

Universidade do Minho
Departamento de Informática

Ferramentas de Aprendizagem por Máquina

(Machine Learning Tools)

LEI @ 2025/2026, 2º sem



- Introdução à plataforma KNIME
- Construção de fluxos de análise de dados
(KNIME *workflows*)
- Experimentação
(*hands on*)

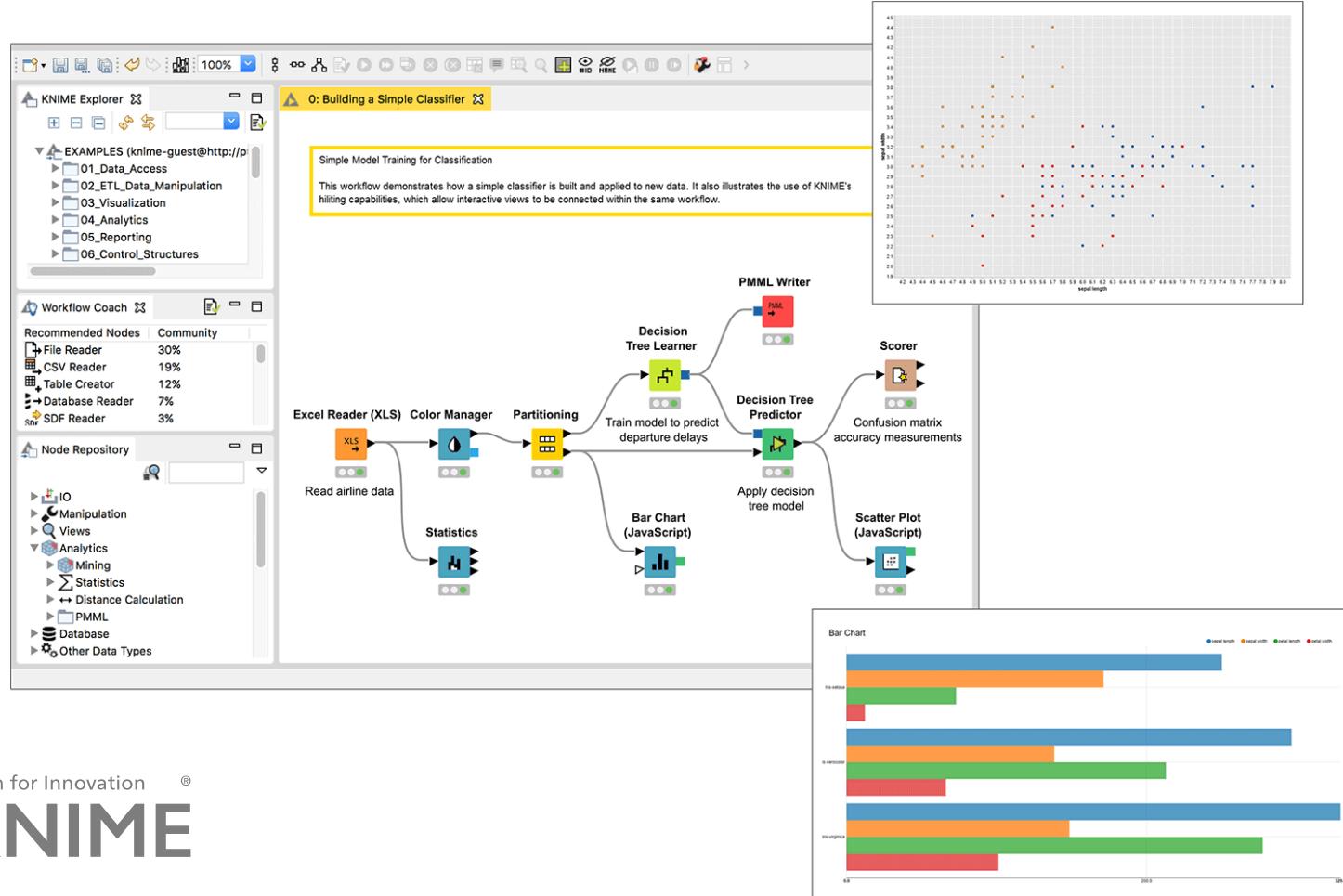
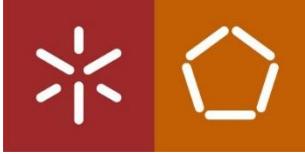


