



University of Minho
School of Engineering



Machine Learning and Decision-Making

ADI @ LEI/3º, MiEI/4º - 2º Semestre
Filipe Gonçalves, Inês Alves, Cesar Analide

Part II – February 2022

Contents

2

Terminology of AI

Intro to Knime

Workflow

Hands On

- Terminology of AI
- Intro to Knime
- Building a Workflow
- Hands On

Technologies of the next decade

3

TERMINOLOGY OF AI

Intro to Knime

Workflow

Hands On

#1 Artificial Intelligence.
AI /Machine Learning / Deep Learning

#2 Internet of Things.
IOT , IIOT, Sensors & Wearables

#3 Mobile/Social Internet
Advancements - Search/Social/ Messaging/Livestreams

#4 Blockchain.
Distributed Ledger Systems, Apps, Infrastructure, Technologies, Cryptocurrencies & DApps

#5 Big Data.
0 1 0 1
1 0 1 1
0 1 1 0
Apps, Infrastructure, Technologies + Predictive Analytics

#6 Automation
Information, Task, Process, Machine, Decision & Action

#7 Robots
Cons./Comm./Indus., Robots, Drones & Autonomous Vehicles

#8 Immersive Media.
- #VR/ #AR/ #MR/ 360°/ Video?Gaming

#9 Mobile Technologies
Infrastructure, networks, standards, services & devices

#10 Cloud Computing.
SaaS, IaaS, PaaS & MESH Apps

#11 3D Printing
Additive Manufacturing & Rapid Prototyping

#12 CX
Customer Journey, Experience Commerce & Personalization

#13 EnergyTech
Efficiency, Energy Storage & Decentralized Grid

#14 Cybersecurity.
Security, Intelligence Detection, Remediation & Adaptation

#15 Voice Assistants
Interfaces, Chatbots & Natural Language Processing

#16 Nanotechnology
Computing, Medicine, Machines + Smart Dust

#17 Collaborative Tech.
Crowd, Sharing, Workplace & Open Source Platforms & Tools

#18 Health Tech.
Advanced Genomics, Bionics & Health Care Tech.

#19 Human-Computer Interaction
Facial/Gesture Recognition, Biometrics, Gaze Tracking

#20 Geo-spatial Tech.
GIS, GPS, Mapping & Remote Sensing, Scanning, Navigation

#21 Advanced Materials
Composites, Alloys, Polymers, Biomimicry, Nanomanufacturing

#22 New Touch Interfaces
Touch Screens, Haptics, 3D Touch, Paper, Feedback & Exoskeletons

#23 Wireless Power

#24 Clean Tech.
Bio-/Enviro-Materials + Solutions, Sustainability, Treatment & Efficiency

#25 Quantum Computing
+ Exascale Computing

#26 Smart Cities
+ Infrastructure & Transport

#27 Edge/Computing
+ Fog Computing

#28 Faster, Better Internet
Broadband incl. Fiber, 5G, Li-Fi, LPN and LoRa

#29 Proximity Tech
Beacons, .RFID, Wi-Fi, Near-Field Communications & Geofencing

#30 New Screens
TVs, Digital Signage, OOH, MicroLEDs & Projections

THE 30 TECHNOLOGIES OF THE NEXT DECADE

Created by: Sean Moffitt @seanmoffitt , Managing Director, @Wikibrands

CC BY NC SA

WIKIBRANDS

Motivation

4

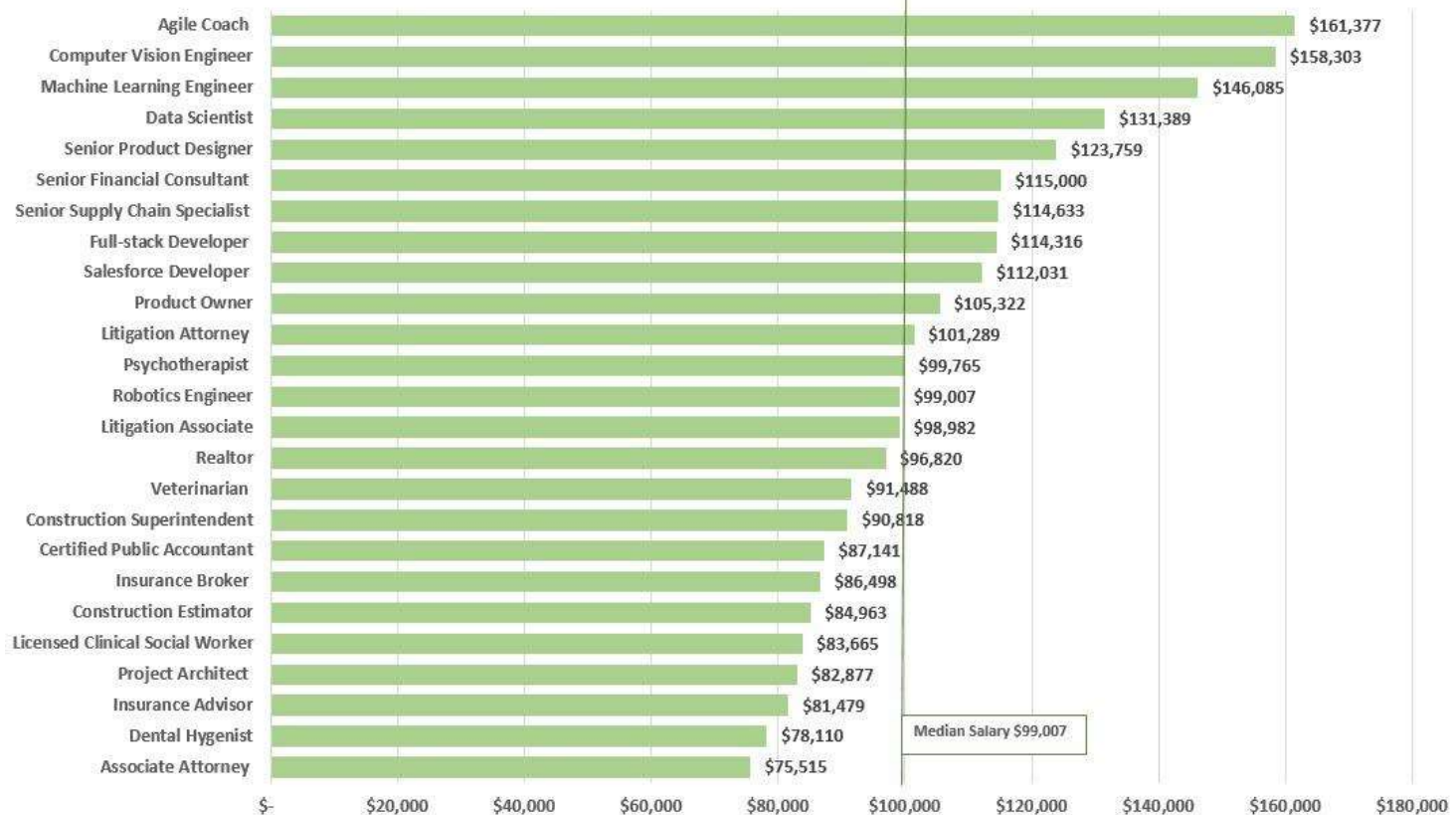
TERMINOLOGY OF AI

Intro to Knime

Workflow

Hands On

Indeed's Best Jobs In the U.S.
Average Base Salary, 2019



(<https://www.forbes.com/sites/louiscolombus/2019/03/17/machine-learning-engineer-is-the-best-job-in-the-u-s-according-to-indeed/#2d134a177bb0>)

Motivation

5

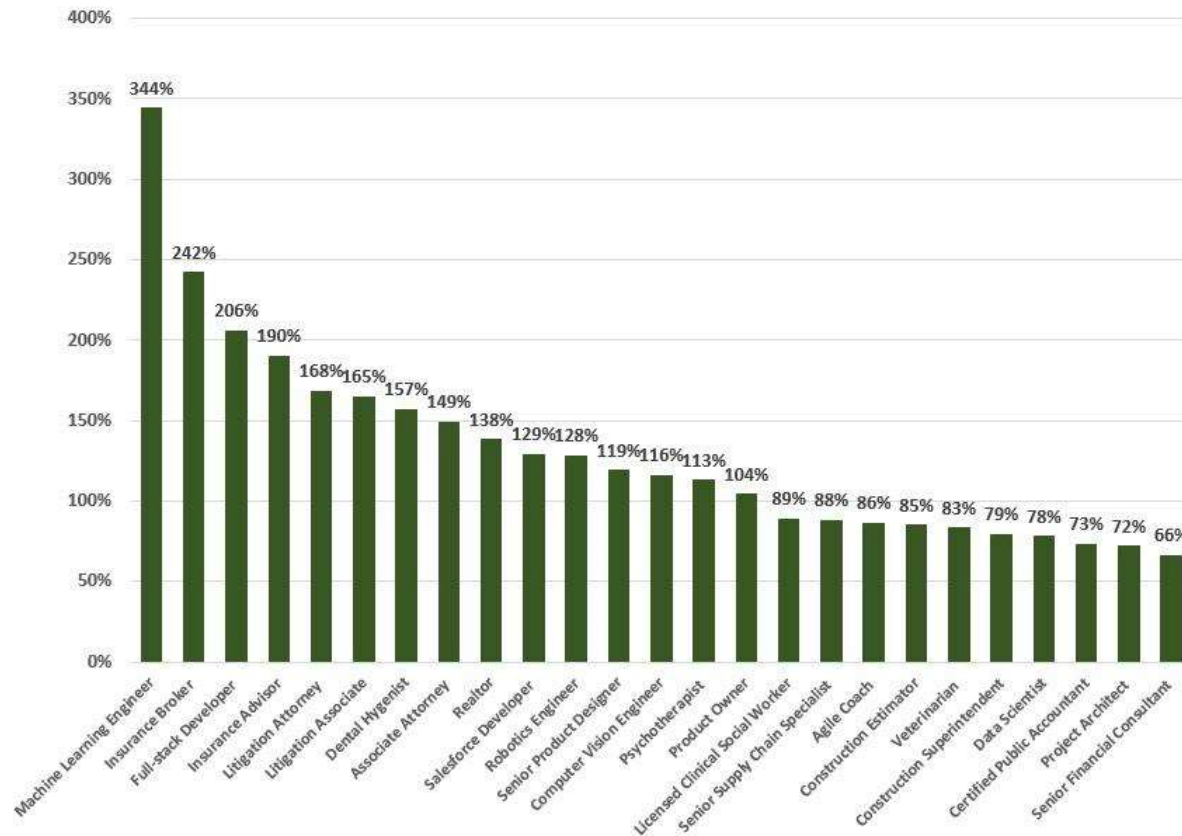
TERMINOLOGY OF AI

Intro to Knime

Workflow

Hands On

Indeed's Best Jobs In The U.S.
% Growth in # of postings, 2015 - 2018
March 14, 2019



(<https://www.forbes.com/sites/louiscolumbus/2019/03/17/machine-learning-engineer-is-the-best-job-in-the-u-s-according-to-indeed/#2d134a177bb0>)

Terminology of AI

6

TERMINOLOGY OF AI

Intro to Knime

Workflow

Hands On

- Artificial Intelligence
- Machine Learning
- Deep Learning
- Data Science
- ...



Terminology of AI

7

TERMINOLOGY OF AI

Intro to Knime

Workflow

Hands On

Machine Learning vs **Data Science**

There is **no universal adherence!!!**

Terminology of AI

8

TERMINOLOGY OF AI

Intro to Knime

Workflow

Hands On

Machine Learning vs

- A -> B system
- PT-PT=**Aprendizagem Automática** (?)
- *“Field of study that gives computer the ability to learn without being explicitly programmed.”*

Arthur Samuel

Usually results in a **software artefact**

Data Science

- Analyse sets of data (datasets)
- PT-PT=**Ciência dos Dados** (?)
- Science of extracting knowledge and insights directly from data

Usually results in **slides and reports**

There is **no universal adherence!!!**

Terminology of AI

9

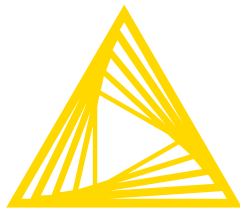
TERMINOLOGY OF AI

Intro to Knime

Workflow

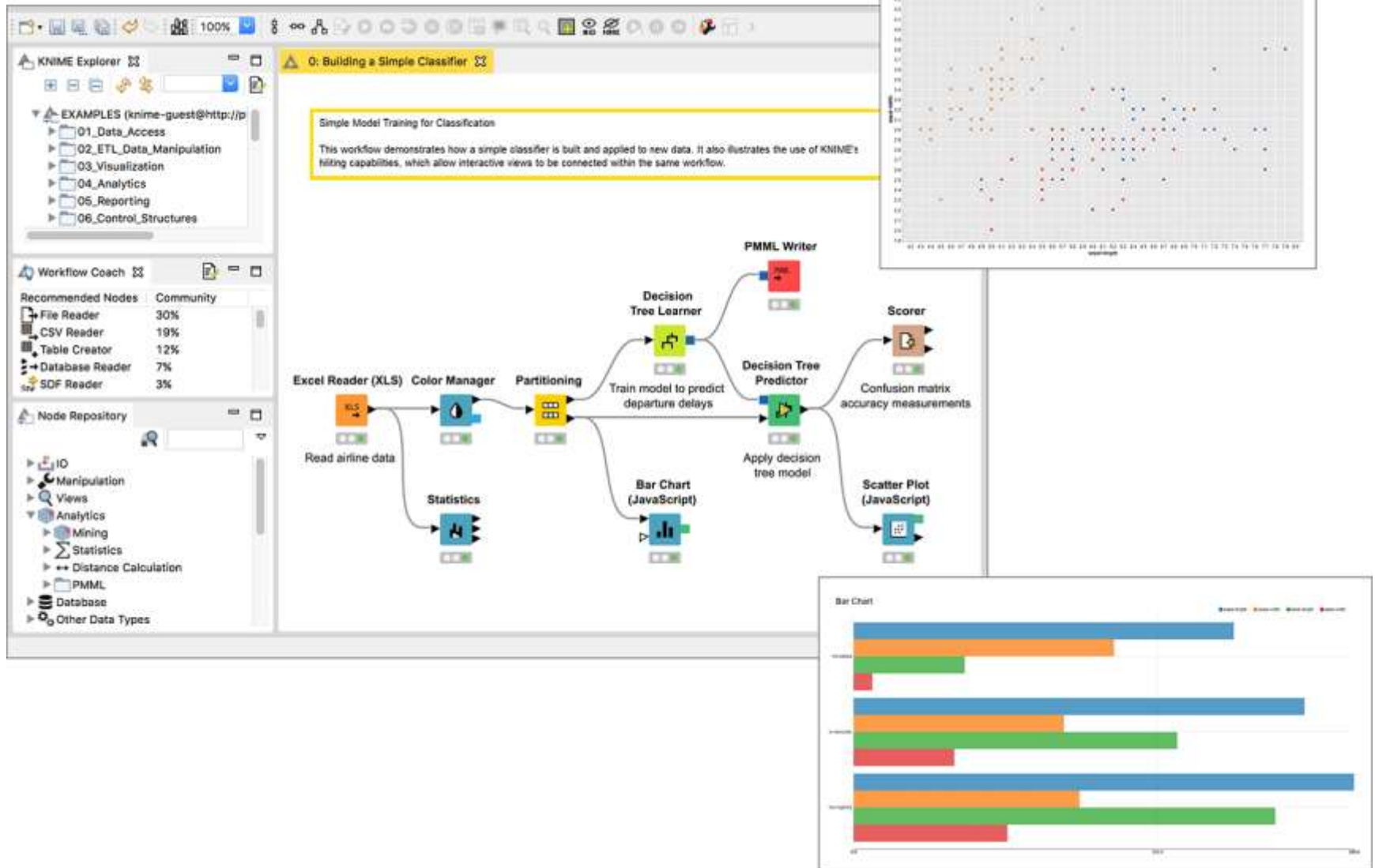
Hands On

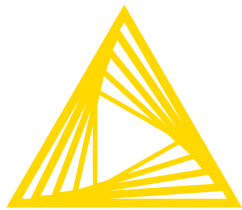




Open for Innovation [®]

