



Universidade do Minho
Departamento de Informática

Ferramentas de Aprendizagem Automática

(Machine Learning Tools)

Aprendizagem e Decisão Inteligentes

Licenciatura em Engenharia Informática/3º ano - 2º semestre

Mestrado [integrado] em Engenharia Informática/4º ano - 2º semestre

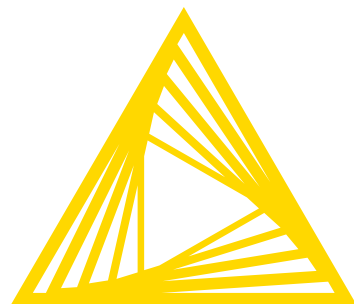
Cesar Analide, Inês Amorim, Pedro Oliveira

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





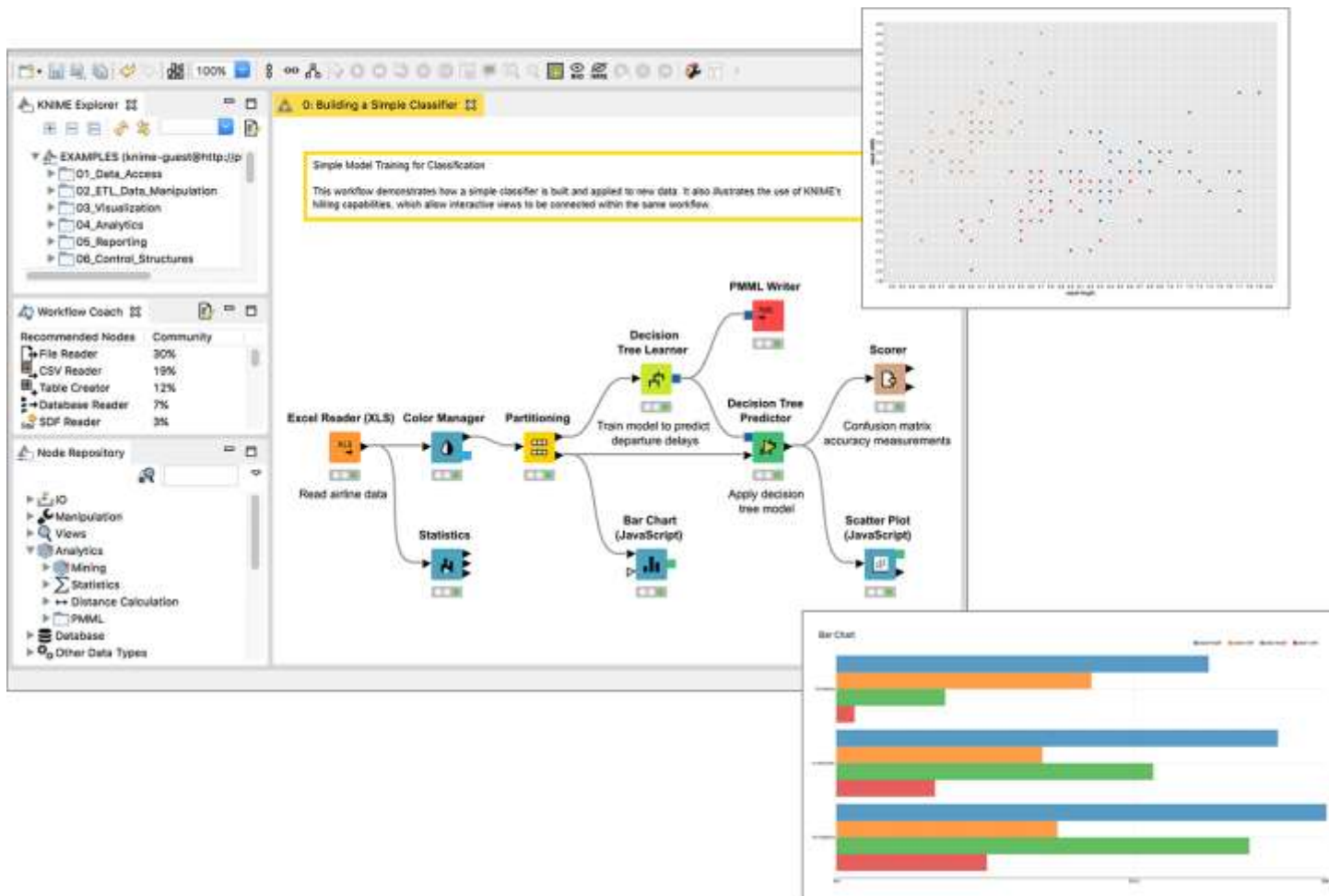
Universidade do Minho
Departamento de Informática

