

Online-Zertifikatslehrgang

Data Analyst IHK

Die neue Generation digitaler
IHK-Weiterbildungen

INHALT

**KNIME Knoten
und ihre Konfiguration**

Modul 3



**Data
Analyst**_(IHK)

The background of the slide features a wireframe globe with glowing nodes and connecting lines, set against a blue gradient. A hand is visible at the bottom left, reaching towards the globe.

Workflow Kontrollstrukturen

Variablen: Beispiel Configuration Knoten

Single Selection Configuration



Control Flow Variables Job Manager Selection

Label: Wählen Sie die Ausgabe

Description: Wählen Sie, wie Sie die Daten ausgeben wollen

Parameter/Variable Name: Ausgabe

Selection Type: Dropdown

Possible Choices: Datei
Tabellenblaetter

Default Value: Datei
Tabellenblaetter

Limit number of visible options: ☐

Number of visible options: 10

Value overwritten by dialog, current value: Tabellenblaetter

Überschrift und Beschreibung für die Komponentensteuerung

Variablenname

Auswahloptionen

Auswahl des Standardwertes

Auswahl in der Komponentensteuerung

Options Flow Variables Memory Policy Job Manager Selection

Wählen Sie die Ausgabe

Tabellenblaetter ▾

OK Apply Cancel ?

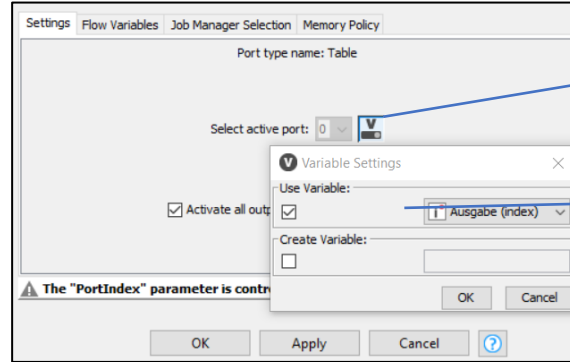
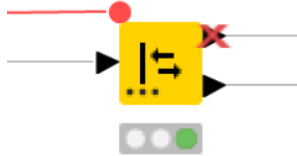
Überschrift und Beschreibung und Auswahl der Variablen

Ergebnis des Variablenkontens:
Variablen und ihre Werte

Flow Variables			
Index	Owner ID	Name	Value
0	3:279:0:48	s Ausgabe	Tabellenblaetter
0	3:279:0:48	i Ausgabe (index)	1
0	3:279	i Component Context	
0		s knime.workspace	C:\knime

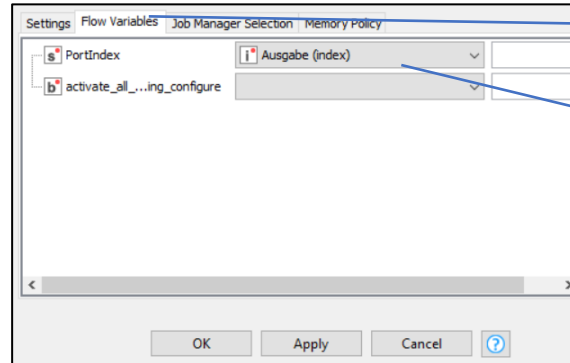
Knoten CASE Switch Start

CASE Switch Start



Aktivierung der Konfiguration einer Einstellung durch eine Variable

Verwendung aktivieren und Variable auswählen



Variablensteuerung im Register „Flow Variables“

Auswahl der Variablen für die verschiedenen Einstellungen

Knoten Create File/Folder Variables

Create File/Folder Variables



Settings | Flow Variables | Job Manager Selection

Base location

Create for: Relative to Current workflow

Folder: ../01_Daten/Output Browse...

File/Folder variables

Variable name	Base location	Value	File extension
base_folder	../01_Daten/Output\		.xlsx

+ Add variable
- Remove variable

⚠ The "path_values" parameter is controlled by a variable.

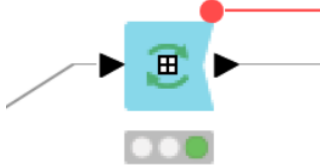
OK Apply Cancel ?

