

**Business Analytics and Data-Driven Decision Making**  
**Hands-on-session Predictive Analytics-2**

# **Decision Trees, Random Forests, Cross Validation with KNIME and Clustering**

*Raghava Mukkamala*

**Associate Professor & Director,  
Centre for Business Data Science**

**Copenhagen Business School, Denmark**

**Email: [rrm.digi@cbs.dk](mailto:rrm.digi@cbs.dk), Centre: <https://cbsbda.github.io/>**



**Some slides based on KNIME Analytics Platform**  
**Source: <https://www.knime.com/sites/default/files/2021-07/slides-l1-ds.pdf>**

# Outline

- KNIME Analytics Platform for Data Scientists: Basics
- Hands-on Sessions
  - Credit Risk Analysis Using Decision Trees
  - Credit Risk Analysis Using Logistic Regression
  - Cross-Validation in KNIME
  - Credit Risk Analysis Using Random Forests with CV
  - Clustering using KNIME

# Link to Datasets

Link to Datasets and hands-on guides:

[https://www.dropbox.com/scl/fo/  
g90s2fj8b9ifqnbqm58xt/  
AMC1Usbx2n3JugmKFRcLCs0?  
rlkey=cszcdyygicm6tk60ib8odfflb&dl=0](https://www.dropbox.com/scl/fo/g90s2fj8b9ifqnbqm58xt/AMC1Usbx2n3JugmKFRcLCs0?rlkey=cszcdyygicm6tk60ib8odfflb&dl=0)

TinyURL: <https://tinyurl.com/predictive-hands-on-02>



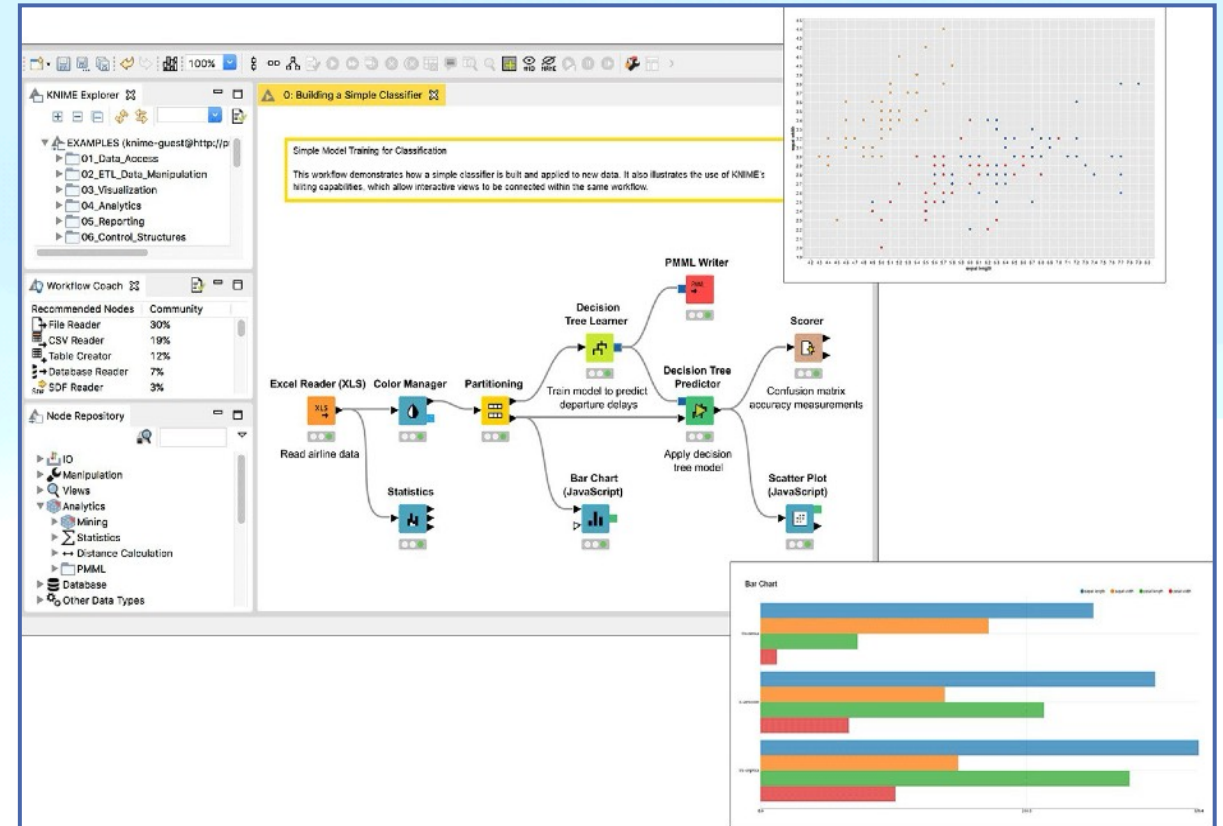
# OVERVIEW: KNIME ANALYTICS PLATFORM



The slides are taken and adapted from:  
KNIME AG: [L1-DS] KNIME Analytics Platform for Data Scientists: Basics  
<https://www.knime.com/sites/default/files/2021-07/slides-l1-ds.pdf>

# What is KNIME Analytics Platform?

- A tool for data analysis, manipulation, visualization, and reporting
- Based on the graphical programming paradigm
- Provides a diverse array of extensions:
  - Machine learning,
  - Text Mining,
  - Network Mining



KNIME Classic look (before 5.0 version)

# KNIME Analytics Platform 5.0 - Modern look

KNIME Analytics Platform 5

## Get started with KNIME Analytics Platform 5

Open for Innovation  
**KNIME**

### Examples

- Combine Clean and ...
- Countf and Sumf
- Non-standard format...

Find more resources on the KNIME Community Hub

### 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 Analytics Platform 5

### Repository

Search Nodes

IO

- Excel Reader
- Excel Writer
- Microsoft Authenticator
- Google Authenticator
- Google Sheets Reader
- Google Sheets Writer
- CSV Reader
- CSV Writer
- Show all

Manipulation

- Row Filter
- Column Filter
- Concatenate
- Value Lookup
- Row Aggregator
- Table Splitter
- Table Cropper
- Cell Extractor
- Show all

Views

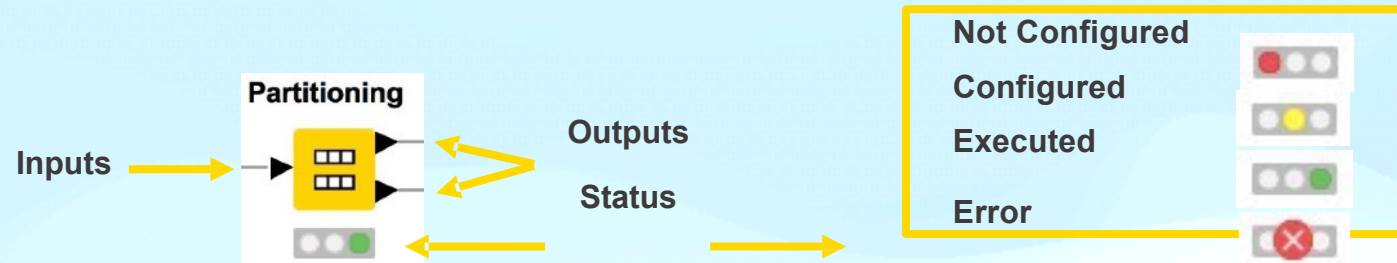
- Bar Chart
- Line Plot
- Pie Chart
- Stacked Area Chart
- Scatter Plot
- Statistics
- Heatmap
- Histogram
- Show all

Start building your workflow by dropping your data or nodes here.

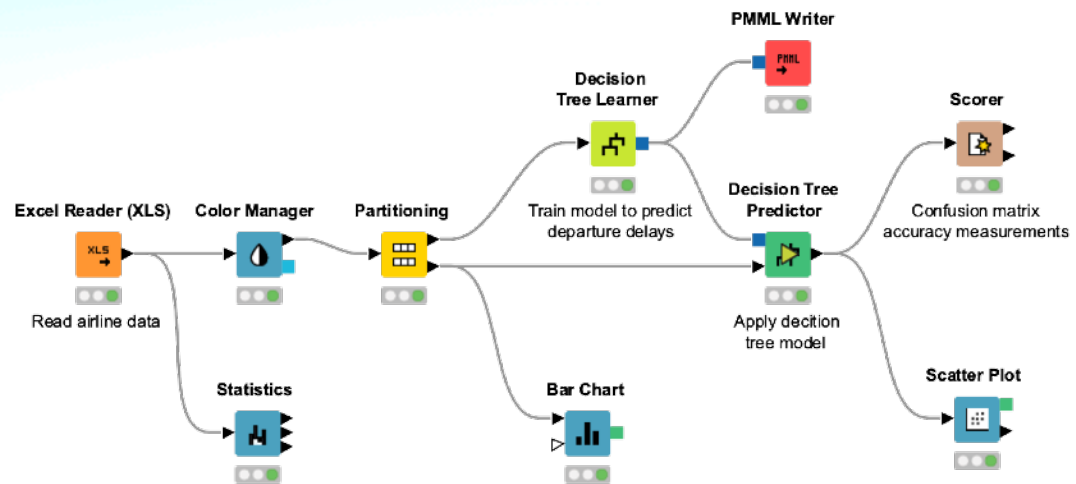
To show the node output, please select a configured or executed node.

# Visual KNIME Workflows

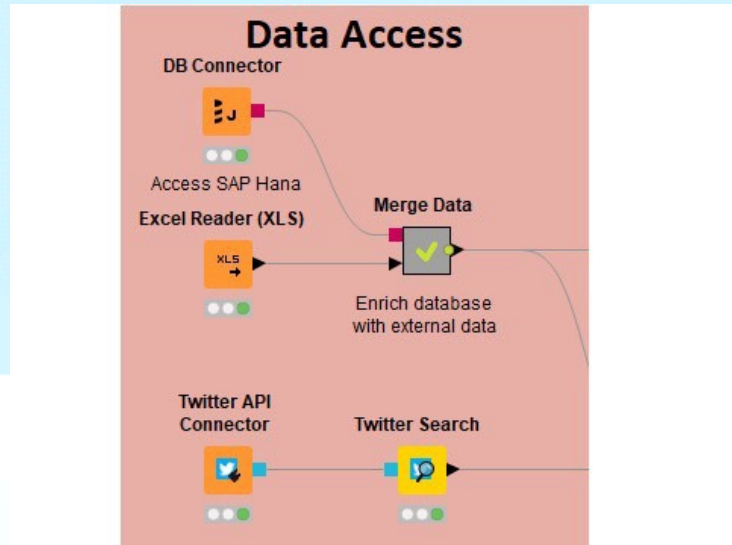
- **NODES** perform individual tasks on data
- Nodes has ports for data in and out



Nodes are combined to create **WORKFLOWS**

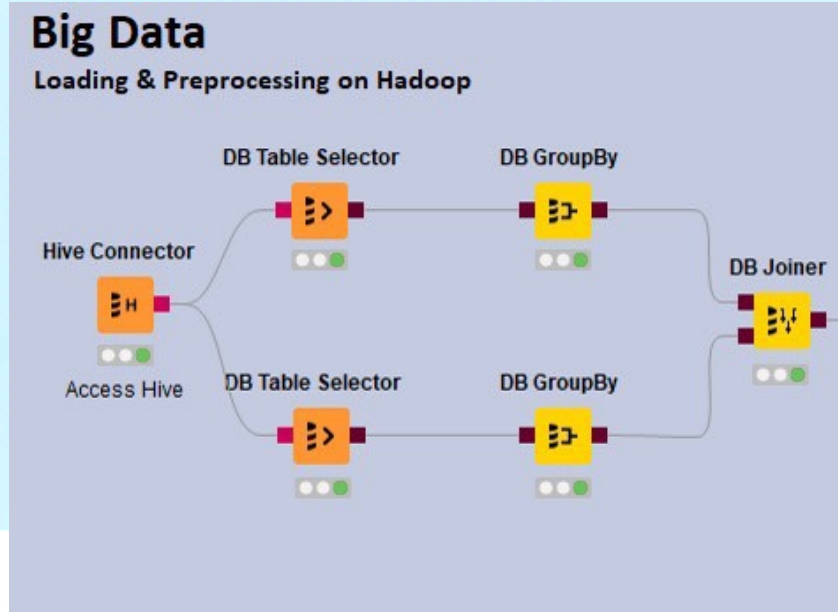


# Data Access



- Databases
  - MySQL, PostgreSQL, Oracle
  - Theobald
  - any JDBC (DB2, MS SQL Server)
  - Amazon DynamoDB
- Files
  - CSV, txt, Excel, Word, PDF
  - SAS, SPSS
  - XML, JSON, PMML
  - Images, texts, networks
- Other
  - Twitter, Google
  - Amazon S3, Azure Blob Store
  - Sharepoint, Salesforce
  - Kafka
  - REST, Web services