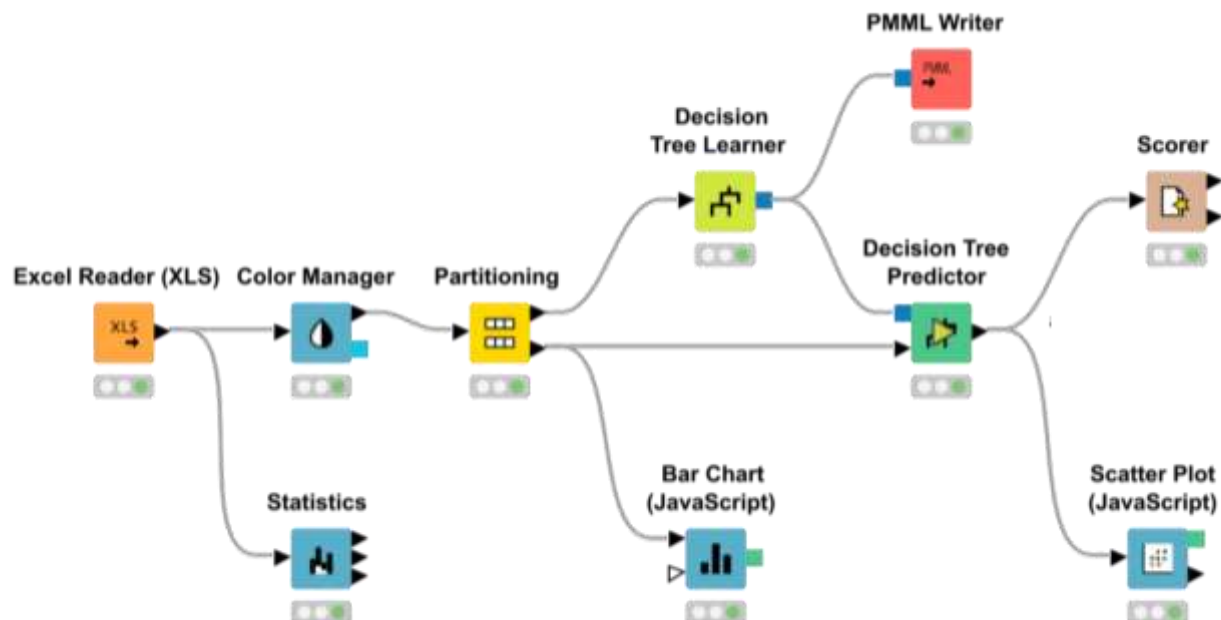


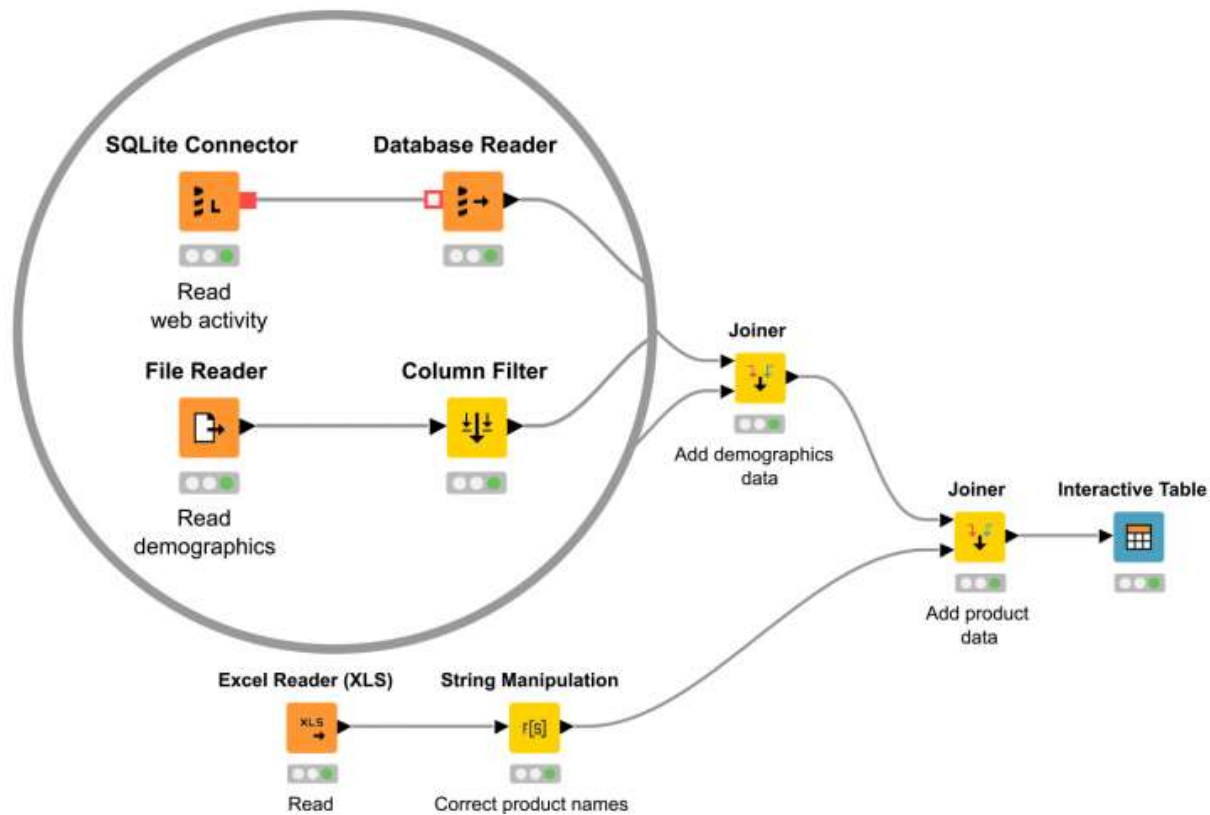
Open for Innovation[®]

