Week 3: Classification

Data Analysis

Leo Hofste & Martin Wesselink

28-02-2020

SAXION UNIVERSITY OF APPLIED SCIENCES

Agenda

1. Recap of week 2

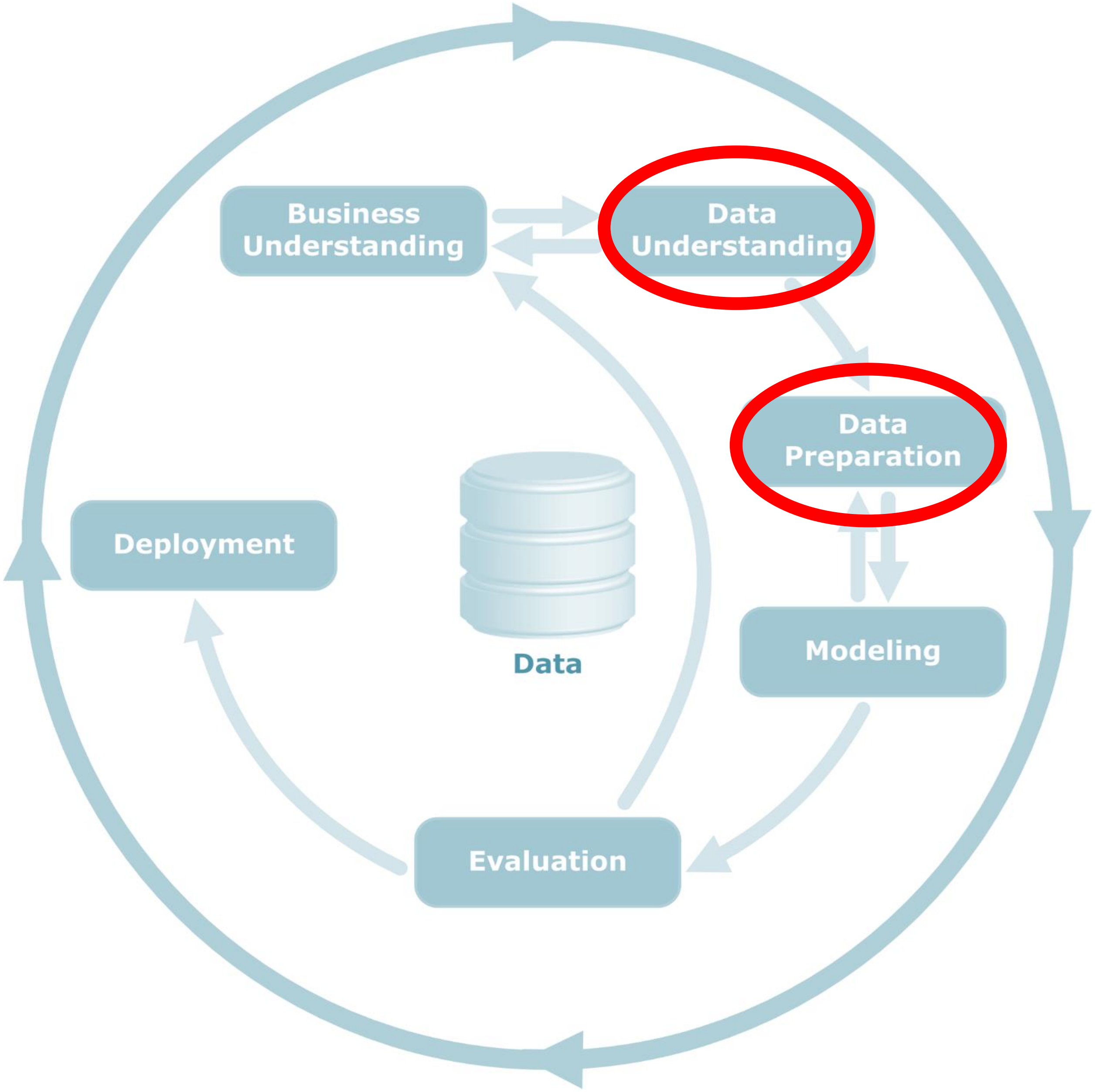
2. Data Modeling

    Supervised learning with classification

3. Exercises

RECAP WEEK 2

CRISP-DM



SAXION UNIVERSITY OF APPLIED SCIENCES

Scania Example

CRISP-DM

Data Understanding

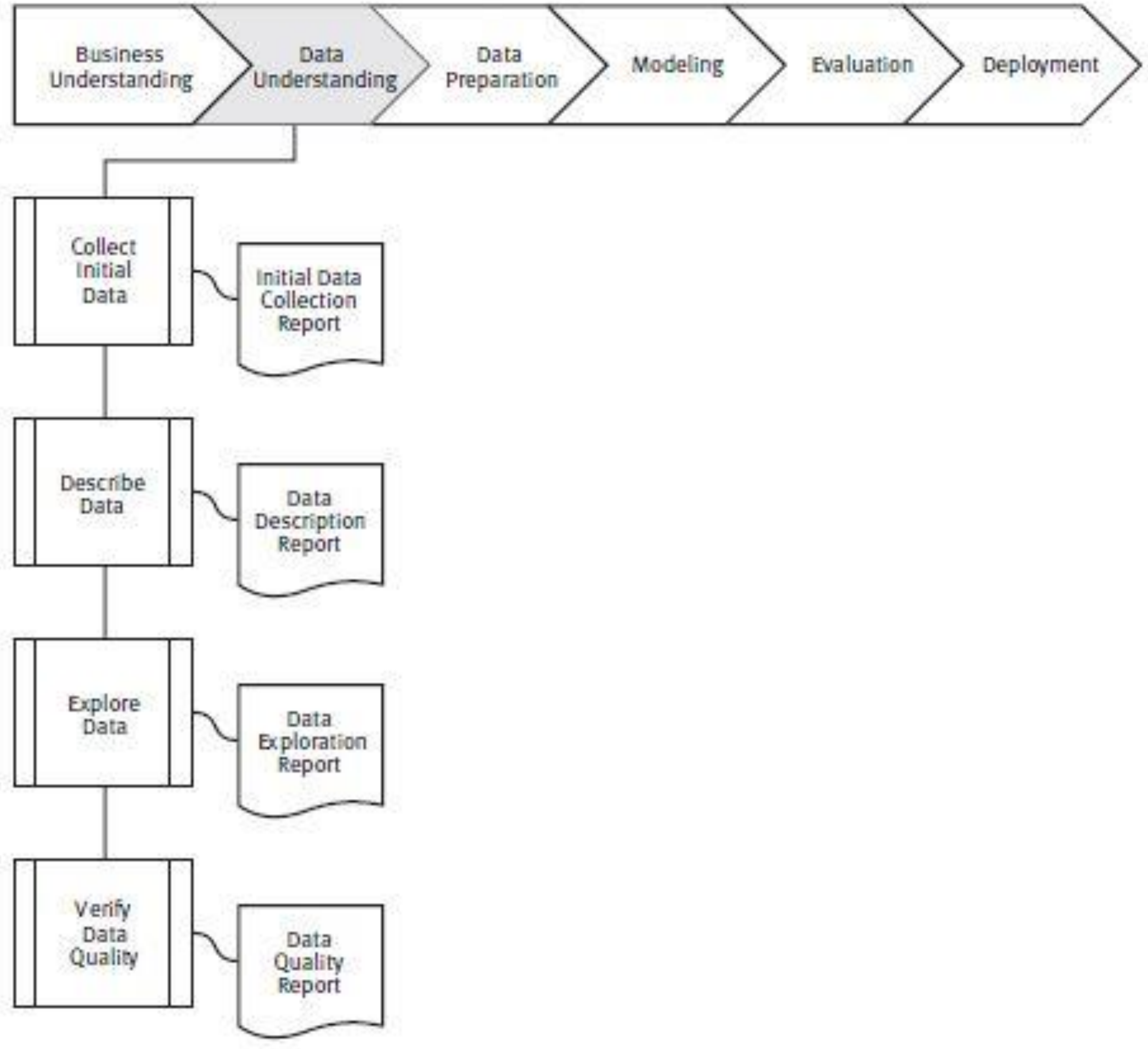
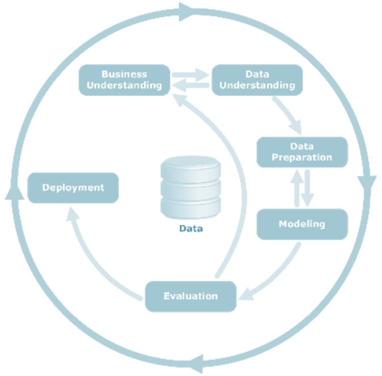


Figure 5: Data understanding

Source: http://crisp-dm.eu/



SAXION UNIVERSITY OF APPLIED SCIENCES

CRISP-DM

Data Understanding

Data exploration

Types of variables

Ratio variable

Nominal variable

Ordinal variable

Interval variable

Name	Age	Gender	Student house	Sport	Interested automation(1-5)	Temperature
Bart	18	M	Yes	Soccer	1	40
Lisa	19	F	Yes	Tennis	2	0
Erik	17	M	No	Soccer	4	-4
Anne	18	F	No	Fitness	5	38
Riek	18	M	No	No	2	-36