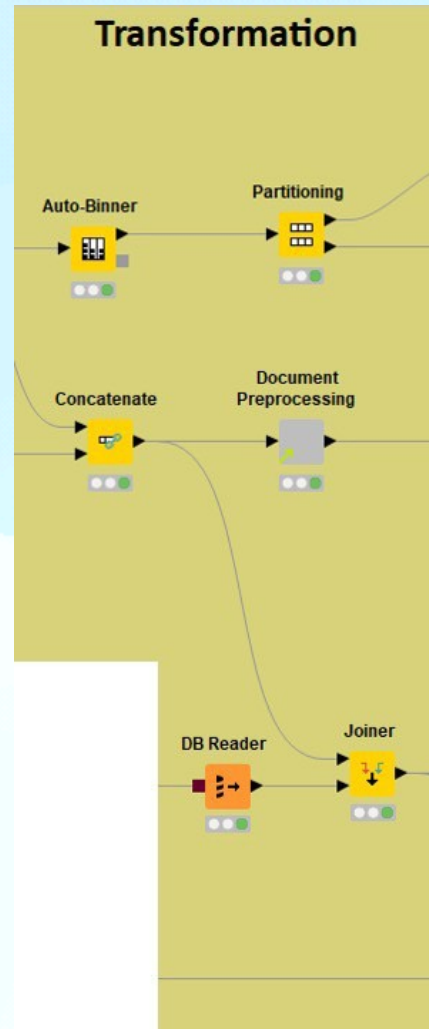
# Big Data



- Spark & Databricks
- HDFS support
- Hive
- Impala
- In-database processing

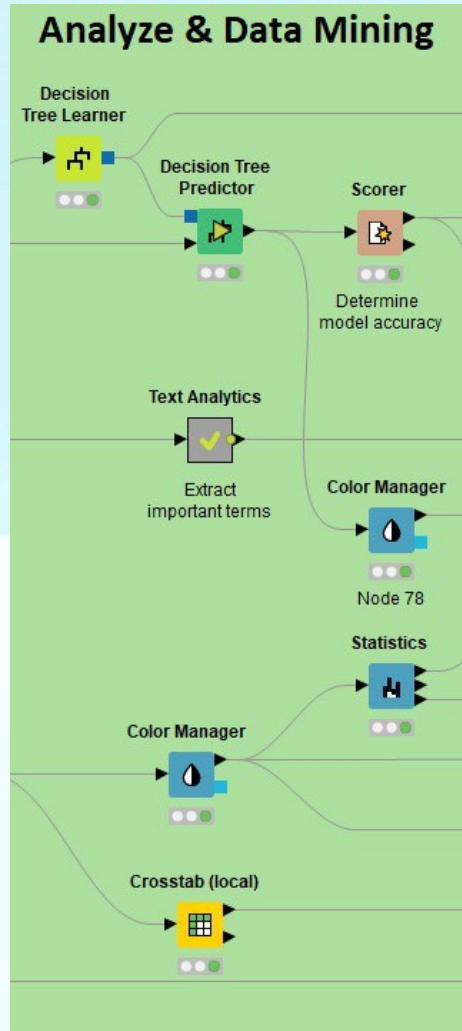


# Data Transformation



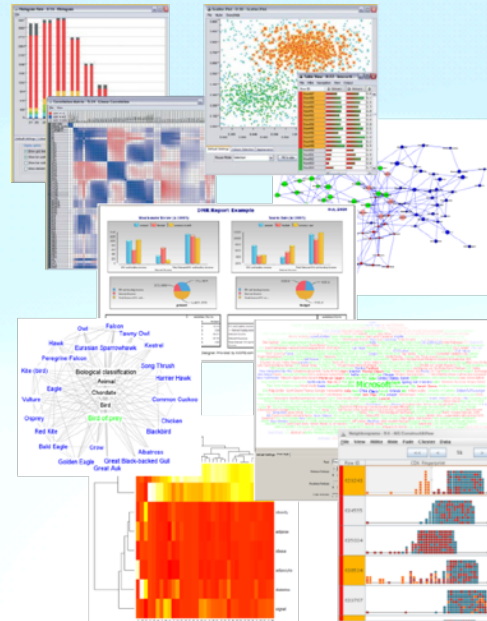
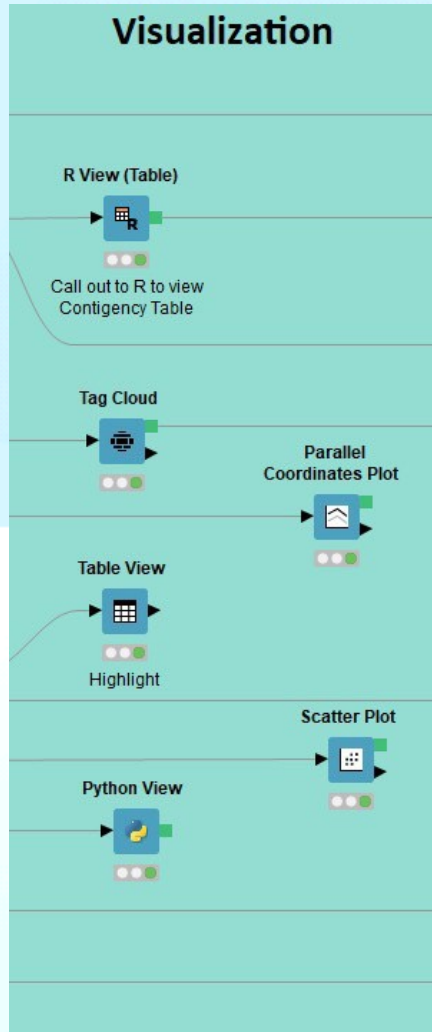
- Preprocessing
  - Row, column, matrix based
- Data blending
  - Join, concatenate, append
- Aggregation
  - Grouping, pivoting, binning
- Feature Creation and Selection

# Analysis & Data Mining



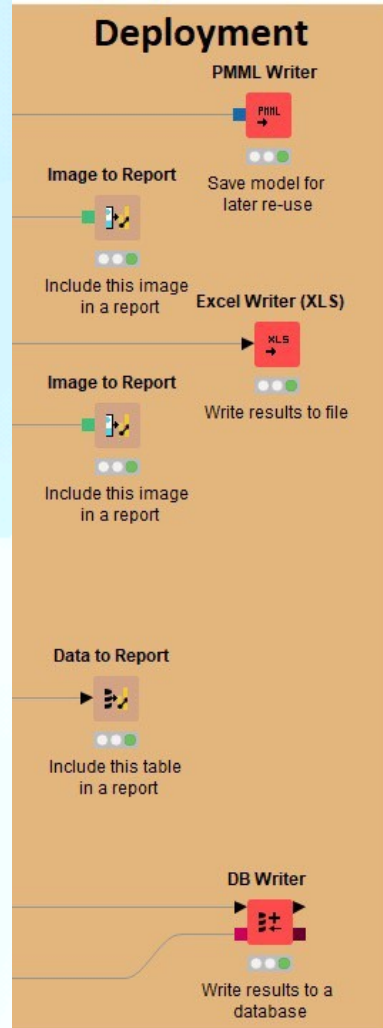
- Regression
  - Linear, logistic
- Classification
  - Decision tree, ensembles, SVM, MLP, Naïve Bayes
- Clustering
  - k-means, DBSCAN, hierarchical
- Validation
  - Cross-validation, scoring, ROC
- Deep Learning
  - Keras, DL4J
- External
  - R, Python, Weka, H2O, Keras

# Visualization



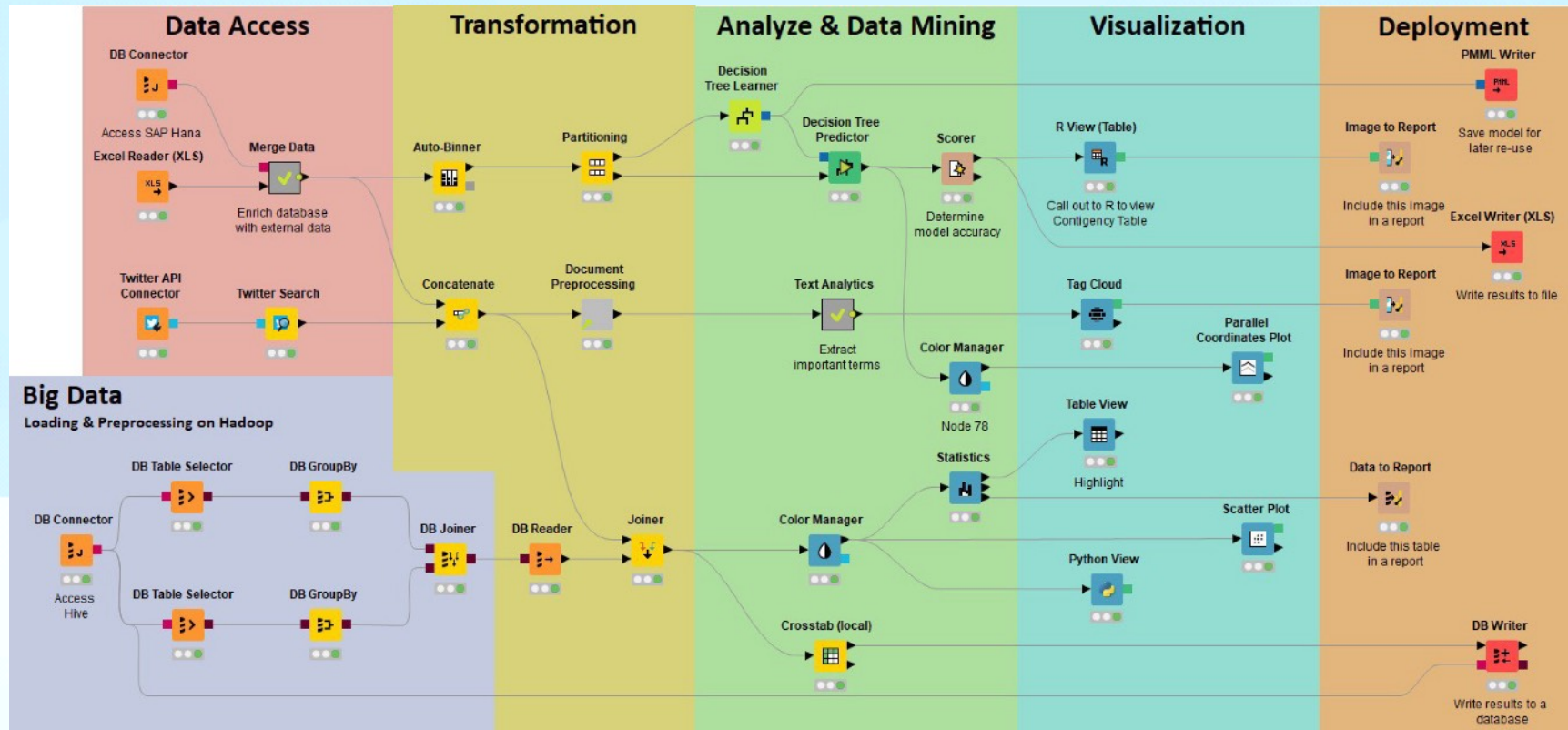
- Interactive Visualizations
- JavaScript-based nodes
  - Scatter Plot, Box Plot, Line Plot
  - Networks, ROC Curve, Decision Tree
  - Plotly Integration
  - Adding more with each release!
- Misc
  - Tag cloud, open street map, molecules
- Script-based visualizations
  - R, Python

# Deployment



- Database
- Files
  - Excel, CSV, txt
  - XML
  - PMML
  - to: local, KNIME Server, Amazon S3, Azure Blob Store
- BIRT Reporting

# Over 2000 Native and Embedded Nodes Included:



## Data Access

MySQL, Oracle, ...  
 SAS, SPSS, ...  
 Excel, Flat, ...  
 Hive, Impala, ...  
 XML, JSON, PMML  
 Text, Doc, Image, ...  
 Web Crawlers  
 Industry Specific  
 Community / 3rd

## Transformation

Row  
 Column  
 Matrix  
 Text, Image  
 Time Series  
 Java Python  
 Community / 3rd

## Analysis & Mining

Statistics  
 Data Mining  
 Machine Learning  
 Web Analytics Text  
 Mining Network  
 Analysis Social  
 Media Analysis  
 R, Weka, Python  
 Community / 3rd

## Visualization

R  
 JFreeChart  
 JavaScript Plotly  
 Community / 3rd

## Deployment

via BIRT  
 PMML  
 XML, JSON  
 Databases Excel,  
 Flat, etc. Text,  
 Doc, Image  
 Industry Specific  
 Community / 3rd

# Install KNIME Analytics Platform

- Select the KNIME version for your computer:
  - Mac
  - Windows – 32 or 64 bit
  - Linux
- Download the archive and extract the file, or download installer package and run it

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

## Windows

KNIME Analytics Platform for Windows (installer) <i>The installer adds an icon to the desktop and suggests suitable memory settings</i>	<a href="#">Download</a> (459 MB)
KNIME Analytics Platform for Windows (self-extracting archive) <i>The self-extracting archive only creates a folder holding the KNIME installation</i>	<a href="#">Download</a> (463 MB)
KNIME Analytics Platform for Windows (zip archive)	<a href="#">Download</a> (547 MB)

## Linux

KNIME Analytics Platform for Linux	<a href="#">Download</a> (583 MB)
------------------------------------	-----------------------------------

## Mac

KNIME Analytics Platform for macOS (10.13 and above)	<a href="#">Download</a> (438 MB)
--	-----------------------------------

Find out what's new in the latest KNIME 4.4 release [here](#).

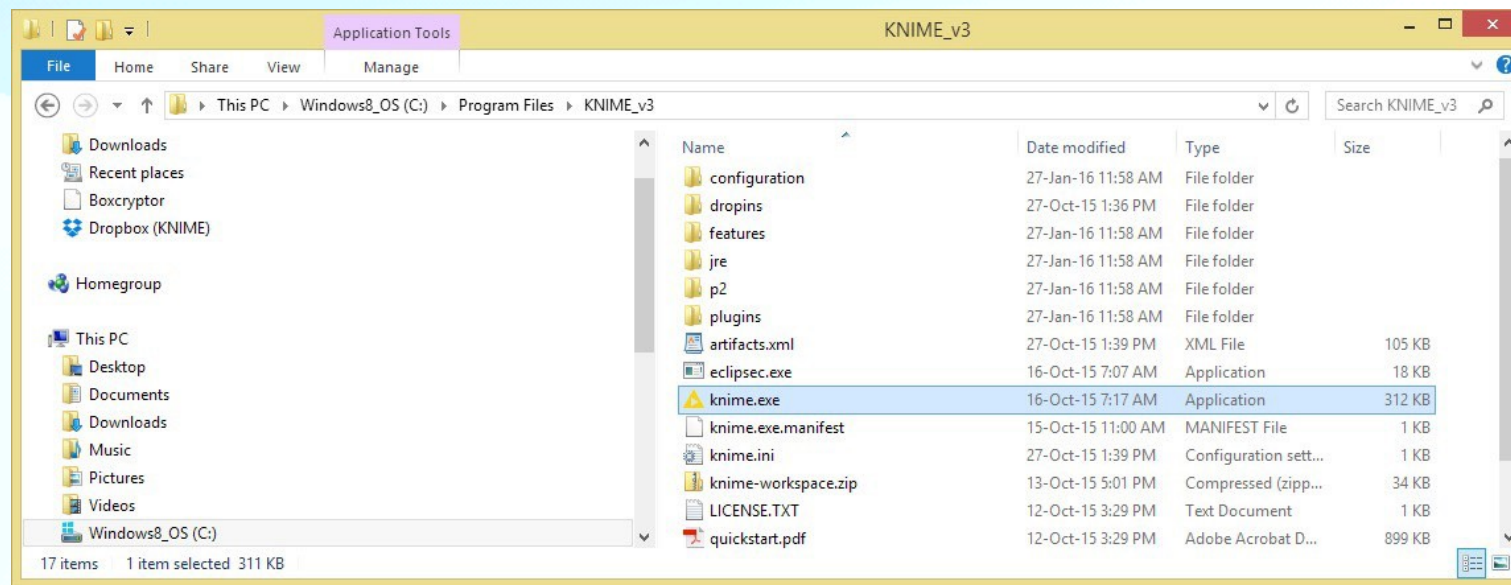
If you are interested in a previous version of KNIME Analytics Platform, please click [here](#).