KNIME

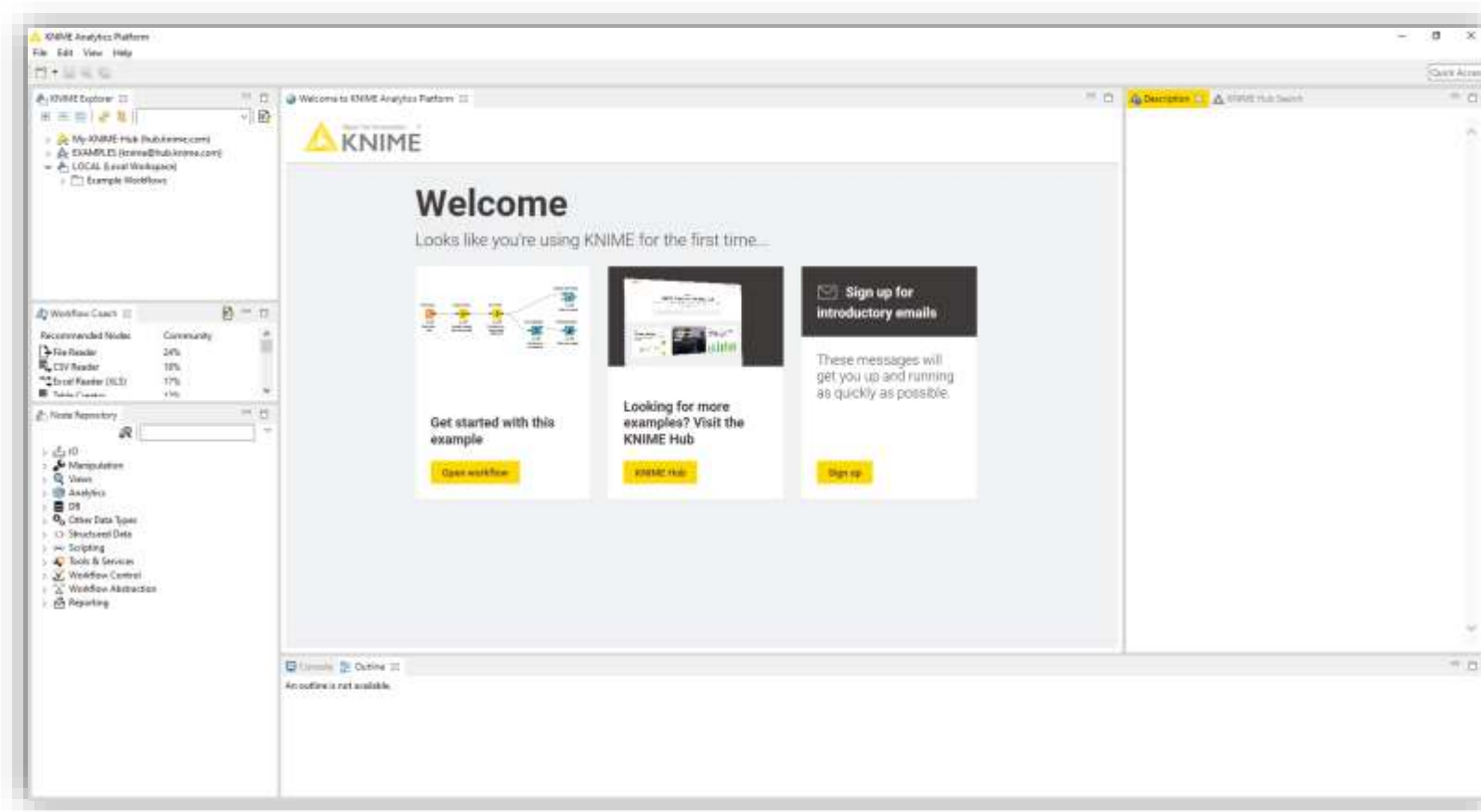
KNIME | Open for Innovation



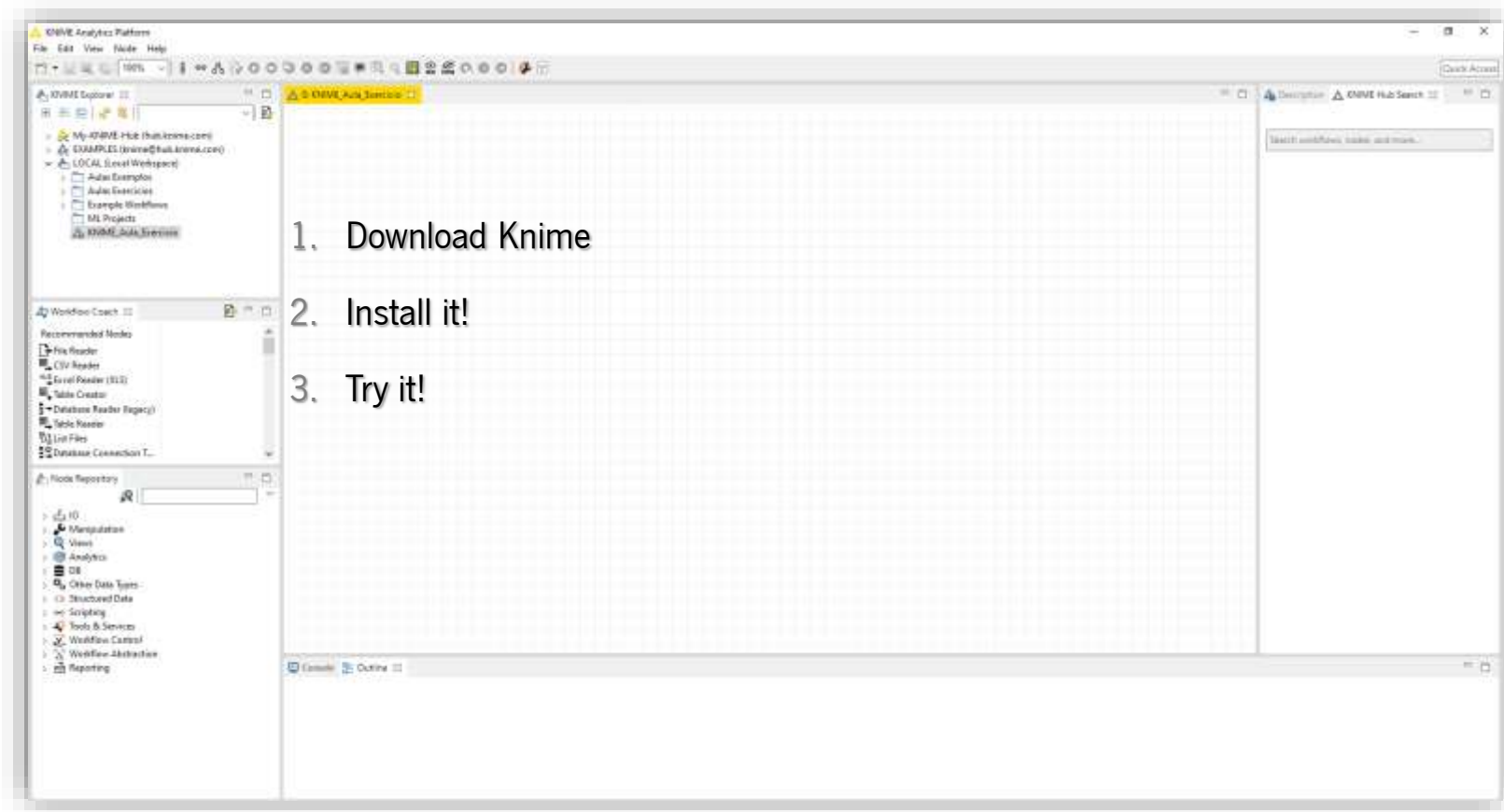










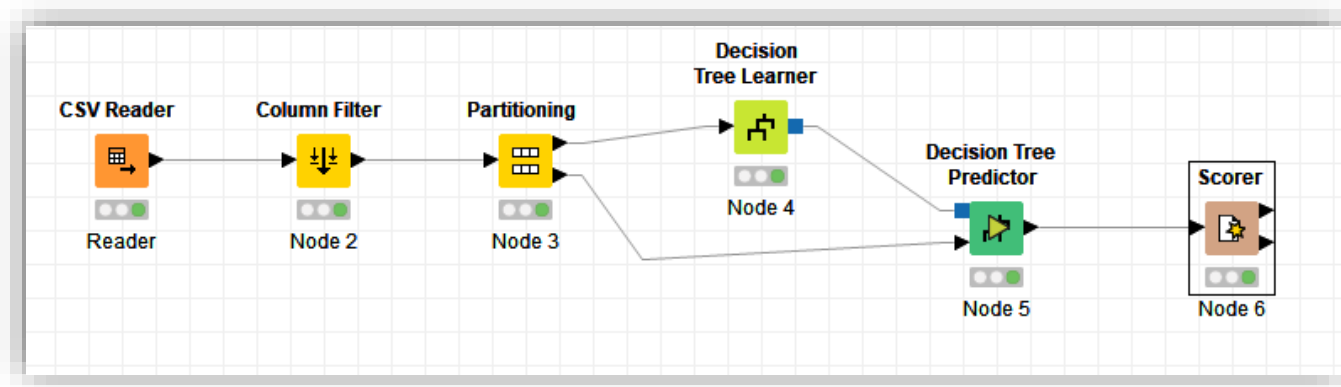


Nodos e Fluxos *Nodes and Workflows*

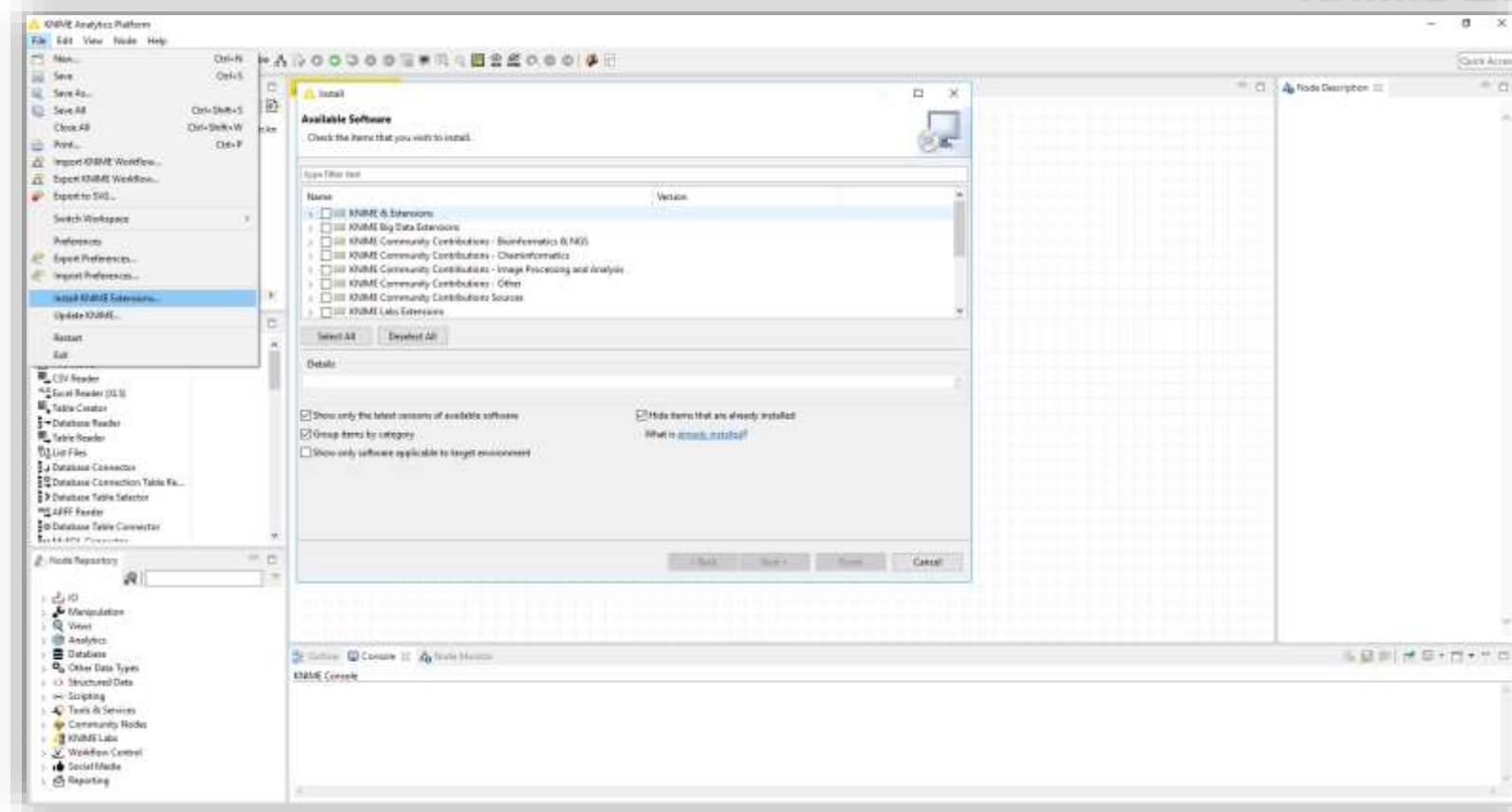
■ Nodos *Nodes*



■ Fluxos *Workflows*



Extensões KNIME *KNIME Extensions*

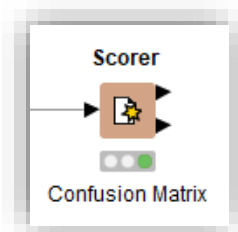


Áreas de trabalho principais *Main Views*

The screenshot displays the KNIME Analytics Platform interface. The central area is the **Workflow Building Area**, which contains a workflow diagram with nodes: CSV Reader, Column Filter, Partitioning, Decision Tree Learner (Node 4), Decision Tree Predictor (Node 5), and Scorer (Node 6). On the left, the **Knime Explorer** panel shows a project tree with 'LOCAL (Local Workspace)' and 'Example Workflows'. Below it, the **Workflow Coach** panel lists recommended nodes like 'Decision Tree Predictor' (85%) and 'Decision Tree To Image' (3%). At the bottom left is the **Node Repository** panel, categorized by 'IO', 'Manipulation', 'View', 'Analytics', 'Database', etc. On the right, the **Node Description** panel provides details for the 'Decision Tree Learner' node, including its purpose, algorithm details, and dialog options. At the bottom right, the **Console Outline Others** panel is visible.

Descrição dos Nodos

Node Description



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

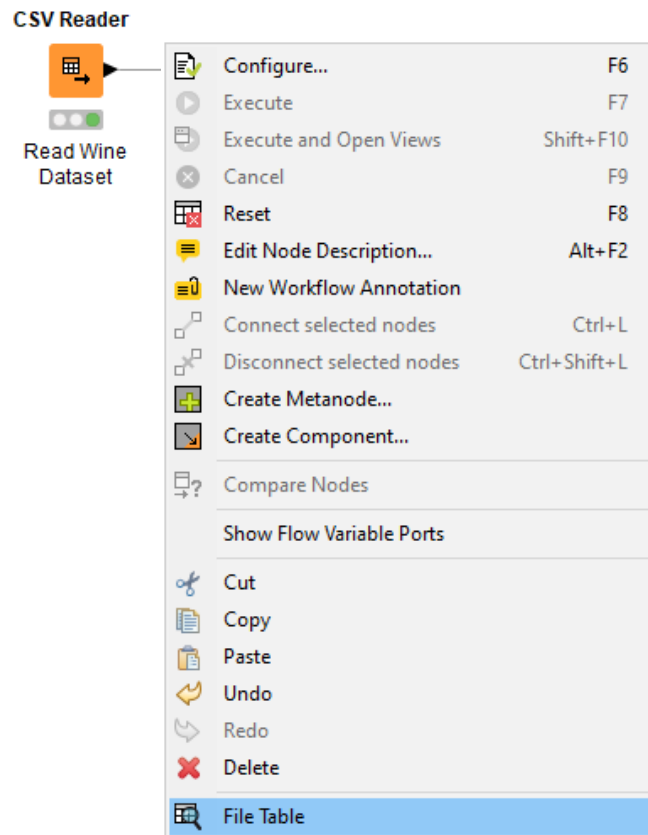
- 0 The confusion matrix.
- 1 The accuracy statistics table.

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.

Visualização do Ficheiro de Dados *Data Table Structure View*



Visualização do Ficheiro de Dados *Data Table Structure View*

File Table - 01 - CSV Reader (Reader)

Table "winequality-red.csv" - Rows: 1599 - Spec - Columns: 12 - Properties - Flow Variables

Column Headers: fixed a..., volatile ..., citric acid, residual...

Data Type (Double): fixed a..., volatile ..., citric acid, residual...

Data Type (String): quality