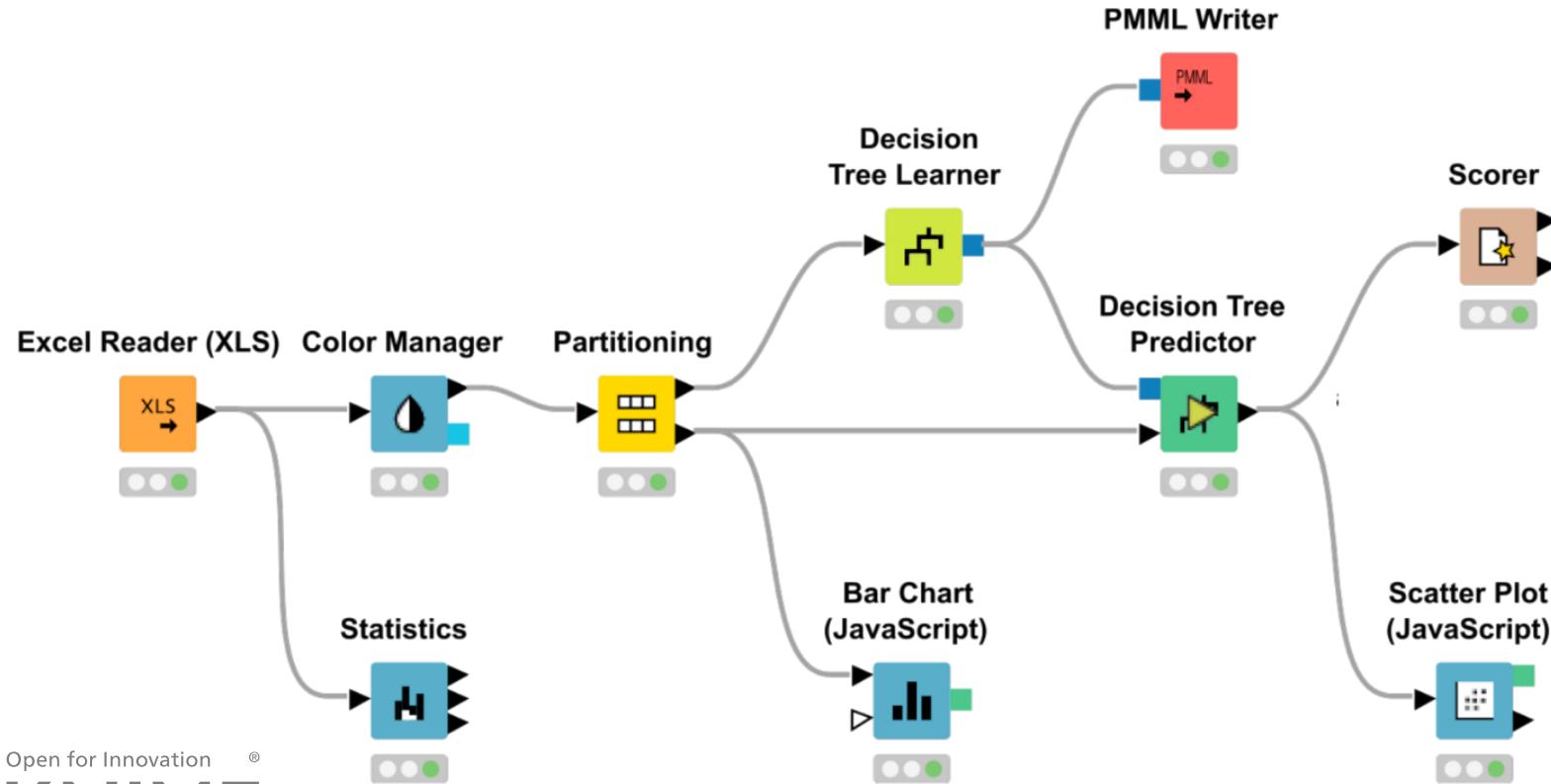
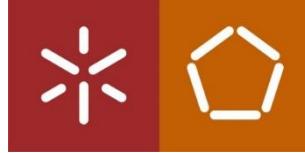


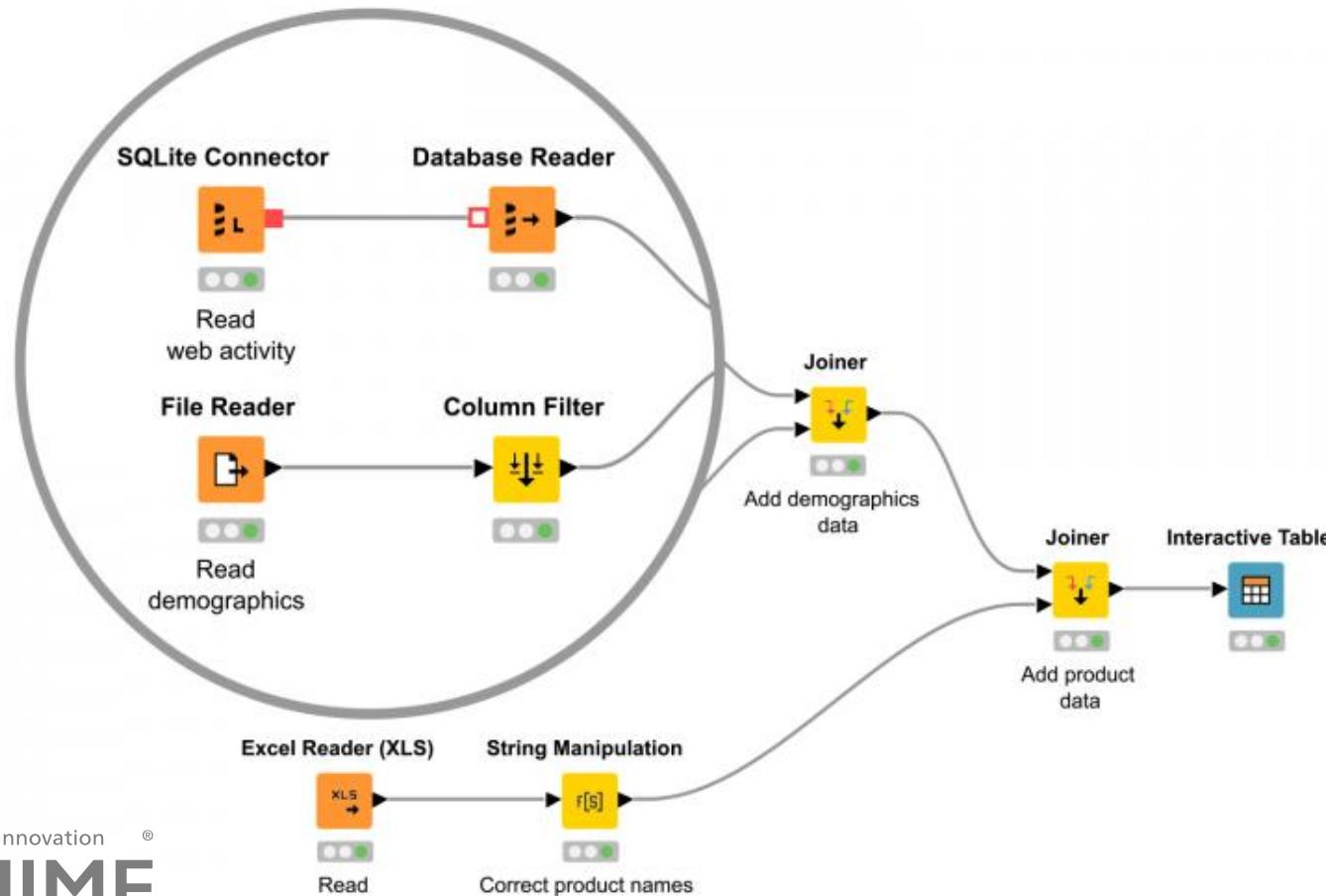
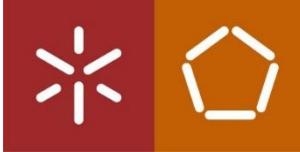
Universidade do Minho
Departamento de Informática