# Start KNIME Analytics Platform

- Use the shortcut created by the installer

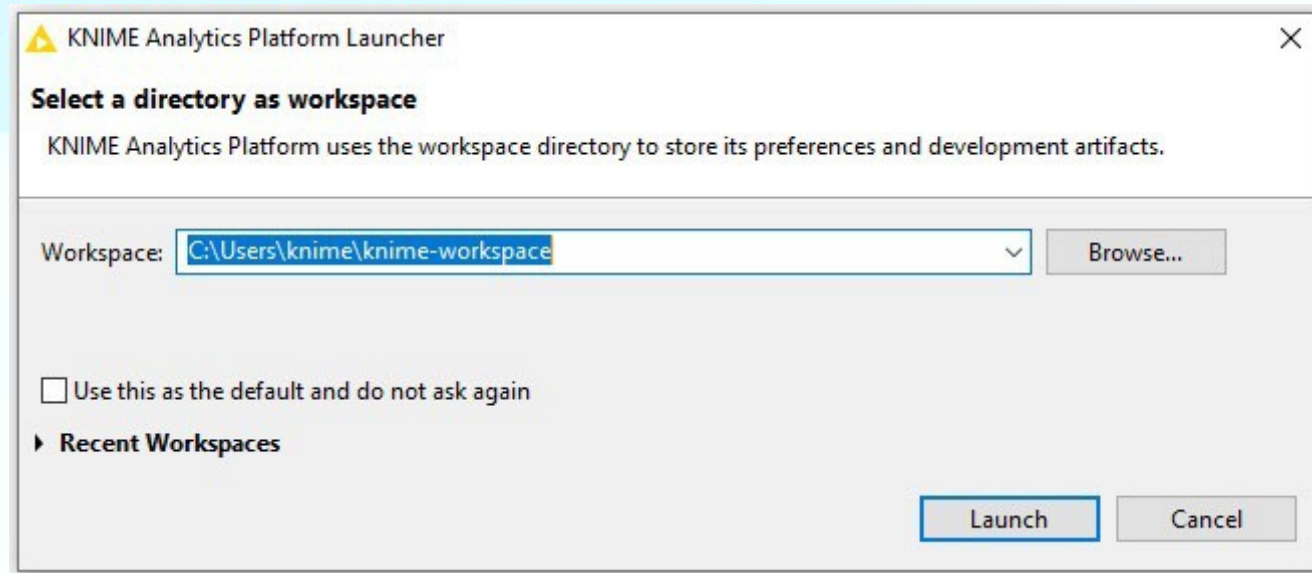


- Or go to the installation directory and launch KNIME via the knime.exe



# The KNIME Workspace

- The workspace is the **folder/directory** in which workflows (and potentially data files) are stored for the current KNIME session.
- Workspaces are portable (just like KNIME)



# The KNIME Analytics Platform Workbench (classic UI)

The screenshot displays the KNIME Analytics Platform Workbench interface. The central area is the **Workflow Editor**, showing a workflow titled "My first Workflow" with four nodes: File Reader (read adult.csv), Row Filter (keep only records born in the US), Column Filter (remove gender), and Table Writer (write table). The **Node Description** panel on the right provides details for the selected Row Filter node. The **Workflow Coach** panel on the left offers recommended nodes, and the **Node Repository** panel at the bottom left lists various node categories. The **Console & Node Monitor** panel at the bottom right shows the execution status of the Row Filter node and its output data.

**KNIME Explorer**

**Workflow Coach**

**Node Repository**

**Workflow Editor**

**Node Description**

**Console & Node Monitor**

**Outline**

**KNIME Hub**

ID	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours
Row0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40
Row1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13
Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40

# The KNIME Analytics Platform Workbench (Modern UI)

**Workflow Description:**

- Data Reading:** CSV Reader (Reading adult.csv)
- Data Partitioning:** Partitioning (Random drawing 80% upper port, 20% lower port)
- Train a Model:** Decision Tree Learner (Train to predict class "income")
- Apply the Model:** Decision Tree Predictor (Apply decision tree model to test set)
- Score the Model:** Scorer (Compute model accuracy)
- Descriptive Statistics:** Statistics (Calculate the statistical properties of the data set attributes)
- Visualize:** Scatter Plot (Create interactive scatter plot)

**Data Table:**

#	Row...	age	workclass	fnlwgt	education	educatio...	marital-s...	occupati...	relations...	race
		Number (inte...)	String	Number (inte...)	String	Number (inte...)	String	String	String	String
1	Row0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White
2	Row1	50	Self-emp-not-i...	83311	Bachelors	13	Married-civ-s...	Exec-manage...	Husband	White
3	Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-clea...	Not-in-family	White
4	Row3	53	Private	234721	11th	7	Married-civ-s...	Handlers-clea...	Husband	Black
5	Row4	28	Private	338409	Bachelors	13	Married-civ-s...	Prof-specialty	Wife	Black
6	Row5	37	Private	284582	Masters	14	Married-civ-s...	Exec-manage...	Wife	White
7	Row6	49	Private	160187	9th	5	Married-spou...	Other-service	Not-in-family	Black

In KNIME 5.0 or later versions

# Switching from Classical UI to Modern UI (in KNIME >5.0)

The screenshot displays the KNIME Analytics Platform interface with a workflow titled "3: Building a Simple Classifier". The workflow consists of the following nodes:

- CSV Reader:** Read the adult data set file. There is one row for each person, plus demographic info and the income group. The file is located in TheData/Basics/.
- Partitioning:** Create two separate partitions from original data set: training set (80%) and test set (20%). The test set deliberately consists of unseen data that will not be used for training.
- Decision Tree Learner:** This node builds a decision tree. Other Learner nodes train other models. Most Learner nodes output a PMML model (blue square output port).
- Decision Tree Predictor:** Apply the model. Predictor nodes apply a specific model to a data set and append the model predictions.
- Scorer:** Compute a confusion matrix between real and predicted class values and calculate the related accuracy measures.

The interface also features a Node Repository on the left, a Workflow Coach, and a Console window at the bottom right showing the following output:

```
KNIME Console
*****
*** Welcome to KNIME Analytics Platform v5.1.2.v202310091347 ***
*** Copyright by KNIME AG, Zurich, Switzerland ***
*****
Log file is located at: /Users/raghava/knime-workspace-50/.metadata/knime/knime.log
WARN Excel Reader 5:266 The node configuration changed and the table spec will be re
WARN Excel Reader 5:266 The node configuration changed and the table spec will be re
```

# Switching from Modern UI to Classic UI (in KNIME >5.0)

Click on icon i

KNIME Analytics Platform - /Users/raghava/knime-workspace-50

Home Building a Simple Classifier

Execute Cancel Reset Create metanode Create component

Repository

Search Nodes

Excel Reader Excel Writer Microsoft Authenticator

Google Authenticator Google Sheets Reader Google Sheets Writer

CSV Reader CSV Writer Show all

Manipulation

Row Filter Column Filter Concatenate

Value Lookup Row Aggregator Table Splitter

Table Cropper Cell Extractor Show all

Single Model Training for Classification

This workflow demonstrates how a simple classifier is built and applied to new data. Find more information on KNIME's Learning page at <http://www.knime.com/learning> (courses, tutorials, cheatsheets, books, and more)

Task

Predict the income group from demographic attributes of the adult data set (census data)

Data Reading

Read the adult data set file. There is one row for each person, plus demographic info, and the income group. The file is located in TheData/Basics/.

Data Partitioning

Create two separate partitions from original data set: training set (80%) and test set (20%). The test set deliberately consists of unseen data that will not be used for training.

Train a Model

This node builds a decision tree. Other Learner nodes train other models. Most Learner nodes output a PMML model (blue square output port).

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 and calculate the related accuracy measures.

1: File Table Flow Variables

Rows: 32561 | Columns: 15

#	Row...	age	workclass	fnlwgt	education	educatio...	marital-s...
		Number (inte...	String	Number (inte...	String	Number (inte...	String
1	Row0	39	State-gov	77516	Bachelors	13	Never-married
2	Row1	50	Self-emp-not-i...	83311	Bachelors	13	Married-civ-s...
3	Row2	38	Private	215646	HS-grad	9	Divorced
4	Row3	53	Private	234721	11th	7	Married-civ-s...
5	Row4	28	Private	338409	Bachelors	13	Married-civ-s...
6	Row5	37	Private	284582	Masters	14	Married-civ-s...
7	Row6	49	Private	160187	9th	5	Married-snou...

KNIME Analytics Platform - /Users/raghava/knime-workspace-50

Home Building a Simple Classifier

Install Extensions

Install Extensions to access additional functionality such as the ability to process complex data types, as well as to use advanced algorithms.

Install Extensions Check for updates

Switch to classic user interface

Switch to the classic KNIME Analytics Platform user interface. To switch back again, click the button "Open KNIME Modern UI" in the top right corner of the classic user interface.

Switch to KNIME classic user interface

Open for Innovation

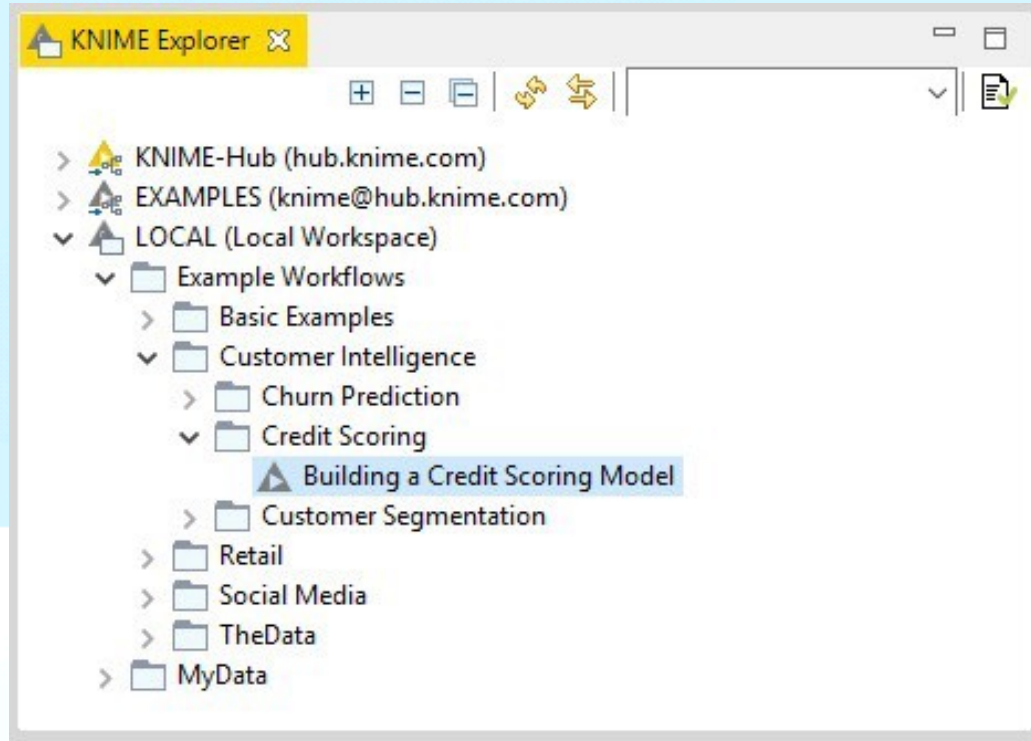
KNIME

Copyright by KNIME AG, Zurich, Switzerland  
contact@knime.com  
<https://www.knime.com>

This software is a bundle of multiple modules, each released under its own license. Please check the individual licenses by clicking "About KNIME" and the "Credits" button.

About KNIME Credits

# KNIME Explorer



- In LOCAL you can access your own workflow projects.