KNIME





Open for Innovation ®

KNIME

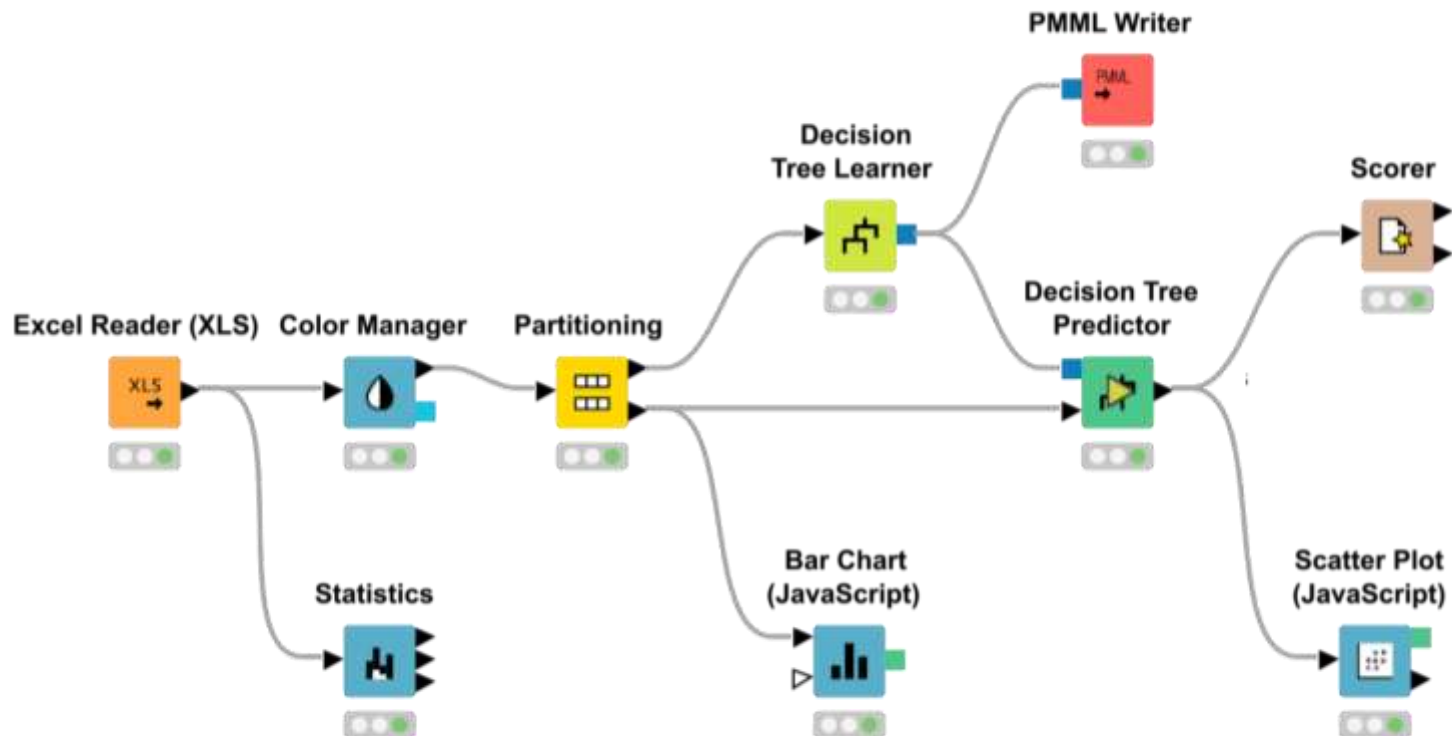
11

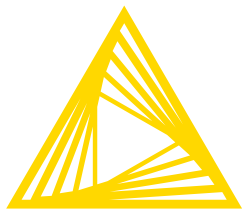
Terminology of AI

INTRO TO KNIME

Workflow

Hands On





Open for Innovation ®

KNIME

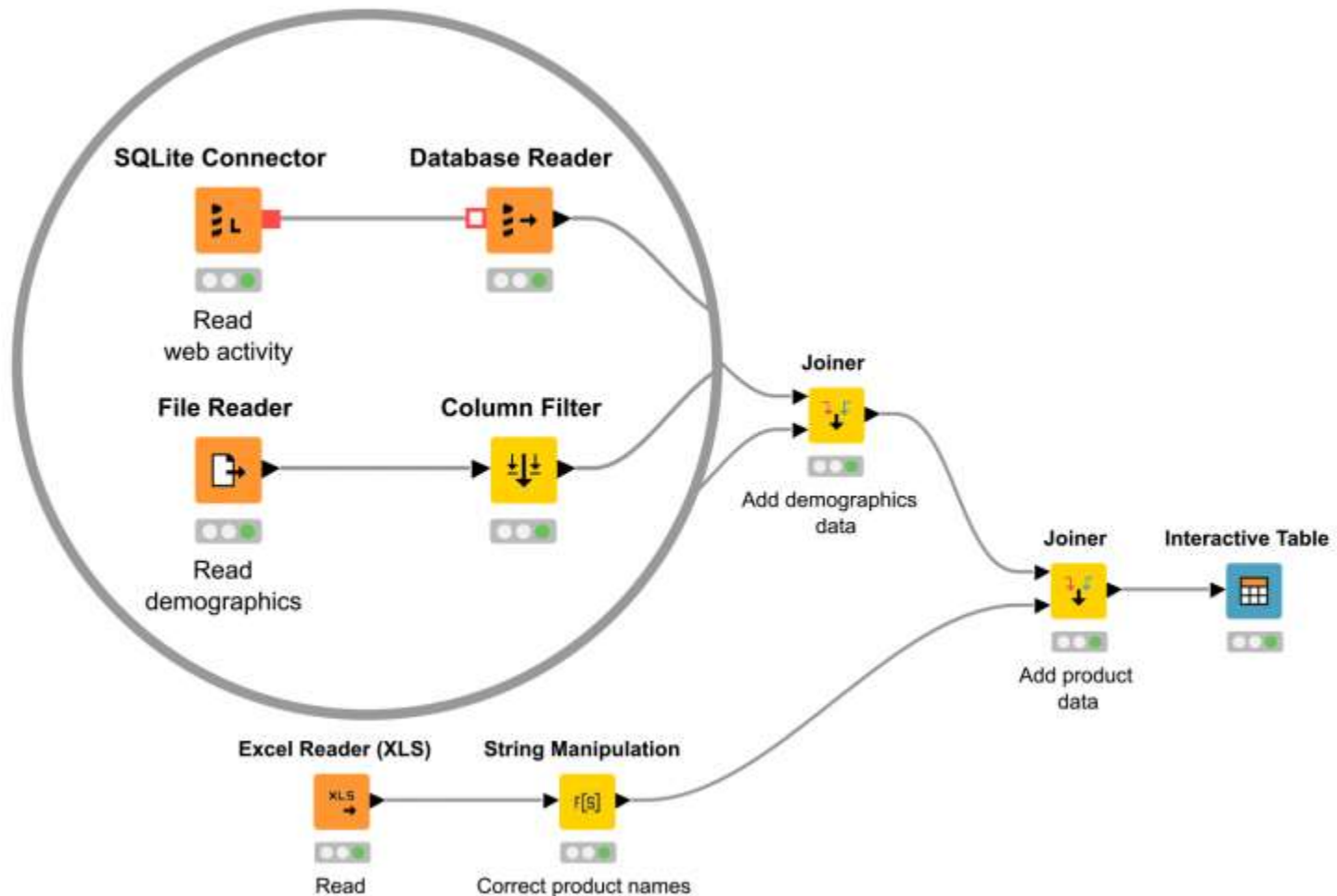
12

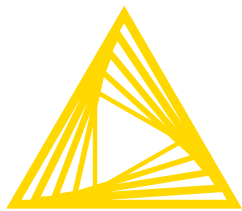
Terminology of AI

INTRO TO KNIME

Workflow

Hands On





Open for Innovation ®

KNIME

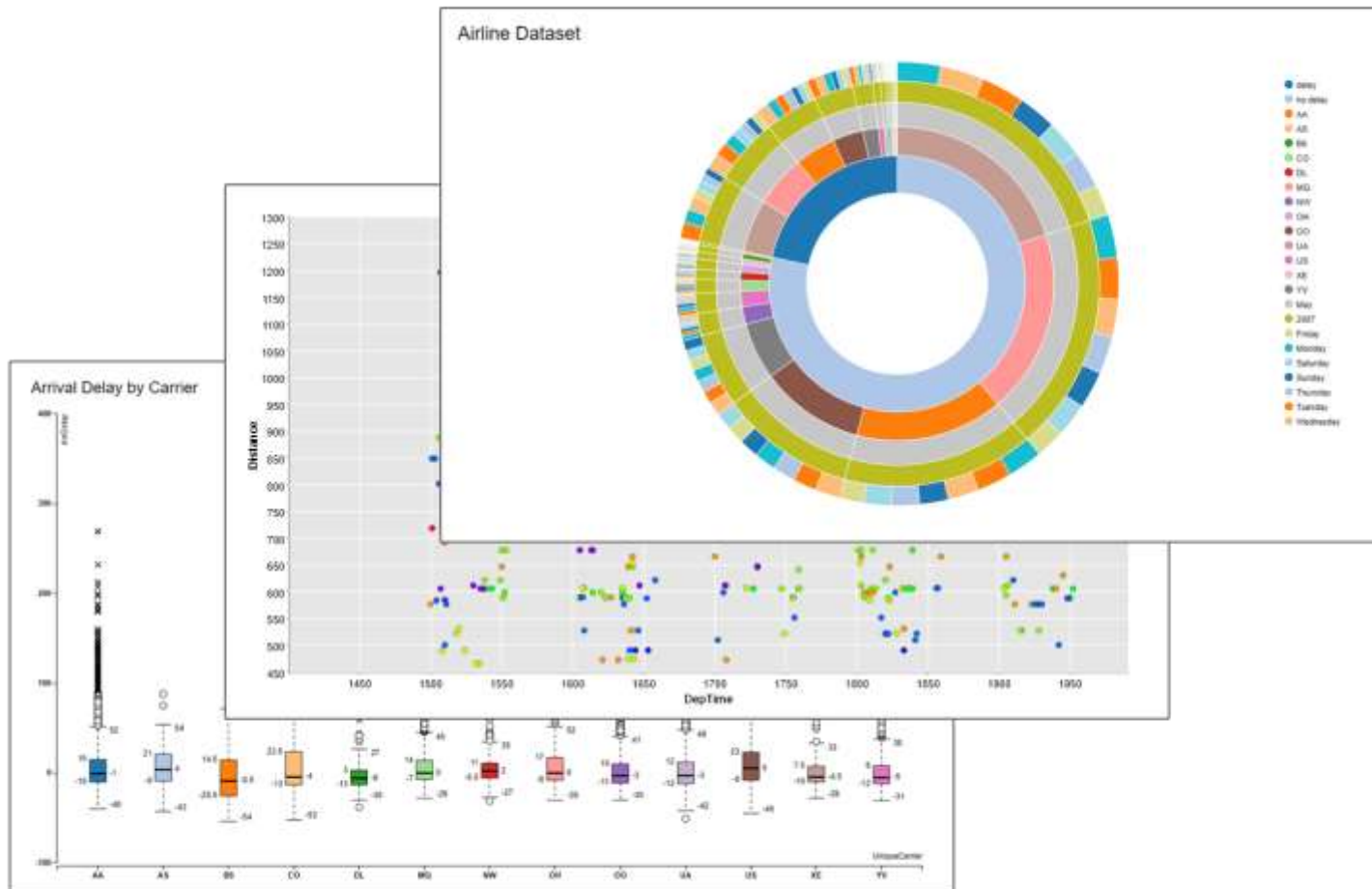
13

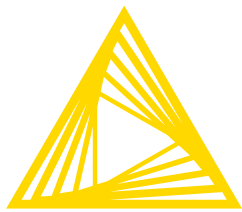
Terminology of AI

INTRO TO KNIME

Workflow

Hands On





Open for Innovation ®

KNIME

14

Terminology of AI

INTRO TO KNIME

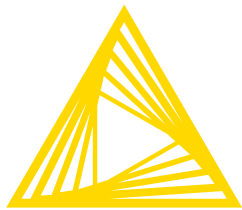
Workflow

Hands On

- **KNIME Analytics Platform** is one of the most popular **open source platforms** used to automate the data science process
- Released in 2006, it is **free** and **open-source**, continuously integrating new developments
- **Additional features** and functionality can be added via **KNIME extensions**
- A **Gartner's leader** for **Data Science and Machine Learning Platforms** for the last six years



Source: Gartner (January 2019)



Open for Innovation [®]