Row ID: Row0, Row1, Row2, Row3, Row4, Row5, Row6, Row7, Row8, Row9, Row10, Row11, Row12, Row13, Row14, Row15, Row16, Row17, Row18, Row19, Row20, Row21, Row22, Row23, Row24, Row25

Data Cells: 7.4, 0.7, 0, 1.9, 0.076, 17, 160, 0.997, 3, 0.57, 9.5, 5

Sort Descending, Sort Ascending, No Sorting

Available Functions: Standard Double, Percentage, Full Precision, Gray Scale, Bars, Standard Complex Number, Default

The screenshot displays the KNIME Analytics Platform interface. On the left, the 'KNIME Explorer' pane shows a project named 'KNIME_Anal.' with a workflow 'KNIME_Anal.knmx' open. The 'Workflow Canvas' shows a workflow with a 'Partitioning' node. The 'Node Repository' pane on the right shows the 'Partitioning' node selected. The 'Node Description' pane on the far right provides details about the 'Partitioning' node, including its purpose, options, and a table of row variables.

Partitioning

The input table is split into two partitions (i.e. rows), e.g. train and test data. The two partitions are available at the two output ports. The following options are available in the dialog:

Dialog Options

Absolute

Specify the absolute number of rows in the first partition. If there are more rows than specified here, all rows are entered into the first table, while the second table contains no rows.

Relative

The percentage of the number of rows in the input table that are in the first partition; it must be between 0 and 100, inclusively.

Take from top

This node puts the top-most rows into the first output table and the remainder in the second table.

Lower sampling

This node always includes the first and the last row and selects the remaining rows (linearly over the whole table (e.g. every third row). This is useful to downsample a sorted column while maintaining minimum and maximum values.

Draw randomly

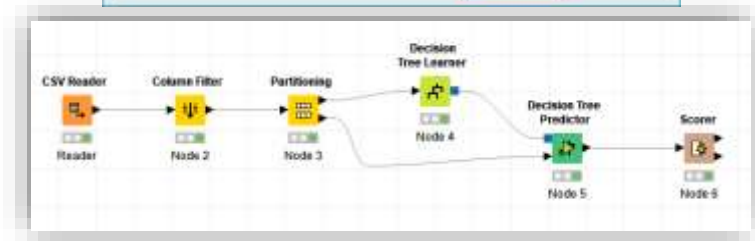
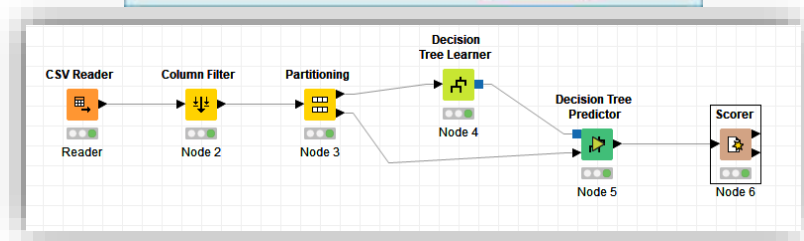
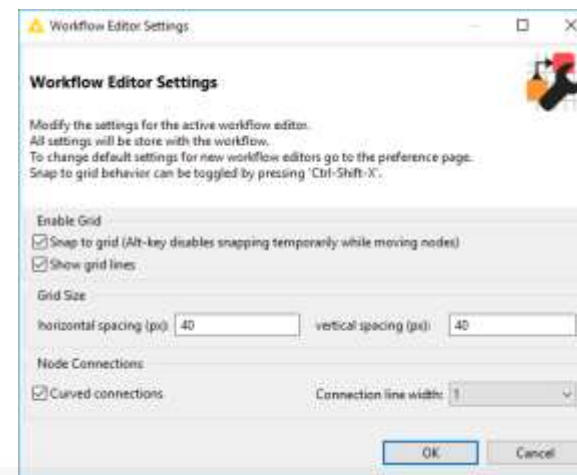
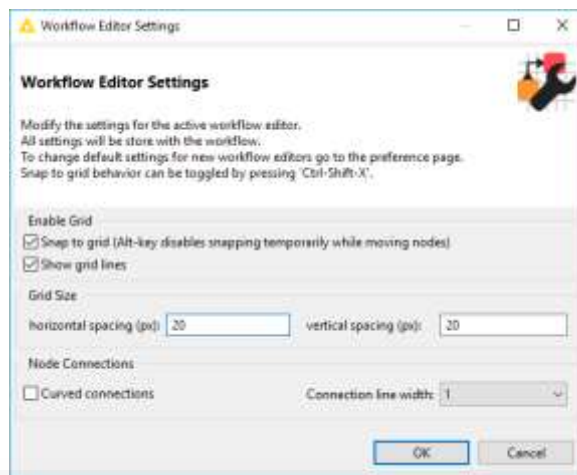
Random sampling of all rows, you may optionally specify a fixed seed (see below).

