



University of Minho
School of Engineering



Machine Learning with Knime

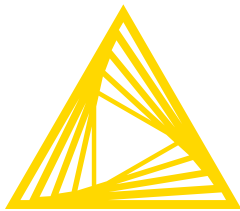
Similarity Based Systems

Perfil ML:FA@MiEI/4º ano - 1º Semestre

@MES/2º ano - 1º Semestre

Bruno Fernandes, Paulo Novais

26/09/2019



Open for Innovation[®]
KNIME

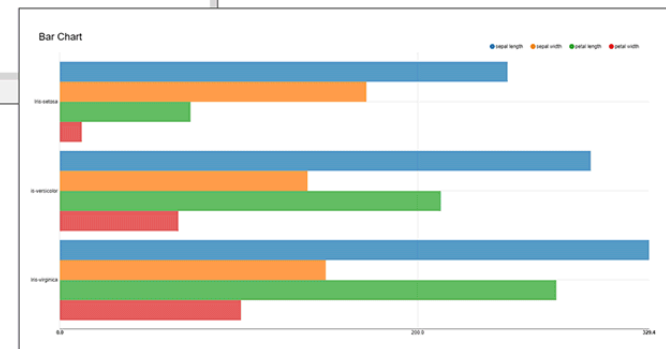
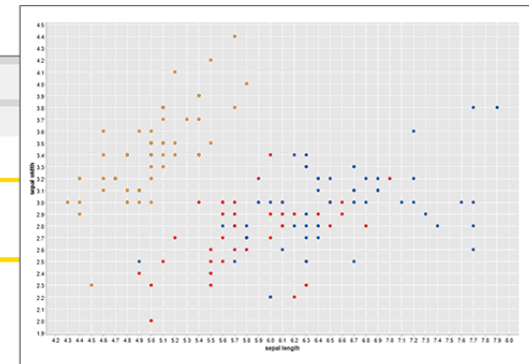
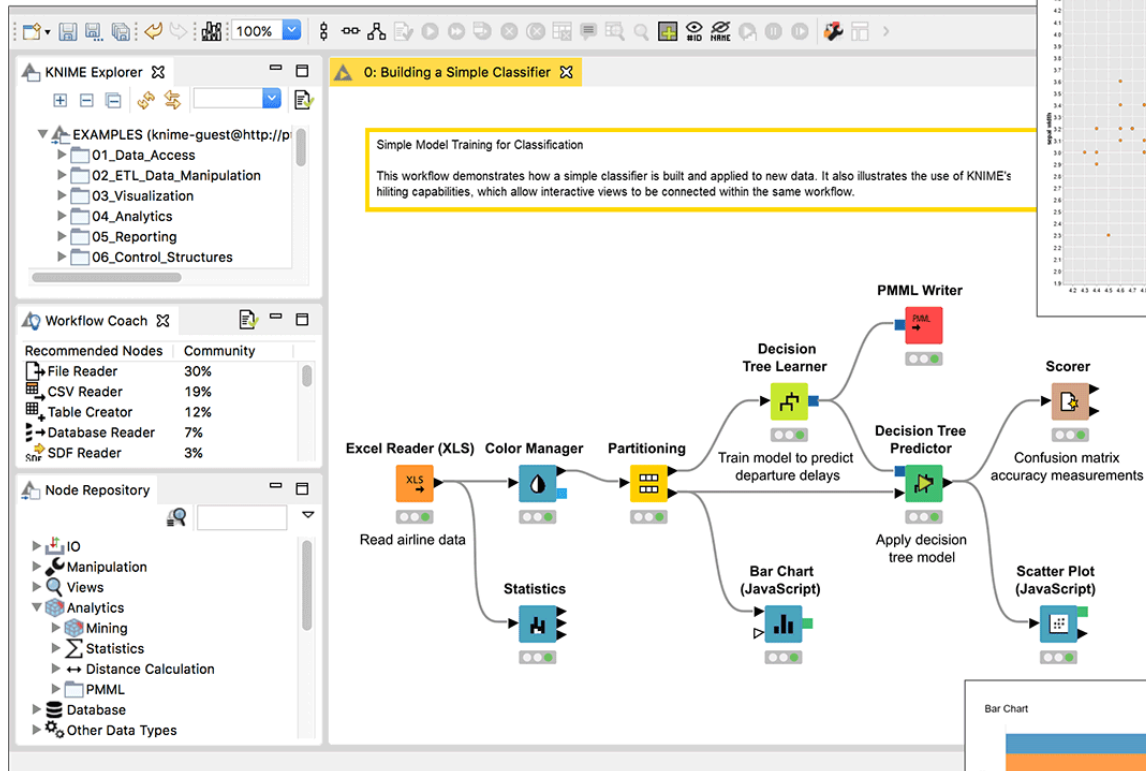
2