- Other mountpoints allow you to connect to

- EXAMPLE Server
- KNIME Hub
- KNIME Server

- The Explorer toolbar on the top has a search box and buttons to

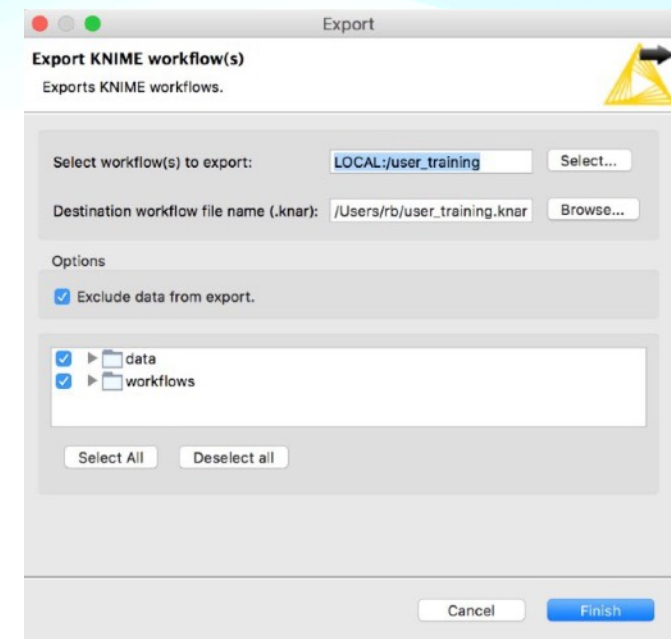
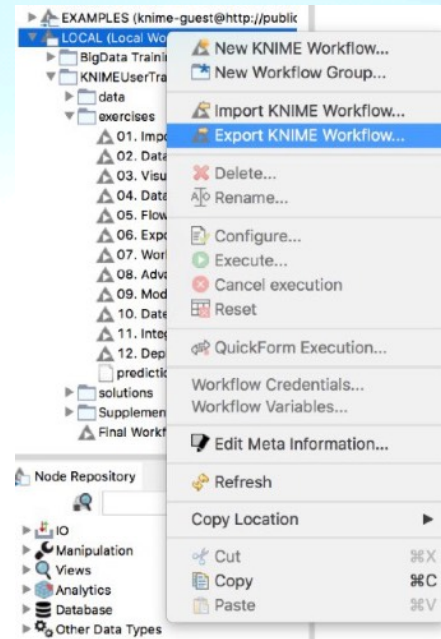
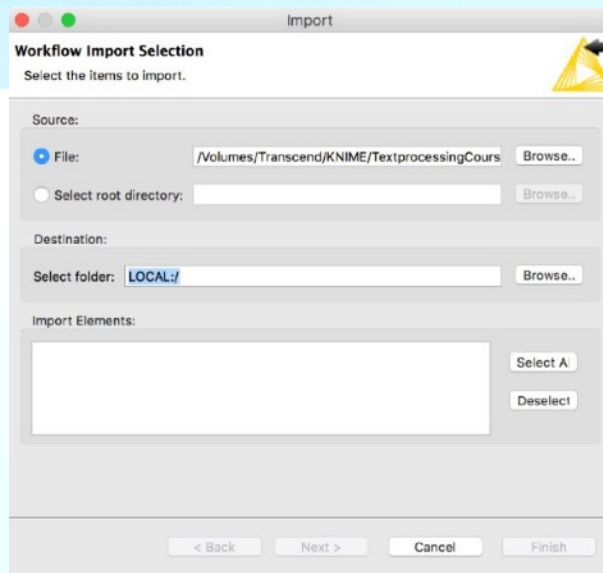
- § select the workflow displayed in the active editor
- § refresh the view

- The KNIME Explorer can contain 4 types of content:

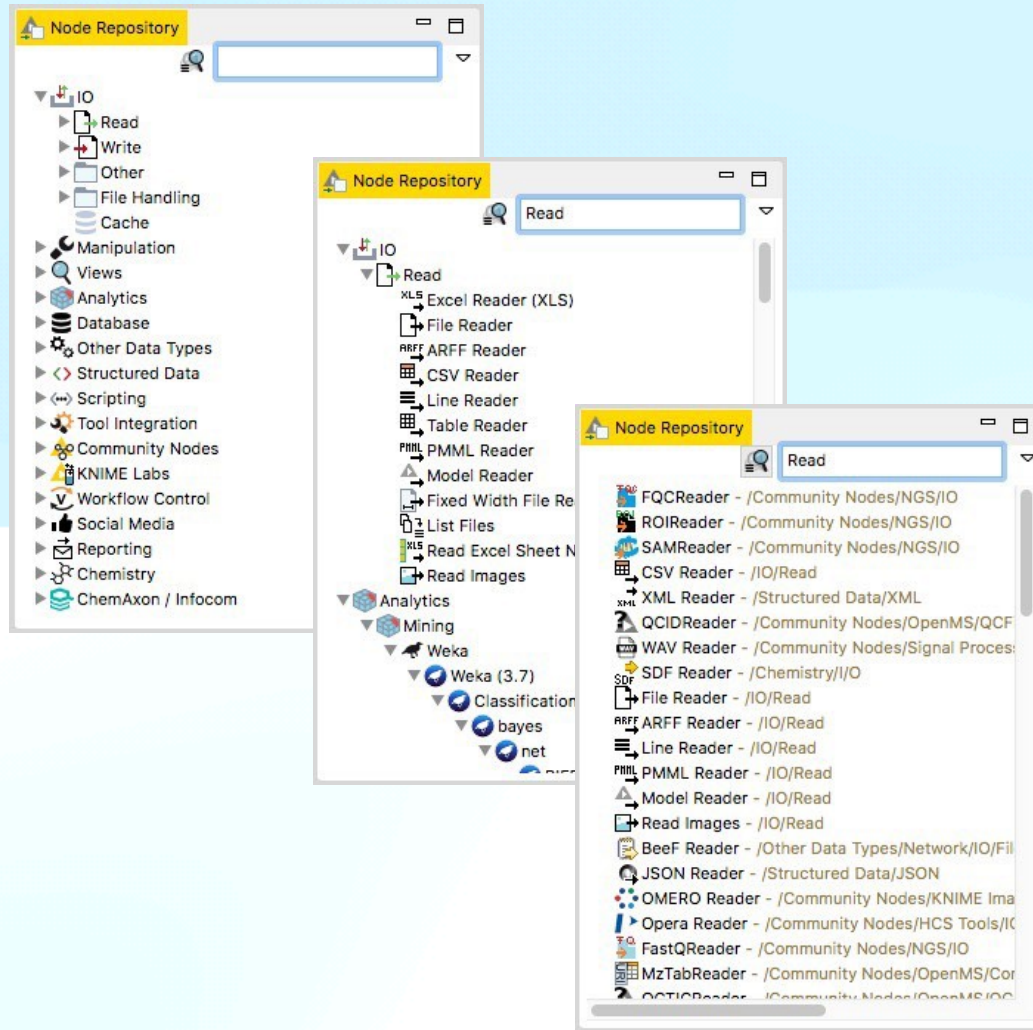
- Workflows
- Workflow groups
- Data files
- Shared Components



# Creating New Workflows, Importing and Exporting

- Right-click inside the KNIME Explorer to create a new workflow or a workflow group, or to import a workflow
- Right-click the workflow or workflow group to export

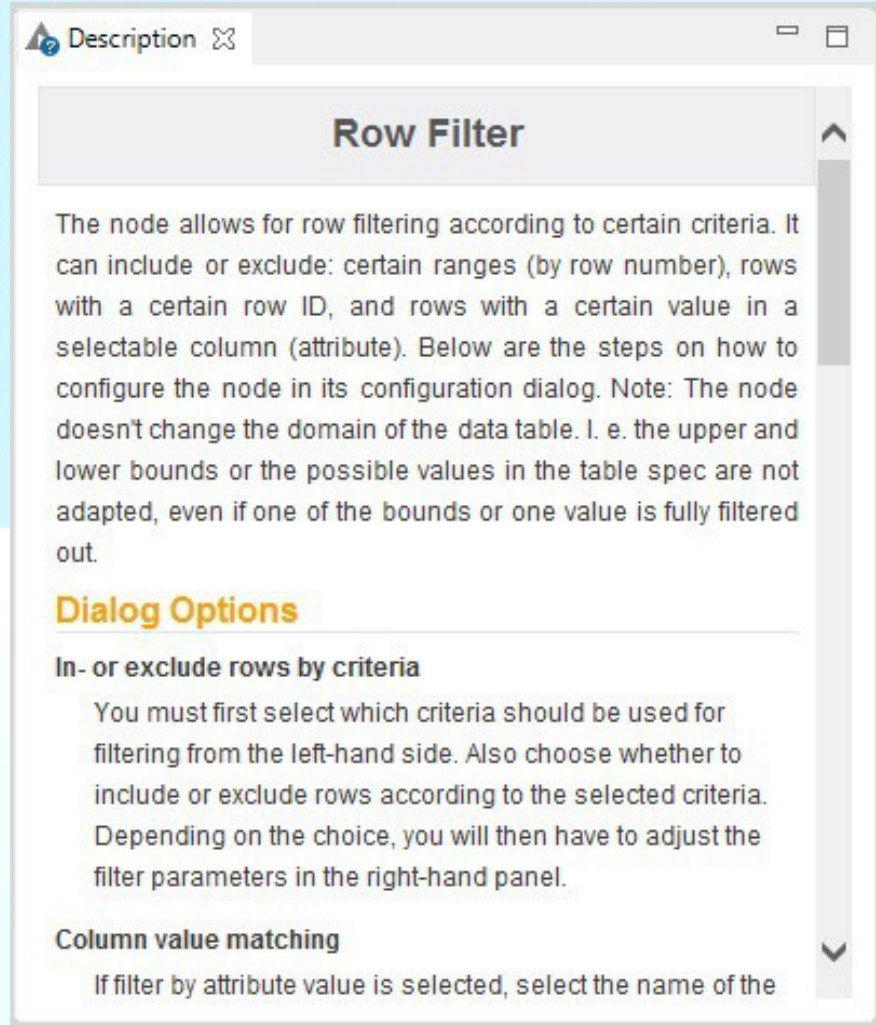


# Node Repository



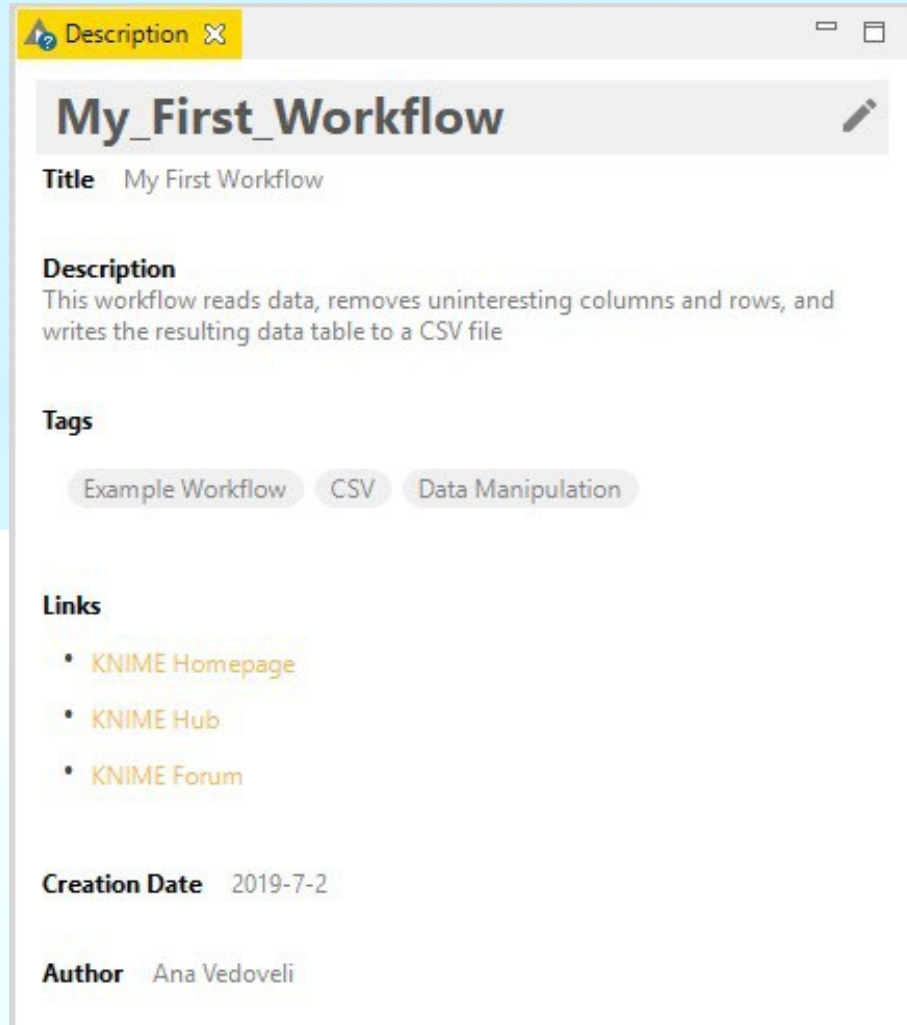
- The Node Repository lists all KNIME nodes
- The search box has 2 modes
  -  Standard Search – exact match of node name
  -  Fuzzy Search – finds the most similar node name

# Description



- The Description window gives information about:
  - Node Functionality
  - Input & Output
  - Node Settings
  - Ports
  - References to literature

# Workflow Description



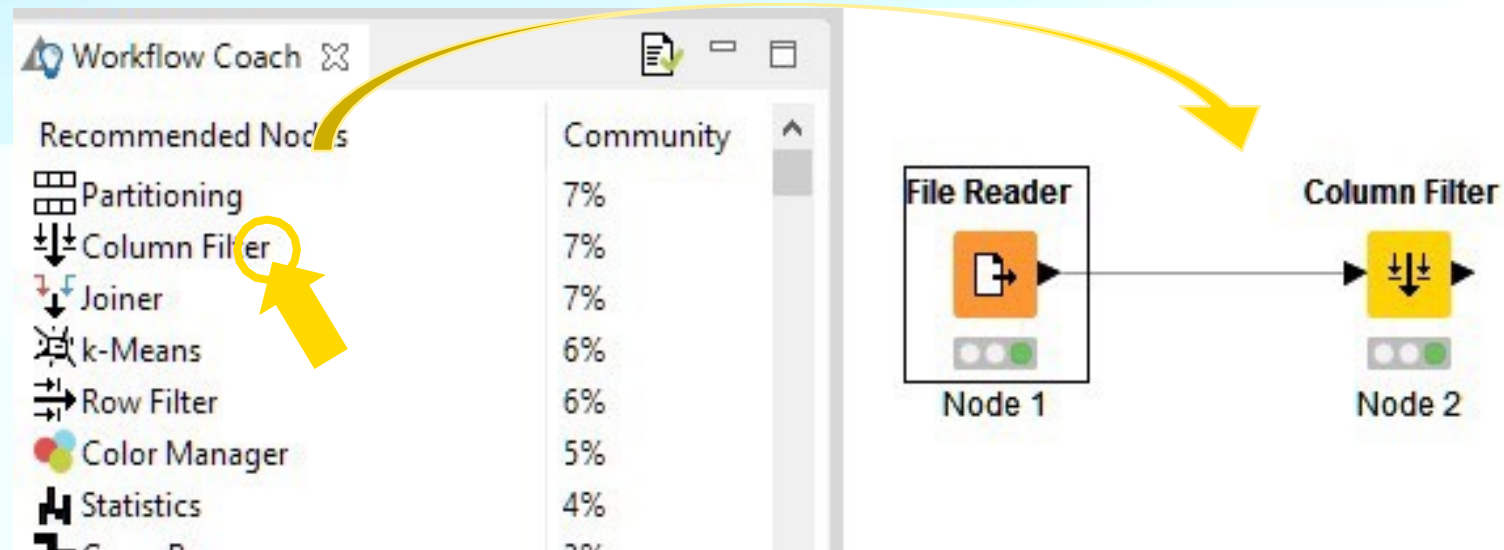
The screenshot shows a window titled "Description" with a close button. The main content area displays the following information:

- Title:** My First Workflow
- Description:** This workflow reads data, removes uninteresting columns and rows, and writes the resulting data table to a CSV file
- Tags:** Example Workflow, CSV, Data Manipulation
- Links:**
  - [KNIME Homepage](#)
  - [KNIME Hub](#)
  - [KNIME Forum](#)
- Creation Date:** 2019-7-2
- Author:** Ana Vedoveli

- When selecting the workflow, the Description window gives information about the workflow's:
  - Title
  - Description
  - Associated Tags and Links
  - Creation Date
  - Author

# Workflow Coach

- Node recommendation engine
  - Gives hints about which node use next in the workflow
  - Based on KNIME communities' usage statistics
  - Based on own KNIME workflows



# Node Monitor

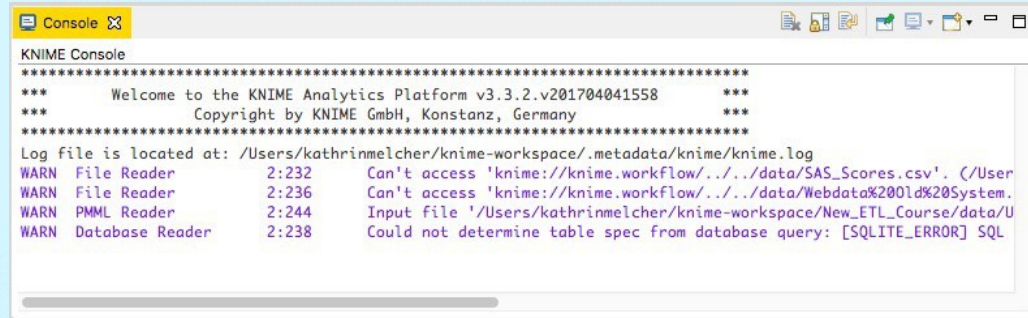
- By default the Node Monitor shows you the output table of the node selected in the workflow editor
- Click on the three dots on the upper right to show the flow variables, configuration, etc.

Node: Get Customers from Database (0:1207)  
State: EXECUTED

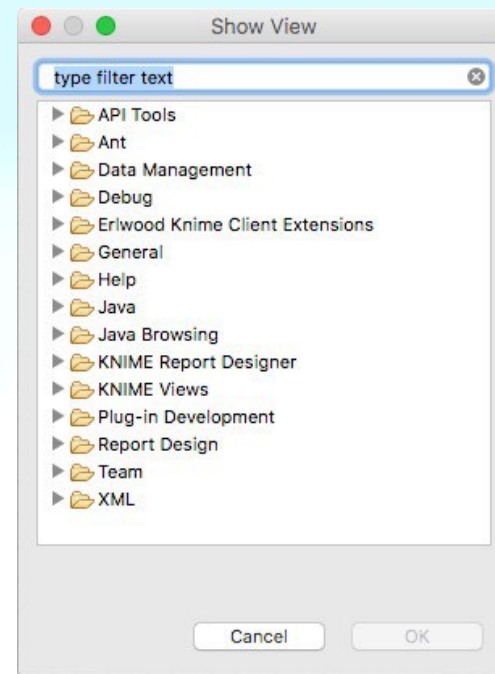
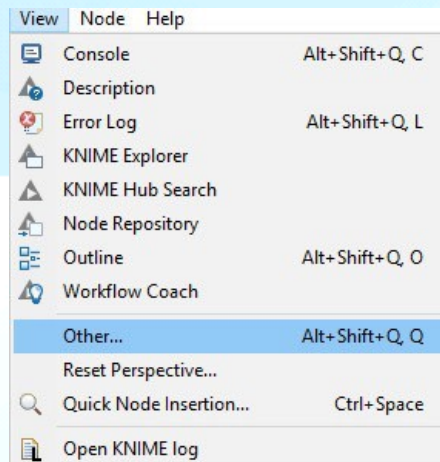
Port Output: Port 0 [Load data]

ID	MaritalStatus	Gender	EstimatedYearlyIncome	NumberOfContracts	Age	Available401K	CustomerV	Products
CustomerID: 722204	S	F	80000	4	42	1	1	4 5 Private Investr
CustomerID: 489847	M	M	60000	2	46	1	1	4 3 Private Investr
CustomerID: 8444723	M	M	40000	1	32	1	2	3 0 P+B Investmer
CustomerID: 1487427	M	M	30000	2	63	1	1	2 2 P+B Investmer
CustomerID: 4693433	M	M	20000	2	63	1	1	3 4 Gold Investme
CustomerID: 7724940	M	M	30000	2	33	1	2	3 0 P+B Investmer
CustomerID: 9784443	M	M	60000	2	34	1	2	3 0 P+B Investmer
CustomerID: 3177757	M	M	70000	2	57	1	1	5 2 Fund Manager

# Console and Other Views



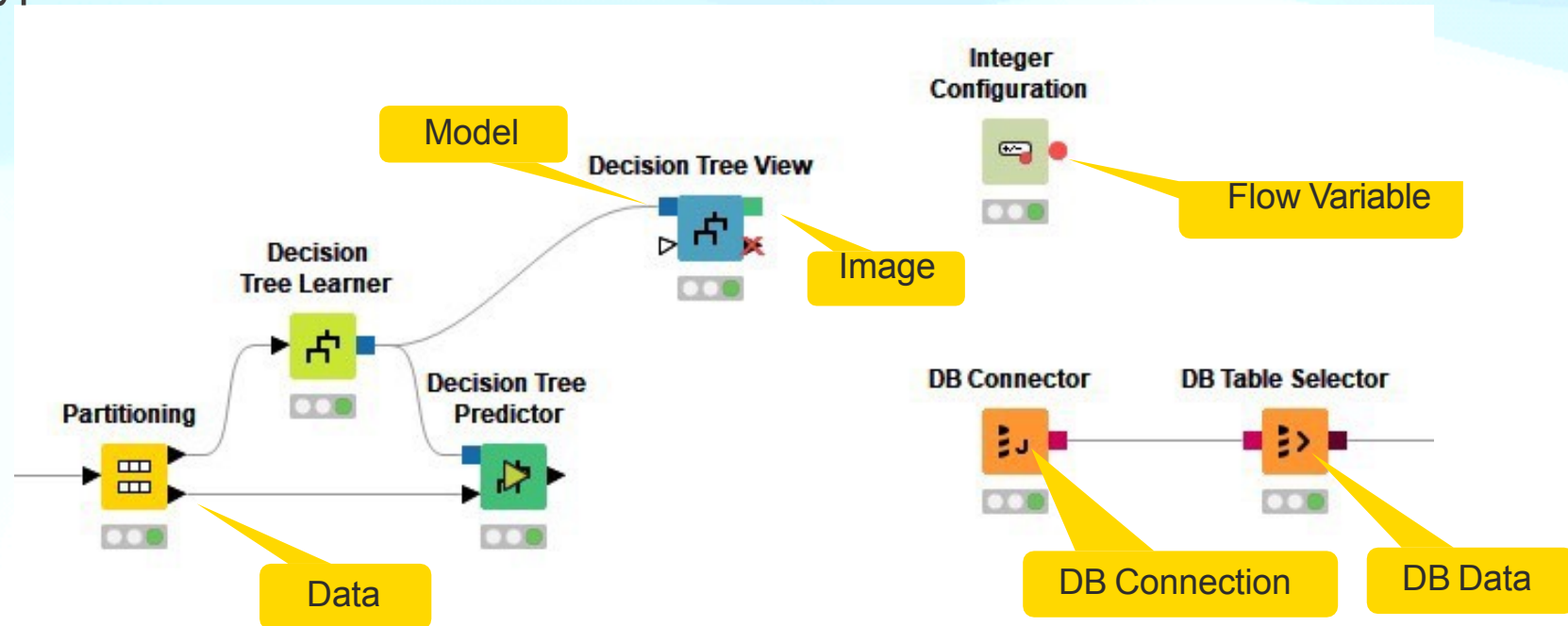
```
KNIME Console
*****
*** Welcome to the KNIME Analytics Platform v3.3.2.v201704041558 ***
*** Copyright by KNIME GmbH, Konstanz, Germany ***
*****
Log file is located at: /Users/kathrinmelcher/knime-workspace/.metadata/knime/knime.log
WARN File Reader 2:232 Can't access 'knime://knime.workflow/../../data/SAS_Scores.csv'. /User
WARN File Reader 2:236 Can't access 'knime://knime.workflow/../../data/Webdata%201d%20System.
WARN PMML Reader 2:244 Input file '/Users/kathrinmelcher/knime-workspace/New_ETL_Course/data/U
WARN Database Reader 2:238 Could not determine table spec from database query: [SQLITE_ERROR] SQL
```



- Console view prints out error and warning messages about what is going on under the hood
- Click on View and select Other... to add different views
  - Node Monitor, Licenses, etc.

# Inserting and Connecting Nodes

- Insert nodes into workspace by dragging them from Node Repository or by double-clicking in Node Repository
- Connect nodes by left-clicking output port of Node A and dragging the cursor to (matching) input port of Node B
- Common port types:



# More on Nodes...

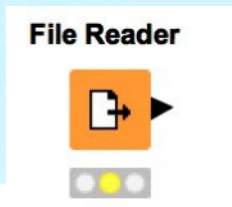
---

- A node can have 4 states:



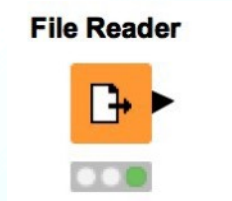
## Not Configured:

The node is waiting for configuration or incoming data.



## Configured:

The node has been configured correctly and can be executed.



## Executed:

The node has been successfully executed. Results may be viewed and used in downstream nodes.

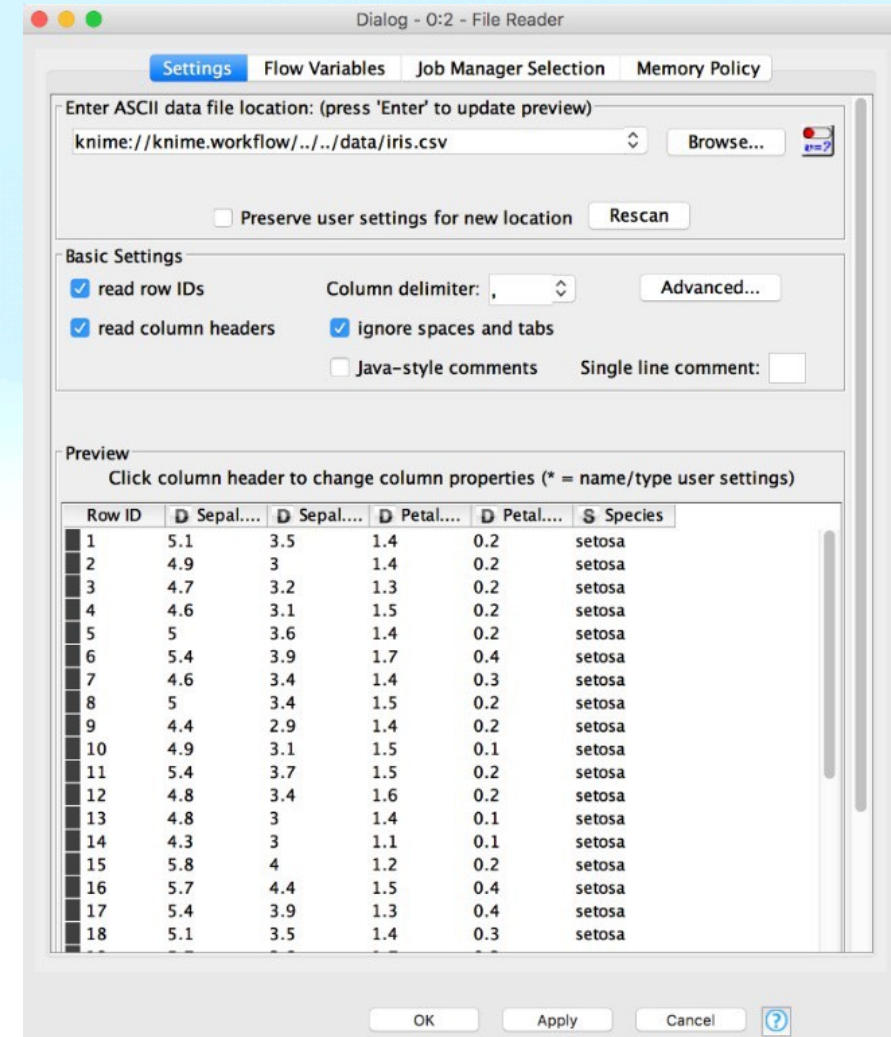


## Error:

The node has encountered an error during execution.