Fixer Pfad für den Ordner oder Datei:
Erzeugt Variable „Base location“

Elemente für den Variablen Pfad
Base location, Dateiname (value),
Dateiendung

Schleifen: Beispiel Group Loop

Group Loop Start



Auswahl der Spalte für die Gruppierung

The screenshot shows the 'Filter' dialog box with the 'Manual Selection' radio button selected. The 'Exclude' list on the left contains the following items: Kunden-ID, Geschlecht, Senior, Partner, Kinder, Telefon-Service, Mehrfachanbindung, Internetservice, Online-Security, and Online-Backup. The 'Include' list on the right contains the item 'Abrechnung'. The 'Enforce exclusion' radio button is selected in the 'Exclude' section, and the 'Enforce inclusion' radio button is selected in the 'Include' section. At the bottom, there is a checkbox labeled 'Input is already sorted by group column(s) [execution fails if not correctly sorted]'. The dialog has 'OK', 'Apply', and 'Cancel' buttons at the bottom right.

The background of the slide features a blue-toned image of a hand reaching out to touch a glowing, wireframe globe. The globe is composed of a network of white lines and dots, symbolizing global connectivity and data. The hand is positioned in the lower-left foreground, with fingers slightly curled as if about to make contact with the globe.

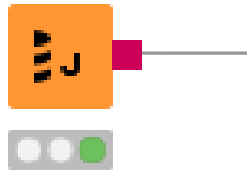
Arbeiten mit Datenbanken

Zugriff auf Datenbanken und Tabellen

Verbindung zur Datenbank herstellen

- Adäquaten Datenbanktreiber auswählen
- URL der DB angeben
- Anmeldedaten angeben (User & Passwort)

DB Connector

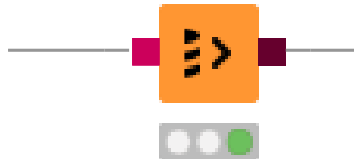


Datenbankverbindung als
Ausgangspunkt

Tabellen in der Datenbank auswählen

- Tabellen in der Datenbank auswählen über *select a Table*
→ Metadaten abfragen
→ Tabelle auswählen

DB Table Selector

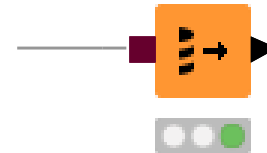


Tabellenselektion

Daten extrahieren

- Tabellen aus der Datenbank nach KNIME exportieren

DB Reader

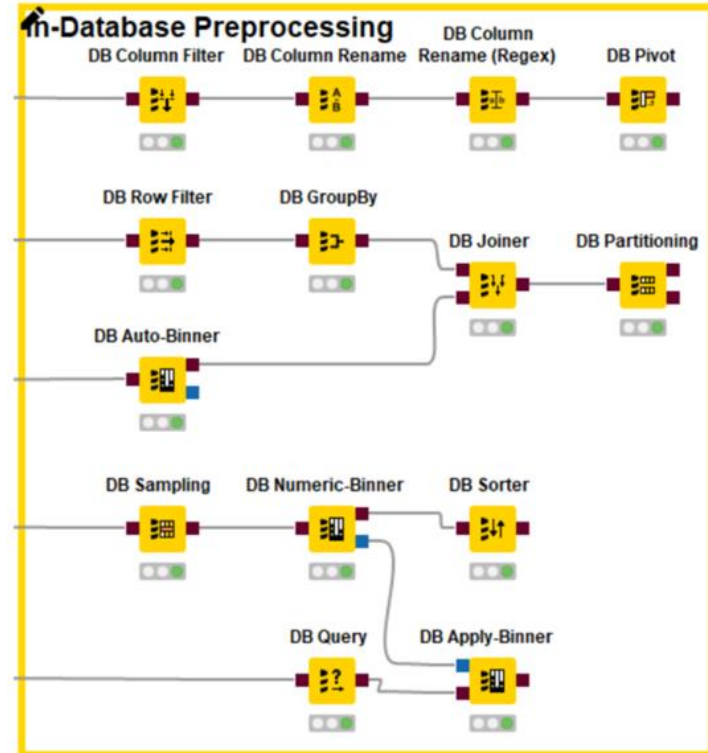


KNIME Table

Query-Knoten in KNIME

Viele Transformationen funktionieren analog wie mit den herkömmlichen Knoten:

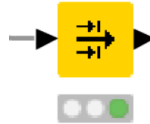
- Zeilen und Spalten filtern
- Join Tabellen / Abfragen
- Stichproben extrahieren
- Binning von numerischen Spalten
- Sortieren
- Aggregieren
- Partitionieren