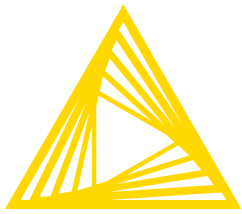
KNIME

Setup

Trying It

Hands On





Open for Innovation ®

KNIME

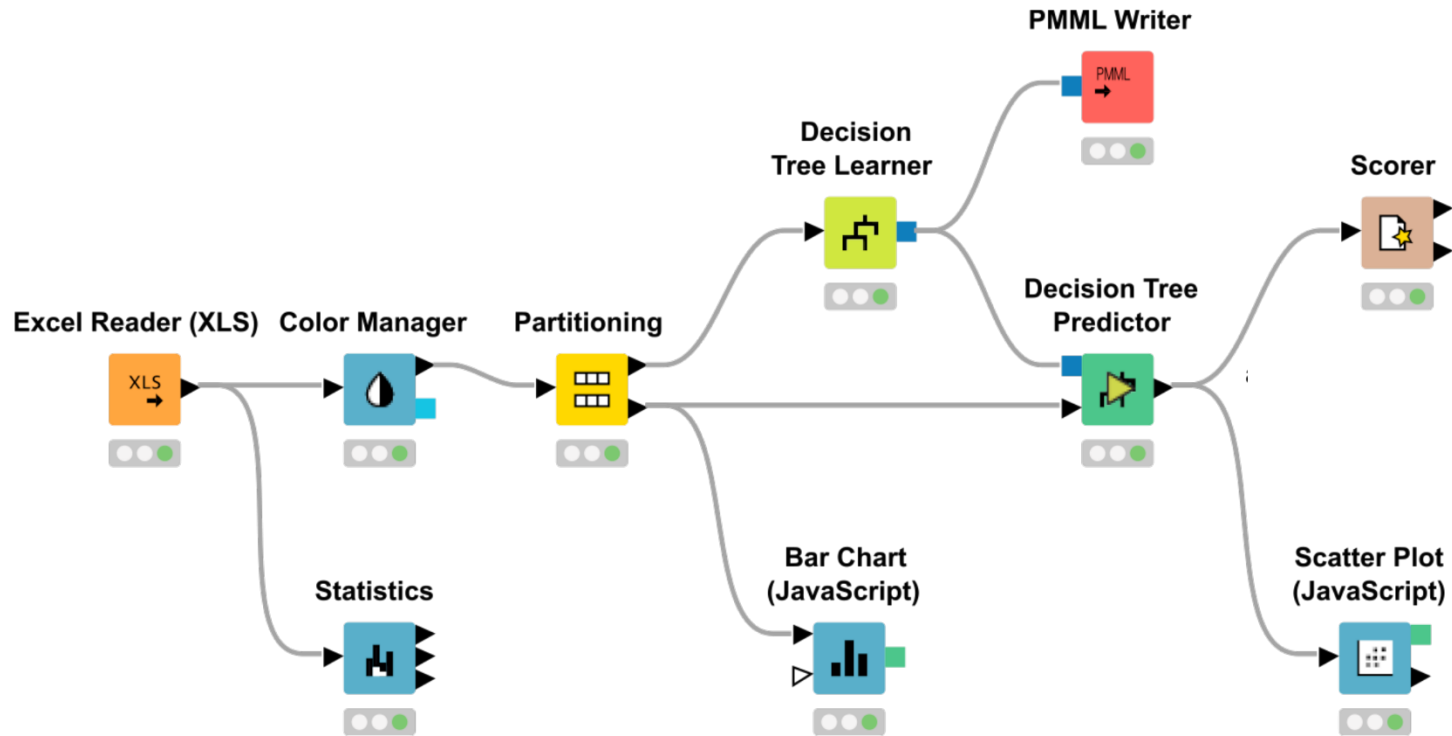
3

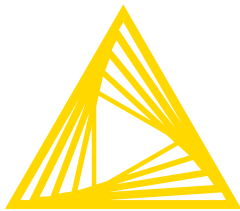
KNIME

Setup

Trying It

Hands On





Open for Innovation ®

KNIME

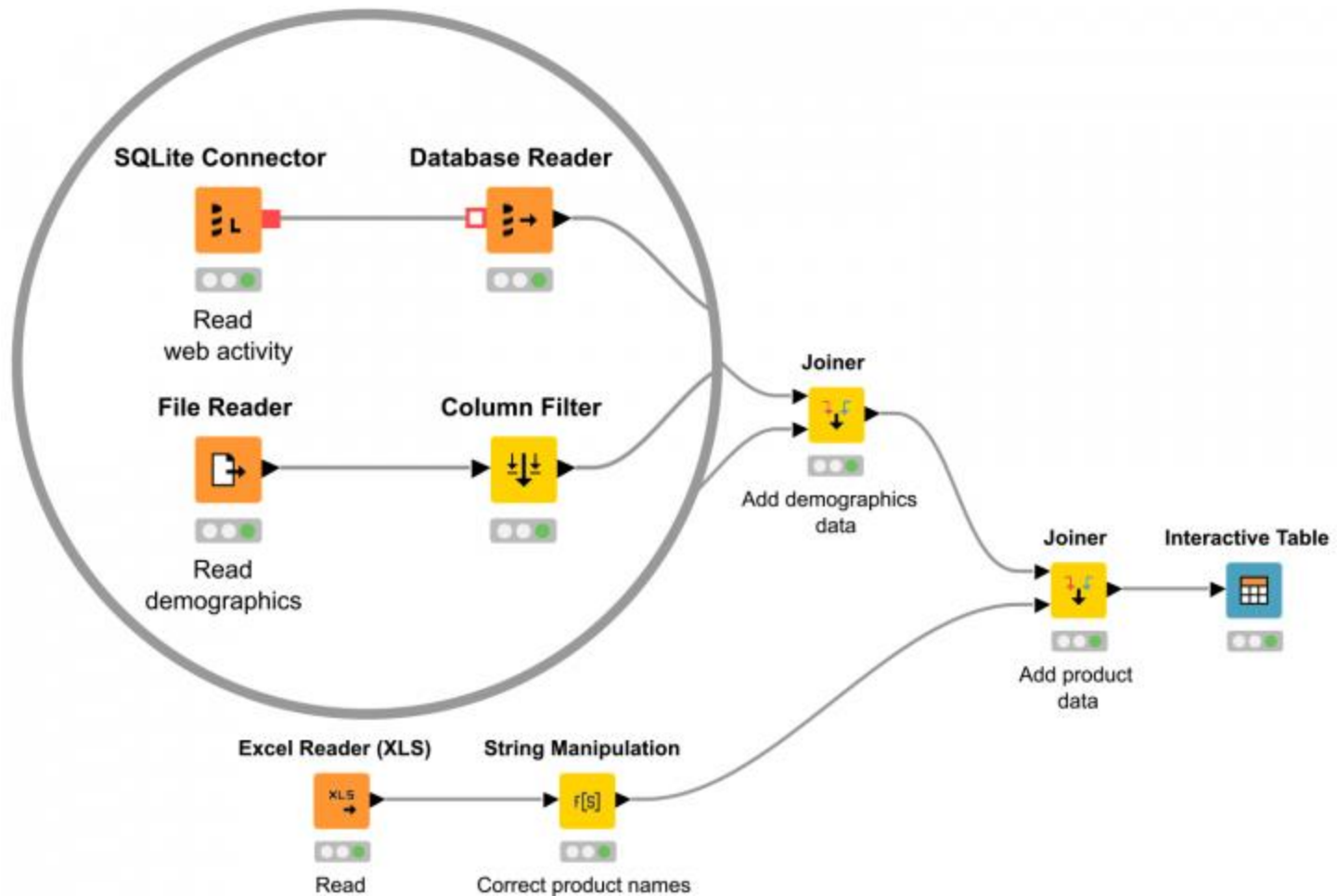
4

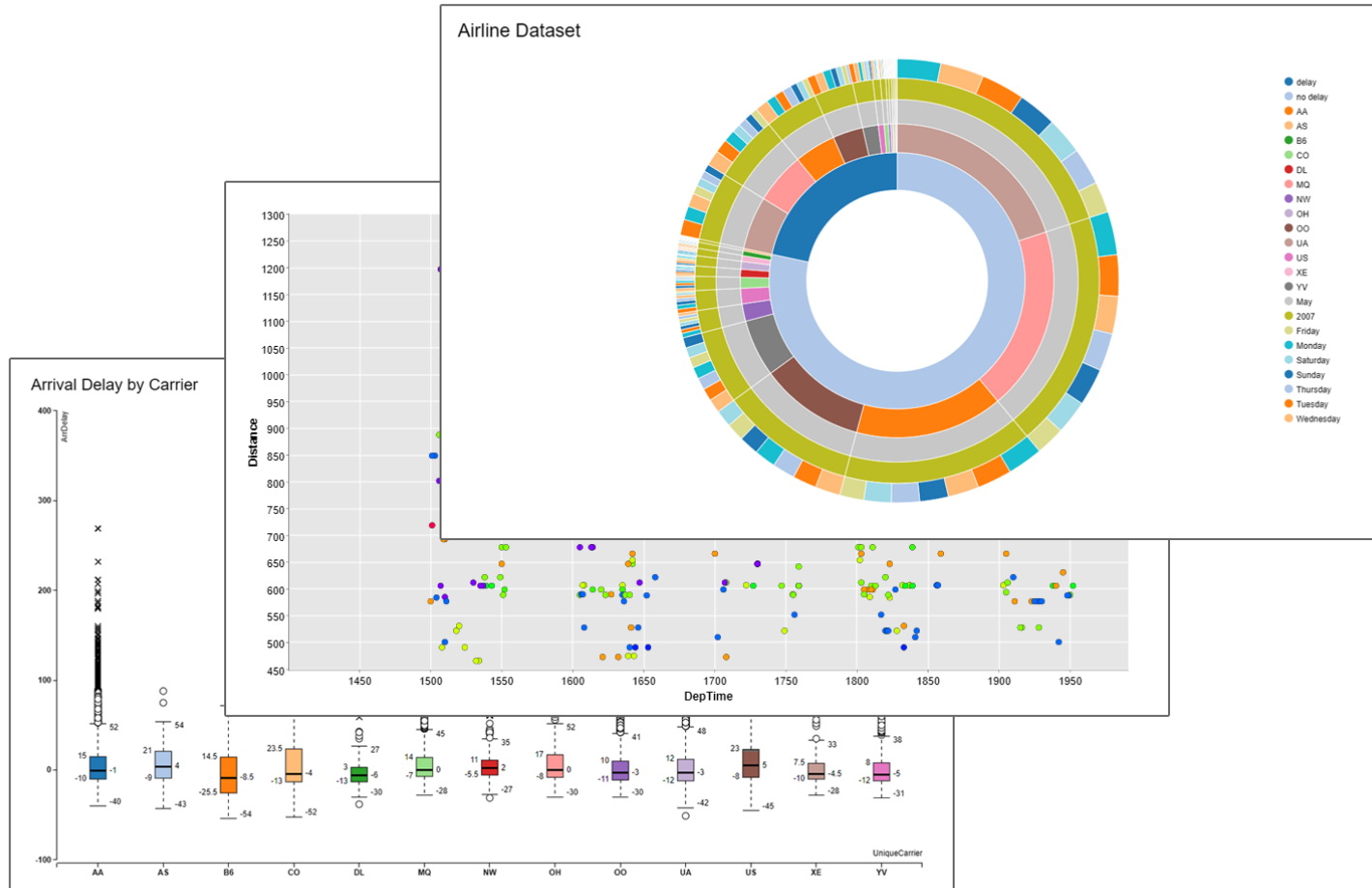
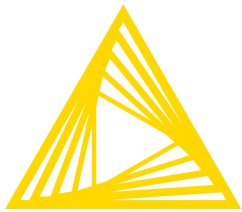
KNIME