KNIME

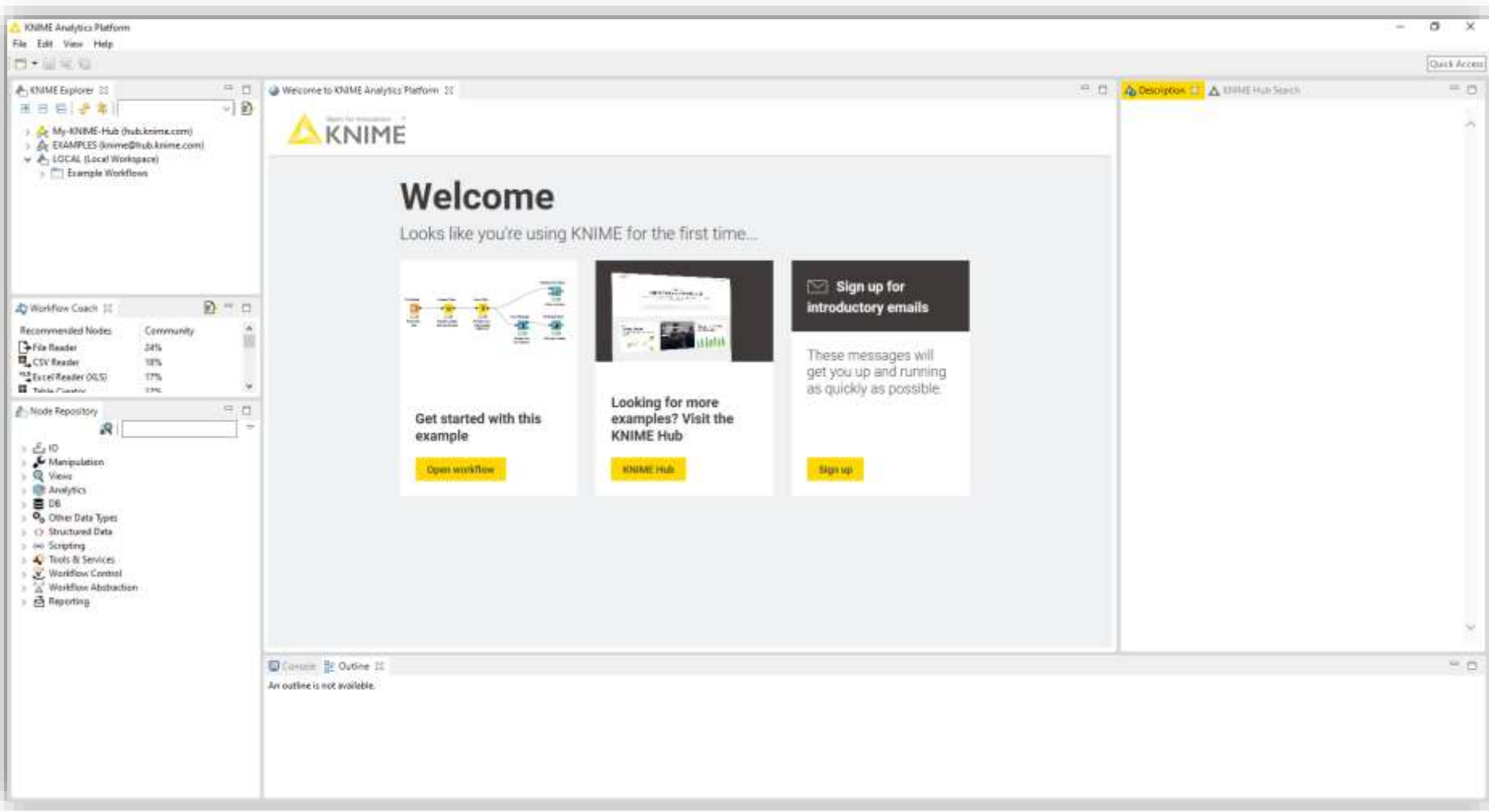
15

Terminology of AI

INTRO TO KNIME

Workflow

Hands On



(Quick) Hands On

16

Terminology of AI

INTRO TO KNIME

Workflow

Hands On

**QUICK
HANDS ON**



(Quick) Hands On

17

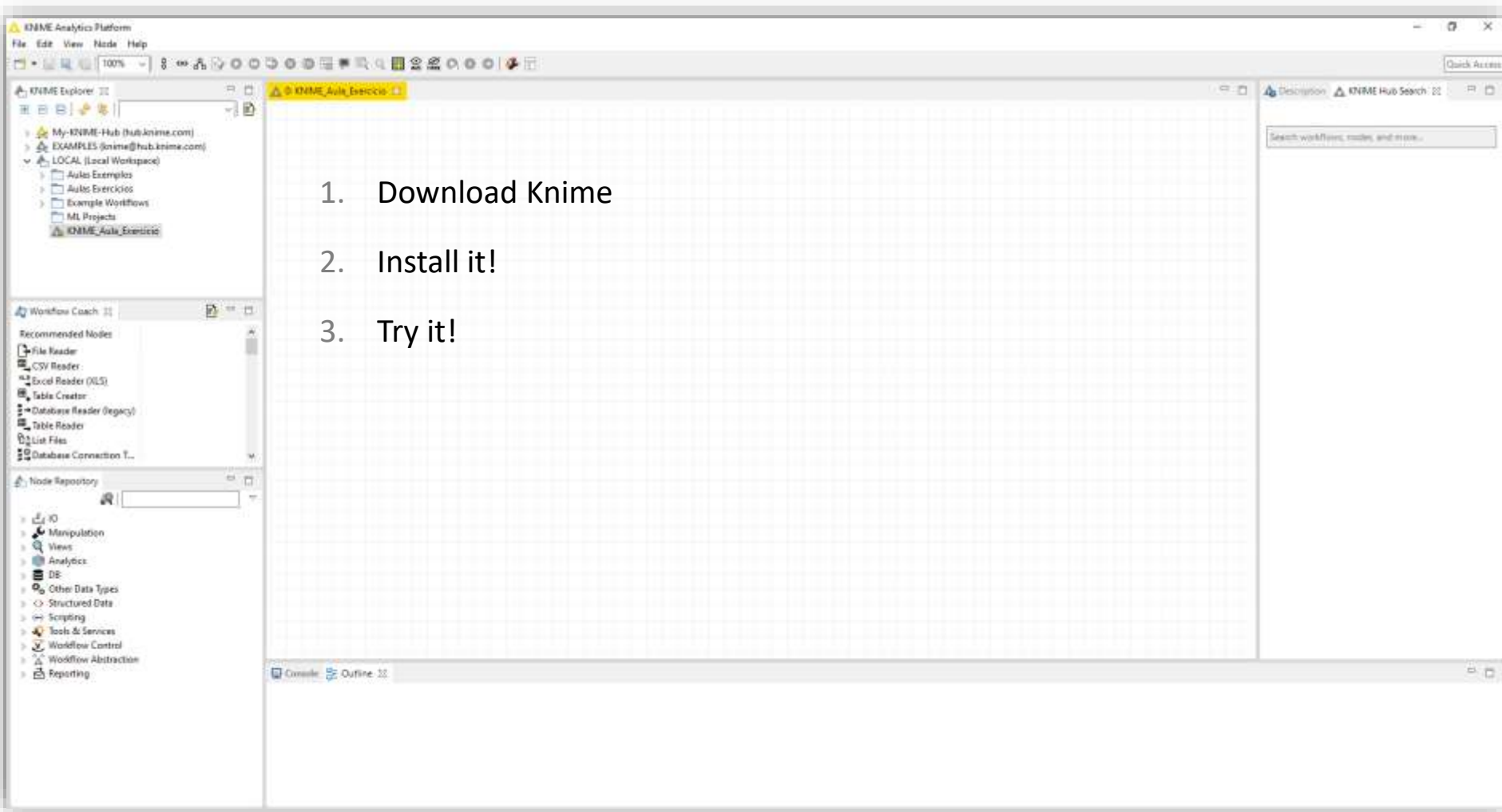
Terminology of AI

INTRO TO KNIME

Workflow

Hands On

1. Download Knime
2. Install it!
3. Try it!



Nodes and Workflows

18

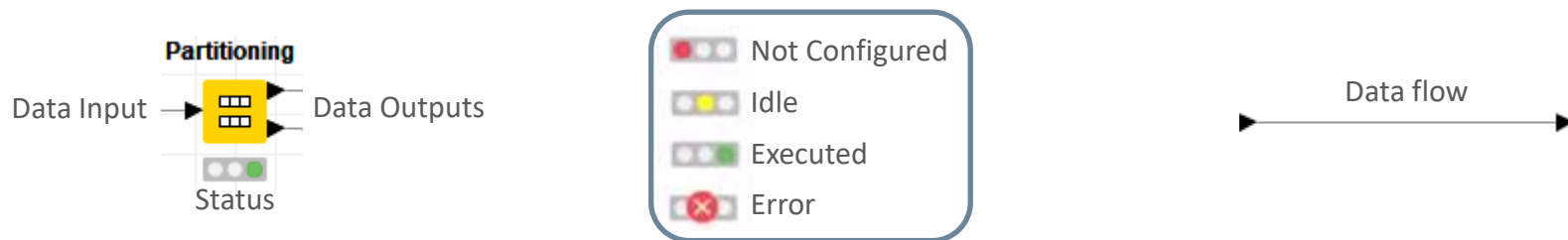
Terminology of AI

INTRO TO KNIME

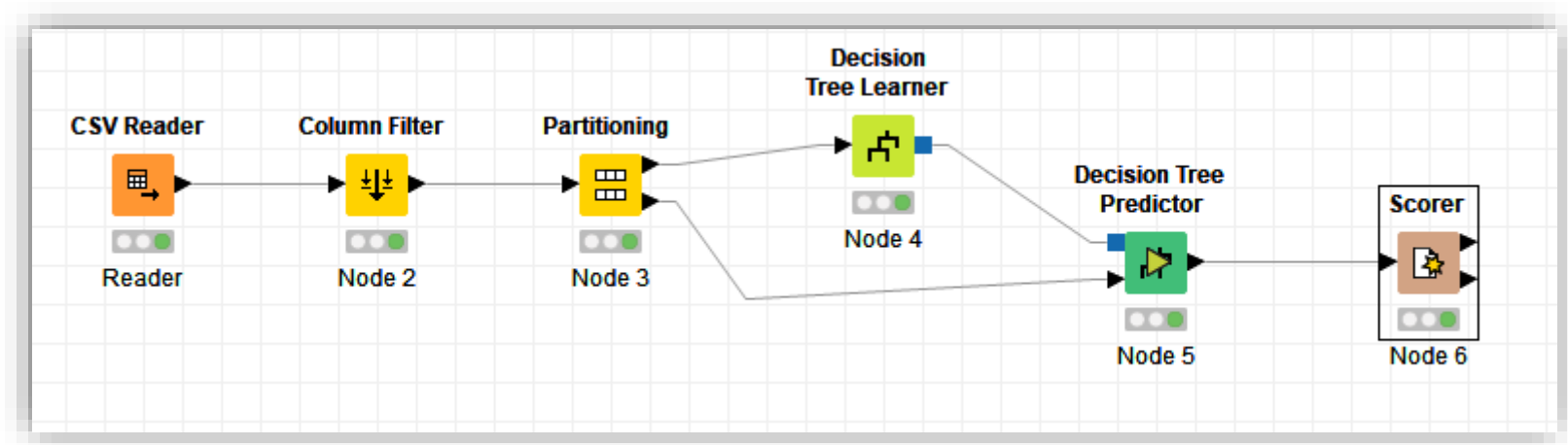
Workflow

Hands On

Nodes



Workflow



KNIME Extensions

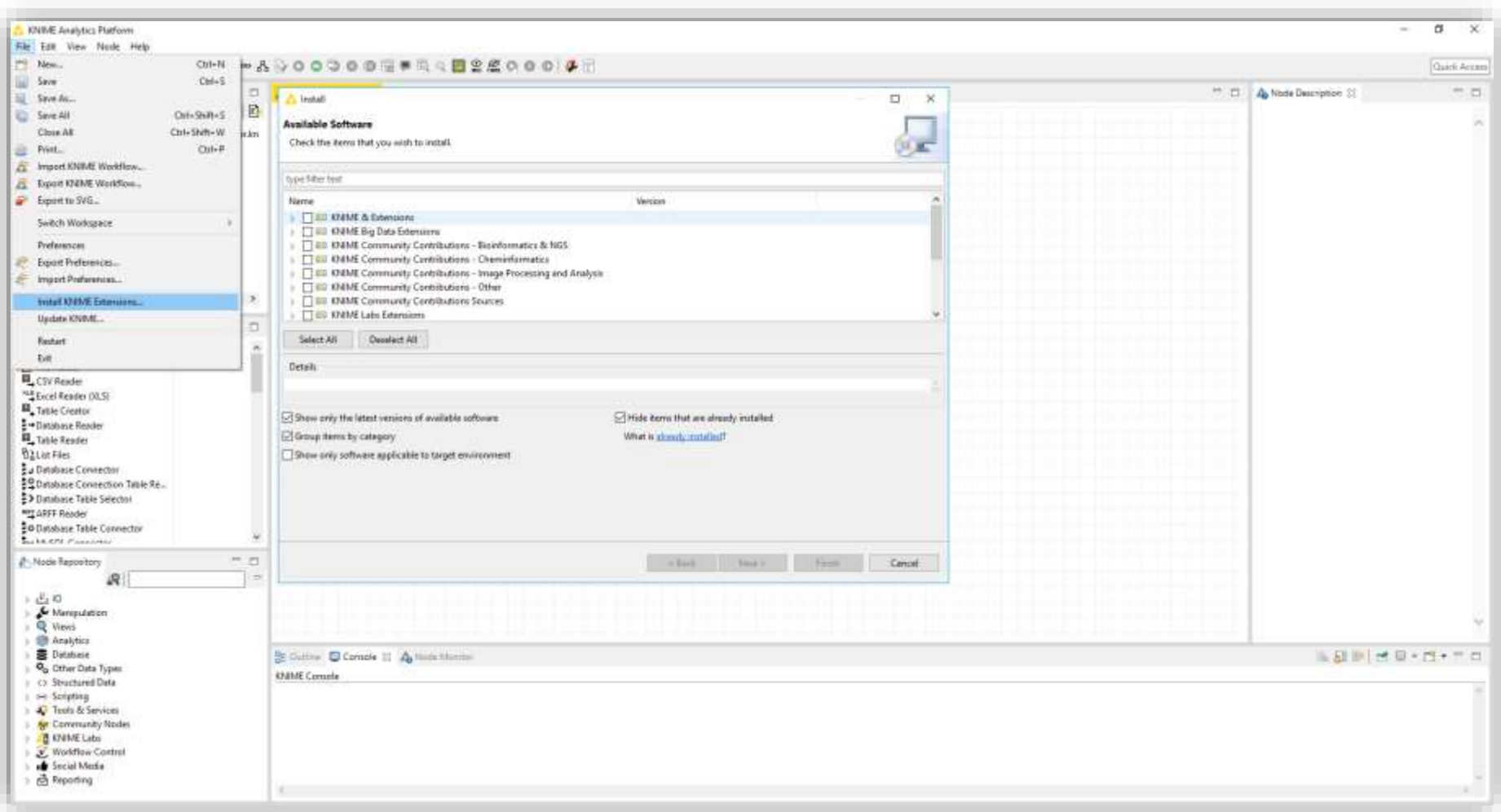
19

Terminology of AI

INTRO TO KNIME

Workflow

Hands On



Main Views

20

Terminology of AI

INTRO TO KNIME

Workflow

Hands On

The screenshot displays the KNIME Analytics Platform interface. The main workspace shows a workflow diagram with the following nodes: CSV Reader (Node 1), Columns Filter (Node 2), Partitioning (Node 3), Decision Tree Learner (Node 4), Decision Tree Predictor (Node 5), and Scorer (Node 6). The interface is divided into several panels:

- KNIME Explorer:** Located on the top left, it shows a tree view of the project structure, including 'EXAMPLES', 'LOCAL (Local Workspace)', 'Example Workflows', and 'Projects'.
- Workflow Coach:** Located on the left side, it displays a list of recommended nodes with their community ratings. For example, 'Decision Tree Predictor' has a rating of 85%.
- Node Repository:** Located at the bottom left, it provides a categorized list of nodes available in the platform.
- Node Description:** Located on the right side, it provides detailed information about the selected node, 'Decision Tree Learner'. It includes a description of the node's function, its parameters, and a section for 'Dialog Options'.
- Console Outline Others:** Located at the bottom, it contains the 'KNIME Console' and 'Node Monitor' tabs.