First interval/ratio then ordinal then nominal

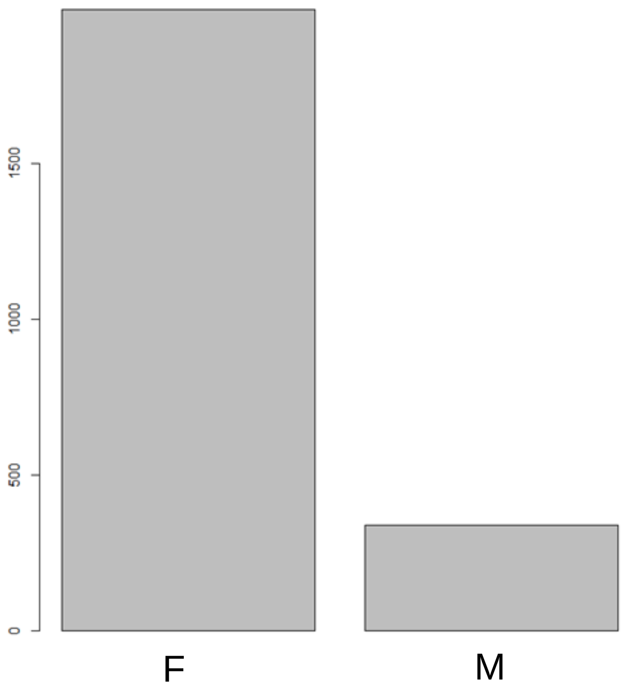
CRISP-DM

Data Understanding

Data exploration

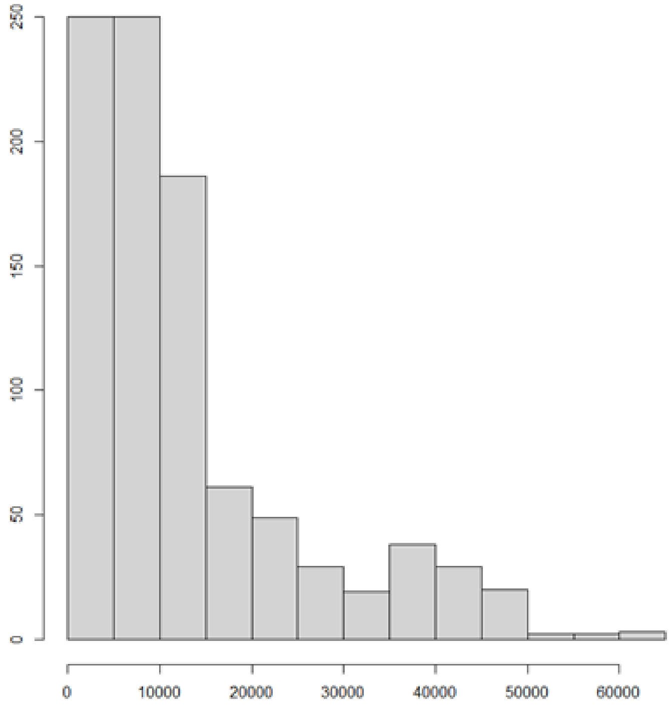
Visualization of the distribution of a variable

Categorical variable



Barchart

Numerical variable



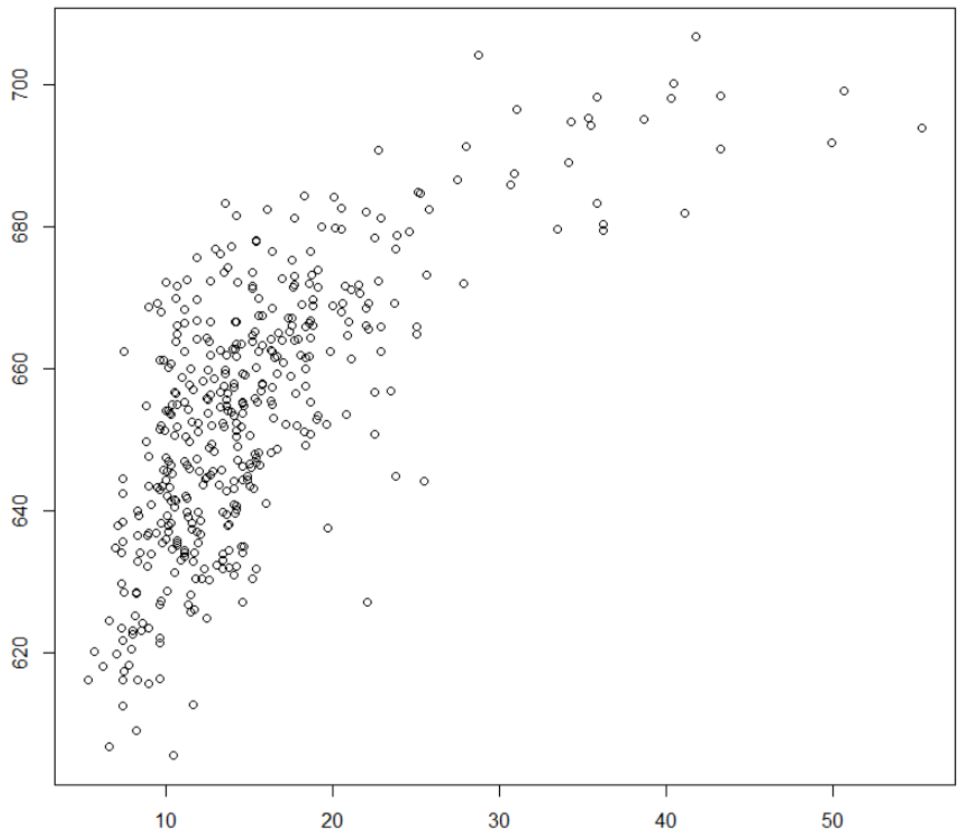
Histogram

CRISP-DM

Data Understanding

Data exploration

Scatterplot and Pearson’s r

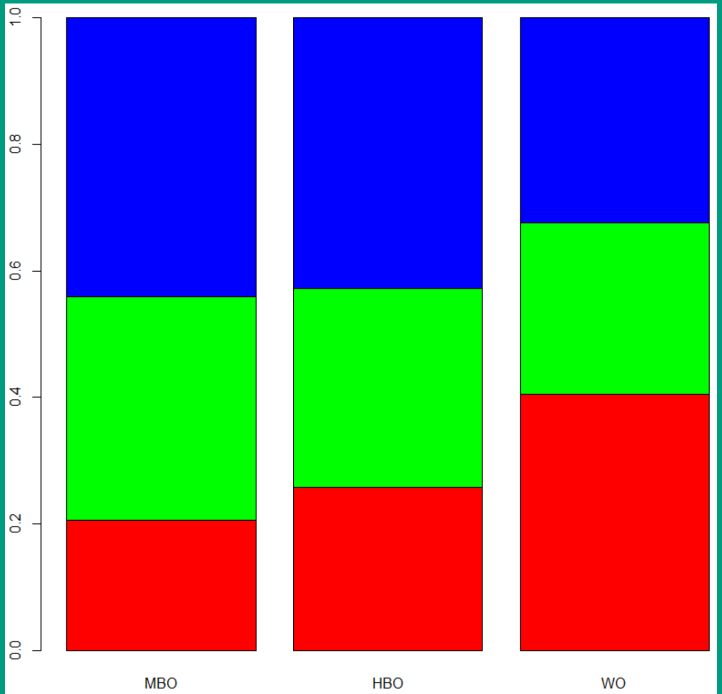


SAXION UNIVERSITY OF APPLIED SCIENCES

Data exploration

CRISP-DM

Data Understanding



Pivot tables/Contingency tables: Correlation

Research among 141 working people
Independent variable: education
Dependent variable: income

Dependent variable

Independent variable

case B

	EDUCATION			
INCOME	**MBO**	**HBO**	**WO**	**TOTAL**
< 3.000	**15**	**30**	**12**	**57**
3.000 - < 5.000	**12**	**22**	**10**	**44**
> 5.000	**7**	**18**	**15**	**40**
TOTAL	**34**	**70**	**37**	**141**

If the independent groups are in the columns, then percentage must be given VERTICALLY.