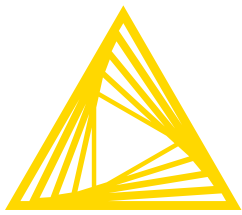
Setup

Trying It

Hands On







Open for Innovation ®

KNIME

6

KNIME

Setup

Trying It

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



@ <https://www.knime.com/>

7

KNIME

SETUP

Trying It

Hands On



[Blog](#) [Forum](#) [Events](#) [Career](#) [Contact](#)

[Download](#)

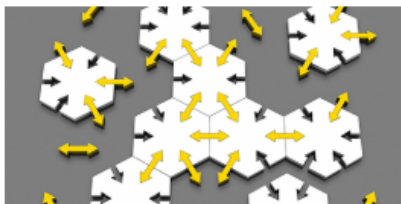
[SOFTWARE](#) / [SOLUTIONS](#) / [LEARNING](#) / [PARTNERS](#) / [COMMUNITY](#) / [ABOUT](#)

KNIME Fall Summit - Austin

November 6 - 9, 2018

[Learn More](#)

... for Developers



... for Data Scientists



... for Decision Makers



Sep 2018
Model Deployment with KNIME Server and Amazon API Gateway

10 Sep 2018
Fun With Tags

03 Sep 2018
Productionizing Data Science with KNIME Server

[more news](#)

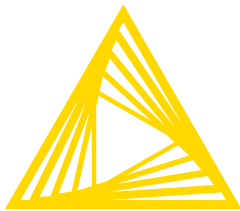
Events

11 Sep 2018 - 09 Oct 2018
KNIME Online Course for Beginners - September/October 2018

24 Sep 2018
KNIME Meetup in Warsaw

25 Sep 2018
Data Science Learnathon: From Raw Data to Deployment, Hamburg

[more events](#)



Open for Innovation [®]

KNIME


8

KNIME

SETUP

Trying It

Hands On

Open for Innovation [®]

KNIME

Hub Blog Forum Events Careers Contact [Download](#)

SOFTWARE / SOLUTIONS / LEARNING / PARTNERS / COMMUNITY / ABOUT

You are here: Home

Download KNIME Analytics Platform

[1 Register for Help & Updates](#) [2 Download KNIME](#) [3 Get Started](#)

Download the latest KNIME Analytics Platform for Windows, Linux, and Mac OS X.

KNIME 4.0.1

Find out [What's New in the new release](#) [here](#).

The KNIME Analytics Platform version is intended for end users and provides everything needed to immediately begin using KNIME as well as extend KNIME with extension packages developed by others.

Windows

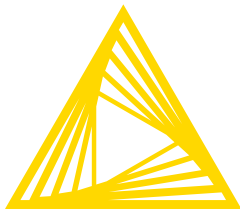
KNIME Analytics Platform for Windows (installer)	64 Bit (441.03 MB)
<i>The installer adds an icon to the desktop and suggests suitable memory settings</i>	32 Bit (437.42 MB)
KNIME Analytics Platform for Windows (self-extracting archive)	64 Bit (444.58 MB)
<i>The self-extracting archive only creates a folder holding the KNIME installation</i>	32 Bit (441.15 MB)
KNIME Analytics Platform for Windows (zip archive)	64 Bit (529.54 MB)
	32 Bit (525.59 MB)

Linux

KNIME Analytics Platform for Linux	64 Bit (554.2 MB)
------------------------------------	-------------------

Mac

KNIME Analytics Platform for Mac OSX (10.11 and above)	64 Bit (522.98 MB)
--	--------------------



Open for Innovation [®]

KNIME

9

KNIME

SETUP

Trying It

Hands On

The screenshot displays the KNIME Analytics Platform interface. The top menu bar includes File, Edit, View, and Help. The left sidebar contains the KNIME Explorer, Workflow Coach, and Node Repository. The main workspace shows a 'Welcome to KNIME Analytics Platform' screen with a 'Welcome' message and three cards: 'Get started with this example', 'Looking for more examples? Visit the KNIME Hub', and 'Sign up for introductory emails'. The bottom status bar indicates 'An outline is not available.'

KNIME Explorer

- My-KNIME-Hub (hub.knime.com)
- EXAMPLES (knime@hub.knime.com)
- LOCAL (Local Workspace)
 - Example Workflows

Workflow Coach

Recommended Nodes	Community
File Reader	24%
CSV Reader	18%
Excel Reader (XLS)	17%
Table Creator	12%

Node Repository

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

Welcome to KNIME Analytics Platform

Welcome

Looks like you're using KNIME for the first time...

Get started with this example

Open workflow

Looking for more examples? Visit the KNIME Hub

KNIME Hub

Sign up for introductory emails

These messages will get you up and running as quickly as possible.

Sign up

Console **Outline**

An outline is not available.

KNIME - Hands On

10

KNIME

SETUP

Trying It

Hands On

- Download KNIME
- Install it!
- Try it!

HANDS ON

Building a Simple Workflow

11

KNIME

Setup

TRYING IT

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)
 - Aulas Exemplos
 - Aulas Exercicios
 - 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 Cell
- 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_Aula_1_Exemplo

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
training set
test set

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

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

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

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

Interactive Table
Display the test data
Interactive Table (local)
Show test data as table

Visualize
Create an interactive scatter plot.
Scatter Plot

Description KNIME Hub Search

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.R. Quinlan and in "SPRINT: A Scalable Parallel Classifier for Data Mining", by J. Shafer, R. Agrawal, M. Mehta (<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.152&rep=rep1&type=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

Console Outline

Node Context Options File Reader

12

KNIME

Setup

TRYING IT

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 (Graphical Properties)
- Data Partitioning:** Partitioning (Create two separate partitions from original data set: training set (80%) and test set (20%)) -> Interactive Table (Display the test data)
- Train a Model:** Decision Tree Learner (Building a decision tree learner) -> Decision Tree Predictor (Apply the Model)
- Score the Model:** Scorer (Compute a confusion matrix between real and predicted class values) -> Visualize (Create an interactive scatter plot)

The File Reader node is selected, and its context menu is open, showing the following options:

- Configure... (F6)
- Execute (F7)
- Execute and Open Views (Shift+F10)
- Cancel (F9)
- Reset (F8)
- Edit Node Description...
- New Workflow Annotation (Alt+F2)
- 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

The right sidebar shows the **File Reader** node description:

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.

Rescan

If clicked, the file content is analyzed again. All settings are reset (unless the "Preserve user settings" option is selected) and the file is read in again to pre-set new settings and the table structure.

