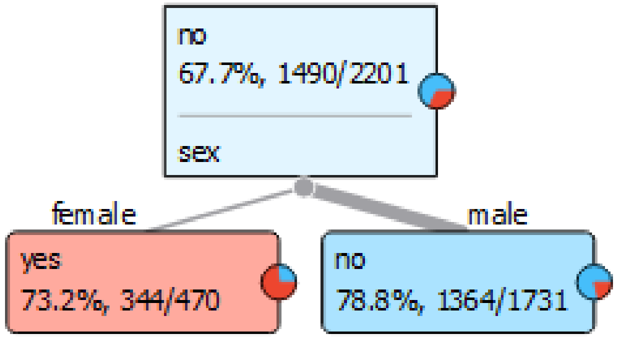
Information gain: example



1. Compute the entropy of the entire set
2. The attribute “sex” splits the dataset into two subsets :
   - **female** with 470 instances (344 survived)
   - **male** with 1731 instances (1364 died)
3. Compute the entropy of each subset
4. Compute the Information gain

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

Information gain: example

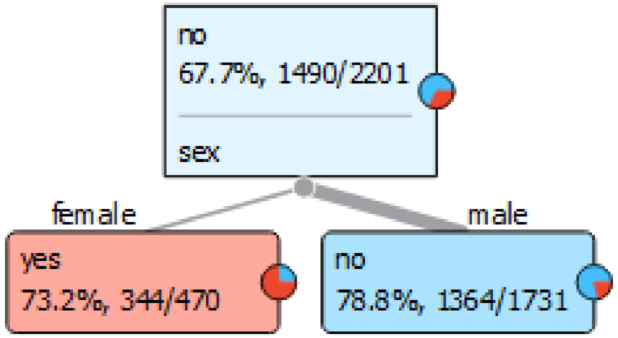


1. Compute the entropy of the entire set
2. The attribute “sex” splits the dataset into two subsets :
   - **female** with 470 instances (344 survived)
   - **male** with 1731 instances (1364 died)
3. Compute the entropy of each subset
4. Compute the Information gain

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

	NO	YES	total
	1490	720	2210
Class probability pi	0,674	0,326	
pi * log (pi, 2)	-0,38	-0,53	
entropy	0,911		

Information gain: example



1. Compute the entropy of the entire set
2. The attribute “sex” splits the dataset into two subsets :
   - **female** with 470 instances (344 survived)
   - **male** with 1731 instances (1364 died)
3. Compute the entropy of each subset
4. Compute the Information gain

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

female	NO	YES	total
	136	334	470
Class probability pi	0,289	0,711	
pi * log (pi, 2)	-0,52	-0,35	
entropy	0,868		

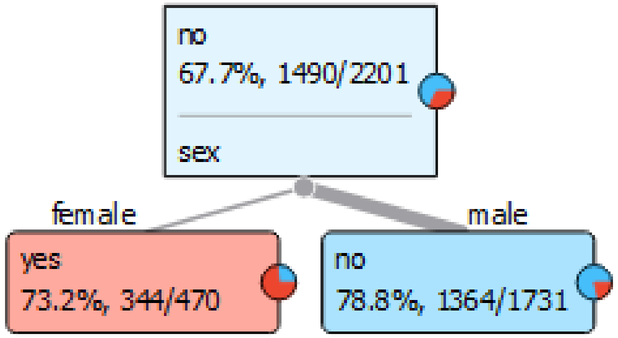
male	NO	YES	total
	1364	367	1731
Class probability pi	0,788	0,212	
pi * log (pi, 2)	-0,27	-0,47	
entropy	0,745		

Information gain: example

1. Compute the entropy of the entire set
2. The attribute "sex" splits the dataset into two subsets :
   - **female** with 470 instances (344 survived)
   - **male** with 1731 instances (1364 died)
3. Compute the entropy of each subset
4. Compute the Information gain

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

$$Gain\ (S, Sex) = 0{,}911\ - \left(\frac{470}{2201} * 0{,}868 + \frac{1731}{2201} * 0{,}745\right) = 0{,}166$$



female	NO	YES	total
	136	334	470
Class probability pi	0,289	0,711	
pi * log (pi, 2)	-0,52	-0,35	
entropy	0,868		

male	NO	YES	total
	1364	367	1731
Class probability pi	0,788	0,212	
pi * log (pi, 2)	-0,27	-0,47	
entropy	0,745		

TDIDT – Top Down Induction of Decision Trees

- We induce decision trees top-down
- There is many possible decision trees for a given dataset
- It is very important which attribute we choose as the root
- Heuristic: we choose the attribute which **best separates** the classes

↑

Information gain

↑

Entropy

Decision tree induction with the ID3 algorithm

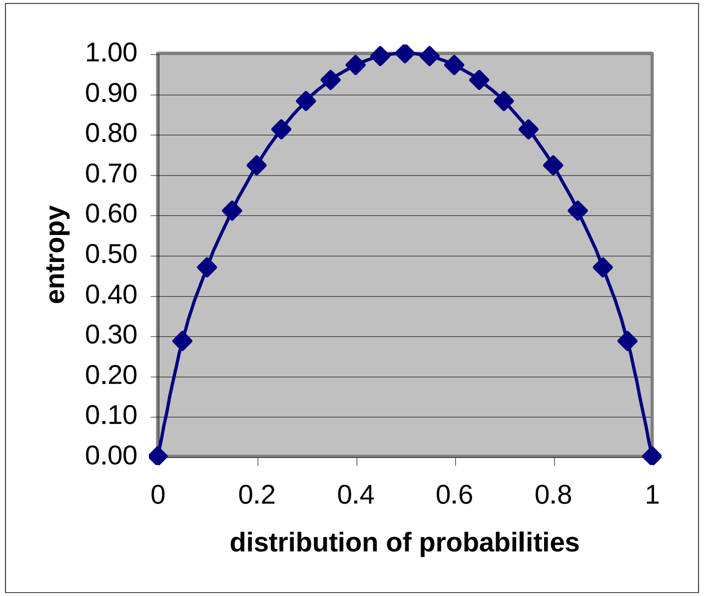
Induce a decision tree on set S:

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 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

Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81-106

Entropy and information gain

probability of class 1	probability of class 2	
p_1	**$p_2 = 1\text{-}p_1$**	**entropy $E(p_1, p_2)$ =** $-p_1 * \log_2(p_1) - p_2 * \log_2(p_2)$
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



number of examples in the subset

probability of the "branch"

attribut A

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

set S

number of examples in set S

Literature

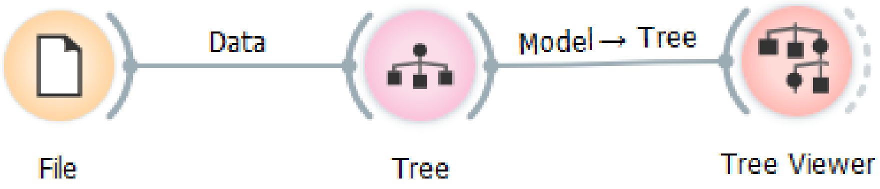
- Max Bramer: Principles of data mining (2007)
    1. Introduction to Data Mining
    2. Data for Data Mining
    3. Using Decision trees for Classification
    4. Decition Tree Induction: Using Entropy for Attribute Selection
    9. More About Entropy

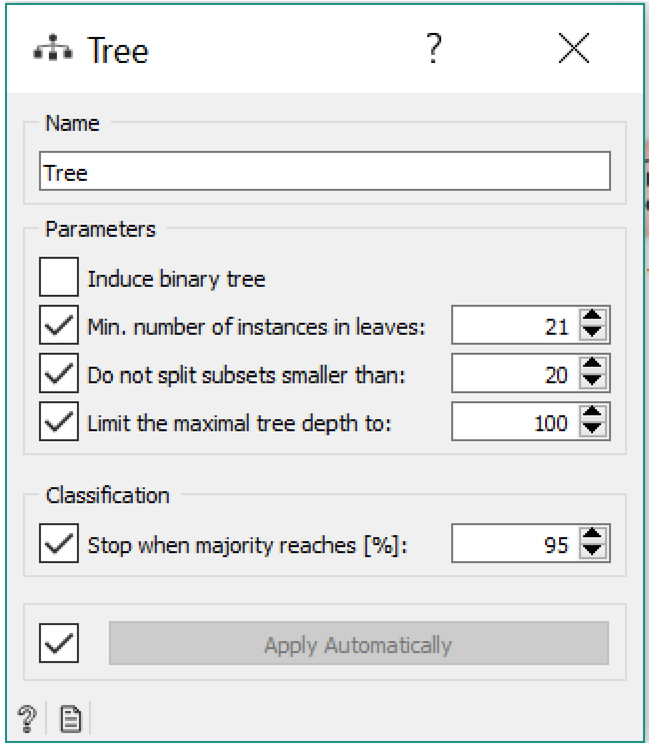
    - Appendix A: Essential Mathematics

Lab exercise 2

Decision trees in Orange

Exercise 1: Induce a decision tree

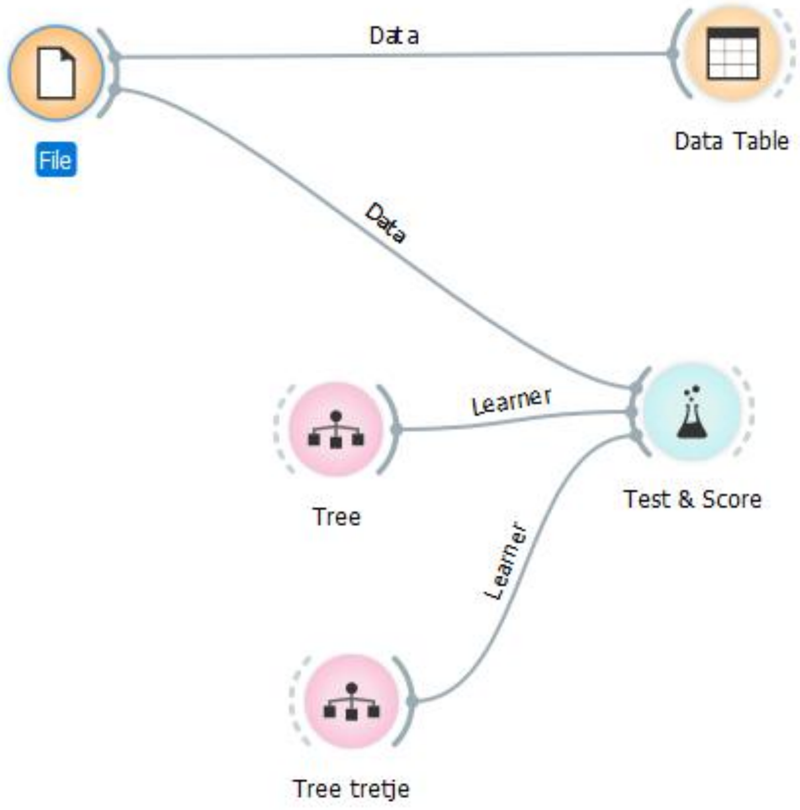




- Dataset: “titanic”
- Play with tree parameters

- Repeat with the “adult” dataset

Exercise 2: Evaluate the decision tree



- Dataset: “zoo”
- Compare tree classifiers with different parameter values