The central area is the **Workflow Building Area**, where the workflow diagram is constructed.

Node Description

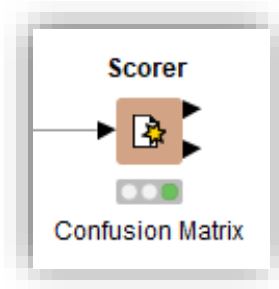
21

Terminology of AI

INTRO TO KNIME

Workflow

Hands On



Description

KNIME Hub Search

Scorer

Compares two columns by their attribute value pairs and shows the confusion matrix, i.e. how many rows of which attribute and their classification match. Additionally, it is possible to highlight cells of this matrix to determine the underlying rows. The dialog allows you to select two columns for comparison; the values from the first selected column are represented in the confusion matrix's rows and the values from the second column by the confusion matrix's columns. The output of the node is the confusion matrix with the number of matches in each cell. Additionally, the second out-port reports a number of **accuracy statistics** such as True-Positives, False-Positives, True-Negatives, False-Negatives, Recall, Precision, Sensitivity, Specificity, F-measure, as well as the overall accuracy and **Cohen's kappa**.

Dialog Options

First column
The first column represents the real classes of the data.

Second column
The second column represents the predicted classes of the data.

Sorting strategy
Whether to sort the labels according to their appearance, or use the lexical/numeric ordering.

Reverse order
Reverse the order of the elements.

Use name prefix
The scores (i.e. accuracy, error rate, number of correct and wrong classification) are exported as flow variables with a hard coded name. This option allows you to define a prefix for these variable identifiers so that name conflicts are resolved.

Missing Values

Choose how to treat missing values in either the reference or prediction column. Default is to ignore them (treat them as if the row did not exist). Alternatively, you can expect the table to not contain missing values in these two columns. If they do, the node will fail during execution.

Ports

Input Ports	
0	Table containing at least two columns to compare.

Output Ports	
0	The confusion matrix.
1	The accuracy statistics table.

Views

Confusion Matrix

Displays the confusion matrix in a table view. It is possible to highlight cells of the matrix which propagates highlighting to the corresponding rows. Therefore, it is possible for example to identify wrong predictions.

Data Table Structure View

22

Terminology of AI

INTRO TO KNIME

Workflow




















Hands On

CSV Reader



Read Wine
Dataset



	Configure...	F6
	Execute	F7
	Execute and Open Views	Shift+F10
	Cancel	F9
	Reset	F8
	Edit Node Description...	Alt+F2
	New Workflow Annotation	
	Connect selected nodes	Ctrl+L
	Disconnect selected nodes	Ctrl+Shift+L
	Create Metanode...	
	Create Component...	
	Compare Nodes	
Show Flow Variable Ports		
	Cut	
	Copy	
	Paste	
	Undo	
	Redo	
	Delete	
	File Table	

Data Table Structure View

23

Terminology of AI

INTRO TO KNIME

Workflow

Hands On

Column Headers

Data Type (Double)

Data Type (String)

Row ID

Data Cells

Row ID	D fixed a...	D volatile ...	D citric acid	D residual...	D chlo...	t...	t...	t...	t...	t...	lcohol	S quality
Row0	7.4	0.7	0	1.9	0.076							=5
Row1	7.8	0.88	0	2.6	0.098							=5
Row2	7.8	0.76	0.04	2.3	0.092							=5
Row3	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3			=6
Row4	7.4	0.7	0	1.9	0.076	11	34	0.998	3			=5
Row5	7.4	0.66	0	1.8	0.075	13	40	0.998	3			=5
Row6	7.9	0.6	0.06	1.6	0.069	15	59	0.996	3			=5
Row7	7.3	0.65	0	1.2	0.065	15	21	0.995	3			=7
Row8	7.8	0.58	0.02	2	0.073	9	18	0.997	3.36	0.57	9.5	=7
Row9	7.5	0.5	0.36	6.1	0.071	17	102	0.998	3.35	0.8	10.5	=5
Row10	6.7	0.58	0.08	1.8	0.097	15	65	0.996	3.28	0.54	9.2	=5
Row11	7.5	0.5	0.36	6.1	0.071	17	102	0.998	3.35	0.8	10.5	=5
Row12	5.6	0.615	0	1.6	0.089	16	59	0.994	3.58	0.52	9.9	=5
Row13	7.8	0.61	0.29	1.6	0.114	9	29	0.997	3.26	1.56	9.1	=5
Row14	8.9	0.62	0.18	3.8	0.176	52	145	0.999	3.16	0.88	9.2	=5
Row15	8.9	0.62	0.19	3.9	0.17	51	148	0.999	3.17	0.93	9.2	=5
Row16	8.5	0.28	0.56	1.8	0.092	35	103	0.997	3.3	0.75	10.5	=7
Row17	8.1	0.56	0.28	1.7	0.368	16	56	0.997	3.11	1.28	9.3	=5
Row18	7.4	0.59	0.08	4.4	0.086	6	29	0.997	3.38	0.5	9	=4
Row19	7.9	0.32	0.51	1.8	0.341	17	56	0.997	3.04	1.08	9.2	=6
Row20	8.9	0.22	0.48	1.8	0.077	29	60	0.997	3.39	0.53	9.4	=6
Row21	7.6	0.39	0.31	2.3	0.082	23	71	0.998	3.52	0.65	9.7	=5
Row22	7.9	0.43	0.21	1.6	0.106	10	37	0.997	3.17	0.91	9.5	=5
Row23	8.5	0.49	0.11	2.3	0.084	9	67	0.997	3.17	0.53	9.4	=5
Row24	6.9	0.4	0.14	2.4	0.085	21	40	0.997	3.43	0.63	9.7	=6
Row25	6.3	0.39	0.16	1.4	0.08	11	23	0.996	3.34	0.56	9.3	=5
Row26	7.6	0.4	0.24	1.8	0.08	4	11	0.996	3.28	0.56	9.5	=5

Log View

24

Terminology of AI

INTRO TO KNIME

Workflow

Hands On

The screenshot displays the KNIME Analytics Platform interface. The left sidebar shows the 'KNIME Explorer' with the 'LOCAL (Local Workspace)' selected, and the 'KNIME_Aula_1' project open. The main area is the 'Log View', showing a detailed log of workflow execution. The log entries are as follows:

```
2018-09-26 11:50:26,149 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:3(1) -> 0:4( 1)]
2018-09-26 11:50:26,149 : DEBUS : main : ConnectionContainerEditPart : : modelling info: null
2018-09-26 11:50:26,150 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:4(1) -> 0:5( 1)]
2018-09-26 11:50:26,150 : DEBUS : main : ConnectionContainerEditPart : : modelling info: null
2018-09-26 11:50:26,150 : DEBUS : main : WorkflowRootEditPart : : part: NodeContainerEditPart( Decision Tree Predictor 0:5 (EXECUTED) )
2018-09-26 11:50:26,151 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:4(1) -> 0:5( 1)]
2018-09-26 11:50:26,151 : DEBUS : main : ConnectionContainerEditPart : : modelling info: null
2018-09-26 11:50:26,151 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:3(2) -> 0:5( 2)]
2018-09-26 11:50:26,151 : DEBUS : main : ConnectionContainerEditPart : : modelling info: bendpoints: 369, 197,
2018-09-26 11:50:26,152 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:5(1) -> 0:6( 1)]
2018-09-26 11:50:26,152 : DEBUS : main : ConnectionContainerEditPart : : modelling info: null
2018-09-26 11:50:26,152 : DEBUS : main : WorkflowRootEditPart : : part: NodeContainerEditPart( Scorer 0:6 (EXECUTED) )
2018-09-26 11:50:26,153 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:5(1) -> 0:6( 1)]
2018-09-26 11:50:26,153 : DEBUS : main : ConnectionContainerEditPart : : modelling info: null
2018-09-26 12:02:30,357 : DEBUS : main : NodeContainer : : Setting dirty flag on KNIME_Aula_1 @
2018-09-26 12:02:30,359 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:1(1) -> 0:2( 1)]
2018-09-26 12:02:30,359 : DEBUS : main : ConnectionContainerEditPart : : modelling info: null
2018-09-26 12:02:30,360 : DEBUS : KNIME-Workflow-Notifier : WorkflowEditor : : Workflow event triggered: WorkflowEvent [Type=WORKFLOW_DIRTY,node=,old=null,new=null]
2018-09-26 12:02:30,364 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:2(1) -> 0:3( 1)]
2018-09-26 12:02:30,364 : DEBUS : main : ConnectionContainerEditPart : : modelling info: null
2018-09-26 12:02:30,364 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:3(2) -> 0:3( 2)]
2018-09-26 12:02:30,364 : DEBUS : main : ConnectionContainerEditPart : : modelling info: bendpoints: 369, 197,
2018-09-26 12:02:30,364 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:3(1) -> 0:4( 1)]
2018-09-26 12:02:30,364 : DEBUS : main : ConnectionContainerEditPart : : modelling info: null
2018-09-26 12:02:30,364 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:4(1) -> 0:5( 1)]
2018-09-26 12:02:30,364 : DEBUS : main : ConnectionContainerEditPart : : modelling info: null
2018-09-26 12:02:30,364 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:5(1) -> 0:6( 1)]
2018-09-26 12:02:30,364 : DEBUS : main : ConnectionContainerEditPart : : modelling info: null
2018-09-26 12:02:40,973 : DEBUS : main : NodeContainerEditPart : : Scorer 0:6 (EXECUTED)
2018-09-26 12:02:40,973 : DEBUS : main : NodeContainerEditPart : : Decision Tree Predictor 0:3 (EXECUTED)
2018-09-26 12:02:42,364 : DEBUS : main : NodeContainerEditPart : : Decision Tree Predictor 0:3 (EXECUTED)
2018-09-26 12:14:04,445 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:1(1) -> 0:2( 1)]
2018-09-26 12:14:04,445 : DEBUS : main : ConnectionContainerEditPart : : modelling info: null
2018-09-26 12:14:04,446 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:2(1) -> 0:3( 1)]
2018-09-26 12:14:04,446 : DEBUS : main : ConnectionContainerEditPart : : modelling info: null
2018-09-26 12:14:04,446 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:3(2) -> 0:5( 2)]
2018-09-26 12:14:04,446 : DEBUS : main : ConnectionContainerEditPart : : modelling info: bendpoints: 369, 197,
2018-09-26 12:14:04,446 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:3(1) -> 0:4( 1)]
2018-09-26 12:14:04,446 : DEBUS : main : ConnectionContainerEditPart : : modelling info: null
2018-09-26 12:14:04,446 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:4(1) -> 0:5( 1)]
2018-09-26 12:14:04,446 : DEBUS : main : ConnectionContainerEditPart : : modelling info: null
2018-09-26 12:14:04,446 : DEBUS : main : ConnectionContainerEditPart : : refreshing visuals for: STD[0:5(1) -> 0:6( 1)]
2018-09-26 12:14:04,446 : DEBUS : main : ConnectionContainerEditPart : : modelling info: null
2018-09-26 12:14:12,053 : DEBUS : main : NodeContainerEditPart : : Partitioning 0:3 (EXECUTED)
```

The right sidebar shows the 'Node Description' for the 'Partitioning' node. The description includes a 'Partitioning' section with a detailed explanation of the node's function: 'The input table is split into two partitions (i.e. row-wise), e.g. train and test data. The two partitions are available at the two output ports. The following options are available in the dialog:'. Below this, there are four sections: 'Absolute', 'Relative', 'Take from top', and 'Linear sampling', each with a brief description of its function. The 'Absolute' section specifies the absolute number of rows in the first partition. The 'Relative' section specifies the percentage of the number of rows in the input table that are in the first partition. The 'Take from top' section specifies the number of rows to take from the top of the input table. The 'Linear sampling' section specifies the number of rows to take from the top of the input table. The 'Draw randomly' section specifies the number of rows to take from the top of the input table. The 'Stratified sampling' section specifies the number of rows to take from the top of the input table.

The bottom of the interface shows the 'Node Repository' and 'Workflow Coach' tabs. The 'Node Repository' tab is active, showing a list of nodes. The 'Workflow Coach' tab is also visible, showing a diagram of the workflow.