Compare HORIZONTAL means calculating VERTICAL percentages

SAXION UNIVERSITY OF APPLIED SCIENCES

Data exploration

CRISP-DM

Data Understanding

Cramers ‘V

- Full coherence➔ After calculating vertical percentage, the differences in the horizontal direction are maximal
- Independence➔ After calculating vertical percentage, there are no differences in the horizontal direction
- Cramer has developed a formula to calculate the strength of cohesion or so called V

V=0	**Indepedence**
V ≈ 0,10	**Weak coherence**
V ≈ 0,25	**Reasonable coherence**
V ≈ 0,50	**Strong coherence**
V ≈ 0,75	**Very strong coherence**
V=1	**Full coherence**

SAXION UNIVERSITY OF APPLIED SCIENCES

CRISP-DM

Data Understanding

Data exploration

Cramers ‘V

Cramer has developed a formula to calculate the strength of cohesion or so called V

INCOME	EDUCATION			
	MBO	HBO	WO	TOTAL
< 3.000	15	30	12	57
3.000 - < 5.000	12	22	10	44
> 5.000	7	18	15	40
TOTAL	34	70	37	141

$$V = \sqrt{\frac{\varphi^2}{\varphi^2_{max}}} = \sqrt{\frac{0,05}{2}} = \sqrt{0.025} = 0,16$$

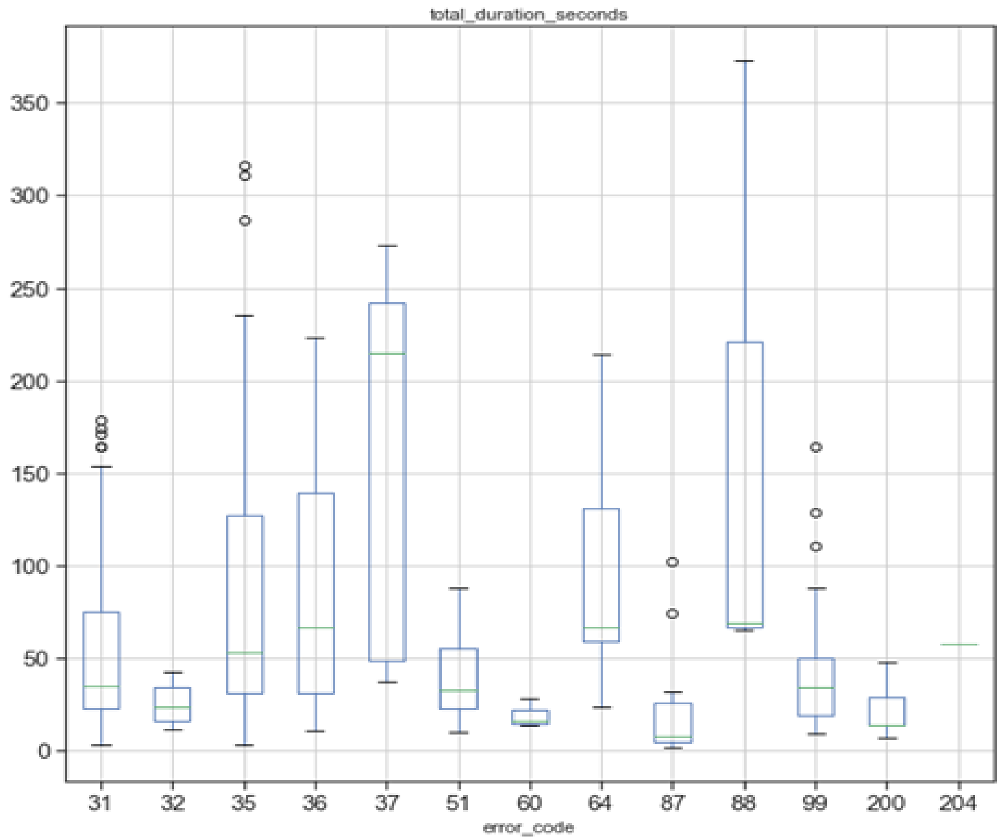
Data exploration

CRISP-DM

Data Understanding

Boxplots

Data distribution of total duration per second of each error



SAXION UNIVERSITY OF APPLIED SCIENCES

SCANIA

Data exploration

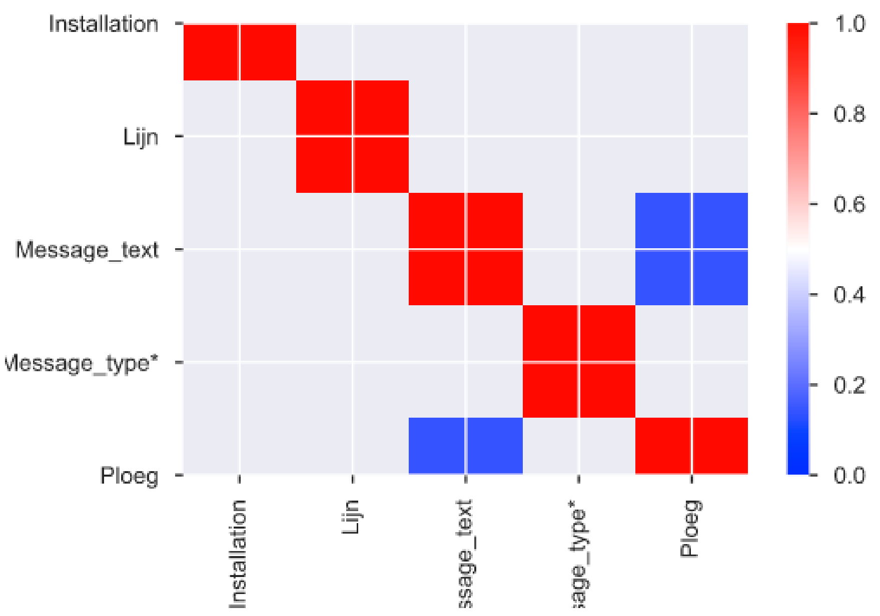
CRISP-DM

Data Understanding

Cramers ‘V

Correlations

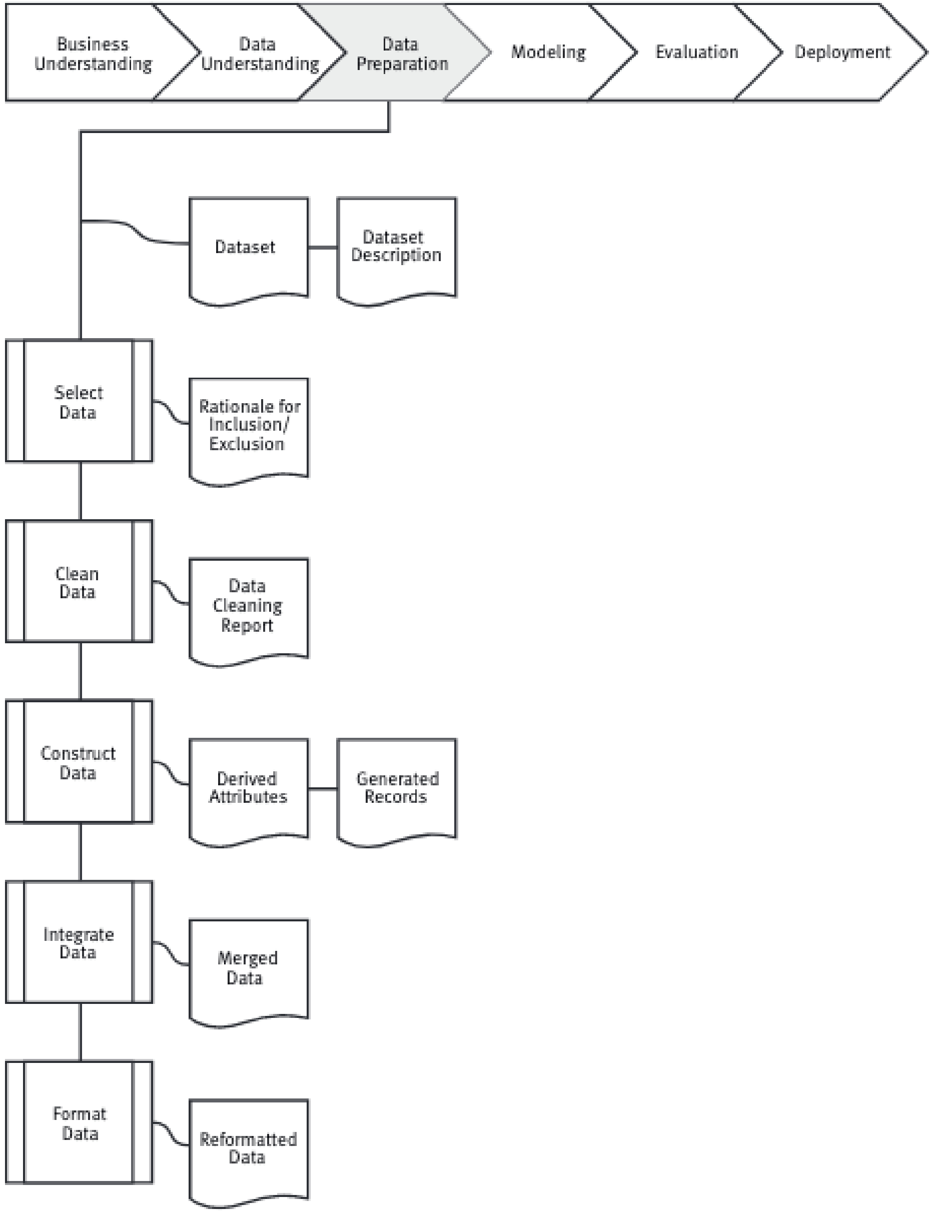
Pearson's r | Spearman's ρ | Kendall's τ | Phik (φk) | Cramér's V (φc) | Recoded



CRISP-DM

Data Preparation

Data preparation



CRISP-DM

Data Preparation

Data preparation

Select data

TABLE 2.1 A Summary of Data Preprocessing Tasks and Potential Methods

Main Task	Subtasks	Popular Methods
Data consolidation	Access and collect the data	SQL queries, software agents, Web services.
	Select and filter the data	Domain expertise, SQL queries, statistical tests.
	Integrate and unify the data	SQL queries, domain expertise, ontology-driven data mapping.
Data cleaning	Handle missing values in the data	Fill in missing values (imputations) with most appropriate values (mean, median, min/max, mode, etc.); recode the missing values with a constant such as "ML"; remove the record of the missing value; do nothing.
	Identify and reduce noise in the data	Identify the outliers in data with simple statistical techniques (such as averages and standard deviations) or with cluster analysis; once identified, either remove the outliers or smooth them by using binning, regression, or simple averages.
	Find and eliminate erroneous data	Identify the erroneous values in data (other than outliers), such as odd values, inconsistent class labels, odd distributions; once identified, use domain expertise to correct the values or remove the records holding the erroneous values.
Data transformation	Normalize the data	Reduce the range of values in each numerically valued variable to a standard range (e.g., 0 to 1 or −1 to +1) by using a variety of normalization or scaling techniques.
	Discretize or aggregate the data	If needed, convert the numeric variables into discrete representations using range- or frequency-based binning techniques; for categorical variables, reduce the number of values by applying proper concept hierarchies.
	Construct new attributes	Derive new and more informative variables from the existing ones using a wide range of mathematical functions (as simple as addition and multiplication or as complex as a hybrid combination of log transformations).
Data reduction	Reduce number of attributes	Principal component analysis, independent component analysis, chi-square testing, correlation analysis, and decision tree induction.
	Reduce number of records	Random sampling, stratified sampling, expert-knowledge-driven purposeful sampling.
	Balance skewed data	Oversample the less represented or undersample the more represented classes.