Stratified sampling

Row Variables

Variable	Value
knime.workflow	/Data/KNIME/Workspace

Views' Customization



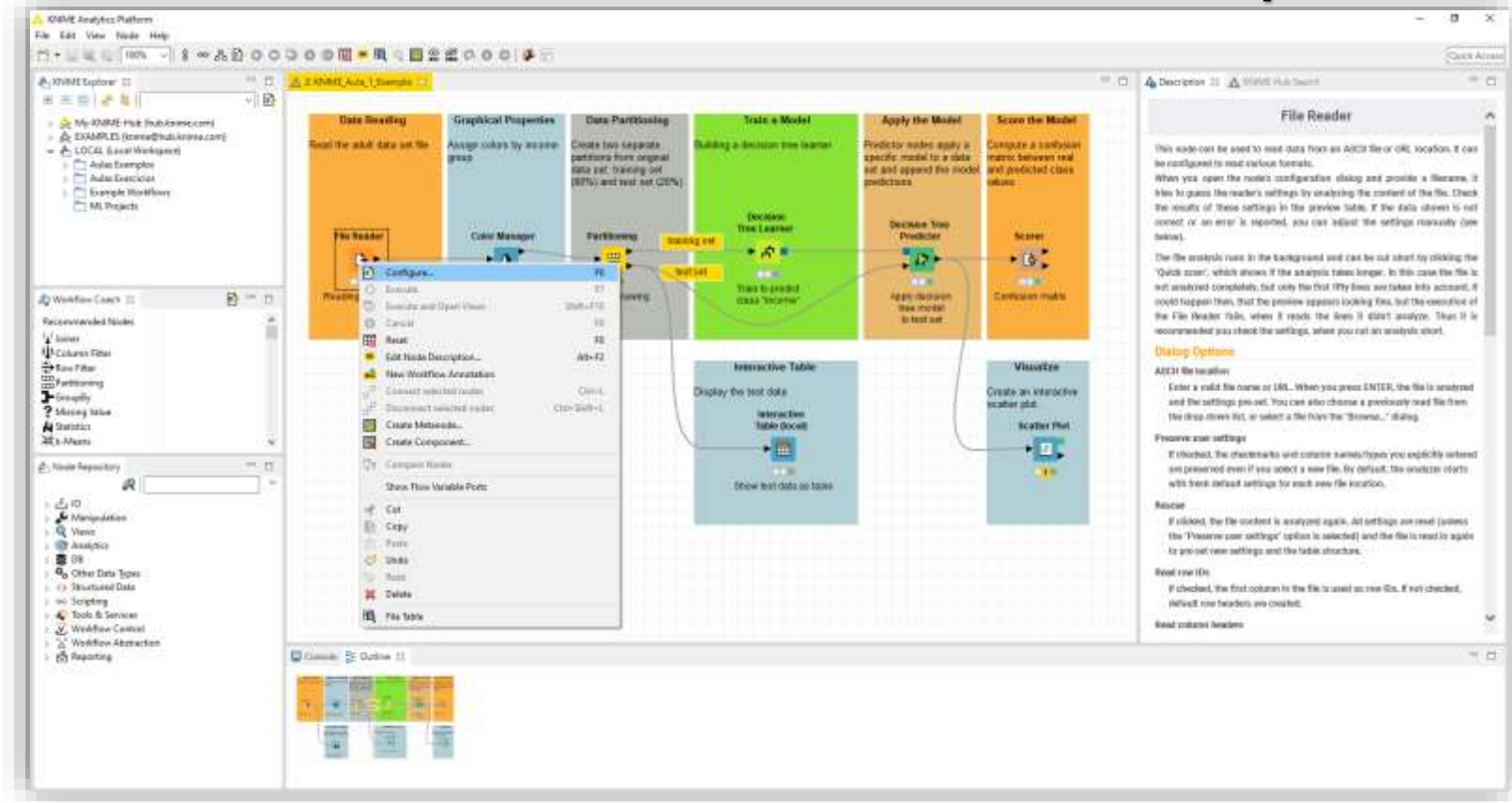
The screenshot displays the KNIME Analytics Platform interface. The main workspace shows a workflow titled "2 KNIME Auto 1 (Sample)". The workflow consists of several nodes connected in a sequence:

- Data Reading:** "File Reader" node, labeled "Providing solution".
- Graphical Properties:** "Color Manager" node, labeled "Red for income <= 50K, Blue for income > 50K".
- Data Partitioning:** "Partitioning" node, labeled "Random Shuffling".
- Train a Model:** "Decision Tree Learner" node, labeled "Train to predict class 'income'".
- Apply the Model:** "Decision Tree Predictor" node, labeled "Apply decision tree model to test set".
- Score the Model:** "Scorer" node, labeled "Compute a confusion matrix between real and predicted class values".
- Descriptive Statistics:** "Statistics" node, labeled "Calculate the statistical properties of the data set".
- Interactive Table:** "Interactive Table (Knock)" node, labeled "Display the test data".
- Visualize:** "Scatter Plot" node, labeled "Create an interactive scatter plot".

The right sidebar shows the configuration for the "Decision Tree Learner" node. It includes a description of the algorithm, a link to the "Using Options" section, and a list of options:

- Class classes:** To select the target attributes. 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 predictive quality (the prediction). See 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: Data Loader



The screenshot displays the KNIME Analytics Platform interface. The central workspace shows a workflow with several nodes: File Reader, Column Manager, Partitioning, Training set, Test set, Decision Tree Learner, Decision Tree Predictor, Scorer, Interactive Table, and Visualize. The File Reader node is selected, and its context menu is open, showing options like Configure..., Execute, Save, and Delete. The right-hand pane displays the 'File Reader' node description, which includes a detailed explanation of the node's function, a list of dialog options, and a 'Quick start' section.

File Reader

This node can be used to read data from an ASCII file or ODBC 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 start", which shows if the analysis takes longer. In this case the file is not analyzed completely, but only the first 100 lines are taken into account. It could happen then, that the preview appears looking fine, but the content of the File Reader fails, when it reads the lines it didn't analyze. Thus it is recommended you check the settings, when you run 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.

Preview row settings

If checked, the checkboxes and column names/types you explicitly selected are preserved even if you select a new file. By default, the analyzer starts with fresh default settings and the table structure.