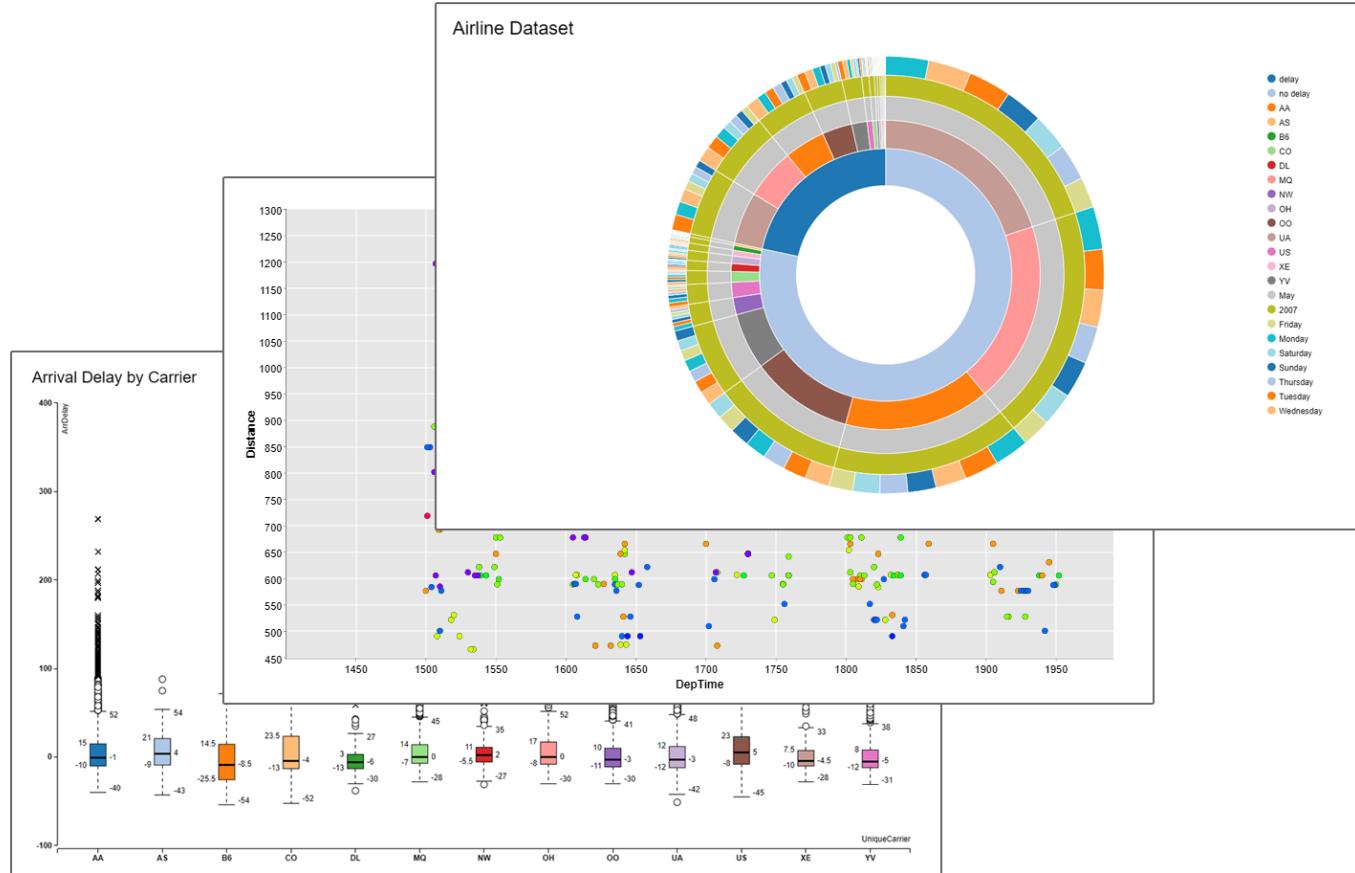


KNIME | Open for Innovation











Get started with KNIME Analytics Platform 5



Local space



The local space is the folder on your computer to store and access KNIME workflows and data produced by your workflows.