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

Node Context Options

File Reader

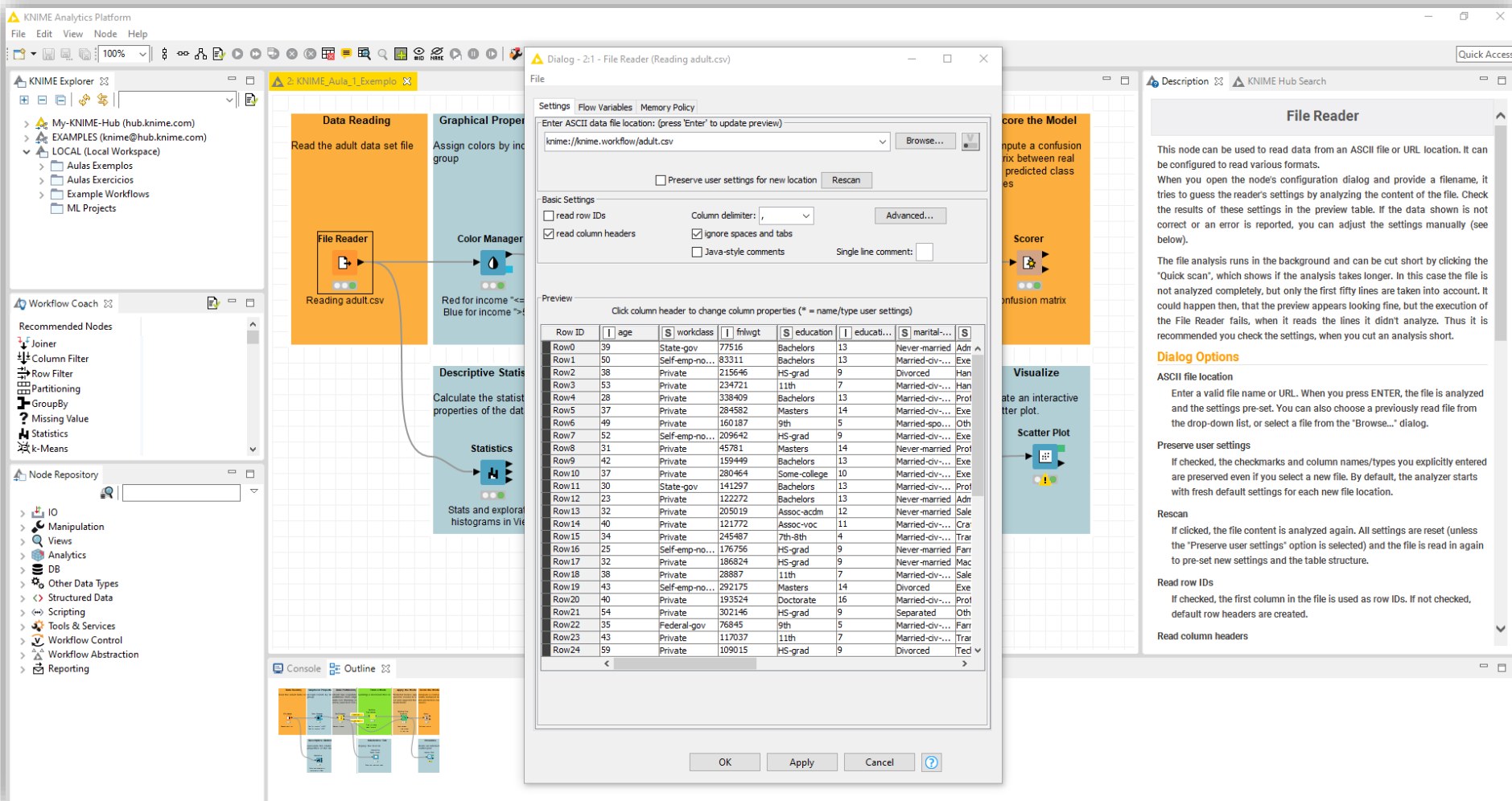
13

KNIME

Setup

TRYING IT

Hands On



The screenshot displays the KNIME Analytics Platform interface with the File Reader node configuration dialog open. The dialog is titled "Dialog - 2:1 - File Reader (Reading adult.csv)" and shows the following settings:

- File:** Enter ASCII data file location: (press 'Enter' to update preview) `knime://knime.workflow/adult.csv`
- Basic Settings:**
 - ☐ read row IDs
 - ☒ read column headers
 - Column delimiter: `,`
 - ☒ ignore spaces and tabs
 - ☐ Java-style comments
 - Single line comment:
- Preview:** Click column header to change column properties (* = name/type user settings)

Row ID	age	workclass	fnlwgt	education	educati...	marital...
Row0	39	State-gov	77516	Bachelors	13	Never-married
Row1	50	Self-emp-no...	83311	Bachelors	13	Married-div...
Row2	38	Private	215646	HS-grad	9	Divorced
Row3	53	Private	234721	11th	7	Married-div...
Row4	28	Private	338409	Bachelors	13	Prof
Row5	37	Private	284582	Masters	14	Married-div...
Row6	49	Private	160187	9th	5	Married-spo...
Row7	52	Self-emp-no...	209642	HS-grad	9	Married-div...
Row8	31	Private	45781	Masters	14	Never-married
Row9	42	Private	159449	Bachelors	13	Married-div...
Row10	37	Private	280464	Some-college	10	Married-div...
Row11	30	State-gov	141297	Bachelors	13	Married-div...
Row12	23	Private	122272	Bachelors	13	Never-married
Row13	32	Private	205019	Assoc-acdm	12	Never-married
Row14	40	Private	121772	Assoc-roc	11	Married-div...
Row15	34	Private	245487	7th-8th	4	Married-div...
Row16	25	Self-emp-no...	176756	HS-grad	9	Never-married
Row17	32	Private	186824	HS-grad	9	Never-married
Row18	38	Private	28887	11th	7	Married-div...
Row19	43	Self-emp-no...	292175	Masters	14	Divorced
Row20	40	Private	193524	Doctorate	16	Married-div...
Row21	54	Private	302146	HS-grad	9	Separated
Row22	35	Federal-gov	76845	9th	5	Married-div...
Row23	43	Private	117037	11th	7	Married-div...
Row24	59	Private	109015	HS-grad	9	Divorced

The background shows the KNIME Explorer with a workflow named "2: KNIME_Aula_1.Exemplo" containing a "File Reader" node. The "Graphical Properties" pane shows the "Color Manager" and "Descriptive Statistics" nodes. The "Console" and "Outline" panes are also visible at the bottom.