SAXION UNIVERSITY OF APPLIED SCIENCES

MODELING: SUPERVISED LEARNING WITH CLASSIFICATION

CRISP-DM

Business Understanding
Data Understanding
Data Preparation
Deployment
Data
Modeling
Evaluation

Machine learning Paradigms

CRISP-DM

Modeling

Data Mining: learning a model from data

Supervised learning:
- Uses labelled data
- The goal is to learn to accurately predict the label from the data

Unsupervised learning:
- Uses un-labelled data
- The goal is to learn natural structure present within data

Reinforcement learning:
- Uses actions and a reward function
- The goal is to learn what sequence of actions maximizes reward
- RL is a special area in data science (not covered in this course!)

CRISP-DM

Data Mining Tasks & Methods	Data Mining Algorithms	Learning Type
Prediction		
Classification	Decision Trees, Neural Networks, Support Vector Machines, kNN, Naïve Bayes, GA	Supervised
Regression	Linear/Nonlinear Regression, ANN, Regression Trees, SVM, kNN, GA	Supervised
Time Series	Autoregressive Methods, Averaging Methods, Exponential Smoothing, ARIMA	Supervised
Association		
Market-basket	Apriory, OneR, ZeroR, Eclat, GA	Unsupervised
Link analysis	Expectation Maximization, Apriory Algorithm, Graph-based Matching	Unsupervised
Sequence analysis	Apriory Algorithm, FP-Growth, Graph-based Matching	Unsupervised
Segmentation		
Clustering	K-means, Expectation Maximization (EM)	Unsupervised
Outlier analysis	K-means, Expectation Maximization (EM)	Unsupervised

Data Mining Tasks & Techniques

Supervised learning

- **Classification** — Logistic regression, Decision tree, Naïve Bayes, Random Forest, Support Vector Machines, Neural Networks, Case-Based Reasoning, Genetic algorithms, Rough sets
- **Regression** — (non-)Linear regression, Regression Trees, Random Forest, Support Vector Machines, Neural Networks
- **Time series** — ARIMA

Unsupervised learning

- **Clustering** — k-Means, Hierarchical
- **Association** — Apriori, Eclat, FP-Growth

Supervised learning with Classification

CRISP-DM

Modeling

Classification

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation.

Explanation in this youtube-film:
https://www.youtube.com/watch?v=q1AwkzJ9IeM

Source: https://medium.com/@sifium/machine-learning-types-of-classification-9497bd4f2e14

Supervised learning with Classification

CRISP-DM

Modeling

Classification

- Most frequently used DM task
- Part of the machine-learning family

- Learn from ‘past’ data, classify new data
- The output variable is categorical (nominal or ordinal) in nature
  - Nominal: {‘man’, ‘woman’}
  - Ordinal: {‘gold’, ‘silver’, ‘bronze’}
  - Ordinal: {‘10.000-20.000’, ‘20.000-40.000’, ‘40.000-100.000’}

- **Better:** the probability of each output value
- For example, Pr(Y=man |..) = .92 and Pr(Y=woman |..) = .08
- ~~Decision trees are probabilistic classifiers (as well as logistic regression, Naïve Bayes, Random Forest)~~

SAXION UNIVERSITY OF APPLIED SCIENCES

Supervised learning with Classification

CRISP-DM

Modeling

Example of Classification: Iris

Tree types of irises

Iris-Setosa; Iris-versicolor, Iris-virginica

Variables:
Sepal length(cm); .....kelk lengte(cm); veldtype numeric
Sepal width(cm); .....kelk breedte(cm); veldtype: numeric
Petal lenghth(cm);....bloemblad lengte(cm); veldtype: numeric
Petal width(cm);......bloemblad breedte(cm); veldtype: numeric

We want to investigate of these four properties are enough to decide if we can categorize an iris in one of these three family types.
This is a ‘Supervised learning process’

- We have data(150 records) in an excel-file:

Sepal lenght	sepal width	petal length	petal width	type
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3,0	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa
5,0	3,6	1,4	0,2	Iris-setosa
5,4	3,9	1,7	0,4	Iris-setosa
4,6	3,4	1,4	0,3	Iris-setosa
5,0	3,4	1,5	0,2	Iris-setosa
4,4	2,9	1,4	0,2	Iris-setosa

CRISP-DM

Modeling

Supervised learning with Classification

Example of Classification: play tennis

We already know from these data if we will play or not.
We will predict if we are playing with certain conditions.

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

The target variable

Supervised learning with Classification

CRISP-DM

Modeling

Example of Classification: photos

- An algorithm that classifies a photo in the right class;
  - Given classes: {“day”, “night”}
  - We have a new photo, and the algorithm gives the right class

Supervised learning with Classification

CRISP-DM

Modeling

Classification Techniques

- Statistical analysis (logistic regression, linear discriminant analysis)
- Decision trees
- Bayesian classifiers
- Support vector machines
- Case-based reasoning (k-Nearest Neighbors)
- Neural networks
- Genetic algorithms
- Rough sets

SAXION UNIVERSITY OF APPLIED SCIENCES

Supervised learning with Classification

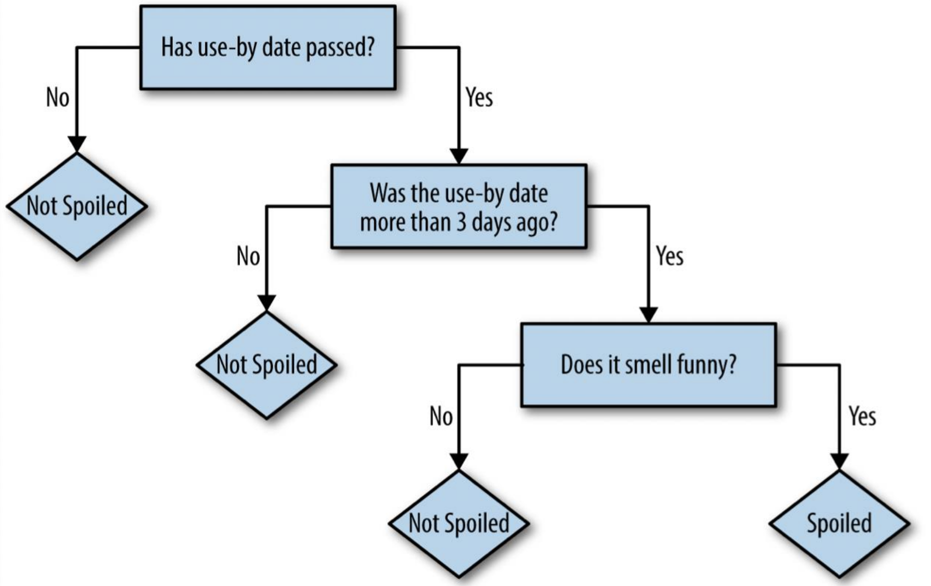
CRISP-DM

Modeling

Decision Trees

To solve a problem, you just have to ask the right questions

Representation of a decision tree:
- Every internal node has a question
- Every branch represents an answer
- Every leaf represents a decision



Supervised learning with Classification

CRISP-DM

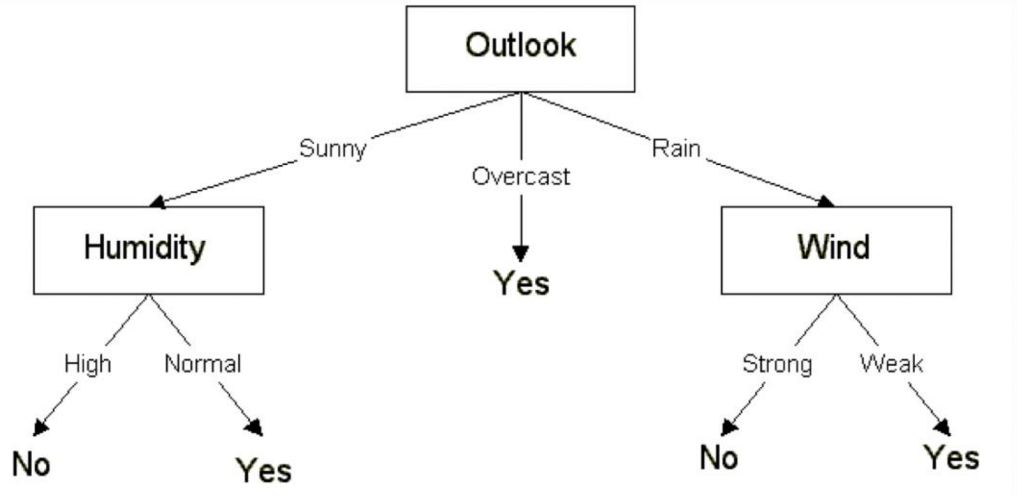
Modeling

Decision Trees

Example: “Will you play tennis?”