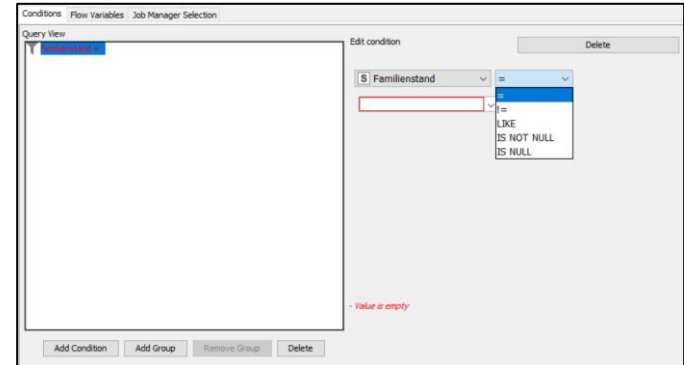
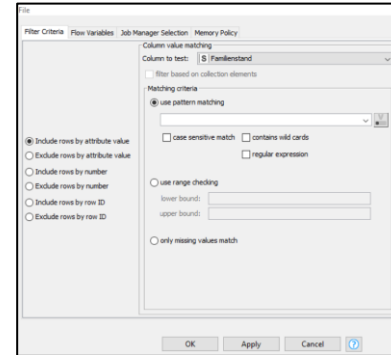
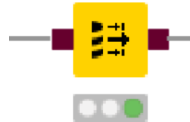
Beispiel Row Filter

Die Konfiguration der Knoten unterscheiden sich leicht von den herkömmlich Knoten, da sie der Logik und der Skriptsprache der Datenbank folgen:

Row Filter



DB Row Filter

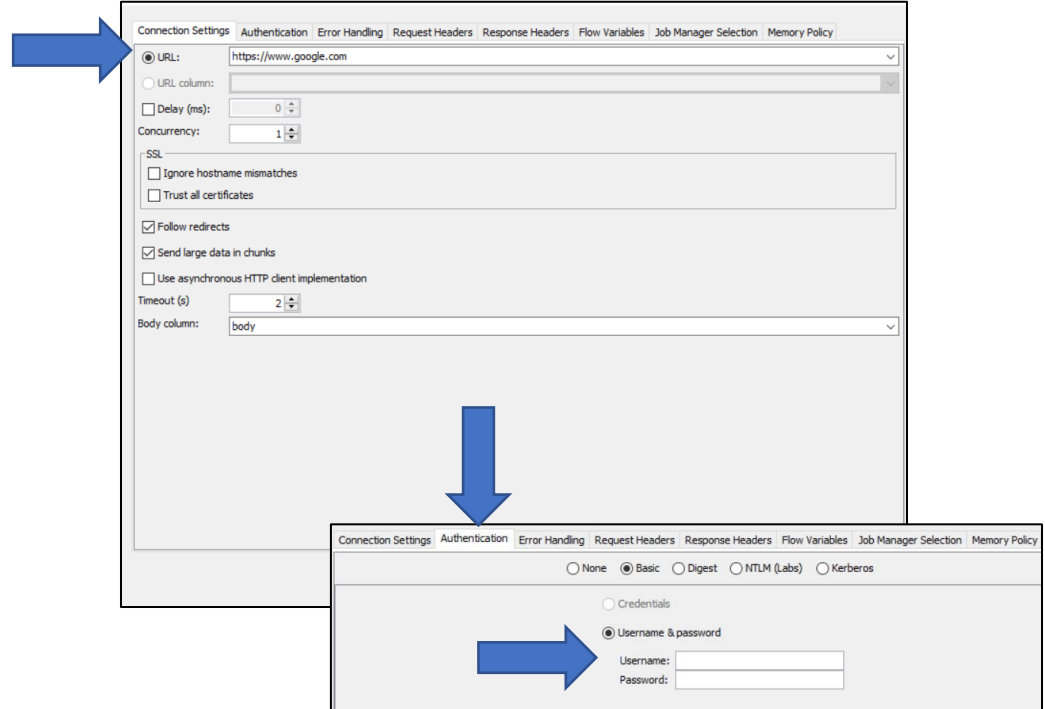


RESTful Web Services in KNMIE

GET Request baut eine Verbindung zur No SQL Datenbank auf:

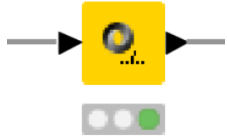
- Den Link setzt man unter URL
- Username und Passwort können im Register „Authentication“ hinterlegt werden