Create workflow
in your local space.

KNIME Community Hub (hub.knime.com)

[Sign in](#)

Connect to the KNIME Community Hub to find workflows, nodes and components, and collaborate in spaces.



Home Building a Simple Classifier X Help Preferences Menu

Execute all Cancel all Reset all 93% ▾

Nodes

Search all nodes

IO

Excel Reader Excel Writer Microsoft Authenticator

CSV Reader CSV Writer Table Creator

SharePoint Online Connector File Reader Show all

Manipulation

Row Filter Column Filter Concatenate

Value Lookup Row Aggregator Table Splitter

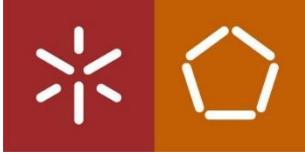
String Cleaner Table Cropper Show all

Views

To show the node output, please select a configured or executed node.

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

knime.com/downloads



Home Building a Simple Classifier X Help Preferences Menu

Execute all Cancel all Reset all

Nodes

Search all nodes

IO

Excel Reader Excel Writer Microsoft Authenticator

CSV Reader CSV Writer Table Creator

SharePoint Online Connector File Reader Show all

Manipulation

Row Filter Column Filter Concatenate

Value Lookup Row Aggregator Table Splitter

String Cleaner Table Cropper Show all

To show the node output, please select a configured or executed node.

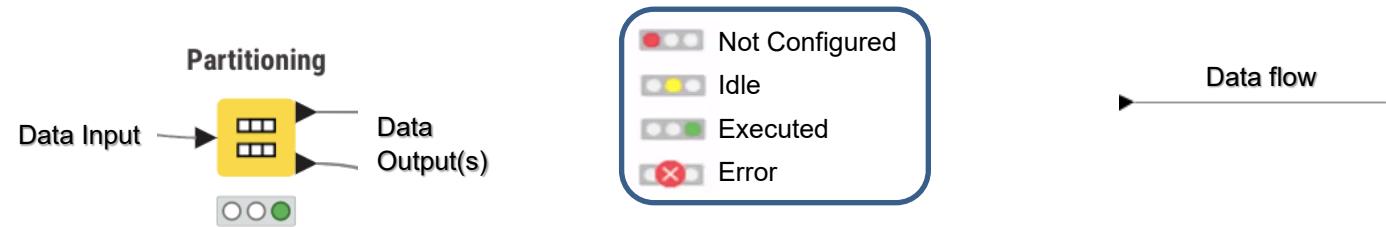
QUICK HANDS ON



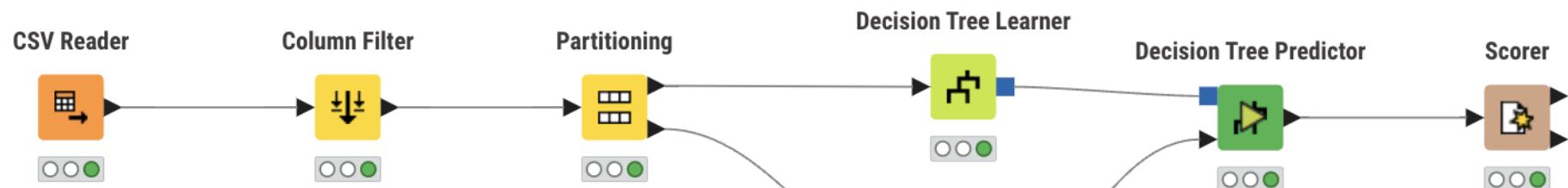
Nodos e Fluxos

Nodes and Workflows

- Nodos
Nodes



- Fluxos
Workflows





Extensões KNIME

KNIME Extensions

Home Building a Simple Classifier •

Execute all Cancel all Reset all

Nodes

Search all nodes

10

Excel Reader Excel Writer Microsoft Authenticator

CSV Reader CSV Writer Table Creator

SharePoint Online Connector File Reader

Show all

Manipulation

Row Filter Column Filter Concatenate

Value Lookup Row Aggregator Table Splitter

```
graph LR; CSVReader[CSV Reader] --> ColumnFilter[Column Filter]; ColumnFilter --> Partitioning[Partitioning]; Partitioning --> DT_Learner[Decision Tree Learner]; Partitioning --> DT_Predictor[Decision Tree Predictor]; DT_Learner --> DT_Predictor;
```

Help Preferences Menu

Check for updates

Show KNIME log in File Explorer

Install extensions ... to access additional functionality, including complex data type processing and advanced algorithms.

Switch workspace ... to access KNIME workflows and data from a different folder on your computer.

Switch to classic user interface ... to use the classic user interface. Switch back again with the button "Open KNIME Modern UI" in the top right corner.