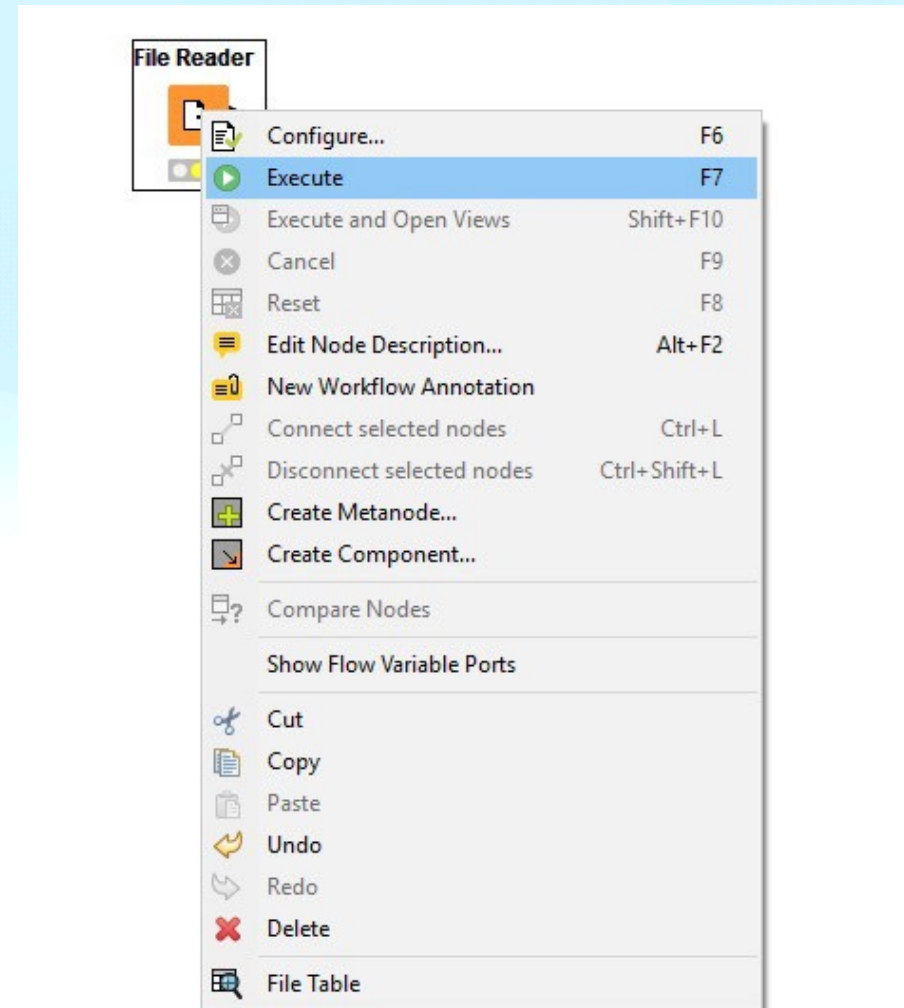
# Node Configuration

- Most nodes require configuration
- To access a node configuration window:
  - Double-click the node
  - Right-click -> Configure

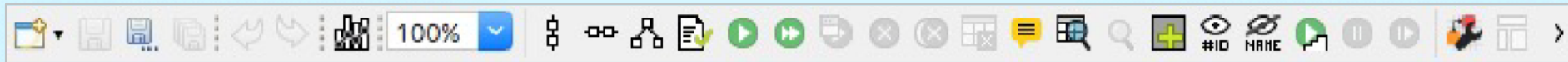


# Node Execution







- Right-click node
- Select Execute in the context menu
- If execution is successful, status shows green light
- If execution encounters errors, status shows red light



# Tool Bar

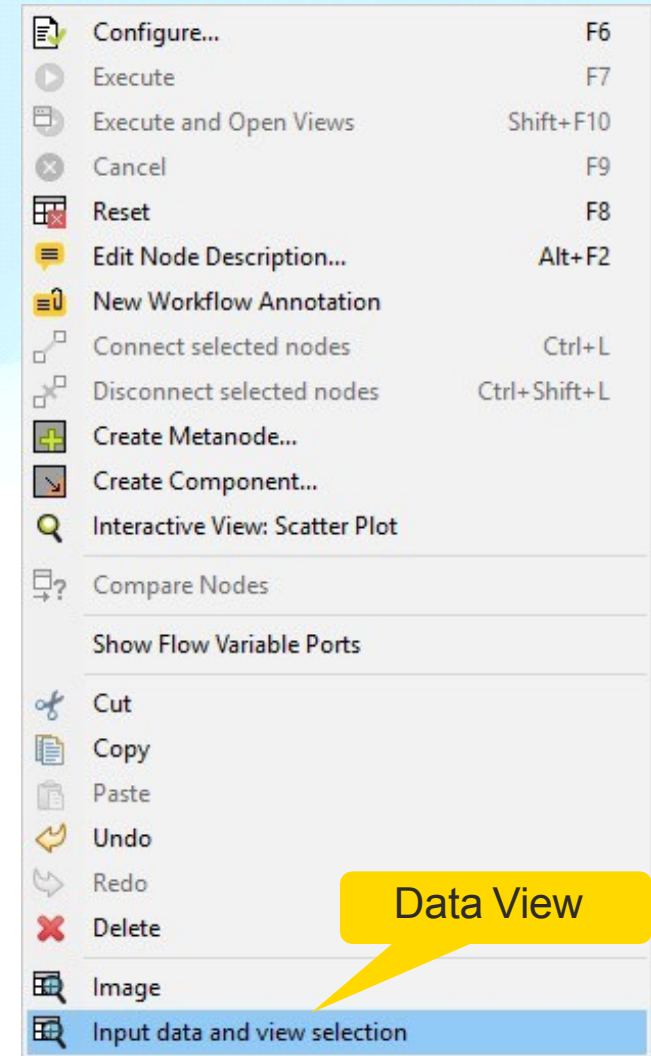
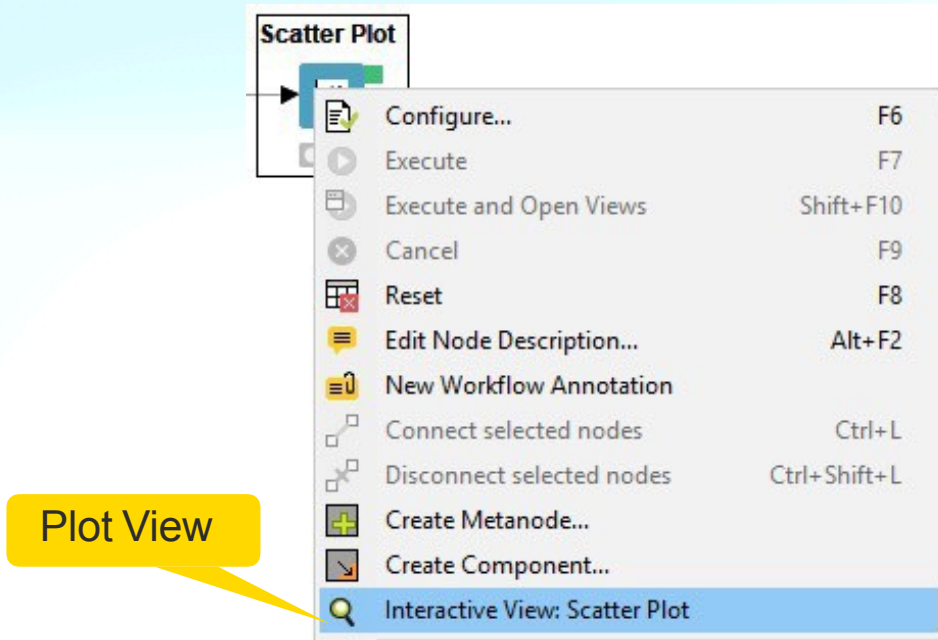
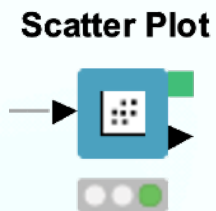


The buttons in the toolbar can be used for the active workflow. The most important buttons are:

-  Execute selected and executable nodes (F7)
-  Execute all executable nodes
-  Execute selected nodes and open first view
-  Cancel all selected, running nodes (F9)
-  Cancel all running nodes
- 

# Node Views

- Right-click node to inspect the execution results by
  - selecting output ports (last option in the context menu) to inspect tables, images, etc.
  - selecting Interactive View to open visualization results in a browser



# KNIME File Extensions

---

Dedicated file extensions for workflows and workflow groups associated with KNIME Analytics Platform

- **\*.knwf** for KNIME Workflow Files

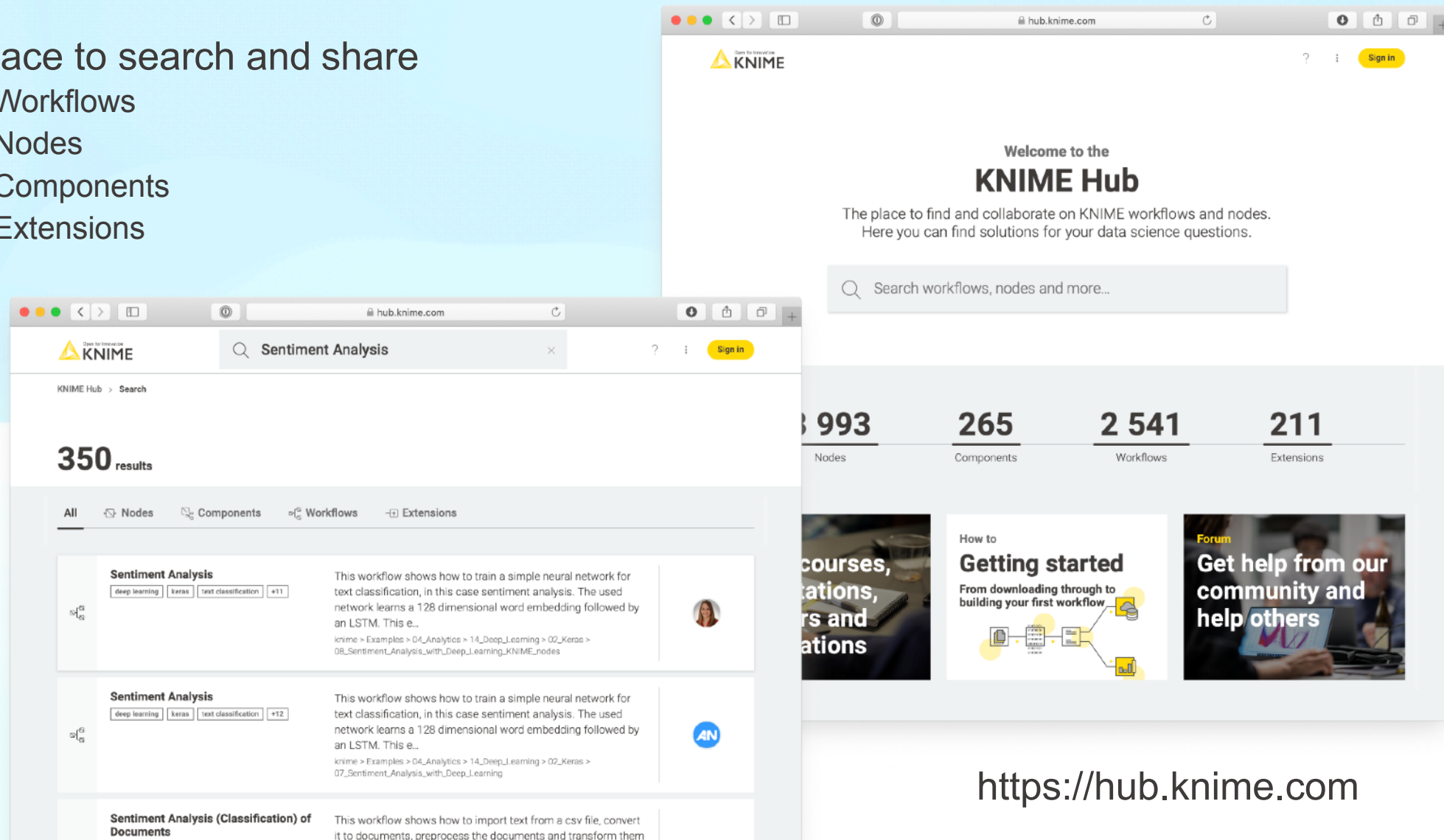


- **\*.knar** for KNIME Archive Files



# Getting Started: KNIME Hub

- Place to search and share
  - Workflows
  - Nodes
  - Components
  - Extensions



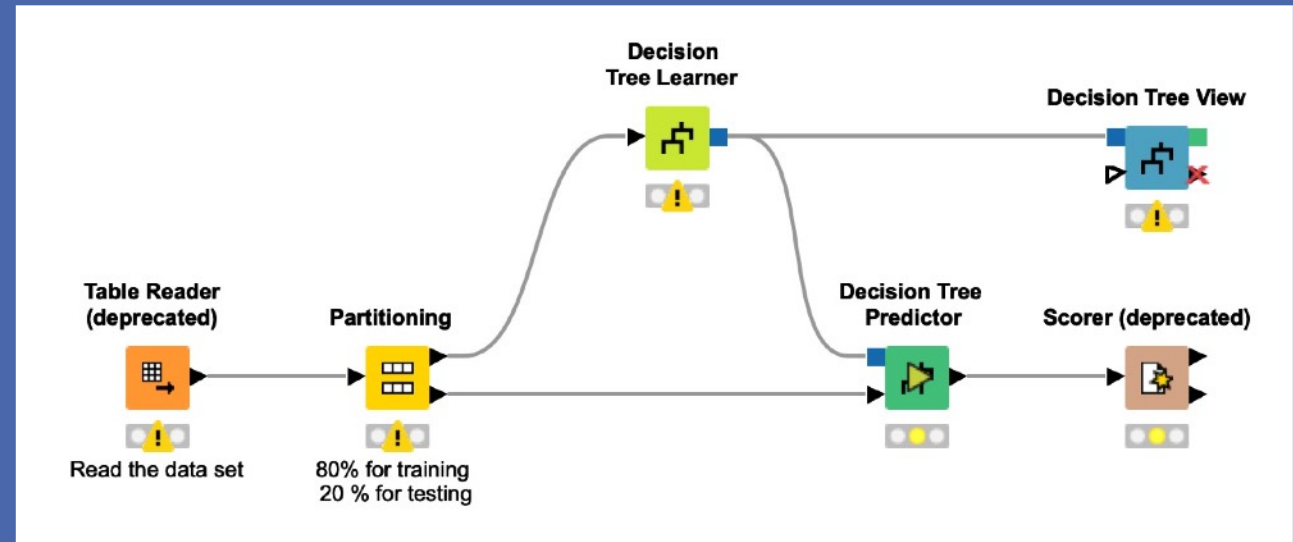
<https://hub.knime.com>

# Getting Started: KNIME Example Server

- Connect via KNIME Explorer to a public repository with large selection of example workflows for many, many applications