GET Request

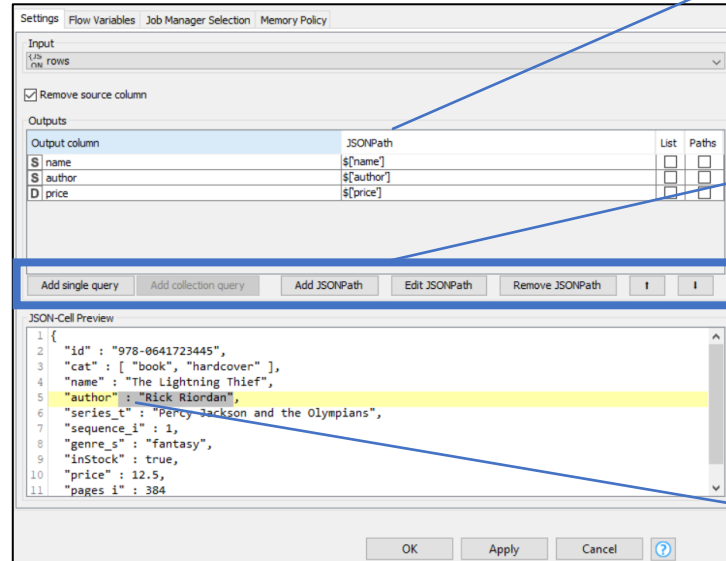


Key-Value Paare extrahieren

JSON Path



Extrahiert die Key-Value Paare aus der geladenen Datenbank



Quers für die Key-Value Paare (Java)

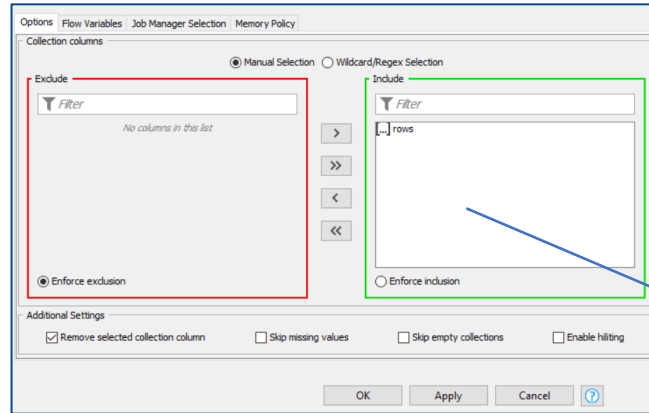
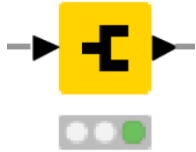
Die Quers können durch Anklicken zusammengestellt werden:

- **Add JSON Path:** extrahiert eine Kategorie mit allen Inhalten
- **Add single query:** extrahiert ein einzelnes Key-Value Paar
- **Add collection query:** extrahiert alle vorhanden Kombinationen zu einem Schlüssel (Key)

Auswahlbereich der Key-Value Paare.
Beispiel. Author: "Rick Riordan"

Listen in Zeilen umwandeln

Ungroup



Mit **Ungroup** werden die gruppierten Kategorien (in Listen zusammengefasst) auf Zeilen verteilt.

Listen sind mit [] zusammengefasst und ihre Elemente durch Trennzeichen (meist Semikolon) getrennt.

Auswahl der gruppierten Spalten

Datentypen umwandeln: String to Date&Time

String to Date&Time



Options | Flow Variables | Job Manager Selection | Memory Policy

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter

- S ID
- S Titel
- S URL-Referenz
- S Ifd. Nr. der Denkmalliste

☐ Enforce exclusion

Include

Filter

- S Eingetragen seit

☒ Enforce inclusion

Replace/Append Selection

☐ Append selected columns Suffix of appended columns: (Date&Time)

☒ Replace selected columns

Type and Format Selection

New type: Date Date format: dd.MM.yyyy

Locale: de-DE Content of the first cell: 22.06.2001

Guess data type and format

Abort Execution

☒ Fail on error

OK Apply Cancel ?

Auswahl der Spalten mit dem Datum im String-Format

Beschreibung des Datumformates

Achtung: Hier muss jedes Zeichen erfasst werden. Fehlende oder falsche Elemente in den Formeln funktionieren nicht.

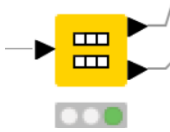
Zur Orientierung steht unter der Formel die erste Zeile der Spalte als Beispiel.

The background of the slide features a blue-toned image of a hand reaching out to touch a wireframe globe. The globe is composed of a network of white lines and dots, symbolizing global connectivity and digital technology. The hand is positioned in the lower-left foreground, with fingers extended towards the globe. The globe itself is centered in the background, showing the outlines of continents. The overall aesthetic is modern and technological.