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

The target variable



Supervised learning with Classification

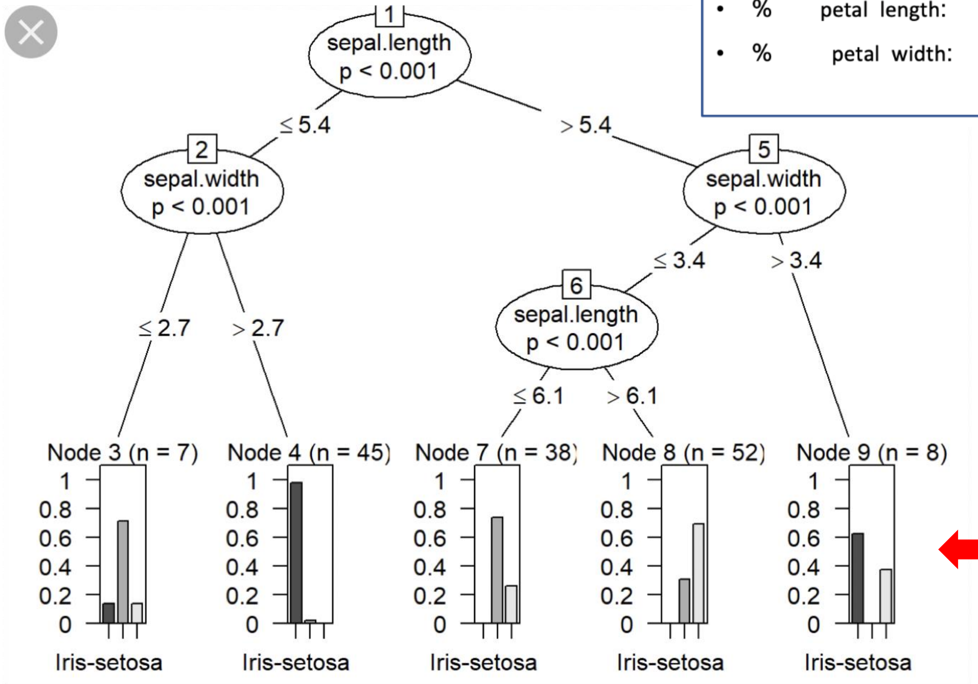
CRISP-DM

Modeling

Decision Trees

Example: Iris flower set
With two variable ‘sepal length’ and ‘sepal width’

	%	Summary Statistics:%	Min	Max	Mean	SD	Class Correlation	
•	%	sepal length:	4.3	7.9	5.84	0.83	0.7826	
•	%	sepal width:	2.0	4.4	3.05	0.43	-0.4194	
•	%	petal length:	1.0	6.9	3.76	1.76	0.9490	(high!)
•	%	petal width:	0.1	2.5	1.20	0.76	0.9565	(high!)



Observations split in pure leaf nodes, i.e. mainly belonging to one category

Supervised learning with Classification

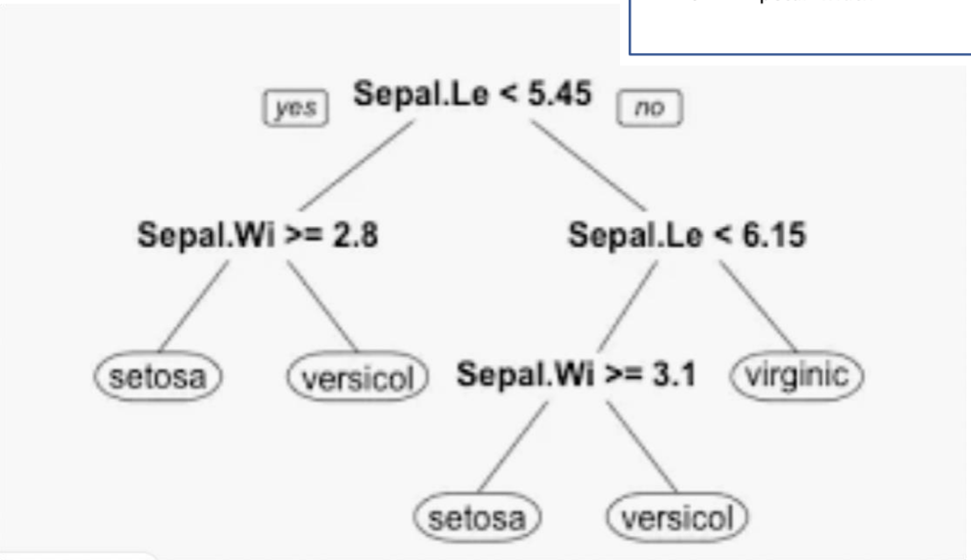
CRISP-DM

Modeling

Decision Trees

Example: Iris flower set
With two variable ‘sepal length’ and ‘sepal width’

% Summary Statistics:%	Min	Max	Mean	SD	Class Correlation	
% sepal length:	4.3	7.9	5.84	0.83	0.7826	
% sepal width:	2.0	4.4	3.05	0.43	-0.4194	
% petal length:	1.0	6.9	3.76	1.76	0.9490	(high!)
% petal width:	0.1	2.5	1.20	0.76	0.9565	(high!)