Extensões KNIME

KNIME Extensions

The screenshot shows the KNIME interface with the 'Available Software' dialog box open. The dialog box is titled 'Install' and contains a list of available extensions under the heading 'Available Software'. A search bar at the top of the list allows for filtering. The list includes categories like KNIME & Extensions, KNIME Big Data Extensions, KNIME Community Extensions - Bioinformatics & NGS, KNIME Community Extensions - Cheminformatics, KNIME Community Extensions - Image Processing and Analysis, KNIME Community Extensions - Other, KNIME Hub & Server Extensions, KNIME Labs Extensions, KNIME Node Development Tools, and KNIME Partner Extensions. Below the list are several filter options: 'Show only the latest versions of available software' (checked), 'Group items by category' (checked), 'Hide items that are already installed' (unchecked), and 'Show only software applicable to target environment' (unchecked). At the bottom of the dialog box are buttons for '< Back', 'Next >', 'Cancel', and 'Finish'.

Available Software

Check the items that you wish to install.

type filter text

Name Version

Select All Deselect All

KNIME & Extensions

KNIME Big Data Extensions

KNIME Community Extensions - Bioinformatics & NGS

KNIME Community Extensions - Cheminformatics

KNIME Community Extensions - Image Processing and Analysis

KNIME Community Extensions - Other

KNIME Hub & Server Extensions

KNIME Labs Extensions

KNIME Node Development Tools

KNIME Partner Extensions

Details

Show only the latest versions of available software

Hide items that are already installed

Group items by category

What is [already installed?](#)

Show only software applicable to target environment

< Back Next > Cancel Finish



Descrição dos Nodos

Node Description

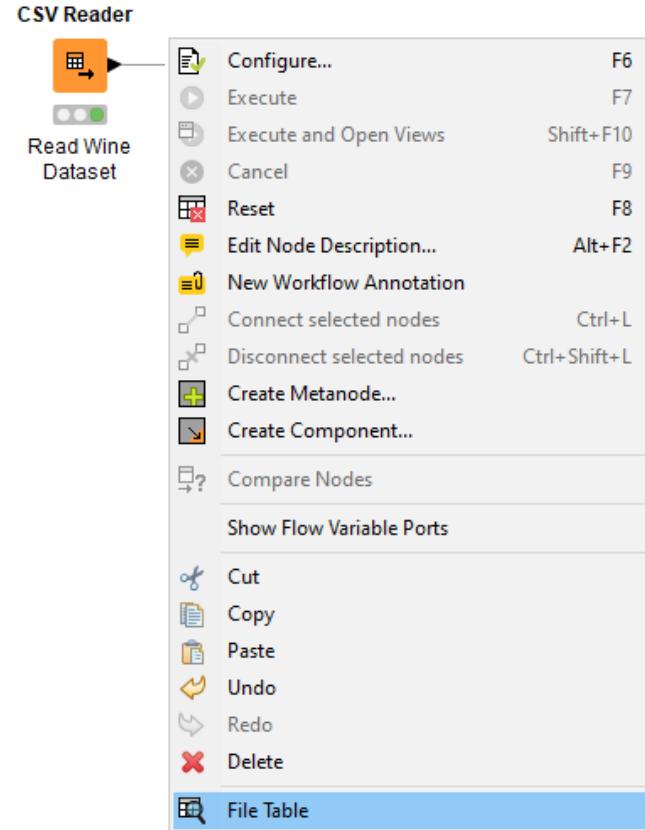
The screenshot shows the KNIME Node Description dialog for the "Scorer" node. The title bar includes the node icon, the title "Scorer", and the KNIME Hub Search button. The main content area contains the following sections:

- Scorer**: A detailed description of the node's function, mentioning it compares two columns by attribute value pairs and outputs a confusion matrix with accuracy statistics like True-Positives, False-Positives, etc.
- Dialog Options**: Configuration options:
 - First column**: The real classes of the data.
 - Second column**: The predicted classes of the data.
 - Sorting strategy**: Whether to sort labels by appearance or lexical/numeric ordering.
 - Reverse order**: Reversing the order of elements.
 - Use name prefix**: Exporting scores as flow variables with a hard-coded prefix.
- Missing Values**: Instructions for handling missing values in reference or prediction columns.
- Ports**: Input and Output ports:
 - Input Ports**: 0 Table containing at least two columns to compare.
 - Output Ports**: 0 The confusion matrix.
1 The accuracy statistics table.
- Views**: A description of the Confusion Matrix view.



Visualização do Ficheiro de Dados

Data Table Structure View





Visualização do Ficheiro de Dados

Data Table Structure View

Column Headers
Data Type (Double)
Data Type (String)

Row ID
Data Cells

File Table - 0:1 - CSV Reader (Reader)
File Hilite Navigation View
Table "winequality-red.csv" - Rows: 1595 Spec - Columns: 12 Properties Flow Variables