The screenshot displays the KNIME Explorer interface. On the left, a file tree shows a hierarchy of folders: My-KNIME-Hub (hub.knime.com) > EXAMPLES (knime@hub.knime.com) > 02\_ETL\_Data\_Manipulation > 00\_Basic\_Examples > 01\_Example\_for\_Standard\_Preprocessing. The main window shows a workflow diagram titled "Simple Preprocessing Example" with the description "Shows the use of different filter nodes." The workflow consists of the following nodes: File Reader, Row Filter (keep rows where "native-country" is missing), Column Filter, Reference Column Filter, Numeric Binner (Replace numeric column age with an attribute column), Nominal Value Row Filter (Filter by this new attribute), Reference Row Filter (Exclude those rows from the original table), Row Filter (filter example with a regular expression), and Concatenate. A yellow box highlights the "Simple Preprocessing Example" title and description. A yellow banner at the top of the workflow window reads: "This is a temporary copy of 'knime://EXAMPLES/Users/knime/Examples/02\_ETL\_Data\_Manipulation/00\_Basic\_Examples/01\_Example\_for\_Standard\_Preprocessing.knime' Use 'Save As...' to save a permanent copy of the workflow to your local workspace, or a mounted KNIME Server." On the right, a smaller window shows the file tree with a tooltip that says "Double-click to see the examples".

# USE CASE I: CREDIT RISK ANALYSIS USING DECISION TREES



# Credit Risk Analysis

- One of the leading banks would like to predict bad customers when customers are applying for loans.
- Credit scoring is one of the predictive modeling applications to predict whether giving credit to an applicant will likely result in profit or loss.
- Many variations and complexities regarding how exactly credit is extended to individuals, businesses, and other organizations for various purposes (e.g., real estate/consumer items).

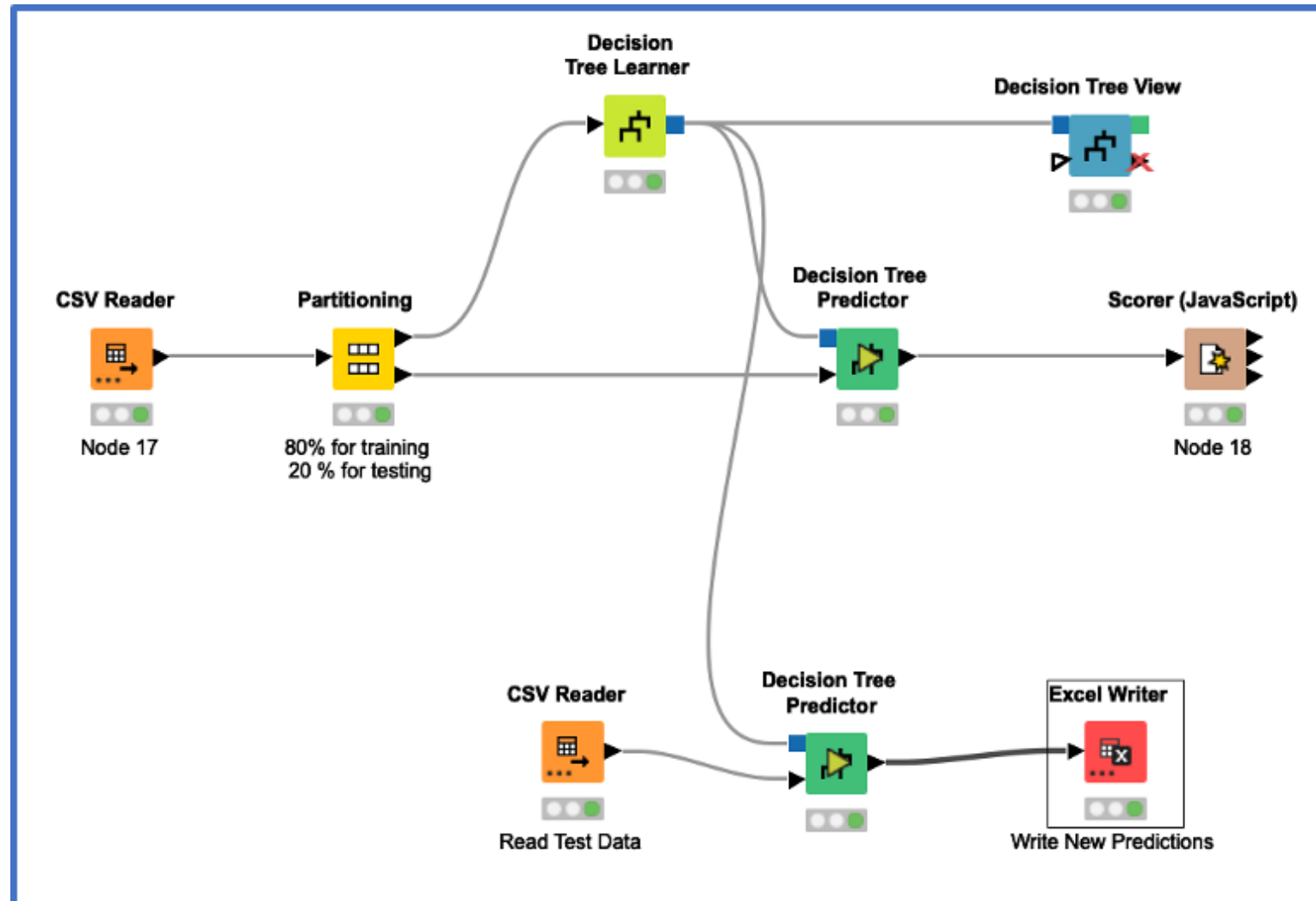
# Credit Risk Analysis

- A lender commonly makes two types of decisions: first, whether to grant credit to a new applicant and second, whether to increase credit limits for existing customers.
- In both cases, we look into a large sample of previous customers with their application details, behavioral patterns, and subsequent credit history available.
- Most of the techniques use this sample to identify the connection between the characteristics of the consumers (annual income, age, number of years in employment with their current employer, etc.) and their subsequent history.
- Typical applications in the consumer market include credit cards, auto loans, home mortgages, mail catalog orders, and a wide variety of personal loan products.

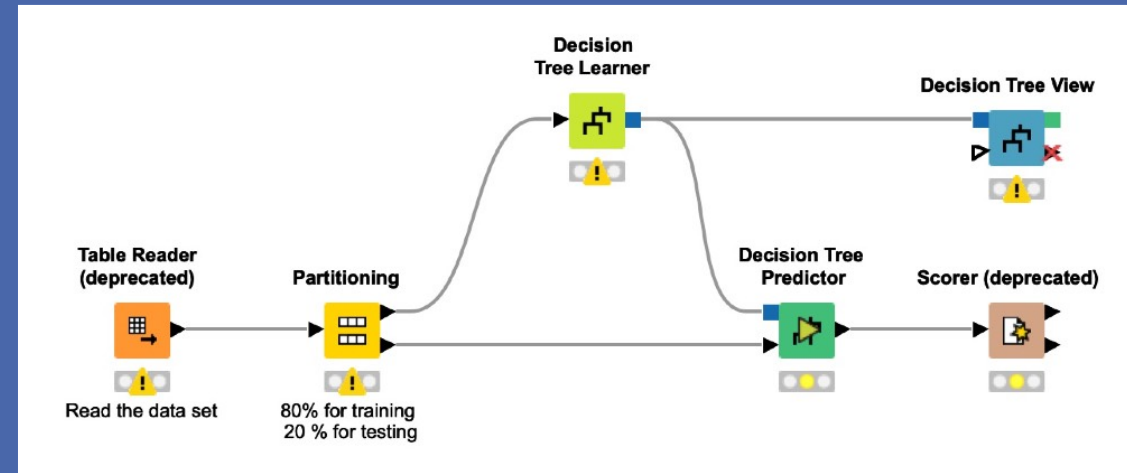
# Dataset

Data Attribute	Explanation
age	Age of Customer
ed	Eductation level of customer
employ	Tenure with current employer (in years)
address	Number of years in same address
income	Customer Income
debtinc	Debt to income ratio
creddebt	Credit to Debt ratio
othdebt	Other debts
Default	Customer defaulted in the past (1= defaulted, 0=Never defaulted) ( <b>dependent variable</b> ). Note that this is a categorical value even though the values are 0 and 1.

# KNIME Workflow for Decision Trees



# USE CASE II: CREDIT RISK ANALYSIS USING LOGISTIC REGRESSION

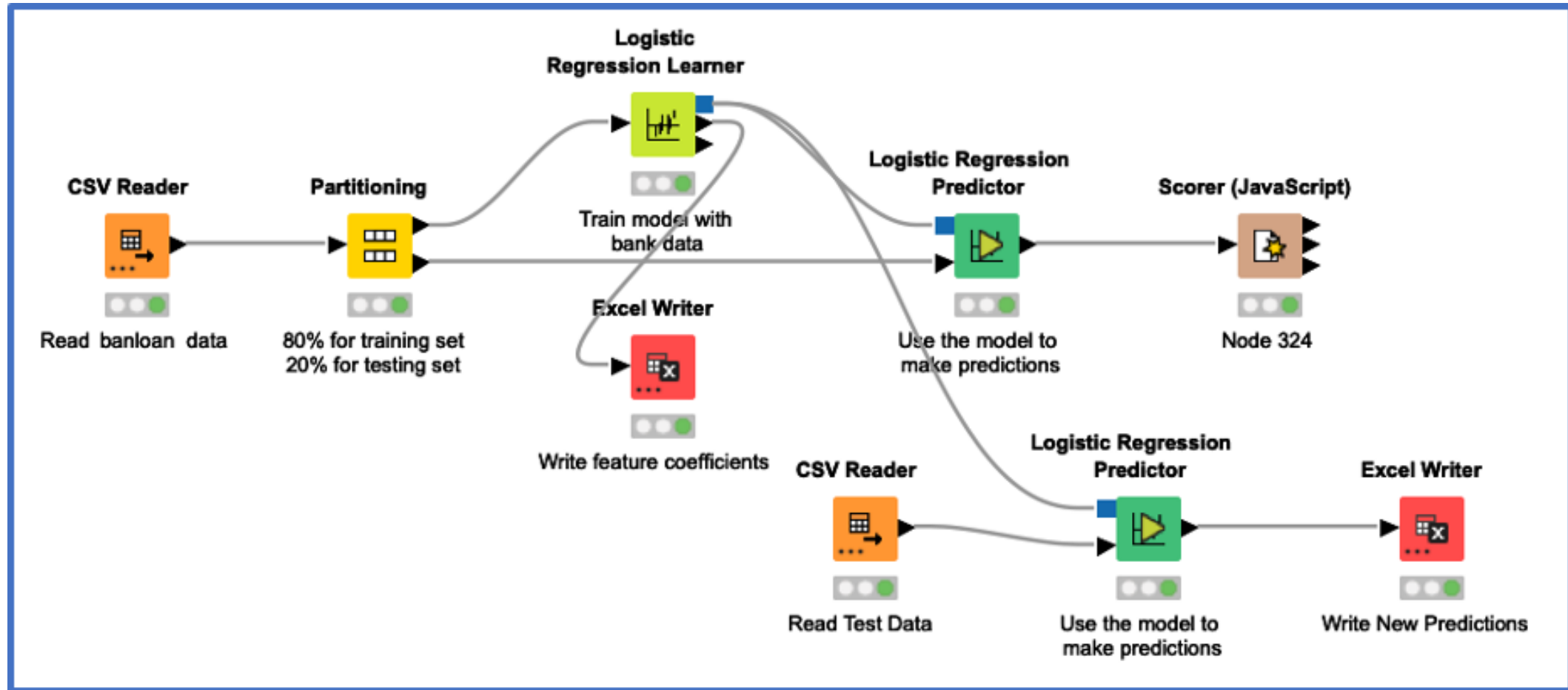


# Credit Risk Analysis

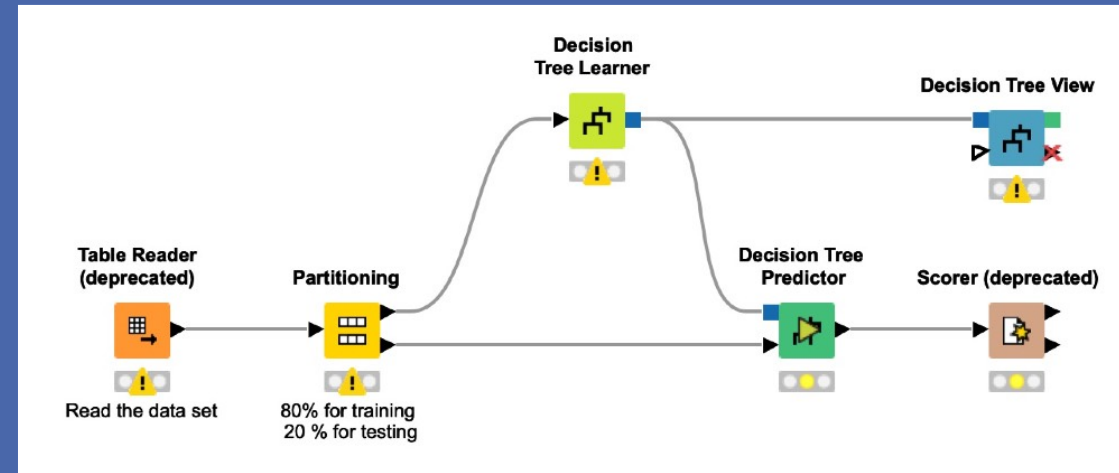
- We use the same dataset as Decision Trees

Data Attribute	Explanation
age	Age of Customer
ed	Education level of customer
employ	Tenure with current employer (in years)
address	Number of years in same address
income	Customer Income
debtinc	Debt to income ratio
creddebt	Credit to Debt ratio
othdebt	Other debts
Default	Customer defaulted in the past (1= defaulted, 0=Never defaulted) ( <b>dependent variable</b> ). Note that this is a categorical value even though the values are 0 and 1.