Views' Customization

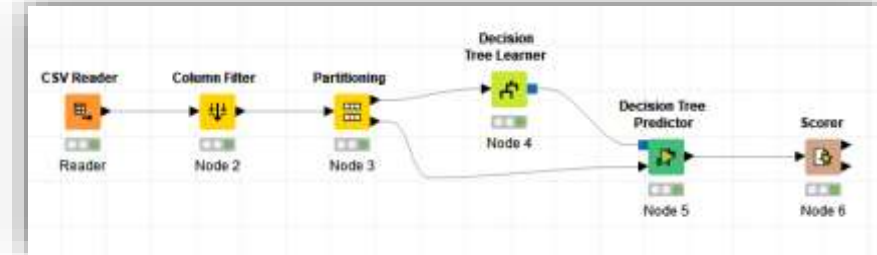
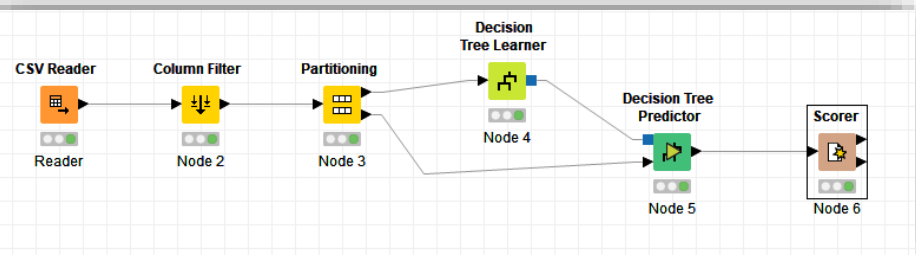
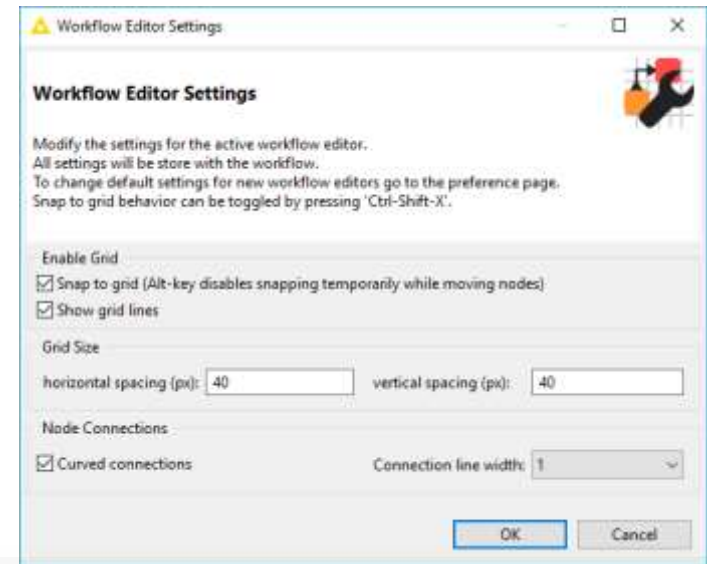
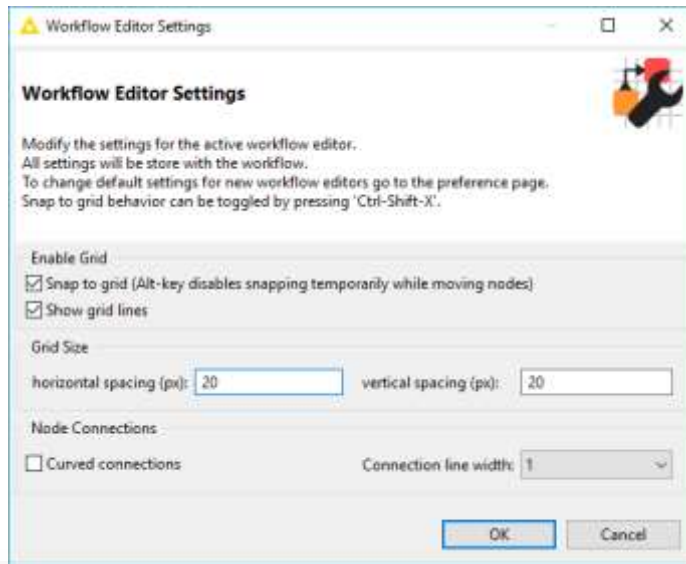
25

Terminology of AI

INTRO TO KNIME

Workflow

Hands On



Building a Simple Workflow

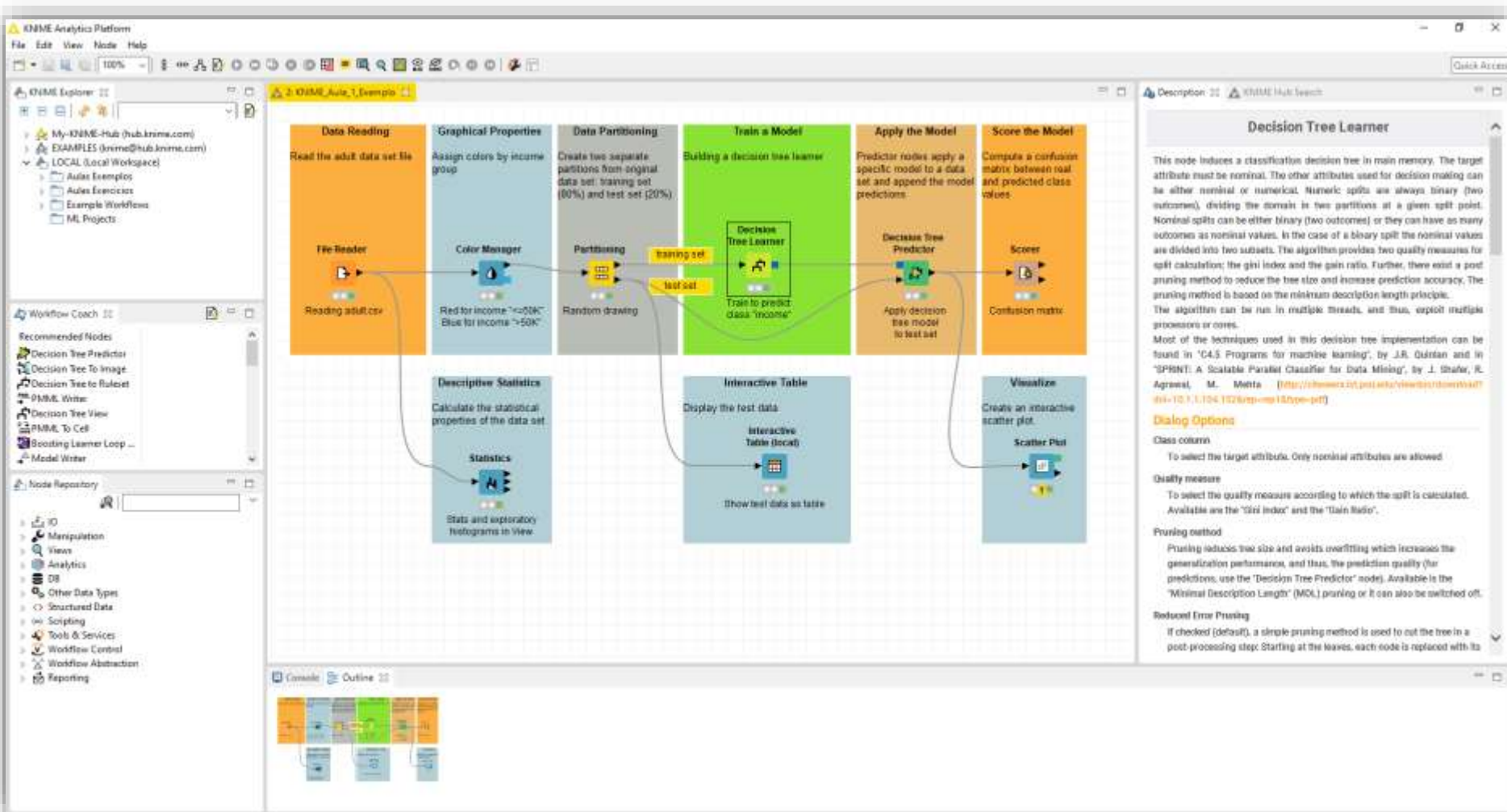
26

Terminology of AI

Intro to Knime

WORKFLOW

Hands On



Node Context Options

Data Loader

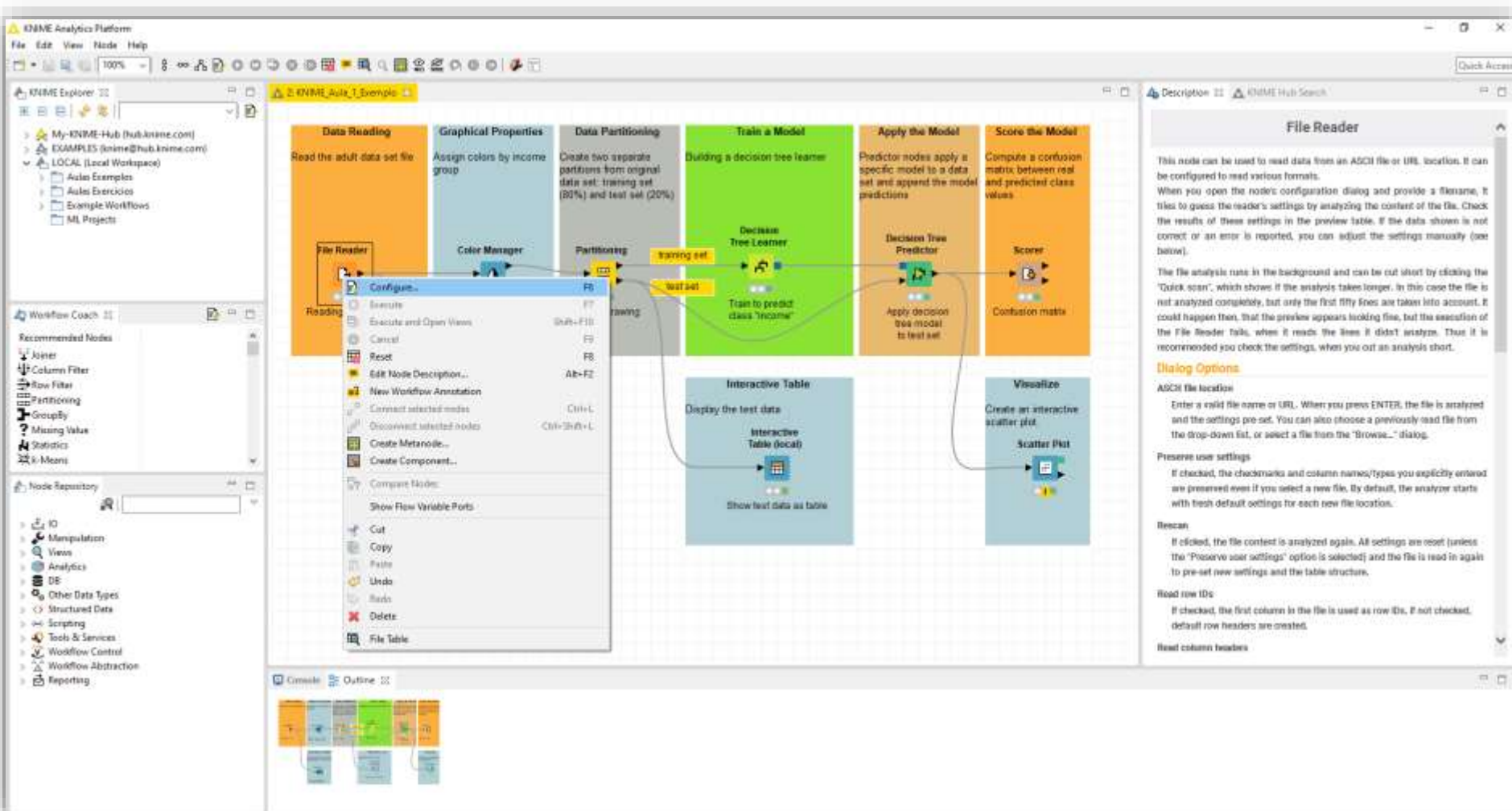
27

Terminology of AI

Intro to Knime

WORKFLOW

Hands On



The screenshot displays the KNIME Analytics Platform interface. The main workspace shows a workflow diagram with the following nodes and connections:

- Data Reading:** File Reader (Reading) -> Color Manager (Assign colors by income group)
- Data Partitioning:** Partitioning (Create two separate partitions from original data set: training set (80%) and test set (20%)) -> Training set -> Test set
- Train a Model:** Decision Tree Learner (Building a decision tree learner) -> Train to predict class "income"
- Apply the Model:** Decision Tree Predictor (Predictor nodes apply a specific model to a data set and append the model predictions) -> Apply decision tree model to test set
- Score the Model:** Scorer (Compute a confusion matrix between real and predicted class values) -> Confusion matrix
- Interactive Table:** Interactive Table (Display the test data) -> Show test data as table
- Visualize:** Scatter Plot (Create an interactive scatter plot)

The **File Reader** node configuration dialog is open, showing the following options:

- File:** /home/user/KNIME/Examples/Local Workspace/ML Projects/Example Workflows/Example Workflow.knime
- Format:** CSV (comma-separated values)
- Encoding:** UTF-8
- Delimiter:** ;
- Header:** 1
- Preview:** 10 rows
- Options:**
 - ☒ Preserve user settings
 - ☒ Read row IDs
 - ☒ Read column headers

The **File Reader** node description is also visible on the right side of the interface:

File Reader

This node can be used to read data from an ASCII file or URL location. It can be configured to read various formats. When you open the node's configuration dialog and provide a filename, it tries to guess the reader's settings by analyzing the content of the file. Check the results of these settings in the preview table. If the data shown is not correct or an error is reported, you can adjust the settings manually (see below).

The file analysis runs in the background and can be cut short by clicking the "Quick scan", which shows if the analysis takes longer. In this case the file is not analyzed completely, but only the first fifty lines are taken into account. It could happen then, that the preview appears looking fine, but the execution of the File Reader fails, when it reads the lines it didn't analyze. Thus it is recommended you check the settings, when you cut an analysis short.

Dialog Options

ASCII file location

Enter a valid file name or URL. When you press ENTER, the file is analyzed and the settings pre-set. You can also choose a previously read file from the drop-down list, or select a file from the "Browse..." dialog.

Preserve user settings

If checked, the checkmarks and column names/types you explicitly entered are preserved even if you select a new file. By default, the analyzer starts with fresh default settings for each new file location.

Read row IDs

If checked, the first column in the file is used as row IDs. If not checked, default row headers are created.

Read column headers

The screenshot shows the 'File Reader (Reading adult.xls)' application window. The 'Settings' tab is active, displaying options for file location, column selection, and data handling. Below the settings is a table of data with columns: Row ID, I, age, S, workclass, I, hlthug, S, education, I, educat, S, marital, S, and S. The table contains 24 rows of data, including fields like 'State-gov', 'Self-emp-inc', 'Private', 'Unemp', 'Married-civ-spouse', 'Divorced', 'Never-married', 'Married-spouse', 'Married-civ-spouse', 'Married-spouse', 'Married-civ-spouse', 'Married-civ-spouse', 'Married-civ-spouse', 'Married-civ-spouse', 'Married-civ-spouse', 'Married-civ-spouse', 'Married-civ-spouse', 'Married-civ-spouse', 'Married-civ-spouse', 'Married-civ-spouse', 'Married-civ-spouse', 'Married-civ-spouse', 'Married-civ-spouse', 'Married-civ-spouse'.