Row ID	ID	fixed a...	D	volatile ...	D	citric acid	D	residual ...	D	chlorides	Sort Descending	Sort Ascending	No Sorting	Available Renderers	Standard Double	Percentage	Full Precision	Gray Scale	Bars	Standard Complex Number	Default	alcohol	S	quality
Row0	7.4	0.7	0	1.9	0.076									0.997	0.997	0.997	0.997	0.997	0.997	0.997	=5	=5	=5	
Row1	7.8	0.88	0	2.6	0.098									0.997	0.997	0.997	0.997	0.997	0.997	0.997	=5	=5	=5	
Row2	7.8	0.76	0.04	2.3	0.092									0.997	0.997	0.997	0.997	0.997	0.997	0.997	=5	=5	=5	
Row3	11.2	0.28	0.56	1.9	0.075									0.998	0.998	0.998	0.998	0.998	0.998	0.998	=6	=6	=6	
Row4	7.4	0.7	0	1.9	0.076									0.998	0.998	0.998	0.998	0.998	0.998	0.998	=5	=5	=5	
Row5	7.4	0.66	0	1.8	0.075									0.998	0.998	0.998	0.998	0.998	0.998	0.998	=5	=5	=5	
Row6	7.9	0.6	0.06	1.6	0.069									0.996	0.996	0.996	0.996	0.996	0.996	0.996	=5	=5	=5	
Row7	7.3	0.65	0	1.2	0.065									0.995	0.995	0.995	0.995	0.995	0.995	0.995	=7	=7	=7	
Row8	7.8	0.58	0.02	2	0.073									0.997	0.997	0.997	0.997	0.997	0.997	0.997	=7	=7	=7	
Row9	7.5	0.5	0.36	6.1	0.071									0.998	0.998	0.998	0.998	0.998	0.998	0.998	=5	=5	=5	
Row10	6.7	0.58	0.08	1.8	0.097									0.996	0.996	0.996	0.996	0.996	0.996	0.996	=5	=5	=5	
Row11	7.5	0.5	0.36	6.1	0.071									0.998	0.998	0.998	0.998	0.998	0.998	0.998	=5	=5	=5	
Row12	5.6	0.615	0	1.6	0.089									0.994	0.994	0.994	0.994	0.994	0.994	0.994	=5	=5	=5	
Row13	7.8	0.61	0.29	1.6	0.114									0.997	0.997	0.997	0.997	0.997	0.997	0.997	=5	=5	=5	
Row14	8.9	0.62	0.18	3.8	0.176									0.999	0.999	0.999	0.999	0.999	0.999	0.999	=5	=5	=5	
Row15	8.9	0.62	0.19	3.9	0.17									0.999	0.999	0.999	0.999	0.999	0.999	0.999	=5	=5	=5	
Row16	8.5	0.28	0.56	1.8	0.092									0.997	0.997	0.997	0.997	0.997	0.997	0.997	=7	=7	=7	
Row17	8.1	0.56	0.28	1.7	0.368									0.997	0.997	0.997	0.997	0.997	0.997	0.997	=5	=5	=5	
Row18	7.4	0.59	0.08	4.4	0.086									0.997	0.997	0.997	0.997	0.997	0.997	0.997	=4	=4	=4	
Row19	7.9	0.32	0.51	1.8	0.341									0.997	0.997	0.997	0.997	0.997	0.997	0.997	=6	=6	=6	
Row20	8.9	0.22	0.48	1.8	0.077									0.997	0.997	0.997	0.997	0.997	0.997	0.997	=6	=6	=6	
Row21	7.6	0.39	0.31	2.3	0.082									0.998	0.998	0.998	0.998	0.998	0.998	0.998	=5	=5	=5	
Row22	7.9	0.43	0.21	1.6	0.106									0.997	0.997	0.997	0.997	0.997	0.997	0.997	=5	=5	=5	
Row23	8.5	0.49	0.11	2.3	0.084									0.997	0.997	0.997	0.997	0.997	0.997	0.997	=5	=5	=5	
Row24	6.9	0.4	0.14	2.4	0.085									0.997	0.997	0.997	0.997	0.997	0.997	0.997	=6	=6	=6	
Row25	6.3	0.39	0.16	1.4	0.08									0.996	0.996	0.996	0.996	0.996	0.996	0.996	=5	=5	=5	
Row26	7.2	0.39	0.14	1.8	0.08									0.996	0.996	0.996	0.996	0.996	0.996	0.996	=5	=5	=5	



Building a Simple Workflow

KNIME Analytics Platform

File Edit View Node Help

KNIME Explorer

- My-KNIME-Hub (hub.knime.com)
- EXAMPLES (knime@hub.knime.com)
- LOCAL (Local Workspace)
 - Aulas Ejemplos
 - Aulas Ejercicios
 - 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

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

Descriptive Statistics: Calculate the statistical properties of the data set

Interactive Table: Display the test data

Visualize: Create an interactive scatter plot

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: Data Loader

The KNIME Analytics Platform interface is shown, illustrating the use of context menus for a 'File Reader' node.

Workflow Overview:

- Data Reading:** 'Read the adult data set file' (File Reader) → '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%)) → 'Decision Tree Learner' (Building a decision tree learner).
- Train a Model:** 'Decision Tree Learner' → 'Decision Tree Predictor' (Train to predict class "Income").
- Apply the Model:** 'Decision Tree Predictor' → 'Scorer' (Apply decision tree model to test set).
- Score the Model:** 'Scorer' → 'Confusion matrix' (Compute a confusion matrix between real and predicted class values).
- Interactive Table:** 'Interactive Table (local)' (Display the test data) → 'Scatter Plot' (Create an interactive scatter plot).

Context Menu (opened on the File Reader node):

- Reading
- Configure...
- Execute
- Execute and Open Views
- Cancel
- Reset
- Edit Node Description...
- New Workflow Annotation
- Connect selected nodes
- Disconnect selected nodes
- Create Metanode...
- Create Component...
- Compare Nodes
- Show Flow Variable Ports
- Cut
- Copy
- Paste
- Undo
- Redo
- Delete
- File Table

File Reader Node Context Options (Description):

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 checkboxes 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 checked, 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: Data Loader

The KNIME Analytics Platform interface showing the Data Reader node configuration dialog and its context options.