Node Context Options

Decision Tree Learner

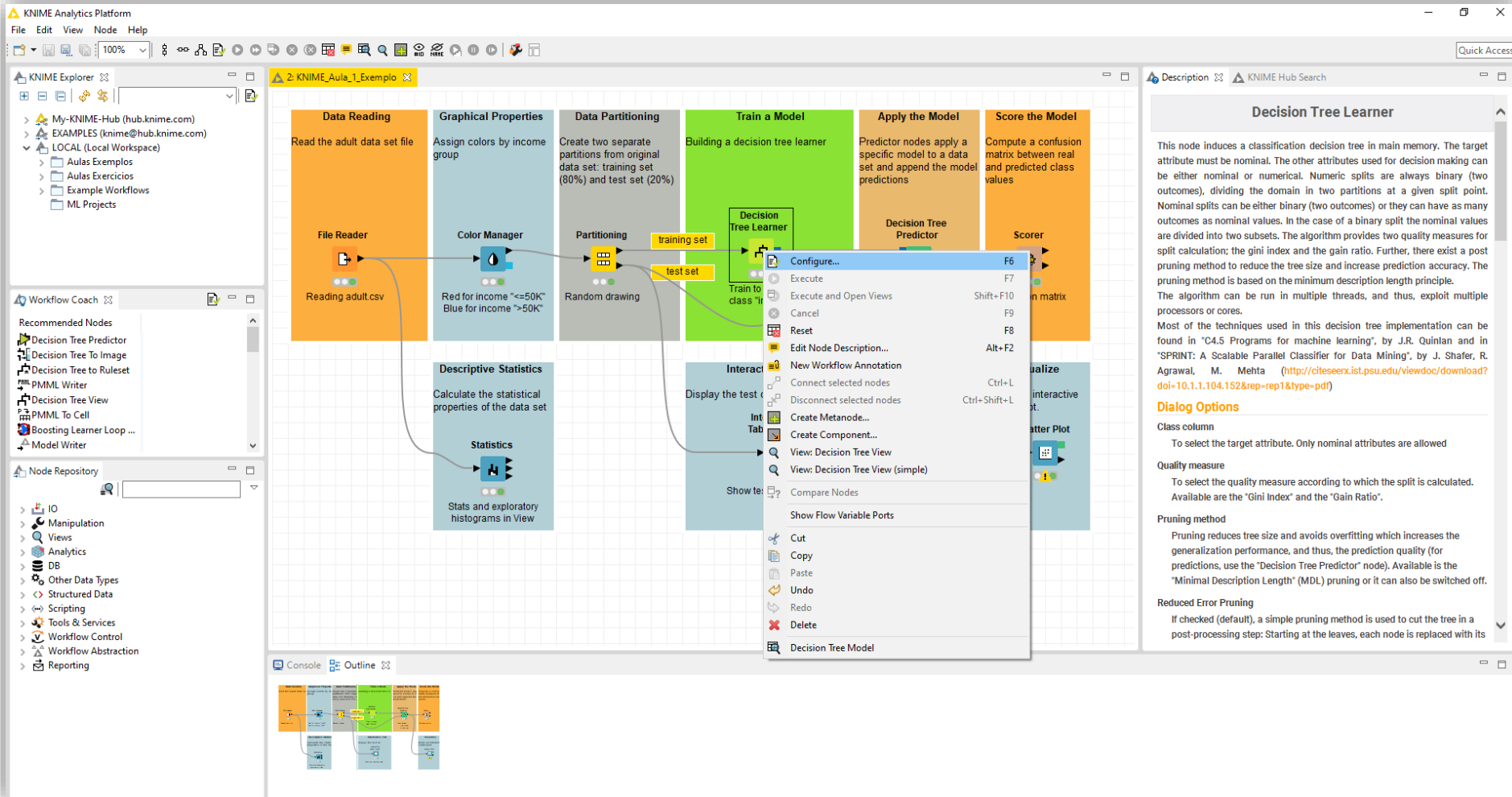
14

KNIME

Setup

TRYING IT

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)
 - Aulas Exemplos
 - Aulas Exercicios
 - 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 Cell
- 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: KNIME_Aula_1_Exemplo

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 1

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 set
Int Tab
Show test set

Context Menu for Decision Tree Learner:

- Configure... (F6)
- Execute (F7)
- Execute and Open Views (Shift+F10)
- Cancel (F9)
- Reset (F8)
- Edit Node Description... (Alt+F2)
- New Workflow Annotation (Ctrl+L)
- Connect selected nodes (Ctrl+Shift+L)
- 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.R. Quinlan and in "SPRINT: A Scalable Parallel Classifier for Data Mining", by J. Shafer, R. Agrawal, M. Mehta (<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.152&rep=rep1&type=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

Decision Tree Learner

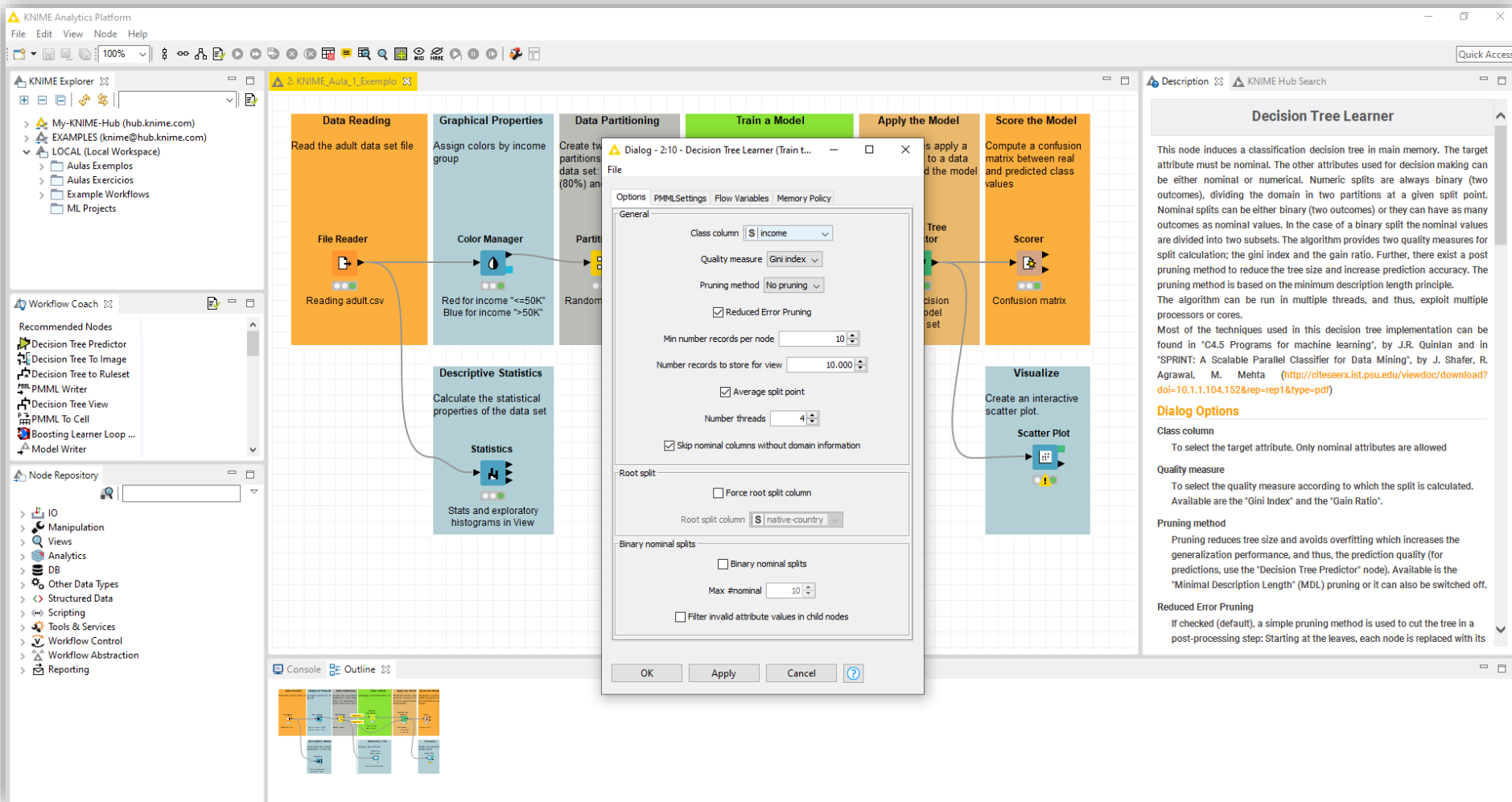
15

KNIME

Setup

TRYING IT

Hands On



The screenshot displays the KNIME Analytics Platform interface. On the left, the 'KNIME Explorer' shows a project structure with 'LOCAL (Local Workspace)' containing 'Aulas Exemplos', 'Aulas Exercicios', 'Example Workflows', and 'ML Projects'. Below it, the 'Workflow Coach' lists recommended nodes like 'Decision Tree Predictor' and 'Decision Tree To Image'. The 'Node Repository' on the bottom left categorizes nodes into IO, Manipulation, Views, Analytics, DB, Other Data Types, Structured Data, Scripting, Tools & Services, Workflow Control, Workflow Abstraction, and Reporting.