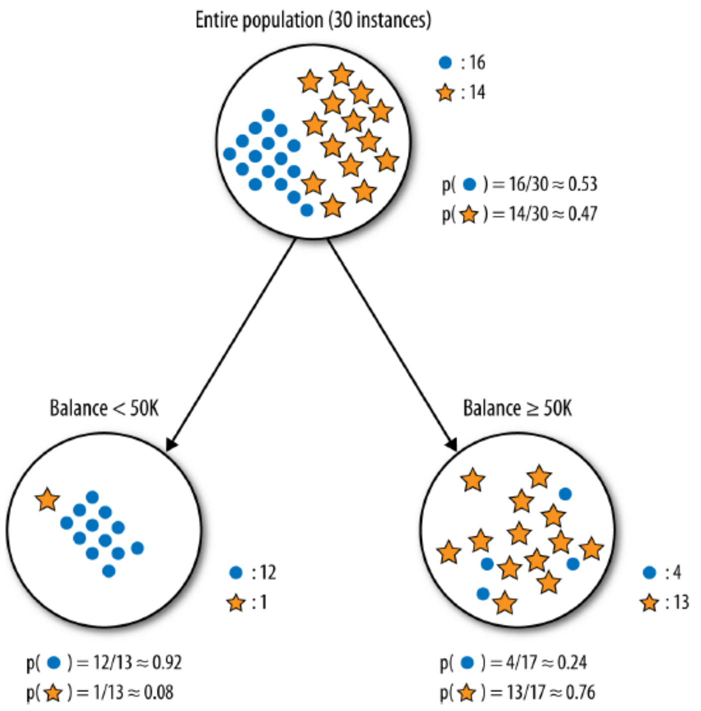
Supervised learning with Classification

CRISP-DM

Modeling

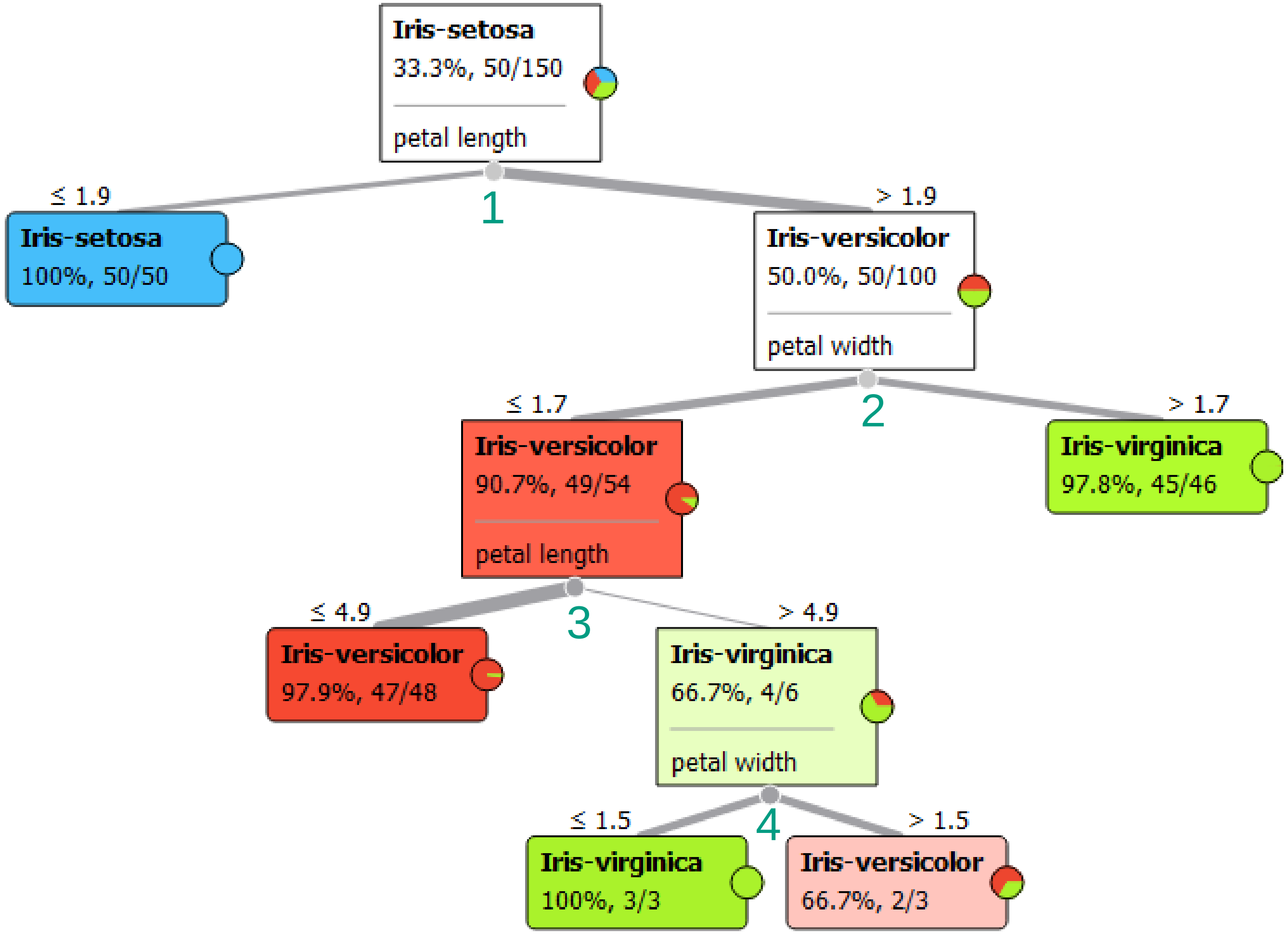
Decision Tree Algorithm

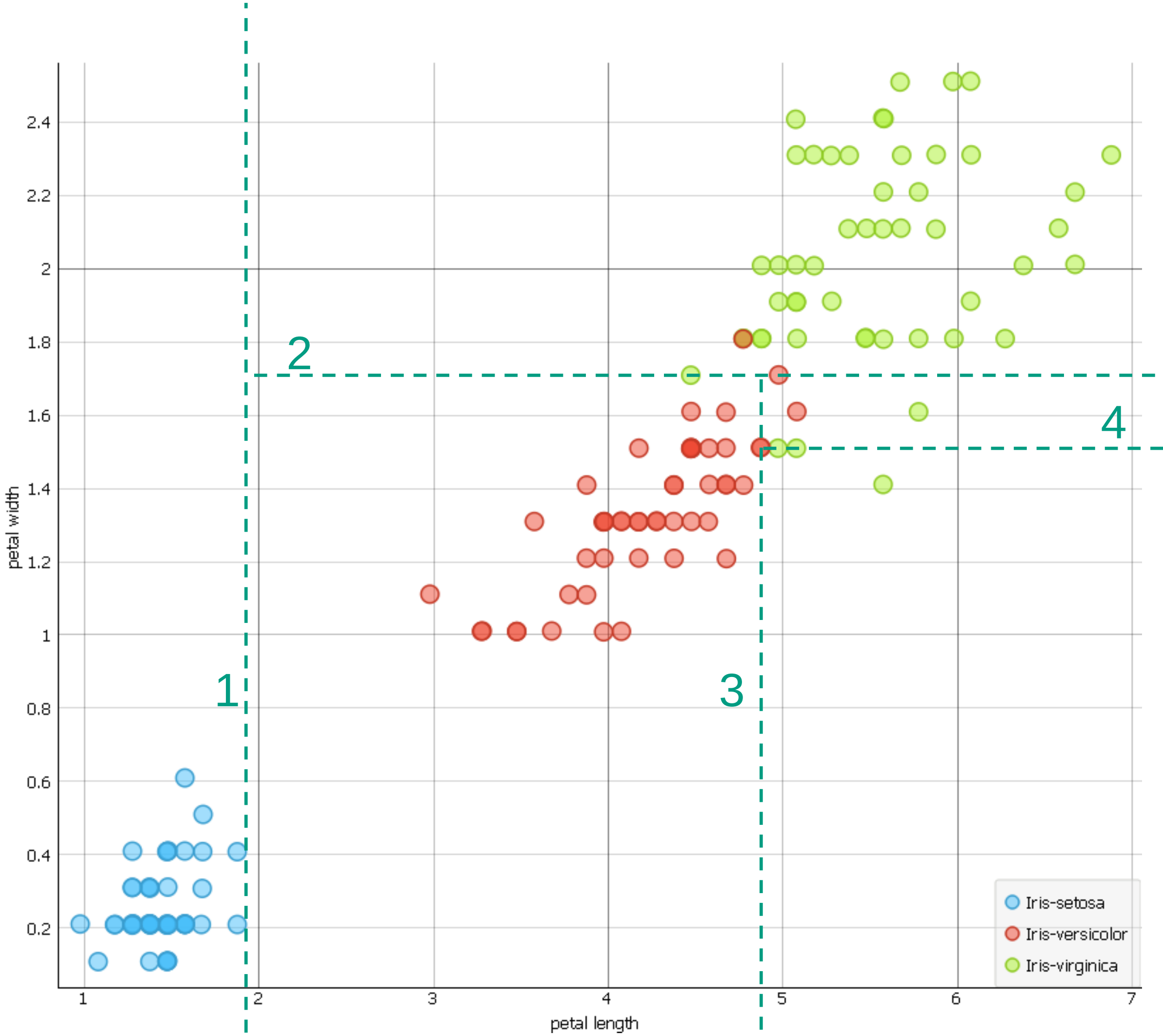
The DT algorithm recursively splits data to improve purity



SAXION UNIVERSITY OF APPLIED SCIENCES

Exercise 1 - Tree





Supervised learning with Classification

CRISP-DM

Modeling

Decision Tree Algorithm

General algorithm (steps) for building a decision tree:

1. Create a root node and assign all training data to it
2. Each leaf node is split into two leaf nodes, choosing the predictor variable and the cutpoint value that most improves purity, if such a split is possible
3. Repeat step 2 until no improvement in purity is possible or a stopping criterion is reached (max. tree height, min. number of observations in a leaf node)

4. To estimate class probabilities the class proportions of the leaf nodes are used

SAXION UNIVERSITY OF APPLIED SCIENCES

Supervised learning with Classification

CRISP-DM

Modeling

Decision Tree Algorithm

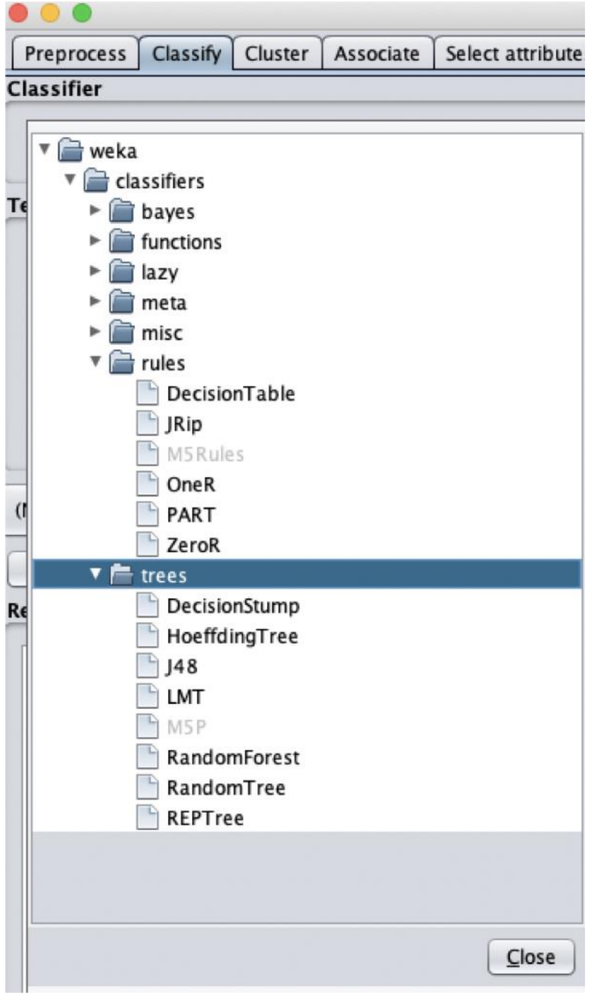
DT algorithms mainly differ on;

1. Splitting criteria
   Which variable, what value, etc.

2. Stopping criteria
   When to stop building the tree

3. Pruning (generalization method)
   Pre-pruning versus post-pruning

Most popular DT algorithms include:
ID3, C4.5, C5; CART; CHAID; M5



Supervised learning with Classification

CRISP-DM

Modeling

Properties of DT model and algorithm

- Trees (unless very large) are easily interpreted, even by non-experts
- Trees can model complex interactions

- The algorithm can deal with missing values
- The algorithm is prone to overfitting (solved by pruning)

- Imbalanced classes (which are common in classification) require special treatment to correctly classify the minority class