Überwachtes Lernen

Knoten Partitioning

Partitioning



Konfiguration der ersten Partition, die zweite enthält den Rest

A screenshot of a software dialog box titled 'First partition'. It has four tabs: 'First partition' (selected), 'Flow Variables', 'Job Manager Selection', and 'Memory Policy'. The main area is titled 'Choose size of first partition'. It contains several radio buttons: 'Absolute' (unselected), 'Relative[%]' (selected), 'Take from top' (unselected), 'Linear sampling' (unselected), 'Draw randomly' (unselected), and 'Stratified sampling' (selected). To the right of 'Absolute' is a spinner box with '100'. To the right of 'Relative[%]' is a spinner box with '80'. To the right of 'Stratified sampling' is a dropdown menu showing 'S überlebt'. At the bottom left is a checkbox 'Use random seed' (unchecked). To its right is a text box containing '1.674.401.222.3'. At the bottom are four buttons: 'OK', 'Apply', 'Cancel', and a help icon (question mark in a circle).

Auswahl der Größe der ersten (oberen) Partition

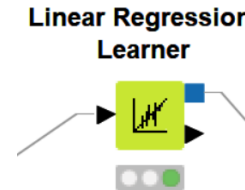
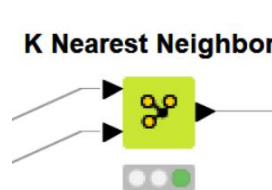
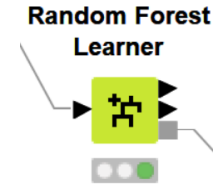
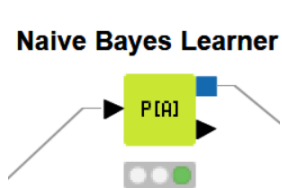
Modus der Zeilenauswahl

Behält bei Stratified Sampling das Verhältnis der Attribute innerhalb einer Spalte in beiden Partitionen bei.

Erstellen von Modellen: Learner

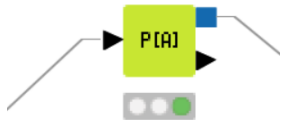
Learner erstellen aus den eingegebenen Daten ein Modell zur Vorhersage.

Das Modell wird über den quadratischen Datenausgang an den Knoten „Predictor“ übergeben.



Naive Bayes

Naive Bayes Learner



The screenshot shows a configuration window for the Naive Bayes Learner. It has four tabs: 'Options' (selected), 'Flow Variables', 'Job Manager Selection', and 'Memory Policy'. The 'Options' tab contains the following settings:

- Classification Column: **S überlebt** (selected from a dropdown)
- Default probability: **0,0001** (spin button)
- Minimum standard deviation: **0,0001** (spin button)
- Threshold standard deviation: **0,0** (spin button)
- Maximum number of unique nominal values per attribute: **20** (spin button)
- ☐ Ignore missing values
- ☐ Create PMML 4.2 compatible model

At the bottom are buttons for 'OK', 'Apply', 'Cancel', and a help icon (question mark in a circle).

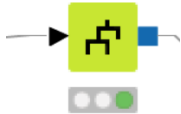
Spalte, die vorhergesagt werden soll

Start- und Grenzwerte für
Wahrscheinlichkeit und Abweichung

Grenze von einzigartigen Wert pro
Spalte. Wird dieser Wert
überschritten, wird die Spalte für die
Erstellung des Modells ignoriert.

Decision Tree

Decision Tree Learner



Options | PMMLSettings | Flow Variables | Job Manager Selection

General

Class column:

Quality measure:

Pruning method:

☒ Reduced Error Pruning

Min number records per node:

Number records to store for view:

☒ Average split point

Number threads:

☒ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column:

Binary nominal splits

☐ Binary nominal splits

Max #nominal:

☐ Filter invalid attribute values in child nodes

OK Apply Cancel ?

Spalte, die vorhergesagt werden soll

Qualitätsmaß mit der die Verteilung der Verzweigungen bestimmt wird

Mindestmenge an Datensätzen für Verzweigungen

Festlegung der ersten Verzweigung auf ein definiertes Merkmal

Random Forest

Random Forest Learner



Options | Flow Variables | Job Manager Selection | Memory Policy

Target Column:

Attribute Selection

☐ Use fingerprint attribute

☒ Use column attributes

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude (Red border):

- ☒ Name
- ☒ PassagierID
- ☒ Ausgangshafen

☒ Enforce exclusion

Include (Green border):

- ☒ Geschlecht
- ☒ Klasse
- ☒ Alter
- ☒ Geschwister
- ☒ Preis
- ☒ Eltern_Kinder

☐ Enforce inclusion

Misc Options

☐ Enable Highlighting (#patterns to store): 2,000

☐ Save target distribution in tree nodes (memory expensive - only important for tree view and PMML export)

Tree Options

Split Criterion:

☐ Limit number of levels (tree depth): 10

☒ Minimum node size: 10

Forest Options

Number of models: 150

☒ Use static random seed: 1622814269727 [New]

[OK] [Apply] [Cancel] [?]