Reader

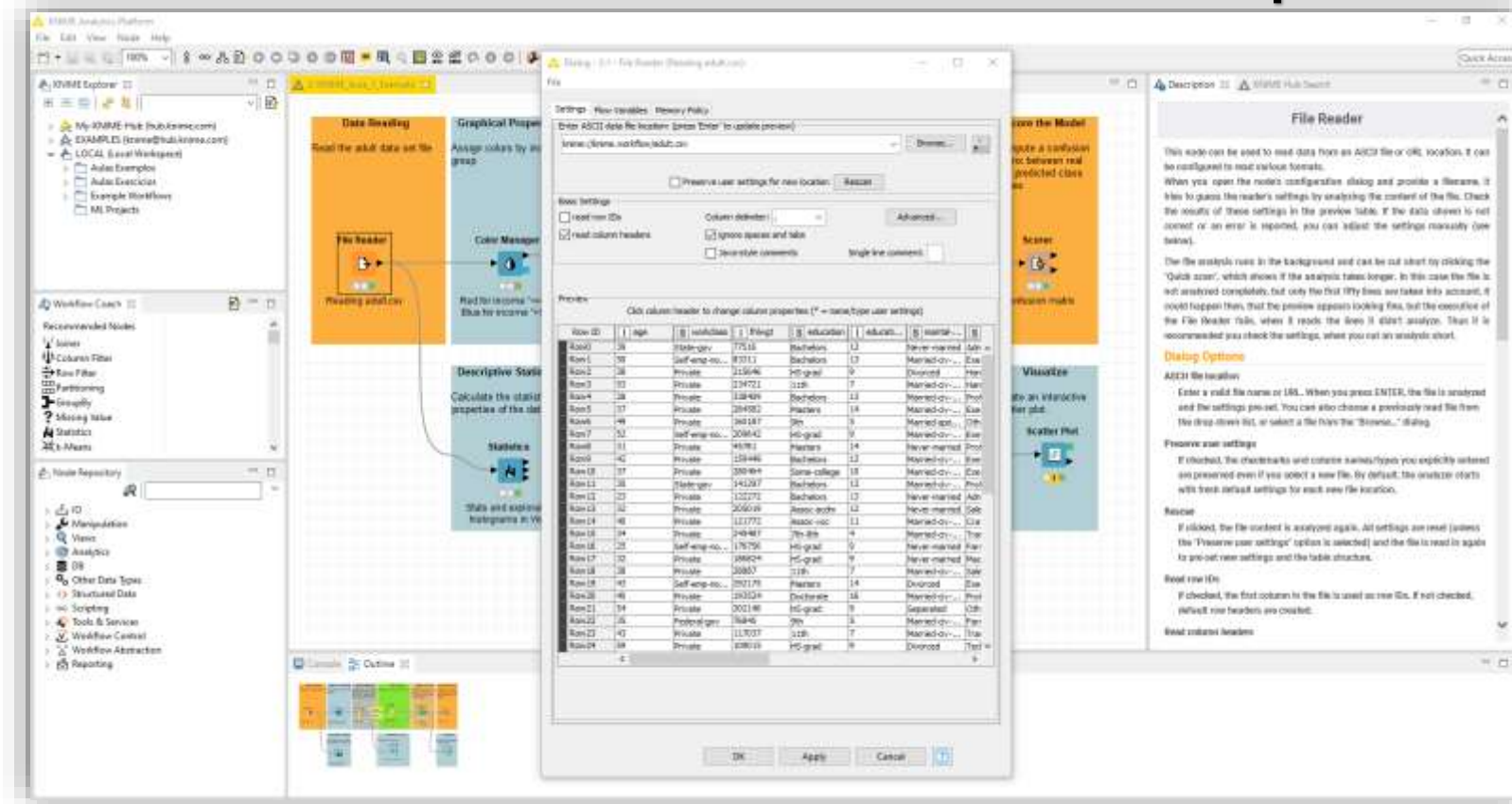
If checked, the file content is analyzed again. All settings are reset (unless the "Preserve row settings" option is selected) and the file is read in again to pre-set new settings and the table structure.

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 screenshot shows the KNIME Data Loader node configuration dialog. The 'File' tab is active, showing the file path and various settings. The 'Preview' tab shows a table of data. The 'Data Revealing' context option is highlighted, showing the 'File Reader' node. The 'Graphical Properties' context option is highlighted, showing the 'Color Manager' node. The 'Descriptive Statistics' context option is highlighted, showing the 'Statistics' node. The 'Visualize' context option is highlighted, showing the 'Scatter Plot' node. The 'File Reader' node is highlighted, showing its description and settings.

Data Revealing
Reveal the actual data set file

Graphical Properties
Assign colors by group

Descriptive Statistics
Calculate the statistical properties of the data

Visualize
Use an interactive bar plot

File Reader
This node can be used to read data from an ASCII file or OLE 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 appears to not connect 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 100 lines are taken into account. It could happen then, that the preview appears looking fine, but the connection of the File Reader fails, when it reads the lines it didn't analyze. Thus it is recommended you check the settings, when you run 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.

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

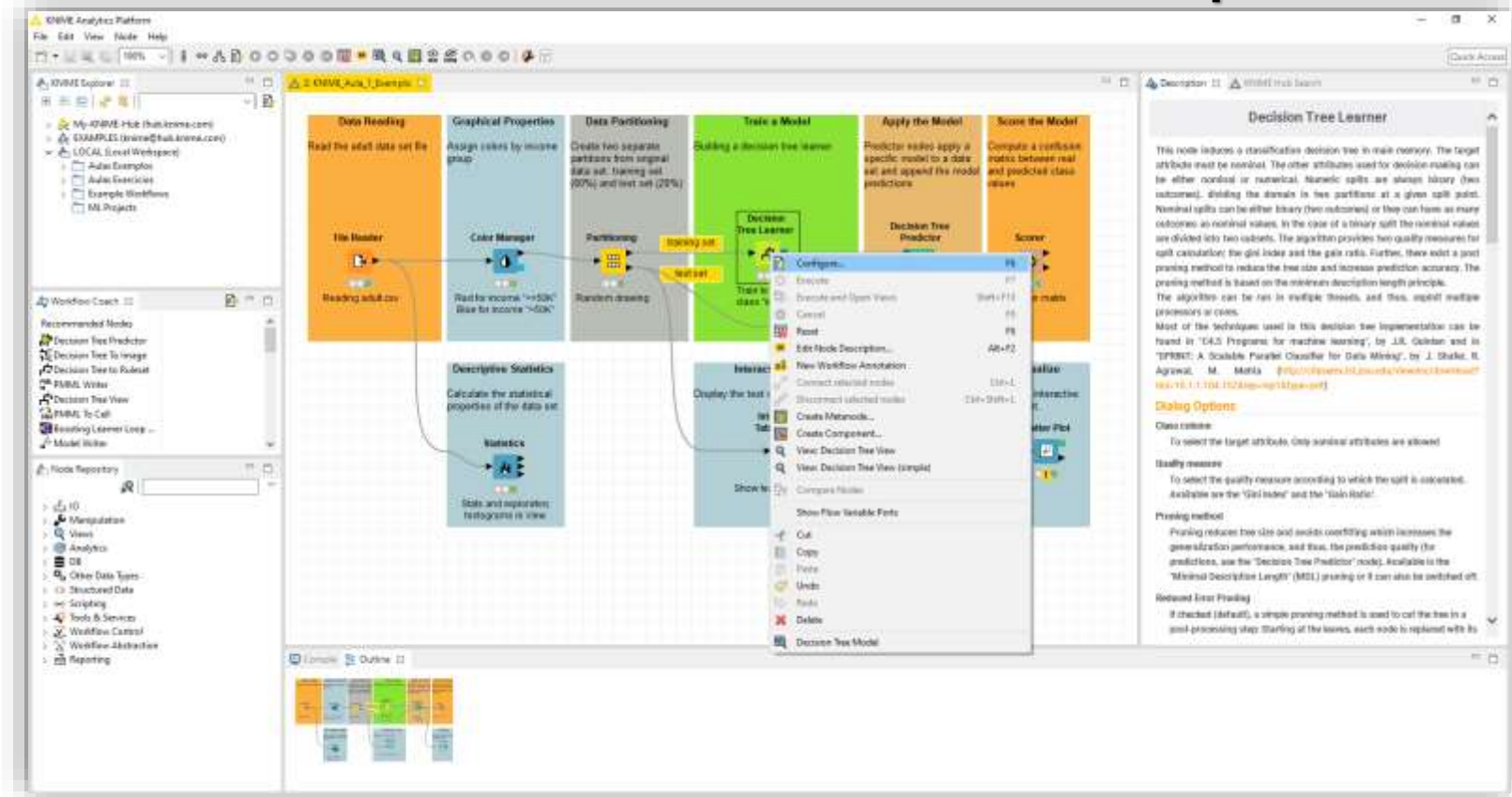
Read row IDs
If checked, the first column in the file is used as row IDs. If not checked, default row headers are created.