Supervised learning with Classification

CRISP-DM

Modeling

Assessment of predictive models

- Predictive performance: ability to correctly classify out-of-sample data
- Interpretability: level of understanding and insight provided by model
- Robustness: ability to 'reasonably accurately' classify noisy data

Assessment of predictive algorithms

- Speed: of model building and usage speed
- Scalability: ability to build and use a model with a growing amount of data

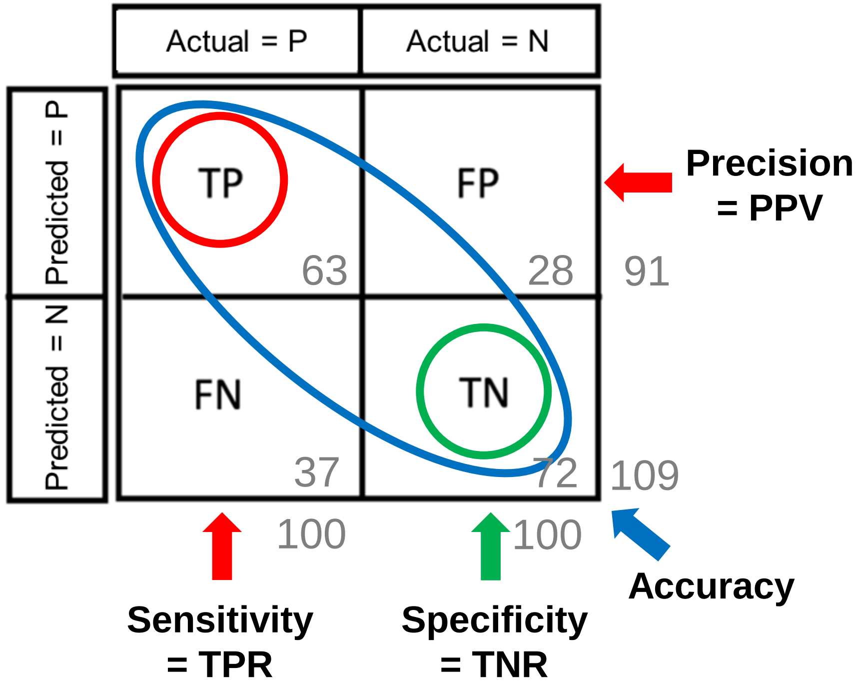
CRISP-DM

Modeling

Supervised learning with Classification

Predictive performance of classification models

In classification problems, the primary source for accuracy estimation is the confusion matrix



$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Supervised learning with Classification

CRISP-DM

Modeling

Examples of confusion matrices

Let us look into four prediction results from 100 positive and 100 negative instances:

A			B			C			C′		
TP=63	FP=28	91	TP=77	FP=77	154	TP=24	FP=88	112	TP=76	FP=12	88
FN=37	TN=72	109	FN=23	TN=23	46	FN=76	TN=12	88	FN=24	TN=88	112
100	100	200	100	100	200	100	100	200	100	100	200
TPR = 0.63			TPR = 0.77			TPR = 0.24			TPR = 0.76		
FPR = 0.28			FPR = 0.77			FPR = 0.88			FPR = 0.12		
PPV = 0.69			PPV = 0.50			PPV = 0.21			PPV = 0.86		
F1 = 0.66			F1 = 0.61			F1 = 0.23			F1 = 0.81		
ACC = 0.68			ACC = 0.50			ACC = 0.18			ACC = 0.82		

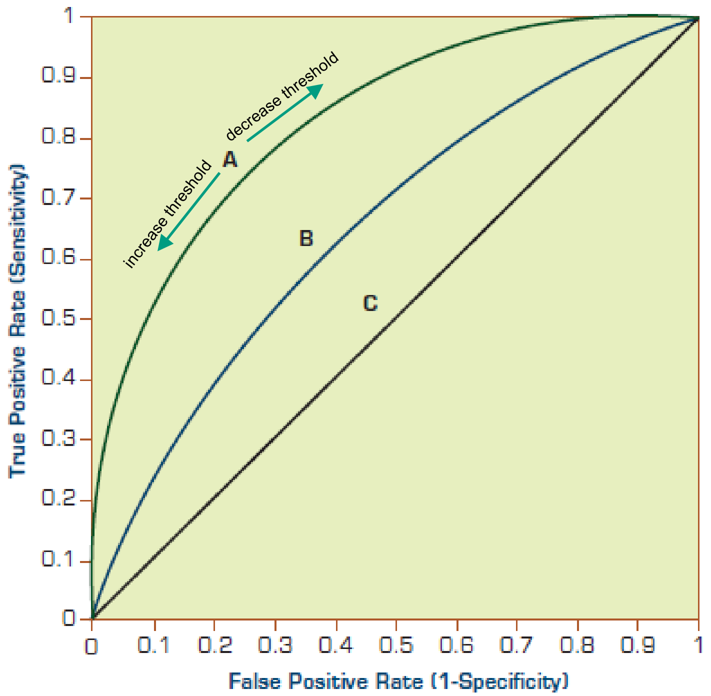
Supervised learning with Classification

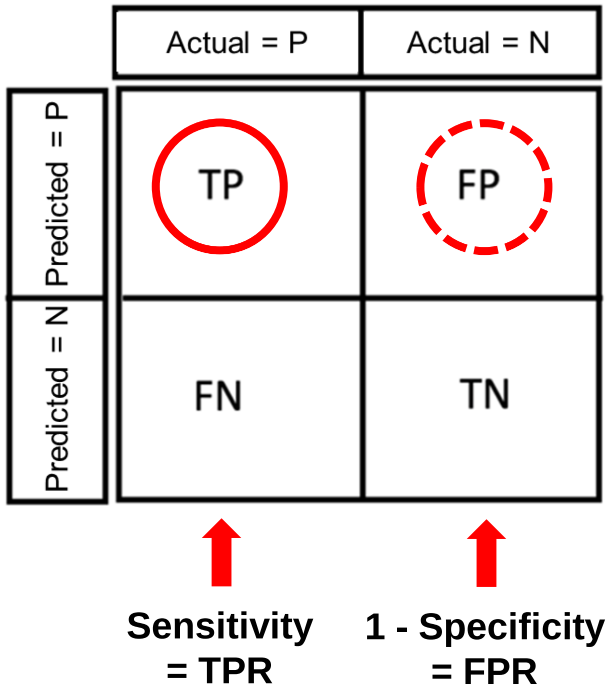
CRISP-DM

Modeling

ROC Curve

- Works with binary probabilistic classifiers
- Created by varying the threshold p-value that determines whether observation are classified as positive (usually 0.5)
- Can be used to find the optimal threshold
- Can be used to compare classifiers (f.i. A and B, C is the random guessing line)





Supervised learning with Classification

CRISP-DM

Modeling

Methods for evaluating model accuracy

- Simple split: split dataset in a training set and a test set
- k-Fold Cross Validation: randomly split dataset in k folds for k times training and testing of a model and averaging accuracy measures
- Leave-one-out: similar to *k*-fold where *k* = number of observations
- Bootstrapping: random sampling (with replacement)
- Jackknifing: similar to leave-one-out

These methods can be used to evaluate the predictive accuracy of both classification and regression models

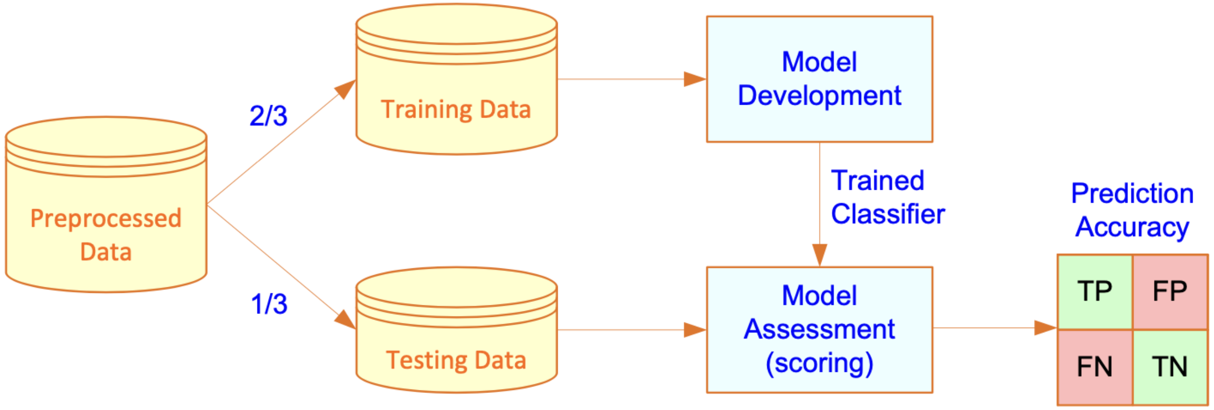
Supervised learning with Classification

CRISP-DM

Modeling

Model accuracy using Simple Split

- Split the data into 2 mutually exclusive sets: training(~70%) and testing (30%)



- For NeuralNetworks, the data is split into three subsets (training [~60%], validation [~20%], testing [~20%])