The main workspace shows a workflow titled '2: KNIME_Aula_1.Exemplo'. It includes nodes for 'Data Reading' (File Reader), 'Graphical Properties' (Color Manager), 'Data Partitioning' (Partitioning), 'Train a Model' (Decision Tree Learner), 'Apply the Model' (Apply Model), 'Score the Model' (Scorer), and 'Visualize' (Scatter Plot). A 'Descriptive Statistics' node is also present.

The 'Decision Tree Learner' dialog box is open, showing the following settings:

- 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

Buttons at the bottom of the dialog are 'OK', 'Apply', 'Cancel', and a help icon.

On the right, the 'Description' pane for the 'Decision Tree Learner' node provides detailed information:

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.R. Quinlan and in "SPRINT: A Scalable Parallel Classifier for Data Mining", by J. Shafer, R. Agrawal, M. Mehta (<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.104.152&rep=rep1&type=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

Decision Tree Learner

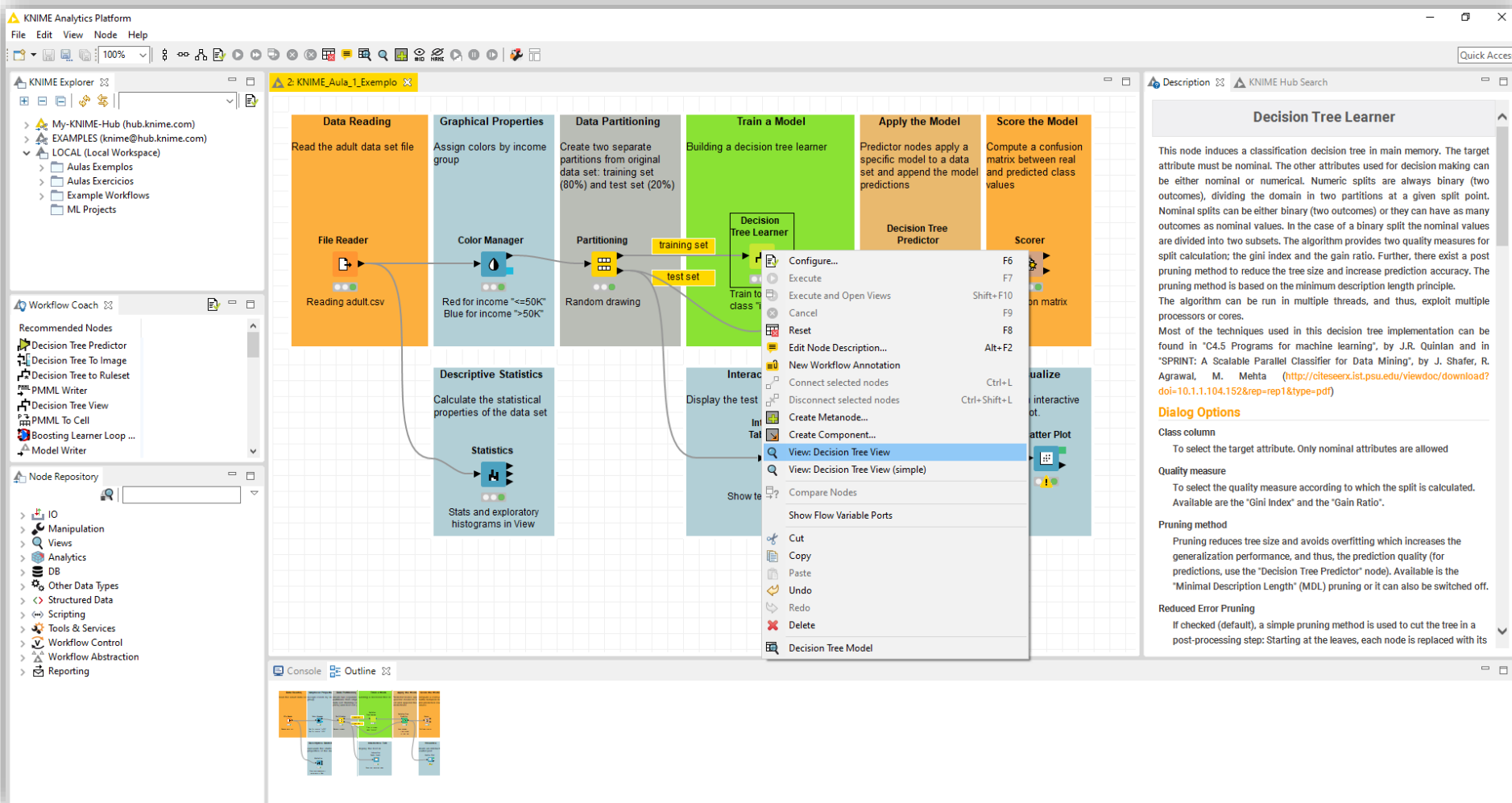
16

KNIME

Setup

TRYING IT

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)
 - Aulas Exemplos
 - Aulas Exercicios
 - 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 Cell
- 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: KNIME_Aula_1_Exemplo

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 1

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
Confusion matrix

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

Interactions
Display the test set
Show test set

Context Menu for Decision Tree Learner:

- Configure...
- 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.R. Quinlan and in "SPRINT: A Scalable Parallel Classifier for Data Mining", by J. Shafer, R. Agrawal, M. Mehta (<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.152&rep=rep1&type=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

Decision Tree Learner

17

KNIME

Setup

TRYING IT

Hands On

KNIME Analytics Platform

File Edit View Node Help

100%

Quick Access

KNIME Explorer

- My-KNIME-Hub (hub.knime.com)
- EXAMPLES (knime@hub.knime.com)
- LOCAL (Local Workspace)
 - Aulas Exemplos
 - Aulas Exercicios
 - 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 Cell
- 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: KNIME_Aula_1.Exemplo

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
training set

Decision Tree Learner

Decision Tree Predictor

Scorer

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

File HiLite Tree

Descriptive Statistics
Calculate the statistical properties of the data set

Statistics
Stats and exploratory histograms in View

Decision Tree View

relationship

Hot-in-family

≤50K (5.912/6.593)

Table:

Category	%	n
≤50K	89,7	5.912
>50K	10,3	681
Total	25,3	6.593

Chart: Color column: income

Husband

≤50K (5.860/10.585)

Table:

Category	%	n
≤50K	55,4	5.860
>50K	44,6	4.725
Total	40,6	10.585

Chart: Color column: income

Wife

≤50K (650/1.244)

Table:

Category	%	n
≤50K	52,3	650
>50K	47,7	594
Total	4,8	1.244

Chart: Color column: income

Own-child

≤50K (4.012/4.065)

Table:

Category	%	n
≤50K	98,7	4.012
>50K	1,3	53
Total	15,6	4.065

Chart: Color column: income

Unmarried

≤50K (2.577/2.754)

Table:

Category	%	n
≤50K	93,6	2.577
>50K	6,4	177
Total	10,6	2.754

Chart: Color column: income

Other-relative

≤50K (776/807)

Table:

Category	%	n
≤50K	96,2	776
>50K	3,8	31
Total	3,1	807

Chart: Color column: income

Zoom: 100.0%

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

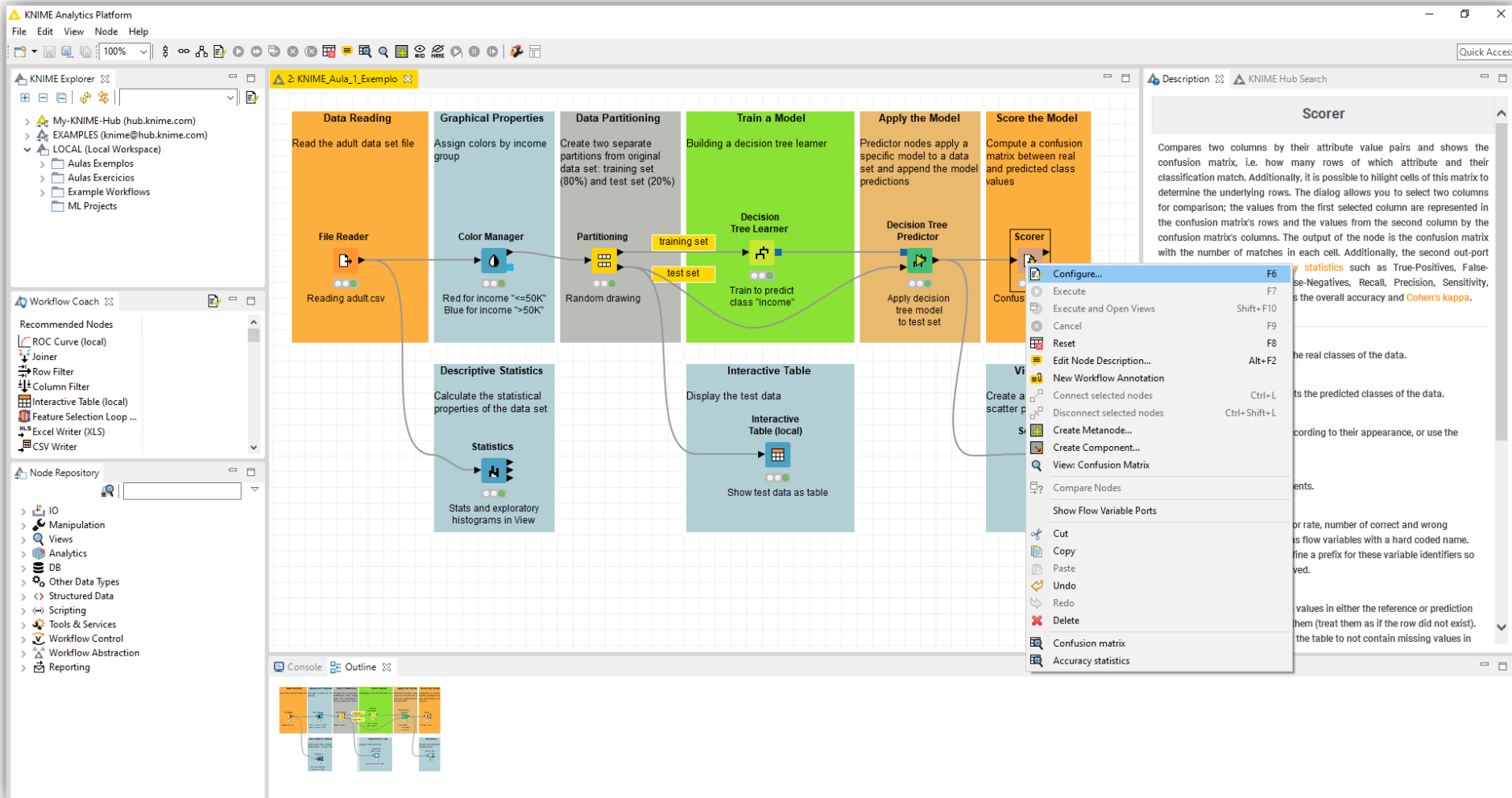
18

KNIME

Setup

TRYING IT

Hands On



The screenshot displays the KNIME Analytics Platform interface with a workflow titled "2: KNIME_Aula_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 workflow description for the Scorer node is as follows:

- Data Reading:** Read the adult data set file. Node: File Reader. Reading adult.csv.
- Graphical Properties:** Assign colors by income group. Node: 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%). Node: Partitioning. Random drawing.
- Train a Model:** Building a decision tree learner. Node: Decision Tree Learner. Train to predict class "Income".
- Apply the Model:** Predictor nodes apply a specific model to a data set and append the model predictions. Node: Decision Tree Predictor. Apply decision tree model to test set.
- Score the Model:** Compute a confusion matrix between real and predicted class values. Node: Scorer. Confusion matrix.

The Scorer node context menu options include:

- Configure... (F6)
- Execute (F7)
- Execute and Open Views (Shift+F10)
- Cancel (F9)
- Reset (F8)
- Edit Node Description... (Alt+F2)
- New Workflow Annotation (Ctrl+L)
- Connect selected nodes (Ctrl+Shift+L)
- Disconnect selected nodes
- Create Metanode...
- Create Component...
- View: Confusion Matrix
- Compare Nodes
- Show Flow Variable Ports
- Cut
- Copy
- Paste
- Undo
- Redo
- Delete
- Confusion matrix
- Accuracy statistics

The Scorer node 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 shows statistics such as True-Positives, False-Negatives, Recall, Precision, Sensitivity, the overall accuracy and Cohen's kappa."

Node Context Options

Scorer

19

KNIME

Setup

TRYING IT

Hands On

The screenshot displays the KNIME Analytics Platform interface. The main workspace shows a workflow titled "2: KNIME_Aula_1.Exemplo" with the following nodes: File Reader (Reading adult.csv), Color Manager (Assign colors by income group), Partitioning (Create two separate partitions from original data set), Decision Tree Learner (Building a decision tree learner), Decision Tree Predictor (Predictor nodes apply a specific model to a data set), and Scorer (Compute a confusion matrix between real and predicted class values). A dialog box for the Scorer node is open, showing the "First Column" as "income" and the "Second Column" as "Prediction (income)". The dialog also includes options for sorting, providing scores as flow variables, and handling missing values.