Spalte, die vorhergesagt werden soll

Auswahl der Merkmale für die Erstellung des Modells

Qualitätsmaß mit der die Verteilung der Verzweigungen bestimmt wird

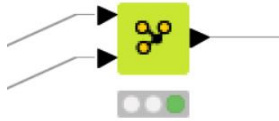
Begrenzung der Anzahl der Verzweigungen

Mindestmenge an Datensätzen für Verzweigungen

Anzahl der erstellten Einzelmodelle (Bäume)

K-Nearest Neighbor

K Nearest Neighbor



Standard settings | Flow Variables | Job Manager Selection | Memory Policy

Column with class labels: S Betrugsfall

Number of neighbours to consider (k): 3

Weight neighbours by distance: ☐

Output class probabilities: ☐

OK Apply Cancel ?

Spalte, die vorhergesagt werden soll

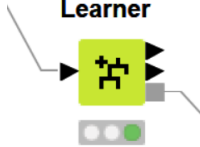
k, Anzahl der Nachbarpunkte, die berücksichtigt werden

Gewichtung nach Entfernung

Ausgabe der berechneten Wahrscheinlichkeiten als zusätzlich Spalte

Linear Regression

Random Forest Learner



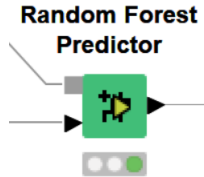
The screenshot shows the 'Linear Regression' configuration window with the following sections:

- Settings**: Flow Variables, Job Manager Selection, Memory Policy.
- Target**: A dropdown menu showing 'D Breite Blütenblatt'.
- Values**: Two panels for feature selection.
 - Exclude** (red border): Contains 'D Länge Kelchblatt' and 'D Breite Kelchblatt'. The 'Enforce exclusion' radio button is selected.
 - Include** (green border): Contains 'D Länge Blütenblatt'. The 'Enforce inclusion' radio button is unselected.
- Regression Properties**: A checkbox for 'Predefined Offset Value' with a value of 0.
- Missing Values in Input Data**: Two radio buttons: 'Ignore rows with missing values.' (unselected) and 'Fail on observing missing values.' (selected).
- Scatter Plot View**: Two input fields: 'First Row:' with value 1 and 'Row Count:' with value 20,000.
- Buttons**: OK, Apply, Cancel, and a help icon.

Spalte, die vorhergesagt werden soll

Auswahl der Merkmale für die
Erstellung des Modells

Vorhersagen Predictor: Beispiel Random Forest



Prediction Settings | Flow Variables | Job Manager Selection | Memory Policy

☐ Change prediction column name

Prediction column name

☒ Append overall prediction confidence

☐ Append individual class probabilities

Suffix for probability columns

☐ Use soft voting

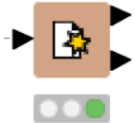
OK Apply Cancel ?

Confidence Value: Maximaler Wert
der einzelnen Wahrscheinlichkeiten

Hängt für jede Klasse eine Spalte mit
den berechneten
Wahrscheinlichkeiten an.

Vorhersage bewerten: Scorer

Scorer



Scorer Flow Variables Job Manager Selection Memory Policy

First Column
[S] überlebt

Second Column
[S] Prediction (überlebt)

Sorting of values in tables
Sorting strategy: Insertion order ☐ Reverse order

Provide scores as flow variables
☐ Use name prefix

Missing values
In case of missing values: ☒ Ignore ☐ Fail

OK Apply Cancel ?

Spalte(Klasse) mit gelernten Werten

Spalte(Klasse) mit vorhergesagten Werten



überlebt \ ...	1	0
1	70	30
0	16	146

Correct classified: 216
Accuracy: 82,443%
Cohen's kappa (κ): 0,618%

Wrong classified: 46
Error: 17,557%

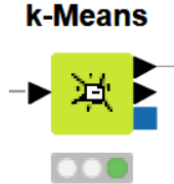
Verteilung der Vorhersagen in Confusion Matrix

Qualitätswerte der Vorsage z.B.: Genauigkeit und Fehler

The background of the slide features a blue-toned image of a hand reaching out towards a wireframe globe. The globe is composed of a network of white lines and dots, symbolizing global connectivity and digital technology. The hand is positioned in the lower-left foreground, with fingers slightly curled as if about to touch the globe.

Unüberwachtes Lernen

K-Means



K-Means Properties Flow Variables Job Manager Selection Memory Policy

Clusters

Number of clusters: 6

Centroid initialization:

☐ First k rows

☒ Random initialization ☒ Use static random seed 0 New

Number of Iterations

Max. number of iterations: 99

Column Selection

Exclude

Filter

Include

Filter

x

y

☐ Always include all columns

Hilite Mapping

☐ Enable Hilite Mapping

OK Apply Cancel ?