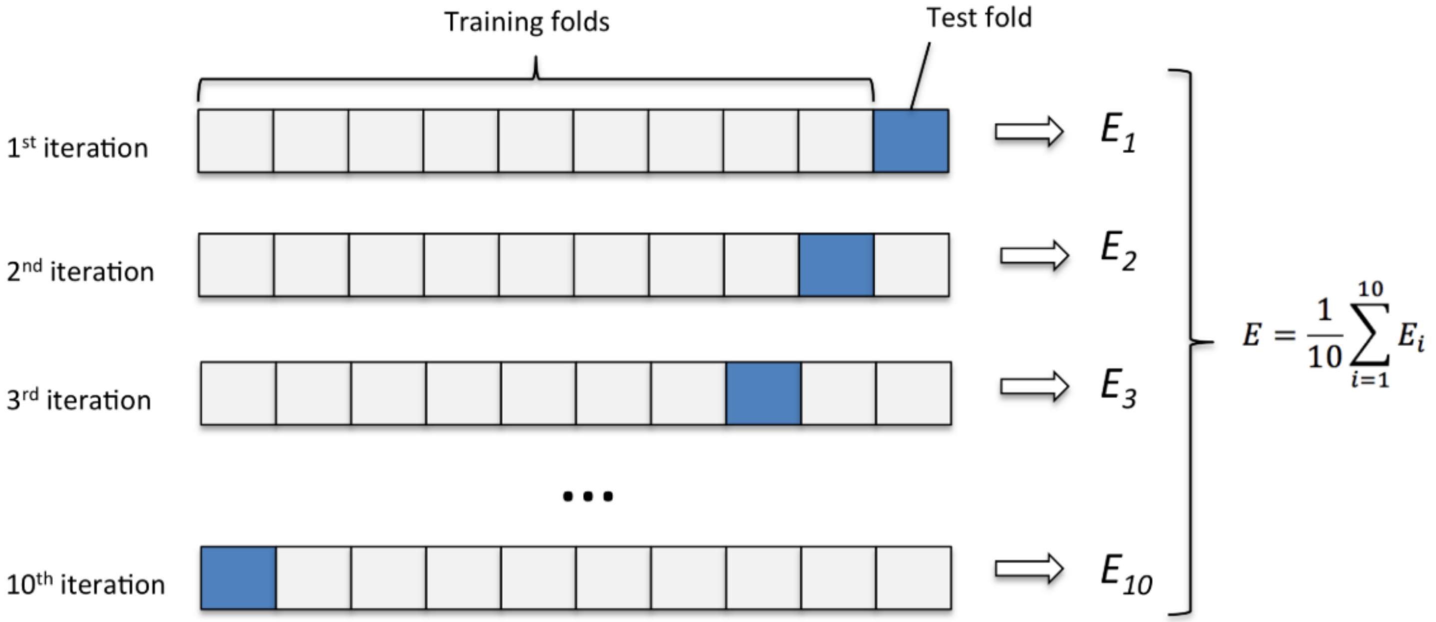
Supervised learning with Classification

CRISP-DM

Modeling

Model accuracy using k-Fold Cross-Validation

- The complete dataset is used as training set to build the model
- The dataset is also randomly partitioned into k folds
- In k iterations, each fold is retained as test set, the remaining k-1 folds are used as training set



SAXION UNIVERSITY OF APPLIED SCIENCES

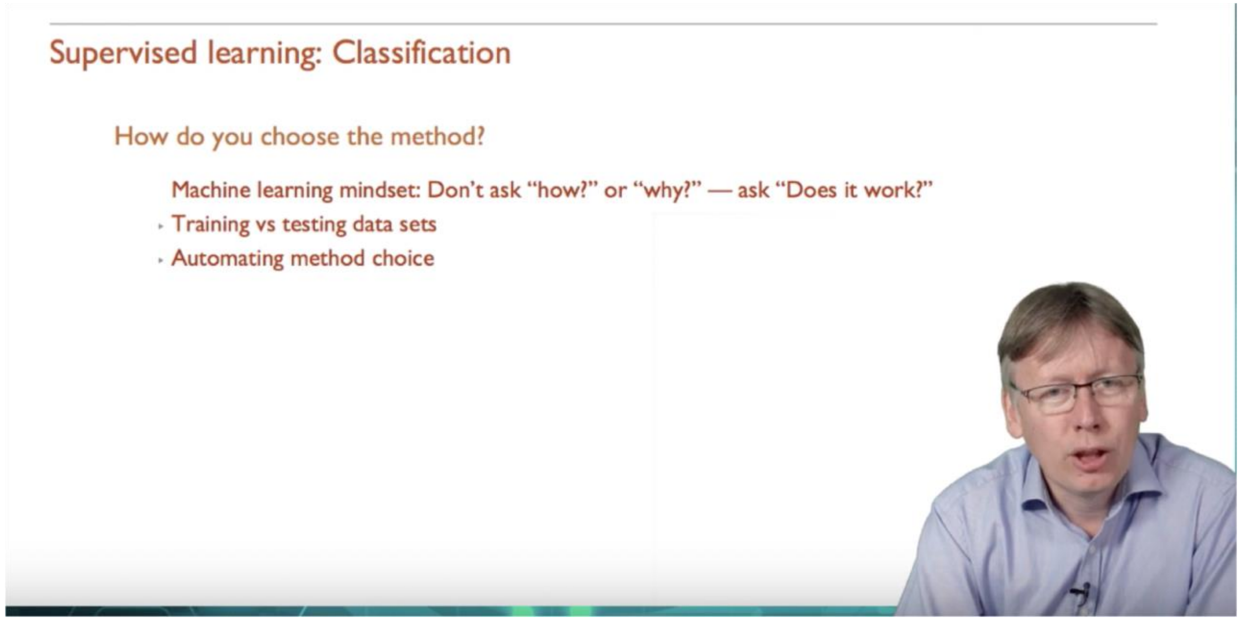
Supervised learning with Classification

CRISP-DM

Modeling

Classification Techniques

Which technique is the best?



https://www.youtube.com/watch?v=q1AwkzJ9IeM

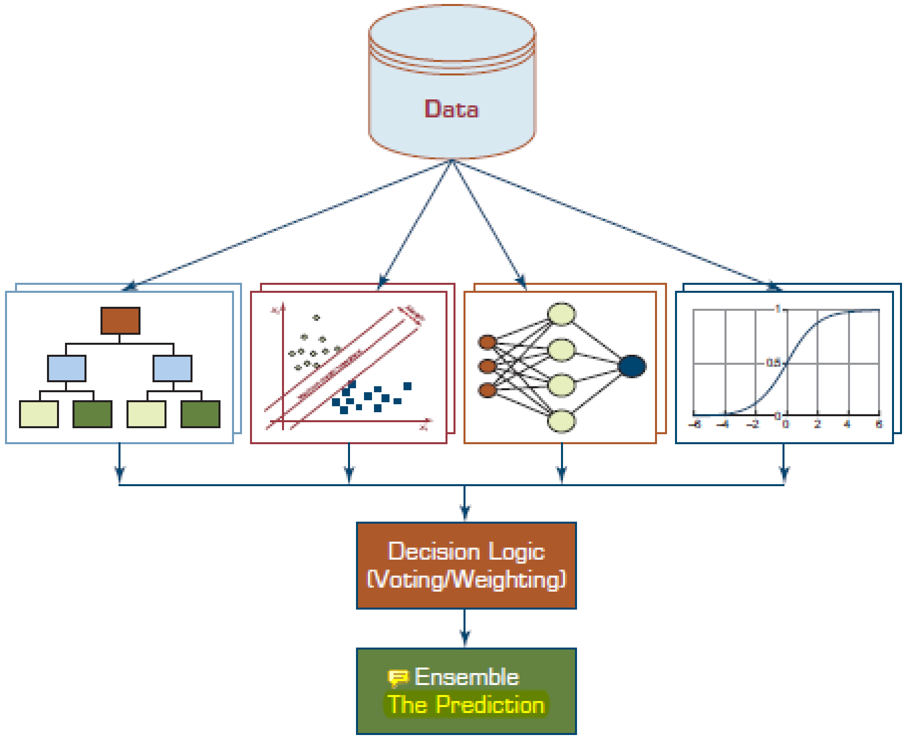
Supervised learning with Classification

CRISP-DM

Modeling

Ensemble method

- Uses multiple prediction models to improve predictive accuracy
- Example: Random Forest is an ensemble of Decision Trees
- Graphical illustration of a heterogeneous ensemble



EXERCISES

Let’s start with the exercises in Orange

Exercise 1

- Go to Blackboard and download the Iris file in week 3 (skip test file!)
- Build a classification model using “Tree, Naive Bayes, kNN and random forest” algorithms
- Check with “**test and score**” which algorithm is the most accurate.

Let’s start with the exercises in Orange

Exercise 2

- Go to Blackboard and download the Iris files in week 3
- Build a classification model using “Tree, Naive Bayes, kNN and random forest” algorithms
- Check with “**test and score**” which algorithm is the most accurate.
- Answer the following questions:
    - Which types of classifiers do we find in Orange?
    - Which type classifier is Naïve Bayes?
    - What is the function of the confusion matrix?
    - How much items are correctly classified in the first run?
    - Visualize the decision tree of the first run.
    - Minimize the leaves of the decision tree(change the ‘min. instances per leave ‘ in 15)
    - Visualize the decision tree of the third run. What is the difference of these two decision trees?

- Use https://www.youtube.com/watch?v=pYXOF0jziGM&t=2s to help you building your model.