Settings | **Plan Variables** | **Memory Policy**

Error ASCII data file location: (press 'Enter' to update preview)

Uri:

☐ Preserve user settings for new location

Basic Settings

☐ read row IDs

☒ read column headers ☒ ignore spaces and tabs ☐ skip-style comments

Preview

Click column header to change column properties (" = name/type user settings)

Row ID	I	age	S	workclass	I	hlthug	S	education	I	educat	S	marital	S
Row0	39	State-gov	77516	Bachelors	13	Never-married	Adv						
Row1	30	Self-emp-inc	83011	Bachelors	13	Married-civ-sp	Exc						
Row2	38	Private	215646	HS-grad	9	Divorced	Exc						
Row3	31	Private	234721	11th	7	Married-civ-sp	Exc						
Row4	38	Private	338409	Bachelors	13	Married-civ-sp	Prot						
Row5	37	Private	384652	Masters	14	Married-civ-sp	Exc						
Row6	49	Private	160183	2th	5	Married-spouse	OTH						
Row7	52	Self-emp-inc	200640	HS-grad	9	Married-civ-sp	Exc						
Row8	31	Private	45761	Masters	14	Never-married	Prot						
Row9	42	Private	159449	Bachelors	13	Married-civ-sp	Exc						
Row10	27	Private	280464	Some-college	30	Married-civ-sp	Exc						
Row11	30	State-gov	141297	Bachelors	13	Married-civ-sp	Prot						
Row12	23	Private	122272	Bachelors	13	Never-married	Adv						
Row13	32	Private	205019	Assoc-acadm	12	Never-married	Sale						
Row14	40	Private	121772	Assoc-voc	11	Married-civ-sp	Civ						
Row15	34	Private	245477	7th-8th	4	Married-civ-sp	Trde						
Row16	25	Self-emp-inc	176756	HS-grad	9	Never-married	Fan						
Row17	32	Private	188824	HS-grad	9	Never-married	Mac						
Row18	38	Private	28887	11th	7	Married-civ-sp	Sale						
Row19	43	Self-emp-inc	252175	Masters	14	Divorced	Exc						
Row20	40	Private	193524	Doctorate	16	Married-civ-sp	Prot						
Row21	34	Private	302146	HS-grad	9	Separated	OTH						
Row22	35	Federal-gov	76840	11th	5	Married-civ-sp	Fan						
Row23	41	Private	117633	11th	7	Married-civ-sp	Trde						
Row24	59	Private	109615	HS-grad	9	Divorced	Trde						

OK Apply Cancel

Node Context Options

Model Learner

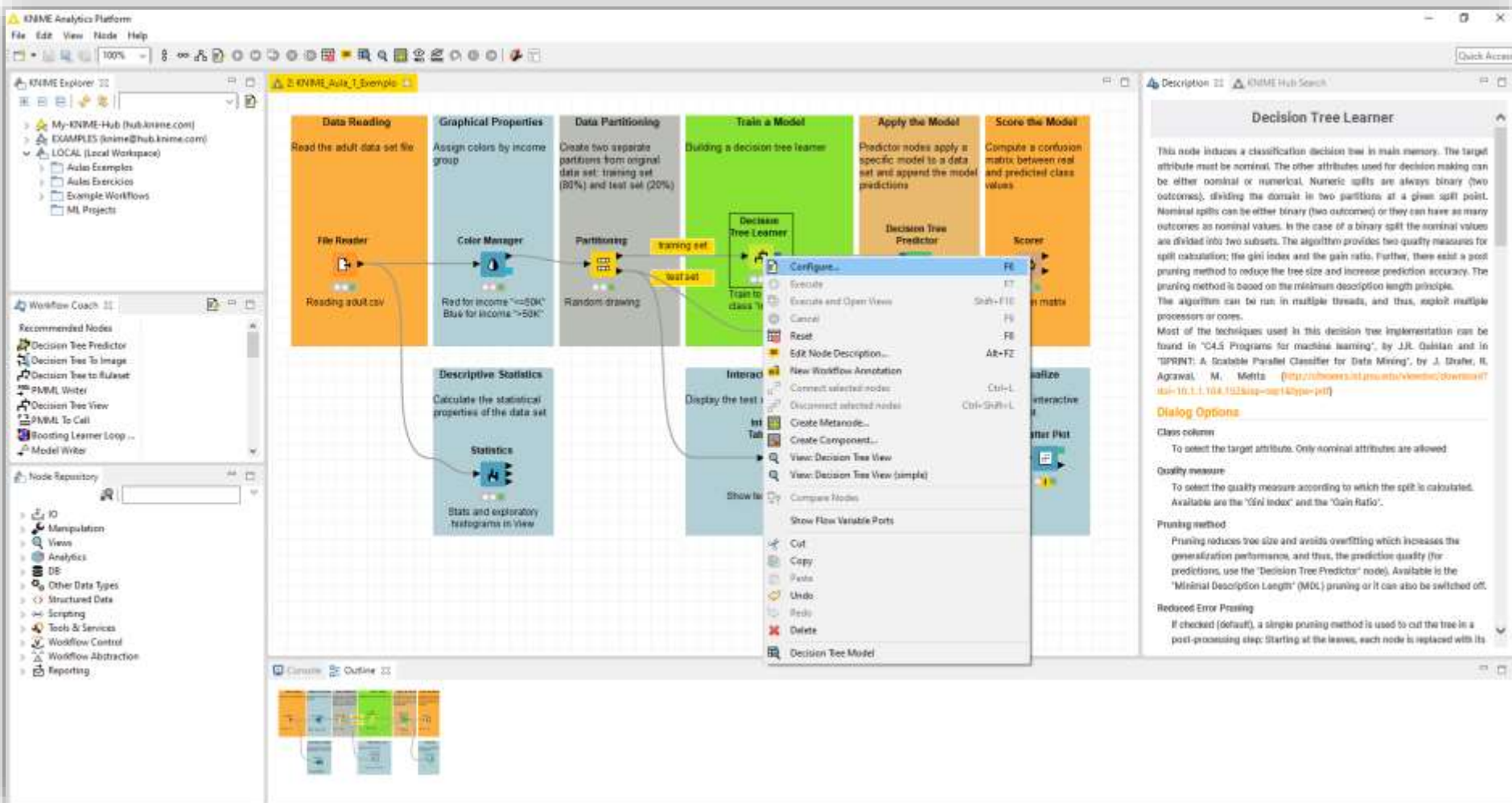
29

Terminology of AI

Intro to Knime

WORKFLOW

Hands On



KNIME Analytics Platform

File Edit View Node Help

100%

KNIME Explorer

- My-KNIME-Hub (hub.knime.com)
- EXAMPLES (knime@hub.knime.com)
- LOCAL (Local Workspace)
 - Auto Examples
 - Auto Exercises
 - Example Workflows
 - ML Projects

Workflow Coach

Recommended Nodes

- Decision Tree Predictor
- Decision Tree to Image
- Decision Tree to Ruleset
- PMML Writer
- Decision Tree View
- PMML To Call
- Boosting Learner Loop...
- Model Writer

Node Repository

- IO
- Manipulation
- Views
- Analytics
- DB
- Other Data Types
- Structured Data
- Scripting
- Tools & Services
- Workflow Control
- Workflow Abstraction
- Reporting

KNIME Auto_1.Example

Data Reading
Read the adult data set file
File Reader
Reading adult.csv

Graphical Properties
Assign colors by income group
Color Manager
Red for income "<=50K"
Blue for income ">50K"

Data Partitioning
Create two separate partitions from original data set: training set (80%) and test set (20%)
Partitioning
Random drawing

Train a Model
Building a decision tree learner
Decision Tree Learner
Train to class

Apply the Model
Predictor nodes apply a specific model to a data set and append the model predictions
Decision Tree Predictor

Score the Model
Compute a confusion matrix between real and predicted class values
Scorer

Descriptive Statistics
Calculate the statistical properties of the data set
Statistics
Stats and exploratory histograms in view

Interact
Display the test results
Int Tab
Show log

Context Menu:

- Execute
- Execute and Open Views
- Cancel
- Reset
- Edit Node Description...
- New Workflow Annotation
- Connect selected nodes
- Disconnect selected nodes
- Create Metanode...
- Create Component...
- View: Decision Tree View
- View: Decision Tree View (simple)
- Compare Nodes
- Show Flow Variable Ports
- Cut
- Copy
- Paste
- Undo
- Redo
- Delete
- Decision Tree Model

Decision Tree Learner

This node induces a classification decision tree in main memory. The target attribute must be nominal. The other attributes used for decision making can be either nominal or numerical. Numeric splits are always binary (two outcomes), dividing the domain in two partitions at a given split point. Nominal splits can be either binary (two outcomes) or they can have as many outcomes as nominal values. In the case of a binary split the nominal values are divided into two subsets. The algorithm provides two quality measures for split calculation: the gini index and the gain ratio. Further, there exist a post pruning method to reduce the tree size and increase prediction accuracy. The pruning method is based on the minimum description length principle. The algorithm can be run in multiple threads, and thus, exploit multiple processors or cores.

Most of the techniques used in this decision tree implementation can be found in "C4.5 Programs for machine learning", by J.J. Quinlan and in "SPRINT: A Scalable Parallel Classifier for Data Mining", by J. Shafer, R. Agrawal, M. Mehta (<http://pubsonline.informaworld.com/doi/10.1.1.108.1523/sep-sep14type=pdf>)

Dialog Options

Class column
To select the target attribute. Only nominal attributes are allowed.

Quality measure
To select the quality measure according to which the split is calculated. Available are the "Gini Index" and the "Gain Ratio".

Pruning method
Pruning reduces tree size and avoids overfitting which increases the generalization performance, and thus, the prediction quality (for predictions, use the "Decision Tree Predictor" node). Available is the "Minimal Description Length" (MDL) pruning or it can also be switched off.

Reduced Error Pruning
If checked (default), a simple pruning method is used to cut the tree in a post-processing step: Starting at the leaves, each node is replaced with its

Node Context Options

Model Learner

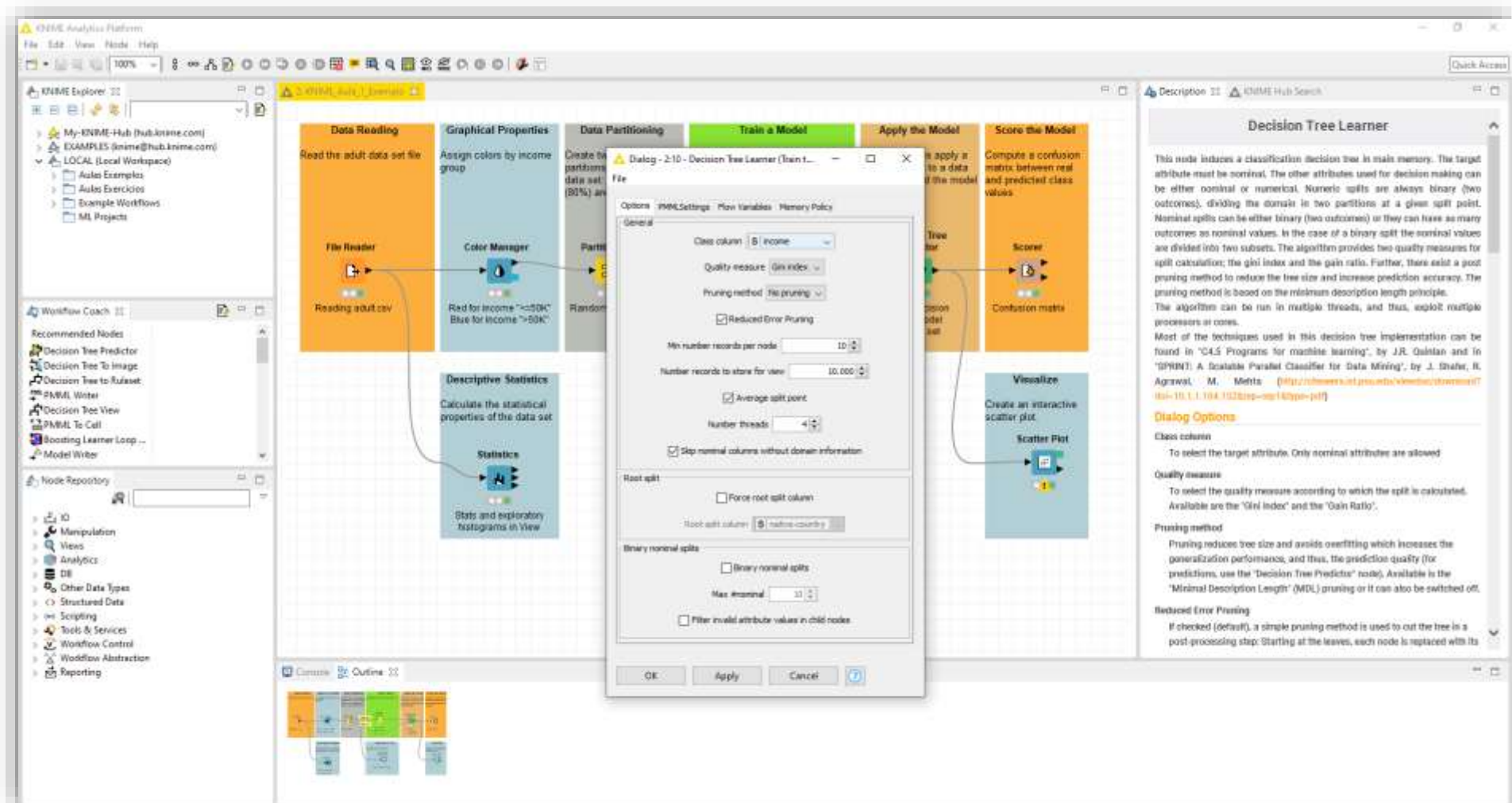
30

Terminology of AI

Intro to Knime

WORKFLOW

Hands On



KNIME Analysis Platform

File Edit View Node Help

100%

KNIME Explorer

- My-KNIME-Hub (hub.knime.com)
- EXAMPLES (knime@hub.knime.com)
- LOCAL (Local Workspace)
 - Auto Examples
 - Auto Exercises
 - Example Workflows
 - ML Projects

Workflow Coach

Recommended Nodes

- Decision Tree Predictor
- Decision Tree To Image
- Decision Tree To Ruleset
- PMML Writer
- Decision Tree View
- PMML To Call
- Boosting Learner Loop ...
- Model Writer

Node Repository

- IO
- Manipulation
- Views
- Analytics
- DB
- Other Data Types
- Structured Data
- Scripting
- Tools & Services
- Workflow Control
- Workflow Abstraction
- Reporting

2-10 - Decision Tree Learner (Train t...

Options: KNIME Settings | Flow Variables | Memory Policy

General

Class column: **Income**

Quality measure: **Gini index**

Pruning method: **No pruning**

☒ Reduced Error Pruning

Min number records per node: **10**

Number records to store for view: **10,000**

☒ Average split point

Number threads: **4**

☒ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column: **native-country**

Binary nominal splits

☐ Binary nominal splits

Max nominal: **32**

☐ Filter invalid attribute values in child nodes

OE Apply Cancel

Decision Tree Learner

This node induces a classification decision tree in main memory. The target attribute must be nominal. The other attributes used for decision making can be either nominal or numerical. Numeric splits are always binary (two outcomes), dividing the domain in two partitions at a given split point. Nominal splits can be either binary (two outcomes) or they can have as many outcomes as nominal values. In the case of a binary split the nominal values are divided into two subsets. The algorithm provides two quality measures for split calculation; the gini index and the gain ratio. Further, there exist a post pruning method to reduce the tree size and increase prediction accuracy. The pruning method is based on the minimal description length principle. The algorithm can be run in multiple threads, and thus, exploit multiple processors at cores.