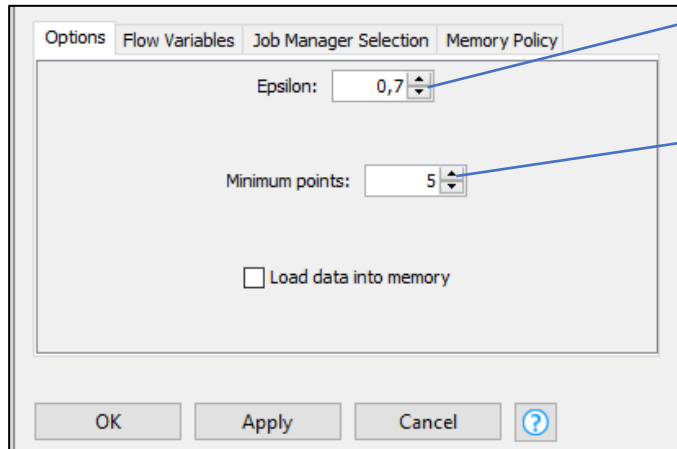
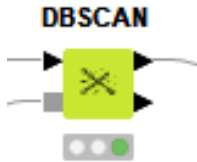
Anzahl der Cluster

Auswahl des ersten Mittelpunktes

Anzahl der Iterationen zur
Optimierung des Modells

Auswahl der Merkmale für die
Erstellung des Modells

DBSCAN

A screenshot of the DBSCAN Options dialog box. It has four tabs: 'Options', 'Flow Variables', 'Job Manager Selection', and 'Memory Policy'. The 'Options' tab is selected. Inside the dialog, there are two spinners: 'Epsilon' set to 0,7 and 'Minimum points' set to 5. Below these is a checkbox labeled 'Load data into memory' which is currently unchecked. At the bottom are buttons for 'OK', 'Apply', 'Cancel', and a help icon (a question mark in a circle).

Options Flow Variables Job Manager Selection Memory Policy

Epsilon: 0,7

Minimum points: 5

☐ Load data into memory

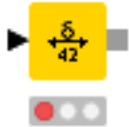
OK Apply Cancel ?

Wert für Epsilon

Mindestzahl an Punkten innerhalb von Epsilon für einen Kernpunkt

Numeric Distances

Numeric Distances



Distance Configuration | Flow Variables | Job Manager Selection | Memory Policy

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter
 No columns in this list

☒ Enforce exclusion

Include

Filter
 No columns in this list

☐ Enforce inclusion

Distance Selection

Standard Distance (Euclidean/ Manhattan)

Configuration

☒ Euclidean
 ☐ Manhattan
 ☐ Maximum
 ☐ Custom 'p' 2.0

☐ Normalize distance (Requires normalized input vectors)

Missing Values

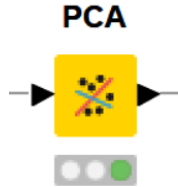
☒ Fail (fails if a missing value cell occurs during computation.)
 ☐ Assume equal (Assume a missing value has the value of the respective counterpart - this will add 0 to the sum of pair wise absolute differences)
 ☐ Average distance (i.e. ignores the missing value and the corresponding value.)

OK Apply Cancel ?

Auswahl der Merkmale für die Berechnung der Entfernungen

Methode der Berechnung

PCA



Settings | Flow Variables | Job Manager Selection | Memory Policy

Target dimensions

☐ Dimension(s) to reduce to

☒ Minimum information fraction

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

No columns in this list

☒ Enforce exclusion

Include

☒ Enforce inclusion

☒ Remove original data columns

☐ Fail if missing values are encountered

OK Apply Cancel ?