Read column headers

Preview

Row ID	Age	Sex	Height	Weight	Education	Married	Divorced
Row0	28	Male	175.15	77.55	High school	Yes	No
Row1	29	Self-emp-inc	171.13	83.13	High school	No	No
Row2	36	Private	175.66	101.90	High school	No	No
Row3	33	Private	174.72	111.8	High school	No	No
Row4	38	Private	178.49	132.97	High school	No	No
Row5	37	Private	178.82	132.97	High school	No	No
Row6	44	Private	180.17	196	High school	No	No
Row7	32	Self-emp-inc	176.42	104.42	High school	No	No
Row8	31	Private	176.71	111.8	High school	No	No
Row9	42	Private	178.46	132.97	High school	No	No
Row10	37	Private	178.46	132.97	High school	No	No
Row11	38	Self-emp-inc	171.13	83.13	High school	No	No
Row12	32	Private	172.72	111.8	High school	No	No
Row13	32	Private	172.72	111.8	High school	No	No
Row14	46	Private	177.72	132.97	High school	No	No
Row15	34	Private	174.72	111.8	High school	No	No
Row16	28	Self-emp-inc	175.66	101.90	High school	No	No
Row17	32	Private	176.42	104.42	High school	No	No
Row18	36	Private	176.42	104.42	High school	No	No
Row19	43	Self-emp-inc	176.42	104.42	High school	No	No
Row20	46	Private	176.42	104.42	High school	No	No
Row21	34	Private	176.42	104.42	High school	No	No
Row22	36	Self-emp-inc	176.42	104.42	High school	No	No
Row23	43	Private	176.42	104.42	High school	No	No
Row24	34	Private	176.42	104.42	High school	No	No

Node Context Options: Model Learner



The screenshot displays the KNIME Analytics Platform interface. The main workspace shows a workflow with several nodes: Data Reading (File Reader), Graphical Properties (Color Manager), Data Partitioning (Partitioning), Train a Model (Decision Tree Learner), Apply the Model (Decision Tree Predictor), and Score the Model (Scorer). A context menu is open over the Decision Tree Learner node, listing various actions such as Configure, Interact, and Delete. The right-hand pane shows the configuration options for the Decision Tree Learner node, including a description of the node's function and various settings like Class column, Quality measure, and Pruning method.

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. Nominal 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, supports multiple processors or cores.

Most of the techniques used in this decision tree implementation can be found in "EAS Programs for machine learning", by J.L. Garbin and in "EPROM: A Scalable Parallel Classifier for Data Mining", by J. Shaker, R. Agrawal, M. Morita (<http://papers.kit.edu/knime-downloads/knime-V8.1.1-104-1024894-eprom.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 improves the generalization performance, and thus, the prediction quality (the predictions, see 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-pruning step starting at the leaves, each node is replaced with its

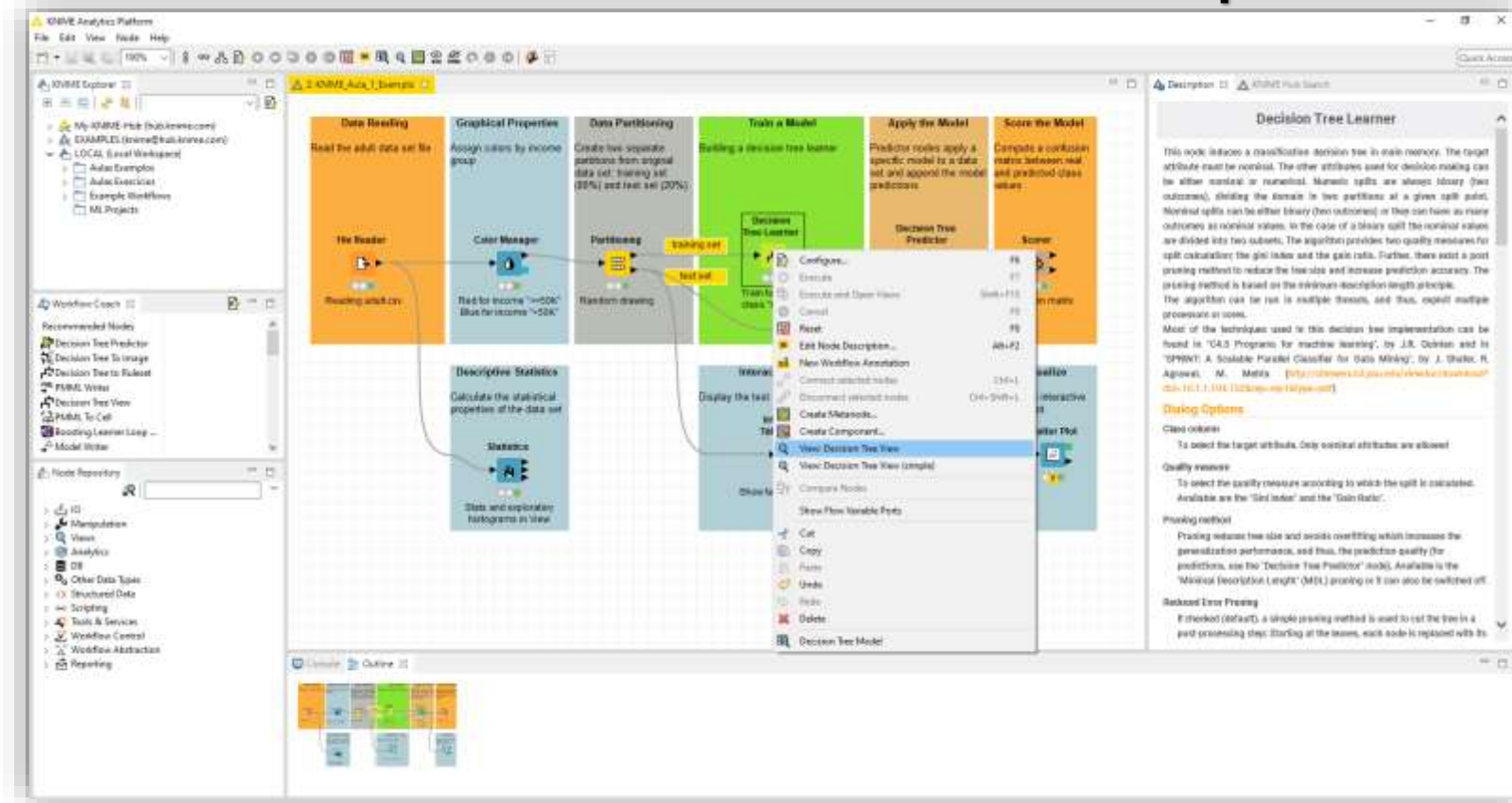
Node Context Options: Model Learner

The screenshot displays the KNIME Analysis Platform interface. The main workspace shows a workflow with several nodes: 'File Reader', 'Color Manager', 'Data Partitioning', 'Train a Model', 'Apply the Model', and 'Score the Model'. A 'Descriptive Statistics' node is also visible. The 'Train a Model' node is selected, and its context options dialog box is open. The dialog box is titled 'Dialog: 210 - Decision Tree Learner (Tree L...)' and contains the following sections:

- 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
- Node split:**
 - ☐ Force root split column
 - Root split column:
- Binary nominal splits:**
 - ☐ Binary nominal splits
 - Max allowed:
 - ☐ Filter invalid attribute values in child nodes

The right-hand pane shows the 'Decision Tree Learner' node description, which includes a detailed explanation of the algorithm and its options.