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

Node Context Options Scorer

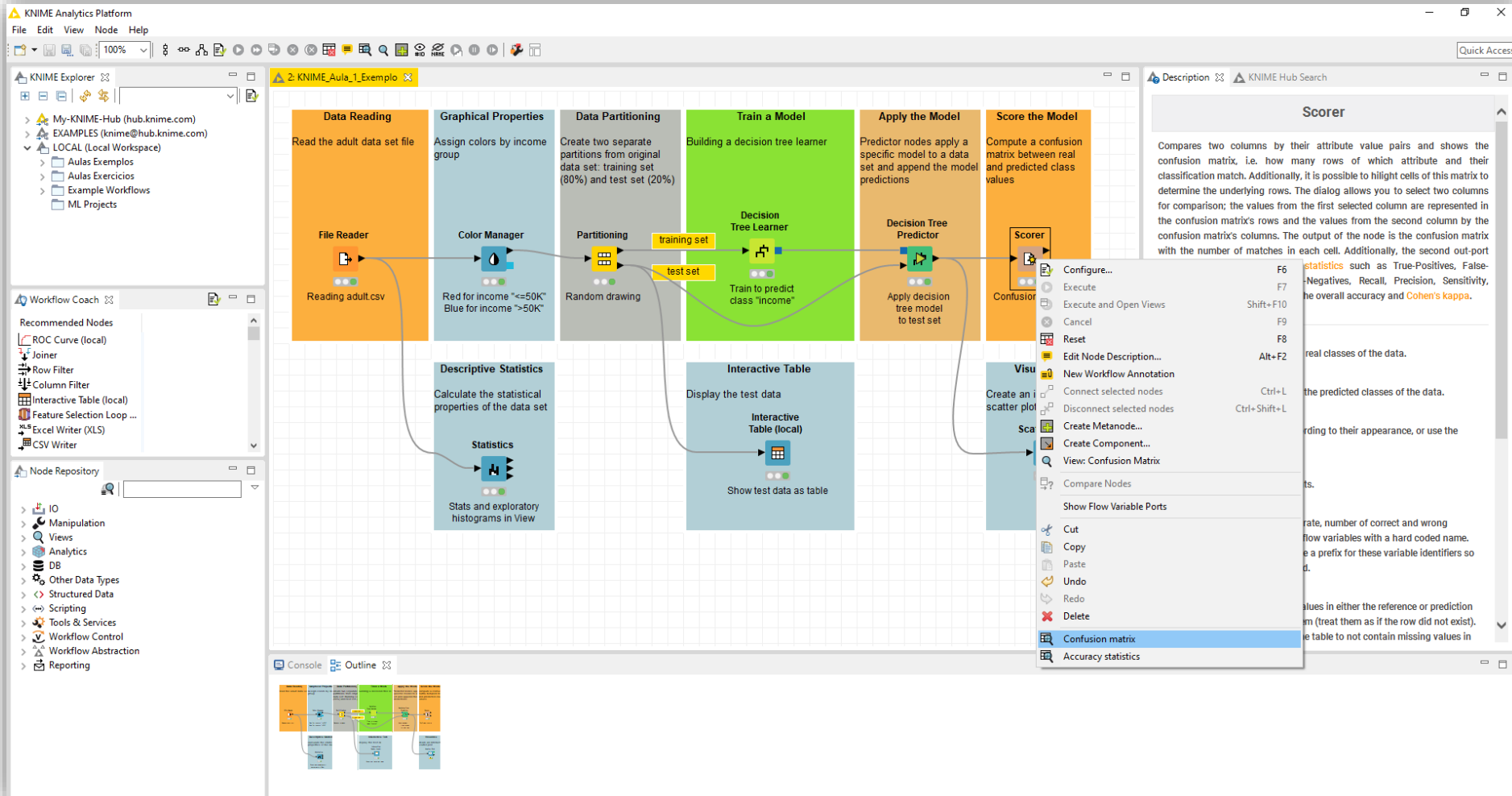
20

KNIME

Setup

TRYING IT

Hands On



The screenshot displays the KNIME Analytics Platform interface with a workflow titled "2: KNIME_Aula_1_Exemplo". The workflow is organized into six main stages:

- Data Reading:** Includes a "File Reader" node reading "adult.csv".
- Graphical Properties:** Includes a "Color Manager" node with settings: "Red for income <=50K", "Blue for income >50K".
- Data Partitioning:** Includes a "Partitioning" node with "Random drawing" settings, splitting data into "training set" and "test set".
- Train a Model:** Includes a "Decision Tree Learner" node with settings: "Train to predict class 'Income'".
- Apply the Model:** Includes a "Decision Tree Predictor" node with settings: "Apply decision tree model to test set".
- Score the Model:** Includes a "Scorer" node with settings: "Compute a confusion matrix between real and predicted class values".

A context menu is open over the "Scorer" node, showing the following options:

- Configure...
- Execute
- Execute and Open Views
- 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
- Accuracy statistics

The "Description" pane on the right provides details for the "Scorer" node:

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

Statistics such as True-Positives, False-Negatives, Recall, Precision, Sensitivity, the overall accuracy and Cohen's kappa.

real classes of the data.

the predicted classes of the data.

according to their appearance, or use the

ts.

rate, number of correct and wrong

flow variables with a hard coded name.

e a prefix for these variable identifiers so

d.

values in either the reference or prediction

m (treat them as if the row did not exist).

e table to not contain missing values in

Node Context Options Scorer

21

KNIME

Setup

TRYING IT

Hands On

Workflow Overview:

- Data Reading:** File Reader (Reading adult.csv)
- Graphical Properties:** Color Manager (Assign colors by income group: Red for income "<=50K", Blue for income ">50K")
- Data Partitioning:** Partition (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:** Scorer (Compute a confusion matrix between real and predicted class values)
- Visualize:** Scatter Plot (Create an interactive scatter plot)

Confusion Matrix - 2:6 - Scorer (Confusion matrix)

Income \ Pr...	<=50K	>50K
<=50K	4557	376
>50K	674	906

Summary Statistics:

- Correct classified: 5,463
- Wrong classified: 1,050
- Accuracy: 83,878 %
- Error: 16,122 %
- Cohen's kappa (κ): 0,531

Scorer Node Description:

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

Hands On

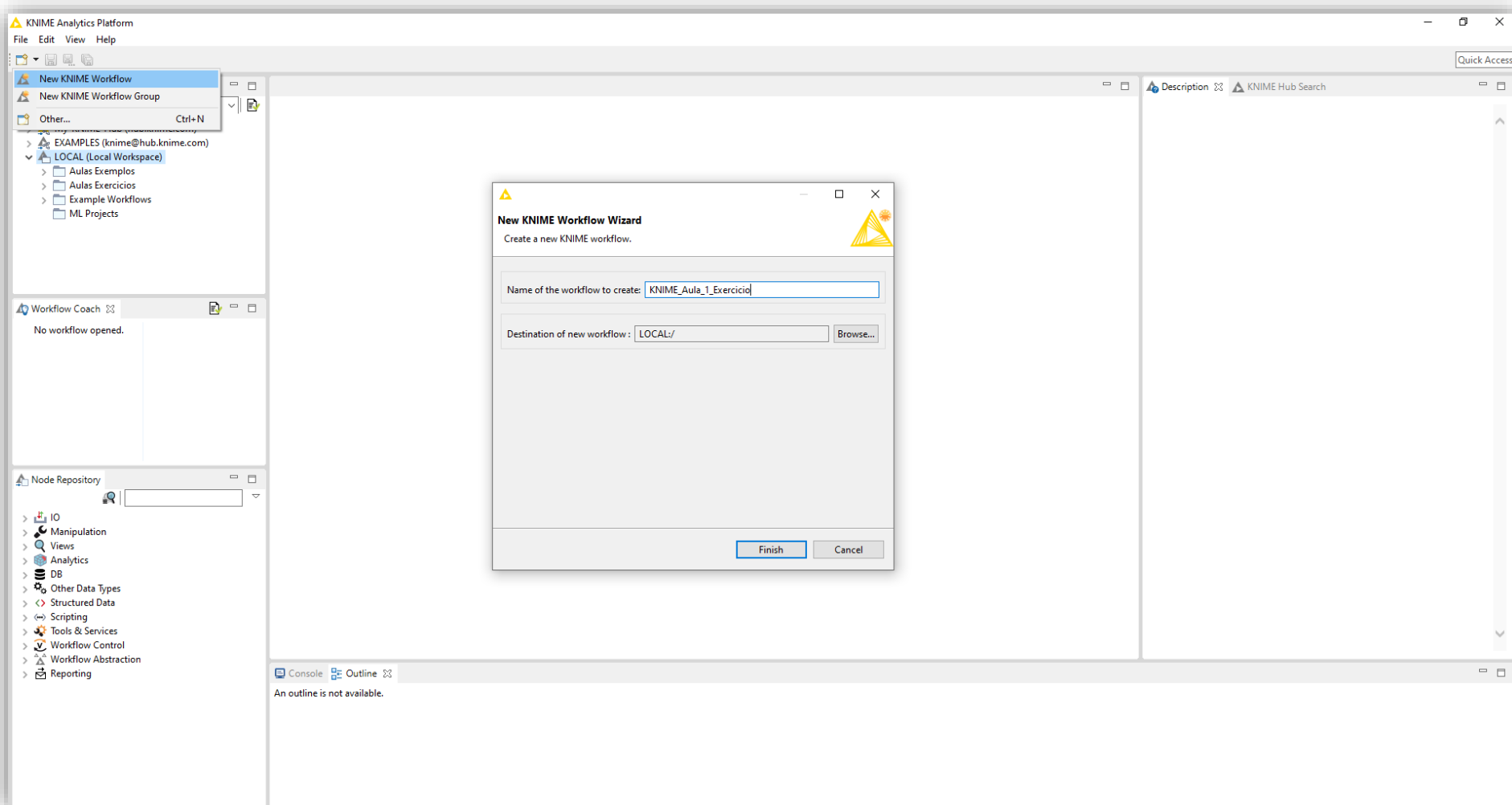
22

KNIME

Setup

Trying It

HANDS ON



Hands On

23

KNIME

Setup

Trying It

HANDS ON