Anzahl von Spalten, auf die die Merkmale reduziert werden sollen

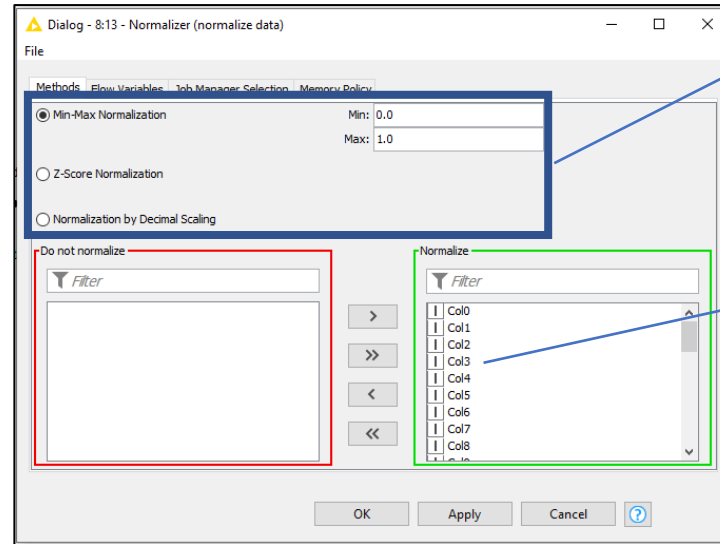
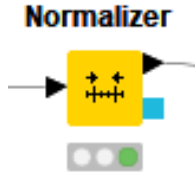
Mindestwert des Informationsgehaltes, der erhalten bleiben soll

Auswahl der Merkmale für die Dimensionsreduktion

The background of the slide features a blue-toned image of a hand reaching out towards a wireframe globe. The globe is composed of a network of white lines and dots, symbolizing global connectivity and digital technology. The hand is positioned in the lower left, with fingers slightly curled as if about to touch the globe.

Neuronale Netze

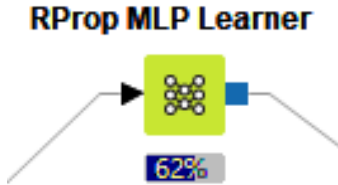
Knoten Normalizer



Methode der Normalisierung

Auswahl der Merkmale für die Normalisierung

Neuronale Netze: Beispiel MLPerceptron



Options | Flow Variables | Job Manager Selection | Memory Policy

Maximum number of iterations: 50

Number of hidden layers: 100

Number of hidden neurons per layer: 10

class column: S Col36

☐ Ignore Missing Values

☐ Use seed for random initialization

Random seed: -416.818.657

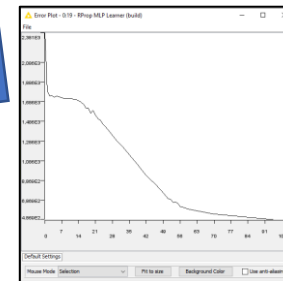
OK Apply Cancel ?

Anzahl der Iterationen zur Verbesserung des Modells

Anzahl der verborgenen Schichten

Anzahl der Neuronen pro verborgener Schicht

Klasse für Vorhersage



Fehlerkurve:
Summe der Fehler
pro Iteration

The background of the slide features a blue-toned image of a hand reaching out to touch a wireframe globe. The globe is composed of a network of white lines and dots, symbolizing global connectivity and digital technology. The hand is positioned in the lower left, with fingers extended towards the globe. The overall aesthetic is modern and technological.

Modelle optimieren

Parameter Optimization Start

Parameter Optimization Loop Start



Parameter	Start value	Stop value	Step size	Integer?	
Tiefe	3	7	1.0	<input checked="" type="checkbox"/>	
AnzBaum	30	150	1.0	<input checked="" type="checkbox"/>	
New parameter	0	1	0.1	<input type="checkbox"/>	

+ Add new parameter

Search strategy: Brute Force

OK Apply Cancel ?

Parameter mit Start- und Stoppwert und Intervallgröße

Wichtig: Die Parameter sind Variablen, die auch an den Learner übergeben werden müssen!

Optimierungsmethode:

- Zufall
- Raster (Brute Force)
- Algorithmisch (Hillclimbing, Bayes)

Parameter Optimization Loop End

Parameter Optimization Loop End



Options | Flow Variables | Job Manager Selection | Memory Policy

Flow variable with objective function value: Accuracy

Function should be...

☒ maximized

☐ minimized

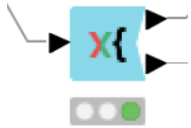
OK Apply Cancel ?

Variable, die optimiert werden soll
(Genauigkeit, Fehler, etc.)

Methode: Maximierung oder
Minimierung der Variablen

Knoten X-Partitioner

X-Partitioner



Standard settings | Flow Variables | Job Manager Selection | Memory Policy

Number of validations: 10

Linear sampling: ☐

Random sampling: ☐

Stratified sampling: ☒

Class column: \$ überlebt

☐ Random seed: 0

Leave-one-out: ☐

OK Apply Cancel ?

Anzahl der Validierungsblöcke (k)

Knoten X-Aggregator

X-Aggregator



Standard settings | Flow Variables | Job Manager Selection | Memory Policy

Target column:

Prediction column:

☐ Add column with fold id

OK Apply Cancel ?

Spalte(Klasse) mit gelernten Werten

Spalte(Klasse) mit vorhergesagten Werten