Most of the techniques used in this decision tree implementation can be found in "C4.5 Programs for machine learning", by J.R. Quinlan and in "SPRINT: A Scalable Parallel Classifier for Data Mining", by J. Shafer, R. Agrawal, M. Mehta (<http://www.cs.cmu.edu/~viewpoints/shafer07/doi-10.1.1.104.1023brep-00p140p00.pdf>)

Dialog Options

Class column

To select the target attribute. Only nominal attributes are allowed

Quality measure

To select the quality measure according to which the split is calculated. Available are the "Gini Index" and the "Gain Ratio".

Pruning method

Pruning reduces tree size and avoids overfitting which increases the generalization performance, and thus, the prediction quality (for predictions, use the "Decision Tree Predicts" node). Available is the "Minimal Description Length" (MDL) pruning or it can also be switched off.

Reduced Error Pruning

If checked (default), a simple pruning method is used to cut the tree in a post-processing step: Starting at the leaves, each node is replaced with its

Node Context Options

Model Learner

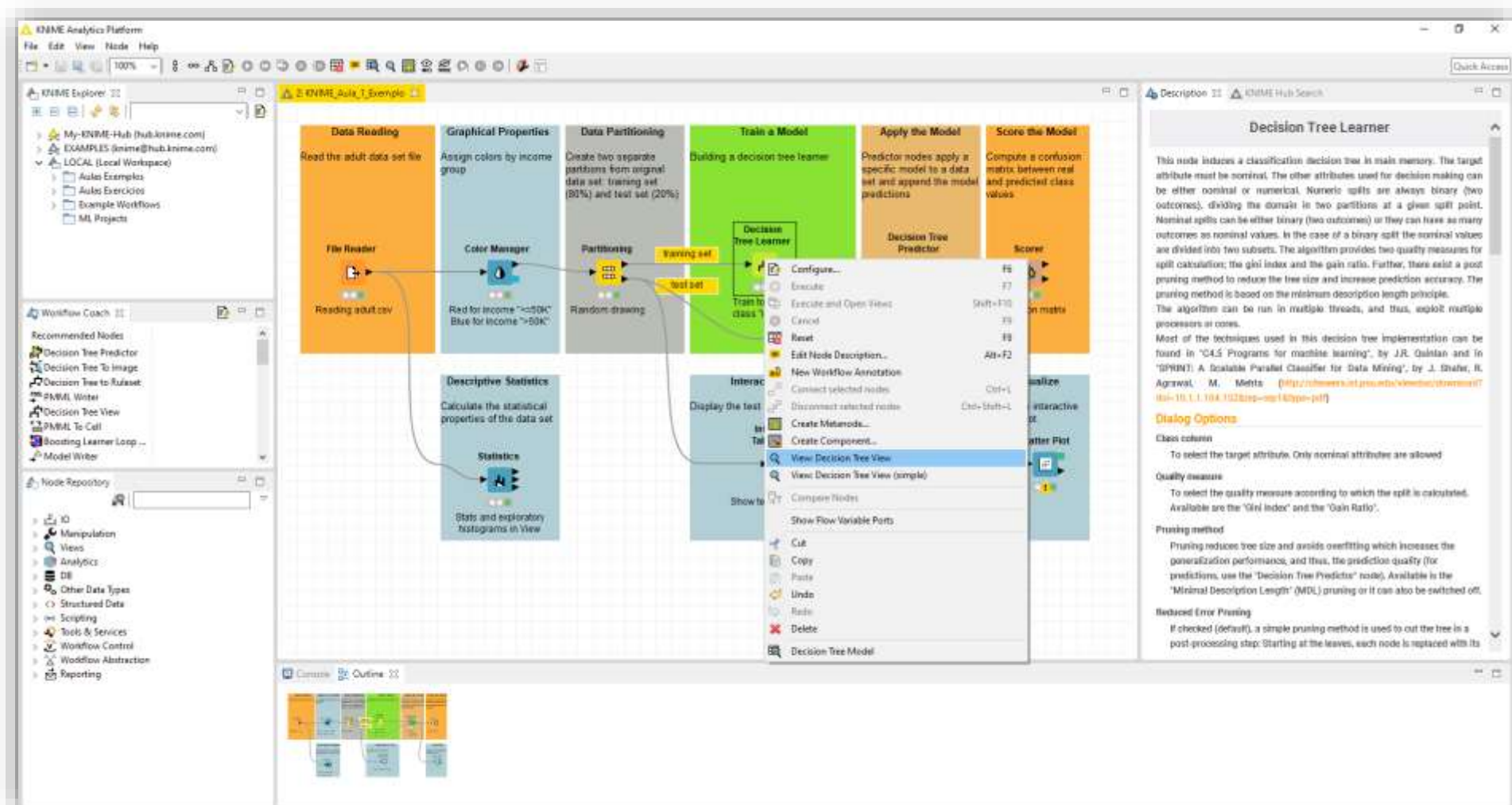
31

Terminology of AI

Intro to Knime

WORKFLOW

Hands On



KNIME Analytics Platform

File Edit View Node Help

100%

KNIME Explorer

- My-KNIME-Hub (hub.knime.com)
- EXAMPLES (knime@hub.knime.com)
- LOCAL (Local Workspace)
 - Auto Examples
 - Auto Exercises
 - Example Workflows
 - ML Projects

Workflow Coach

Recommended Nodes

- Decision Tree Predictor
- Decision Tree To Image
- Decision Tree To Ruleset
- PMML Writer
- Decision Tree View
- PMML To Call
- Boosting Learner Loop...
- Model Writer

Node Repository

- IO
- Manipulation
- Views
- Analytics
- DB
- Other Data Types
- Structured Data
- Scripting
- Tools & Services
- Workflow Control
- Workflow Abstraction
- Reporting

KNIME_Rule_1_Example

Data Reading
Read the adult data set file

Graphical Properties
Assign colors by income group

Data Partitioning
Create two separate partitions from original data set: training set (80%) and test set (20%)

Train a Model
Building a decision tree learner

Apply the Model
Predictor nodes apply a specific model to a data set and append the model predictions

Score the Model
Compute a confusion matrix between real and predicted class values

File Reader
Reading adult.csv

Color Manager
Red for income "<=50K"
Blue for income ">50K"

Partitioning
Random drawing

Decision Tree Learner
Configure...
Execute
Execute and Open View...
Cancel
Reset
Edit Node Description...
New Workflow Annotation
Connect selected nodes
Disconnect selected nodes
Create Metanode...
Create Component...
View Decision Tree View
View Decision Tree View (simple)
Compare Nodes
Show Flow Variable Ports
Cut
Copy
Paste
Undo
Redo
Delete
Decision Tree Model

Descriptive Statistics
Calculate the statistical properties of the data set

Statistics
Stats and exploratory histograms in View

Interact
Display the test

Scorer
Confusion matrix

Decision Tree Learner

This node induces a classification decision tree in main memory. The target attribute must be nominal. The other attributes used for decision making can be either nominal or numerical. Numeric splits are always binary (two outcomes), dividing the domain in two partitions at a given split point. Nominal splits can be either binary (two outcomes) or they can have as many outcomes as nominal values. In the case of a binary split the nominal values are divided into two subsets. The algorithm provides two quality measures for split calculation: the gini index and the gain ratio. Further, there exist a post pruning method to reduce the tree size and increase prediction accuracy. The pruning method is based on the minimum description length principle. Most of the techniques used in this decision tree implementation can be found in "C4.5 Programs for machine learning", by J.R. Quinlan and in "SPRINT: A Scalable Parallel Classifier for Data Mining", by J. Shalek, R. Agrawal, M. Mehta (<http://chewers.cs.toronto.edu/~mehta/chronos/DT-10-11-11-104-1023top-asp16/open.pdf>)

Dialog Options

Class column
To select the target attribute. Only nominal attributes are allowed

Quality measure
To select the quality measure according to which the split is calculated. Available are the "Gini Index" and the "Gain Ratio".

Pruning method
Pruning reduces tree size and avoids overfitting which increases the generalization performance, and thus, the prediction quality (for predictions, use the "Decision Tree Predicts" node). Available in the "Minimal Description Length" (MDL) pruning or it can also be switched off.

Reduced Error Pruning
If checked (default), a simple pruning method is used to cut the tree in a post-processing step: Starting at the leaves, each node is replaced with its

Node Context Options Model Learner

32

Terminology of AI

Intro to Knime

WORKFLOW

Hands On

Decision Tree View - 2/10 - Decision Tree Learner (Train to predict)

File H1Like Tree

Decision Tree Learner

This node induces a classification decision tree in main memory. The target attribute must be nominal. The other attributes used for decision making can be either nominal or numerical. Numeric splits are always binary (two outcomes), dividing the domain in two partitions at a given split point. Nominal splits can be either binary (two outcomes) or they can have as many outcomes as nominal values. In the case of a binary split the nominal values are divided into two subsets. The algorithm provides two quality measures for split calculation; the gini index and the gain ratio. Further, there exist a post pruning method to reduce the tree size and increase prediction accuracy. The

Decision Tree View - 2/10 - Decision Tree Learner (Train to predict)

File H1Like Tree

Decision Tree Learner

This node induces a classification decision tree in main memory. The target attribute must be nominal. The other attributes used for decision making can be either nominal or numerical. Numeric splits are always binary (two outcomes), dividing the domain in two partitions at a given split point. Nominal splits can be either binary (two outcomes) or they can have as many outcomes as nominal values. In the case of a binary split the nominal values are divided into two subsets. The algorithm provides two quality measures for split calculation; the gini index and the gain ratio. Further, there exist a post pruning method to reduce the tree size and increase prediction accuracy. The

Decision Tree View - 2/10 - Decision Tree Learner (Train to predict)

File H1Like Tree

Decision Tree Learner

This node induces a classification decision tree in main memory. The target attribute must be nominal. The other attributes used for decision making can be either nominal or numerical. Numeric splits are always binary (two outcomes), dividing the domain in two partitions at a given split point. Nominal splits can be either binary (two outcomes) or they can have as many outcomes as nominal values. In the case of a binary split the nominal values are divided into two subsets. The algorithm provides two quality measures for split calculation; the gini index and the gain ratio. Further, there exist a post pruning method to reduce the tree size and increase prediction accuracy. The

Node Context Options

Scorer

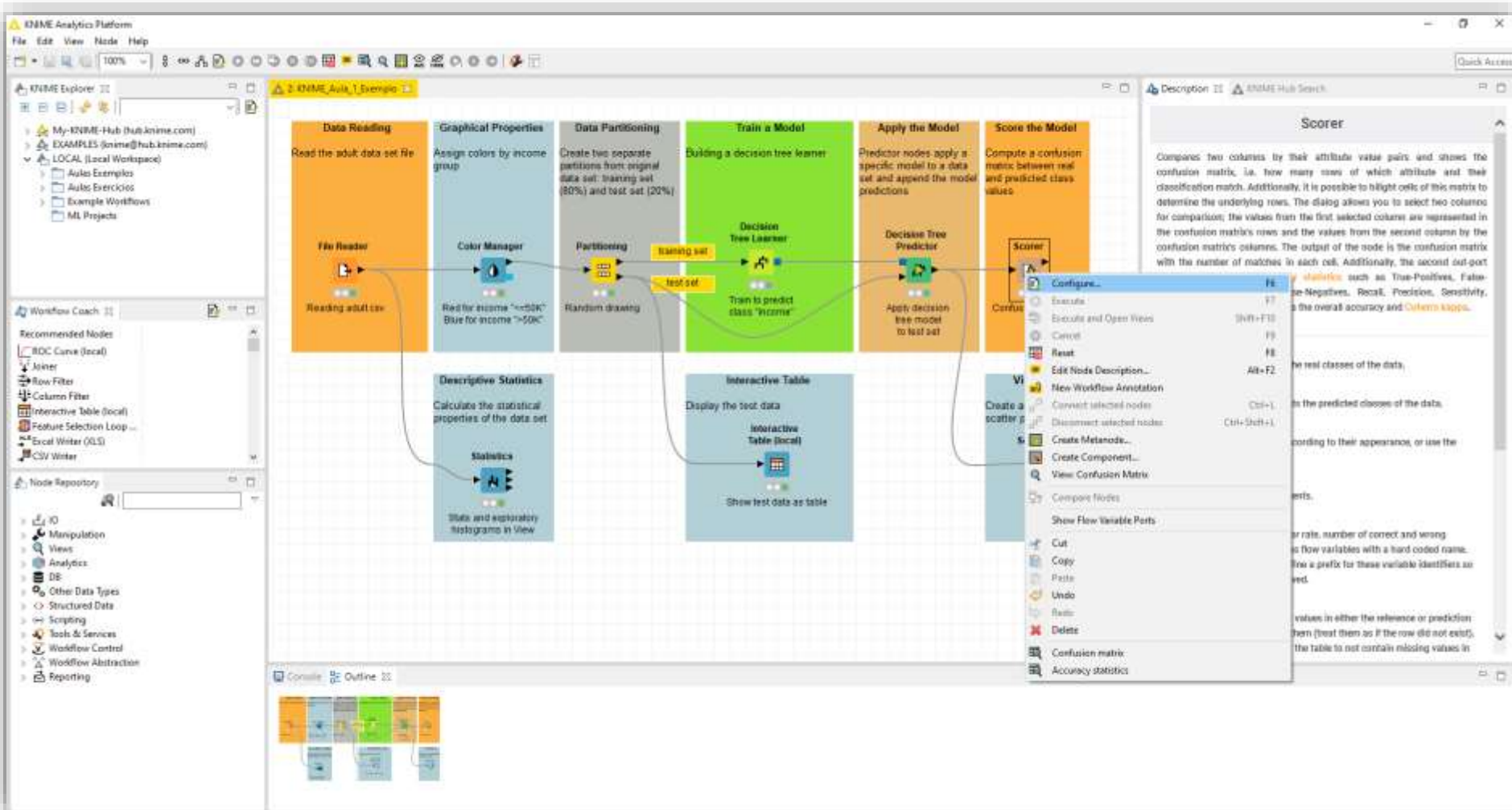
33

Terminology of AI

Intro to Knime

WORKFLOW

Hands On



The screenshot displays the KNIME Analytics Platform interface. The main workspace shows a workflow titled "2. KNIME_Auto_1.Example". The workflow consists of several nodes: "File Reader" (Data Reading), "Color Manager" (Graphical Properties), "Partitioning" (Data Partitioning), "Decision Tree Learner" (Train a Model), "Decision Tree Predictor" (Apply the Model), and "Scorer" (Score the Model). The "Scorer" node is selected, and its context menu is open, showing options like "Configure...", "Execute", "Cancel", "Reset", "Edit Node Description...", "New Workflow Annotation", "Connect selected nodes", "Disconnect selected nodes", "Create Metanode...", "Create Component...", "View Confusion Matrix", "Compare Nodes", "Show Flow Variable Ports", "Cut", "Copy", "Paste", "Undo", "Redo", "Delete", "Confusion matrix", and "Accuracy statistics".