Node Context Options: Model Learner



The screenshot displays the KNIME Analytics Platform interface. On the left, the 'KNIME Explorer' shows a project named 'EXAMPLE1' with subfolders for 'Auto Examples', 'Auto Exercises', 'Example Workflows', and 'ML Projects'. The 'Workflow Coach' on the left lists recommended nodes: 'Decision Tree Predictor', 'Decision Tree to Image', 'Decision Tree to Rule Set', 'FPMML Writer', 'Decision Tree View', 'FPMML to Cell', 'Boosting Learner Loop', and 'Model Store'. The 'Node Repository' on the bottom left shows a tree structure of nodes categorized by 'Manipulation', 'Visual', 'Analytics', 'DB', 'Other Data Types', 'Scripting', 'Tools & Services', 'Workflow Control', 'Workflow Abstraction', and 'Reporting'.

The main workspace shows a workflow with the following 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 'Decision Tree Learner' node is selected, and its context menu is open, showing options such as '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 (sample)', 'Compare Nodes', 'Show Flow Variable Ports', 'Cut', 'Copy', 'Paste', 'Undo', 'Redo', and 'Delete'. The 'View Decision Tree View' option is highlighted.

The right pane shows the 'Description' of the 'Decision Tree Learner' node. It explains that this node induces a classification decision tree in main memory, where the target attribute must be nominal. It details the splitting criteria (Gini index and gain ratio) and the pruning method (minimum description length principle). It also mentions that the algorithm can be run in multiple threads and supports multiple preprocessing options. The 'Display Options' section lists 'Class labels' and 'Quality measure'. The 'Pruning method' section describes the 'Minimal description length (MDL) pruning' and the 'Reduced Error Pruning' method.

Node Context Options: Model Learner

The screenshot displays the KNIME Analytics Platform interface. The main workspace shows a workflow with the following nodes: File Reader, Color Manager, Partitioning, Training set, Decision Tree Learner, Decision Tree Predictor, and Scorer. The 'Decision Tree Learner' node is selected, and its context menu is open, showing options like 'Data Reading', 'Graphical Properties', 'Data Partitioning', 'Train a Model', 'Apply the Model', and 'Score the Model'. The 'Decision Tree Learner' node is highlighted in green. The 'Decision Tree Learner' node context menu is open, showing options like 'Data Reading', 'Graphical Properties', 'Data Partitioning', 'Train a Model', 'Apply the Model', and 'Score the Model'. The 'Decision Tree Learner' node is highlighted in green.

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. Nominal 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 decision and increase prediction accuracy. The

Decision Tree View - 218 - Decision Tree Learner (train to predict)

File Hide Tree

Table: % n

Category	%	n
<=50K	75.8	18,787
>50K	24.2	6,051
Total	100.0	24,838

Chart: Color column: income

relationship

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50K	52.3	12,999
>50K	47.7	11,839
Total	100.0	24,838

Chart: Color column: income

Table: % n

Category	%	n
<=50		

Node Context Options: Scorer

The screenshot displays the KNIME Analytics Platform interface. A workflow is visible in the main workspace, consisting of several nodes: File Reader, Column Manager, Partitioning, Training set, Test set, Decision Tree Learner, Decision Tree Predictor, and Scorer. The Scorer node is selected, and its context menu is open, showing various options such as Configure..., Execute, Execute and Open View, Cancel, Reset, Edit Node Description..., New Workflow Annotation, Convert selected nodes, Disconnect selected nodes, Create Metadata..., Create Component..., View Confusion Matrix, Compare Nodes, Show Flow Variable Ports, Out, Copy, Paste, Undo, Redo, Delete, Confusion matrix, and Accuracy statistics.

The Scorer node's description panel on the right provides detailed information about its function: "Compare 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 output port provides a list of the underlying rows." The description also includes a table showing the confusion matrix structure and a list of available attributes for comparison.

Node Context Options: Scorer

The screenshot displays the KNIME Analytics Platform interface. The main workspace shows a workflow with several nodes: 'File Reader', 'Color Manager', 'Partitioner', 'Decision Tree (Learner)', 'Decision Tree (Predictor)', 'Scorer', and 'Visualization'. A context menu is open for the 'Scorer' node, showing options for 'First Column', 'Second Column', 'Sorting of values in tables', 'Sorting strategy', 'Provide scores as flow variables', and 'Missing values'. The 'First Column' is set to 'income' and the 'Second Column' is set to 'Predicted (score)'. The 'Sorting strategy' is set to 'Shannon order'. The 'Provide scores as flow variables' checkbox is unchecked. The 'Missing values' section shows 'Ignore' selected.

Scorer

Compare two columns by their attribute value pairs and shows the confusion matrix, i.e. how many rows of which attribute and their classification matches. 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 in 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 output 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 **Chatters-Supps**.

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 profile:
The counts (i.e. accuracy, error rate, number of correct and wrong classifications) are reported as flow variables with a fixed coded name. This option allows you to define a prefix for these variable identifiers so that names conflicts are resolved.

Missing Values:
Choose how to treat missing values in either the reference or prediction columns. 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

The screenshot displays the KNIME Analytics Platform interface. A workflow is visible in the central workspace, consisting 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 such as 'Configure...', 'Execute', 'Execute and Open View', 'Cancel', 'Reset', 'Edit Node Description...', 'New Workflow Annotation', 'Connect selected nodes', 'Disconnect selected nodes', 'Create Notebook...', 'Create Component...', 'View Confusion Matrix', 'Compare Models', 'Show Flow Variable Ports', 'Cut', 'Copy', 'Paste', 'Undo', 'Redo', 'Delete', 'Confusion matrix', and 'Accuracy statistics'. The 'Confusion matrix' option is highlighted. The right-hand pane shows the 'Scorer' node's description, which explains that it compares two columns by their attribute value pairs and shows the confusion matrix, including how many rows of which attribute and their classification match. It also mentions that it is possible to highlight cells of this matrix to determine the underlying rows. The bottom status bar shows 'Console' and 'Outline' tabs.

Node Context Options: Scorer

The screenshot displays the KNIME Analytics Platform interface. A workflow is visible in the main workspace, consisting of several nodes: 'File Reader', 'Color Manager', 'Data Partitioning', 'Train a Model', 'Apply the Model', and 'Scorer'. The 'Scorer' node is highlighted, and its context options dialog is open. The dialog shows a confusion matrix and various performance metrics.

Confusion Matrix - 2d - Score (Confusion matrix)

	income <= 50K	> 50K
income <= 50K	216	316
> 50K	174	365

Statistics

- Correct classified: 5.463
- Wrong classified: 1.036
- Accuracy: 83.878 %
- Error: 16.122 %
- Cohen's kappa (k): 0.551

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 counts (i.e. accuracy, error rate, number of correct and wrong classifications) are reported as flow variables with a fixed named 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 columns. 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