Workflow Overview:

```
graph LR; A[File Reader] --> B[Descriptive Statistics]; A --> C[Scatter Plot];
```

Data Reader Node Configuration Dialog:

Settings Tab:

- Enter ASCII data file location: `knime://knime.workflow/adult.csv`
- Preserve user settings for new location
- Browse...
- read row IDs
- Column delimiter: `,`
- read column headers
- ignore spaces and tabs
- Java-style comments
- Single line comment: `;`

Preview Tab:

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

Context Options Dialog:

File Tab:

- Input a confusion matrix between real predicted classes

Scorer Tab:

- Score the Model
- Scorer
- confusion matrix

Visualize Tab:

- 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 Tab:

- 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: Model Learner

The screenshot shows the KNIME Analytics Platform interface with a workflow titled "2-KNIME_Aula_1_Exemplo". The workflow consists of several nodes:

- Data Reading:** Read the adult data set file (File Reader) and Assign colors by income group (Color Manager).
- Graphical Properties:** 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%).
- Train a Model:** Building a decision tree learner (Decision Tree Learner node selected).
 - Context menu options for the Decision Tree Learner node include: 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), Show test, Compare Nodes, Show Flow Variable Ports, Cut, Copy, Paste, Undo, Redo, Delete, and Decision Tree Model.
- 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).
- Output:** Stats and exploratory histograms in View.

The "Decision Tree Learner" node is highlighted in green, indicating it is the active node. A context menu is open over this node, listing various options for configuration and interaction. The "Configure..." option is currently selected.

Decision Tree Learner Node Description:

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 processor 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://citeseer.ist.psu.edu/viewdoc/download?doi=10.1.1.104.152&rep=rep1&type=pdf>)

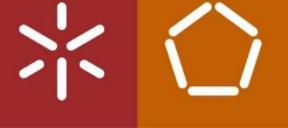
Dialog Options:

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: Model Learner

KNIME Analytics Platform

File Edit View Node Help

KNIME Explorer

LOCAL (Local Workspace)

Aulas Exemplos

Aulas Ejercicios

Example Workflows

ML Projects

Workflow Coach

Recommended Nodes

- Decision Tree Predictor
- Decision Tree To Image
- Decision Tree to Ruleset
- PMMI Writer
- Decision Tree View
- PMMI 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 Aulas_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 partitions of data sets (80% and 20%)

Random

Train a Model

Dialog - 2:10 - Decision Tree Learner (Train t...)

File

Options PMMLSettings Flow Variables Memory Policy

General

Class column S 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 S native-country

Binary nominal splits

Binary nominal splits

Max #nominal 10

Filter invalid attribute values in child nodes

Apply the Model

Tree Predictor

Decision model set

Scorer

Score the Model

Compute a confusion matrix between real and predicted class values

Scorer

Confusion matrix

Visualize

Create an interactive scatter plot.

Scatter Plot

Description

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



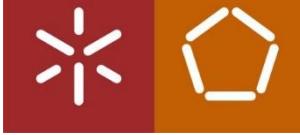
Node Context Options: Model Learner

The screenshot shows the KNIME Analytics Platform interface with a workflow titled "2: KNIME_Aula_1_Exemplo". The workflow consists of several nodes:

- Data Reading:** File Reader (Reading adult.csv) → File Reader (Reading adult.csv)
- Graphical Properties:** Color Manager (Red for income "<=50K", Blue for income ">50K")
- Data Partitioning:** Partitioning (Random drawing) → Partitioning (Random drawing) (training set, test set)
- Train a Model:** Decision Tree Learner (Configure..., Train to class "1")
- Apply the Model:** Decision Tree Predictor (Execute, Execute and Open Views, Cancel, Reset, Configure..., 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), Show tree, Show Flow Variable Ports, Compare Nodes, Cut, Copy, Paste, Undo, Redo, Delete, Decision Tree Model)
- Score the Model:** Scorer (F6, Shift+F10, F9, F8, Alt+F2)
- Descriptive Statistics:** Statistics (Stats and exploratory histograms in View)

A context menu is open over the "Decision Tree Predictor" node, showing options like "Execute", "Execute and Open Views", "Cancel", "Reset", "Configure...", "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)", "Show tree", "Show Flow Variable Ports", "Compare Nodes", "Cut", "Copy", "Paste", "Undo", "Redo", "Delete", and "Decision Tree Model".

The right side of the interface displays the "Decision Tree Learner" node details, including its description, citation information (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 settings, quality measure selection, pruning methods, and reduced error pruning information.



Node Context Options: Model Learner

KNIME Analytics Platform

File Edit View Node Help

100% Quick Access

KNIME Explorer

LOCAL (Local Workspace)

- Aulas Ejemplos
- Aulas Ejercicios
- 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 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

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.