The "Scorer" node description is visible on the right side of the interface:

Scorer

Compares two columns by their attribute value pairs and shows the confusion matrix, i.e. how many rows of which attribute and their classification match. Additionally, it is possible to highlight cells of this matrix to determine the underlying rows. The dialog allows you to select two columns for comparison; the values from the first selected column are represented in the confusion matrix's rows and the values from the second column by the confusion matrix's columns. The output of the node is the confusion matrix with the number of matches in each cell. Additionally, the second out-port shows statistics such as True-Positives, False-Positives, Recall, Precision, Sensitivity, and the overall accuracy and Cohen's kappa.

Node Context Options Scorer

34

Terminology of AI

Intro to Knime

WORKFLOW

Hands On

The screenshot displays the KNIME Analysis Platform interface. The main workspace shows a workflow with the following nodes: File Reader (Data Reading), Color Manager (Graphical Properties), Partitioner (Data Partitioning), Decision Tree Learner (Train a Model), Decision Tree (Apply the Model), and Scorer (Score the Model). A 'Dialog - 2d - Scorer (Confusion matrix)' is open, showing the configuration for the Scorer node. The dialog includes fields for 'First Column' (Income) and 'Second Column' (Prediction (Income)), a 'Sorting strategy' dropdown (set to 'Insertion order'), and checkboxes for 'Provide scores as flow variables' and 'Missing values' (set to 'Ignore').

The right-hand pane shows the 'Scorer' node context options. The description states: 'Compares two columns by their attribute value pairs and shows the confusion matrix, i.e. how many rows of which attribute and their classification match. Additionally, it is possible to highlight cells of this matrix to determine the underlying rows. The dialog allows you to select two columns for comparison; the values from the first selected column are represented in the confusion matrix's rows and the values from the second column by the confusion matrix's columns. The output of the node is the confusion matrix with the number of matches in each cell. Additionally, the second out-port reports a number of **accuracy statistics** such as True-Positives, False-Positives, True-Negatives, False-Negatives, Recall, Precision, Sensitivity, Specificity, F-measure, as well as the overall accuracy and **Cohen's kappa**.'

The 'Dialog Options' section provides further details:

- First column:** The first column represents the real classes of the data.
- Second column:** The second column represents the predicted classes of the data.
- Sorting strategy:** Whether to sort the labels according to their appearance, or use the lexical/numeric ordering.
- Reverse order:** Reverse the order of the elements.
- Use name prefix:** The scores (i.e. accuracy, error rate, number of correct and wrong classification) are exported as flow variables with a hard coded name. This option allows you to define a prefix for these variable identifiers so that name conflicts are resolved.
- Missing Values:** Choose how to treat missing values in either the reference or prediction column. Default is to ignore them (treat them as if the row did not exist). Alternatively, you can expect the table to not contain missing values in

Node Context Options Scorer

35

Terminology of AI

Intro to Knime

WORKFLOW

Hands On

The screenshot displays the KNIME Analytics Platform interface. The main workspace shows a workflow titled "2. KNIME_Auto_1.Example". The workflow consists of several nodes: "File Reader" (Data Reading), "Color Manager" (Graphical Properties), "Partitioning" (Data Partitioning), "Decision Tree Learner" (Train a Model), "Decision Tree Predictor" (Apply the Model), and "Scorer" (Score the Model). The "Scorer" node is selected, and its context menu is open, showing options like "Configure...", "Execute", "Execute and Open View", "Cancel", "Reset", "Edit Node Description...", "New Workflow Annotation", "Connect selected nodes", "Disconnect selected nodes", "Create Metanode...", "Create Component...", "View Confusion Matrix", "Compare Models", "Show Flow Variable Ports", "Cut", "Copy", "Paste", "Undo", "Redo", "Delete", "Confusion matrix", and "Accuracy statistics".

The "Scorer" node description is visible on the right side of the interface:

Scorer

Compares two columns by their attribute value pairs and shows the confusion matrix, i.e. how many rows of which attribute and their classification match. Additionally, it is possible to highlight cells of this matrix to determine the underlying rows. The dialog allows you to select two columns for comparison; the values from the first selected column are represented in the confusion matrix's rows and the values from the second column by the confusion matrix's columns. The output of the node is the confusion matrix with the number of matches in each cell. Additionally, the second out-port

The context menu options for the "Scorer" node are:

- Configure...
- Execute
- Execute and Open View
- Cancel
- Reset
- Edit Node Description...
- New Workflow Annotation
- Connect selected nodes
- Disconnect selected nodes
- Create Metanode...
- Create Component...
- View Confusion Matrix
- Compare Models
- Show Flow Variable Ports
- Cut
- Copy
- Paste
- Undo
- Redo
- Delete
- Confusion matrix
- Accuracy statistics

Node Context Options

Scorer

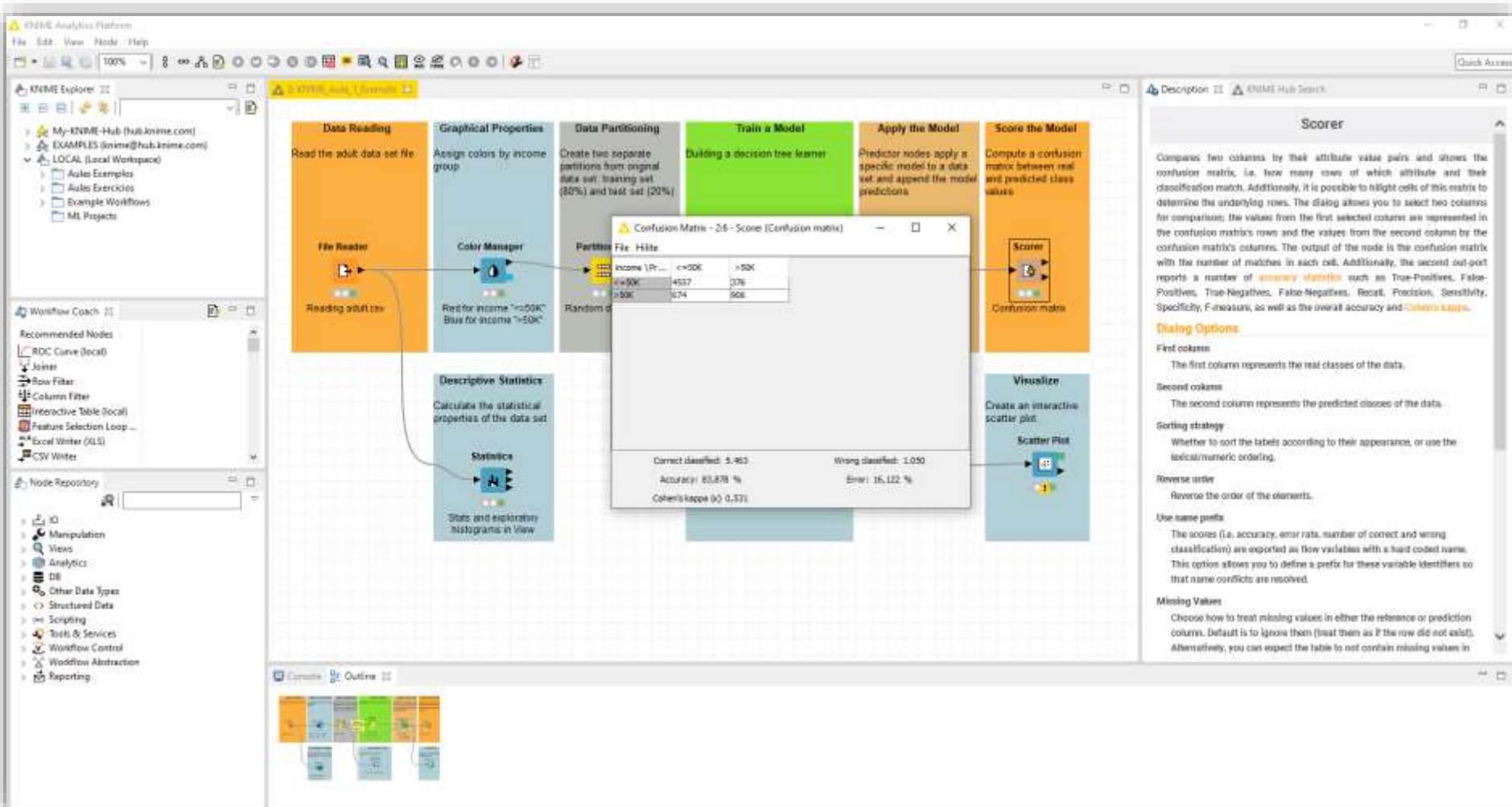
36

Terminology of AI

Intro to Knime

WORKFLOW

Hands On



The screenshot displays the KNIME Analysis Platform interface. The main workspace shows a workflow with several nodes: File Reader, Color Manager, Partition File, Train a Model, Apply the Model, and Scorer. A detailed view of the Scorer node context options is shown on the right side of the interface.

Scorer

Compares two columns by their attribute value pairs and shows the confusion matrix, i.e. how many rows of which attribute and their classification match. Additionally, it is possible to highlight cells of this matrix to determine the underlying rows. The dialog allows you to select two columns for comparison; the values from the first selected column are represented in the confusion matrix's rows and the values from the second column by the confusion matrix's columns. The output of the node is the confusion matrix with the number of matches in each cell. Additionally, the second out-port reports a number of **accuracy statistics** such as True-Positives, False-Positives, True-Negatives, False-Negatives, Recall, Precision, Sensitivity, Specificity, F-measure, as well as the overall accuracy and **Cohen's kappa**.

Dialog Options

First column:
The first column represents the real classes of the data.

Second column:
The second column represents the predicted classes of the data.

Sorting strategy:
Whether to sort the labels according to their appearance, or use the lexical/numeric ordering.

Reverse order:
Reverse the order of the elements.

Use name prefix:
The scores (i.e. accuracy, error rate, number of correct and wrong classification) are exported as flow variables with a hard coded name. This option allows you to define a prefix for these variable identifiers so that name conflicts are resolved.

Missing Values:
Choose how to treat missing values in either the reference or prediction column. Default is to ignore them (treat them as if the row did not exist). Alternatively, you can expect the table to not contain missing values in

Confusion Matrix - 2x2 - Score (Confusion matrix)

Income (P...)	<=50K	>50K
<=50K	4537	376
>50K	374	968

Correct classified: 5,963
Accuracy: 83,878 %
Cohen's kappa (κ): 0,571

Wrong classified: 1,030
Error: 16,122 %

Hands On

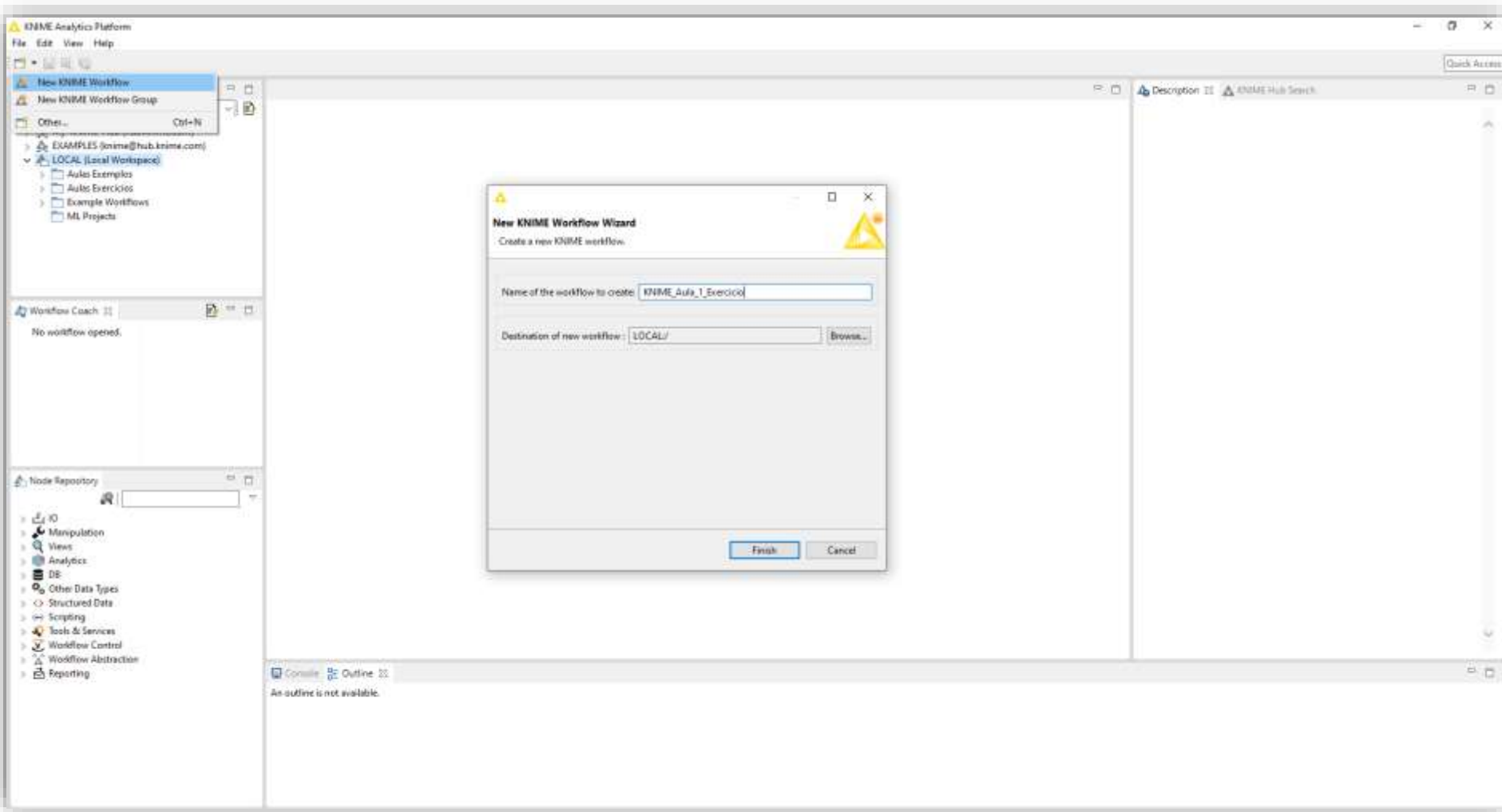
37

Terminology of AI

Intro to Knime

Workflow

HANDS ON



Hands On

38

Terminology of AI

Intro to Knime

Workflow

HANDS ON

HANDS ON