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

Table: Category % n

Category	%	n
<=50K	75,0	19.787
>50K	24,0	6.261
Total	100,0	26.048

Chart: Color column: income

relationship

Not-in-family

Table: Category % n

Category	%	n
50K	59,12	6.593
10K	39,7	5.512
25K	10,3	681
Total	25,3	6.593

Chart: Color column: income

Husband

Table: Category % n

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

Chart: Color column: income

Wife

Table: Category % n

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

Chart: Color column: income

Own-child

Table: Category % n

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

Chart: Color column: income

Unmarried

Table: Category % n

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

Chart: Color column: income

Other-relative

Table: Category % n

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

Chart: Color column: income

Console Outline

Zoom: 100.0%

```
graph LR; FR[File Reader] --> R[Reading adult.csv]; CM[Color Manager] --> P[Partitioning]; P --> TS[training set]; DTL[Decision Tree Learner] --> DP[Decision Tree Predictor]; DP --> S[Scorer];
```



Node Context Options: Scorer

The KNIME Analytics Platform interface showing a workflow titled "2: KNIME_Aula_1_Exemplo".

Workflow Components:

- 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 " $\leq 50K$ ", 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 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).
- 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) - Show test data as table.

Scorer Node Context Options (Right-clicked):

- 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
- Disconnect selected nodes (Ctrl+Shift+L)
- Create Metanode...
- Create Component...
- View: Confusion Matrix
- Compare Nodes
- Show Flow Variable Ports
- Cut
- Copy
- Paste
- Undo
- Redo
- Delete
- Confusion matrix
- Accuracy statistics

Console and Outline panes:

Console and Outline panes are visible at the bottom of the interface.



Node Context Options: Scorer

The screenshot shows the KNIME Analytics Platform interface with a workflow titled "2: KNIME_Aulas_1.Example".

Workflow Overview:

- Data Reading:** A File Reader node reads the adult data set file ("adult.csv").
- Graphical Properties:** A Color Manager node assigns colors by income group, mapping red for income " $<=50K$ " and blue for income " $>50K$ ".
- Data Partitioning:** A Partitioning node creates two separate partitions from the original data set: training set (80%) and test set (20%).
- Train a Model:** A Decision Tree Learner node builds a decision tree learner.
- Apply the Model:** A Decision Tree Predictor node applies the specific model to the data set and appends the model predictions.
- Score the Model:** A Scorer node computes a confusion matrix between real and predicted class values.
- Descriptive Statistics:** A Statistics node calculates the statistical properties of the data set.
- Visualize:** A Scatter Plot node creates an interactive scatter plot.

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

This dialog box allows for comparing two columns by their attribute value pairs. It includes fields for First Column (set to \$income) and Second Column (set to \$Prediction(income)). It also provides options for sorting values in tables, providing scores as flow variables, and handling missing values (with radio buttons for Ignore or Fail). Buttons at the bottom include OK, Apply, Cancel, and Help.

Scorer Node Documentation:

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

Console and **Outline** panes are visible at the bottom of the interface.



Node Context Options: Scorer

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 provides 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. Use a prefix for these variable identifiers so that.

values in either the reference or prediction column (treat them as if the row did not exist). Use the table to not contain missing values in

F6 F7 F8 F9 F10 Alt+F2 Ctrl+L Ctrl+Shift+L

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

KNIME Analytics Platform

File Edit View Node Help

KNIME Explorer

EXAMPLES (knime@hub.knime.com)

LOCAL (Local Workspace)

Aulas Exemplos

Aulas Ejercicios

Example Workflows

ML Projects

Workflow Coach

Recommended Nodes

Joiner

Row Filter

Column Filter

Interactive Table (local)

Feature Selection Loop ...

Excel Writer (XLS)

CSV 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

Interactive Table

Display the test data

Interactive Table (local)

Show test data as table

Descriptive Statistics

Calculate the statistical properties of the data set

Statistics

Stats and exploratory histograms in View

Console Outline

Quick Access

Description KNIME Hub Search



Node Context Options: Scorer

The screenshot shows the KNIME Analytics Platform interface with a workflow titled "2: KNIME_Aulas_1.Example".

Workflow Overview:

- Data Reading:** A File Reader node reads the "adult.csv" file.
- Graphical Properties:** A Color Manager node assigns colors by income group, mapping Red for income " $<=50K$ " and Blue for income " $>50K$ ".
- Data Partitioning:** A Partition node creates two separate partitions from the original dataset: training set (80%) and test set (20%).
- Train a Model:** A Train a Model node, specifically a "Building a decision tree learner", is used to build a model.
- Apply the Model:** Predictor nodes apply a specific model to the data set and append the model predictions.
- Score the Model:** A Scorer node compares two columns by their attribute value pairs and shows the confusion matrix. The dialog for this node is open, showing the following data:

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

Descriptive Statistics: A Statistics node calculates the statistical properties of the data set, resulting in Stats and exploratory histograms in View.

Visualize: A Scatter Plot node creates an interactive scatter plot.

Scorer Node Dialog (Open):

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

Console and **Outline** panes are also visible at the bottom of the interface.



Building a Simple Classifier

Nodes

Search all nodes

IO

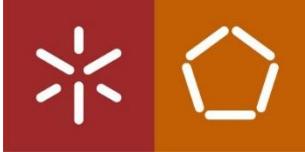
- Excel Reader
- Excel Writer
- Microsoft Authenticator
- CSV Reader
- CSV Writer
- Table Creator
- SharePoint Online Connector
- File Reader
- Show all

Manipulation

- Row Filter
- Column Filter
- Concatenate
- Value Lookup
- Row Aggregator
- Table Splitter
- String Cleaner
- Table Cropper
- Show all

To show the node output, please select a configured or executed node.

QUICK HANDS ON



Universidade do Minho
Departamento de Informática

Ferramentas de Aprendizagem por Máquina

(Machine Learning Tools)

LEI @ 2025/2026, 2º sem