# KNIME Workflow for Logistics Regression

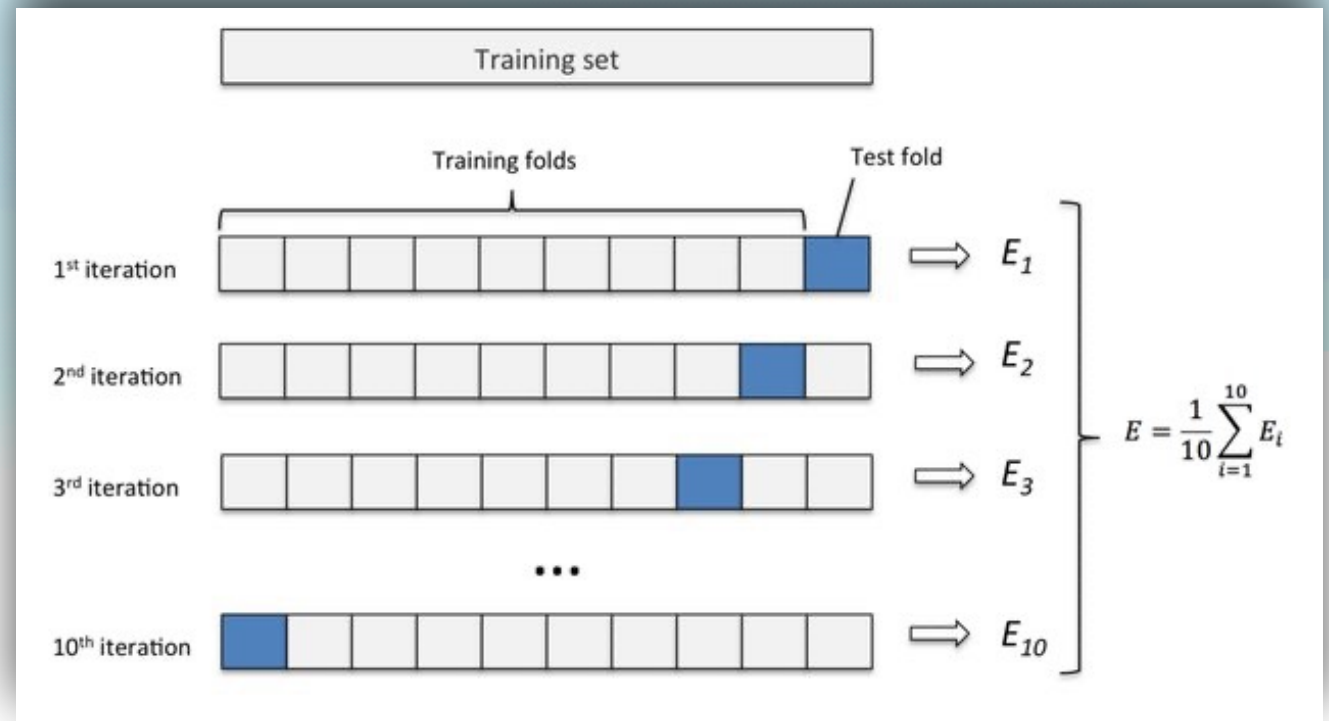


# CROSS VALIDATION IN ML MODELS



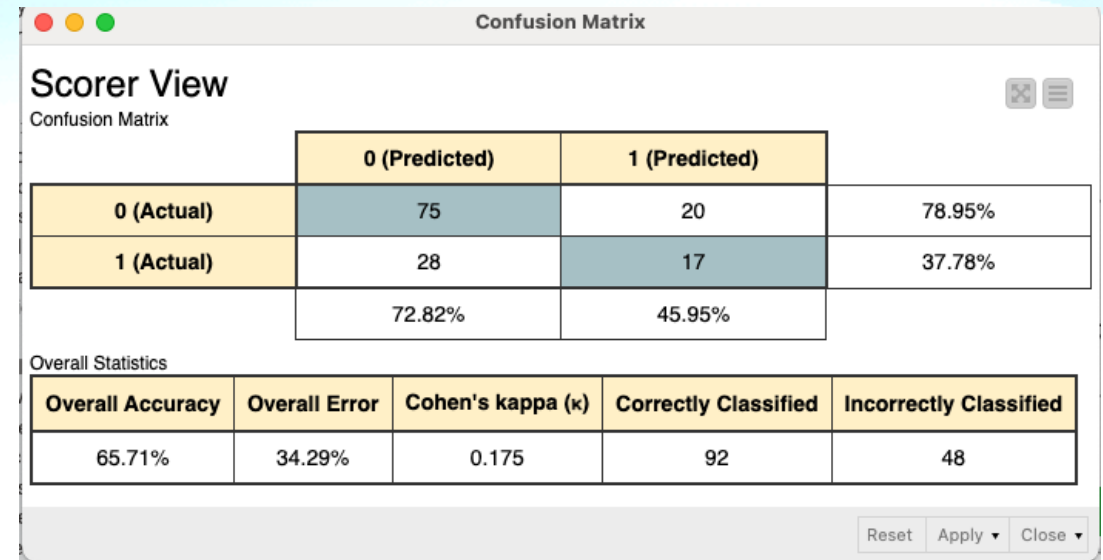
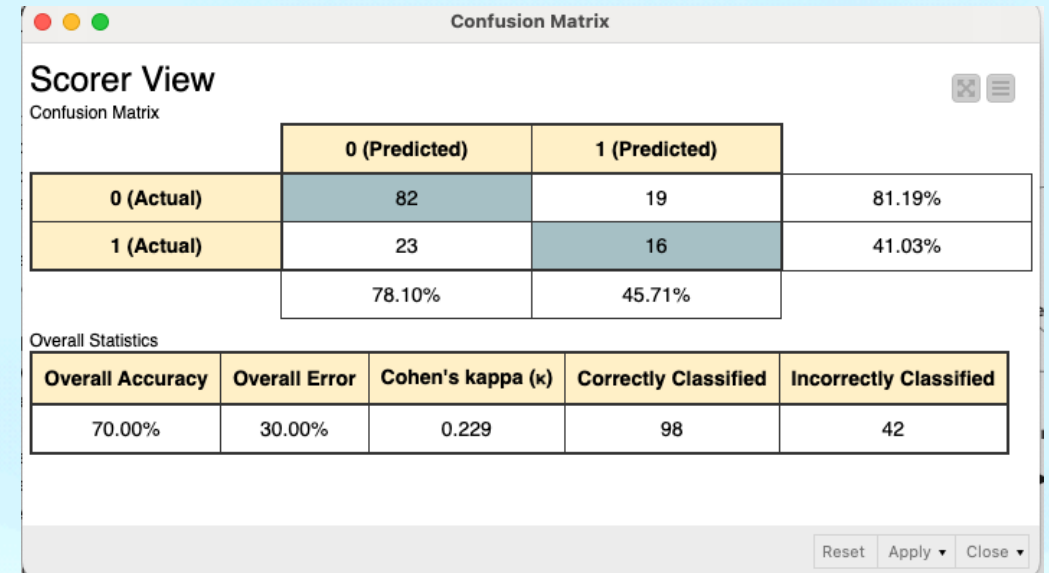
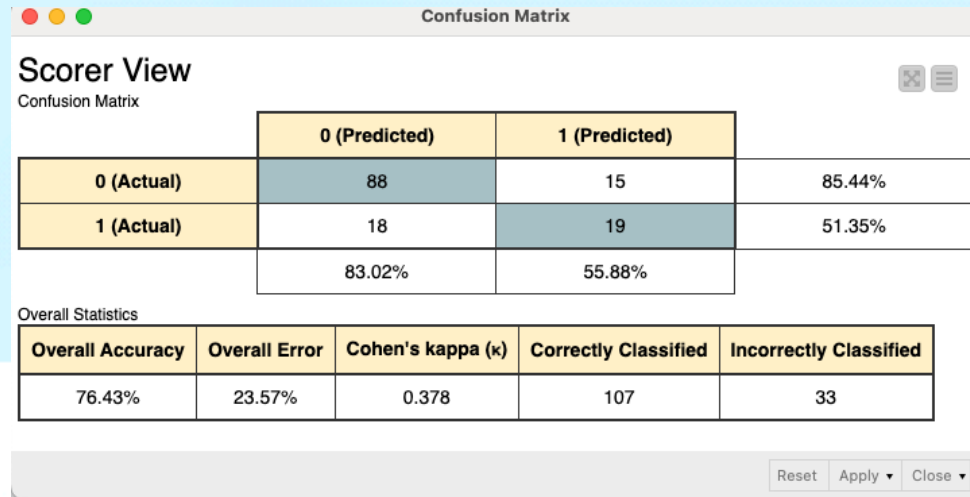
# Cross-validation

- A less biased or less optimistic estimate of the model than the train/test split
- k-folds cross-validation, where k is the number of splits (e.g., 10)



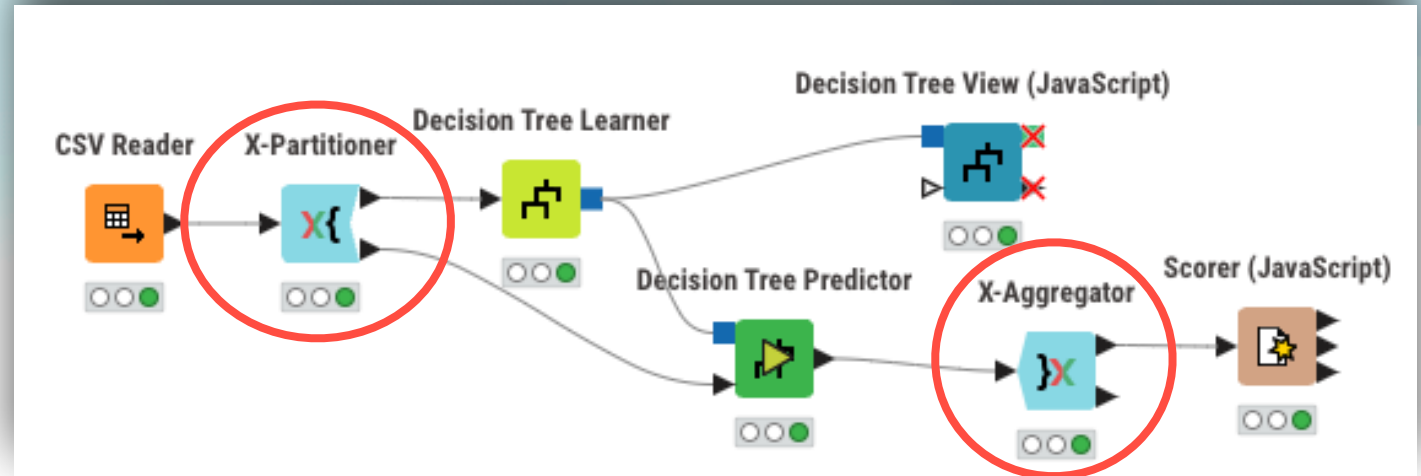
```
KFold(n_splits=10, random_state=None, shuffle=False)
```

# Why Cross-validation?

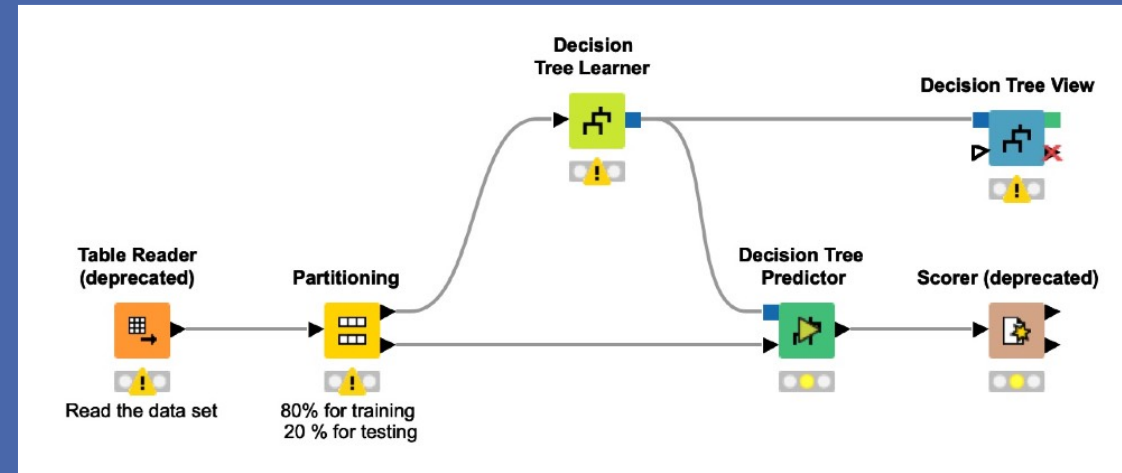


# How to do Cross-validation in KNIME?

- Replace the *Partitioning* node with the *x-Partitioner* node
- Set the number of validations to 10 and choose the stratified sampling
- Add an *x-Aggregator* node before the *Scorer* node



# USE CASE II: CREDIT RISK ANALYSIS USING RANDOM FORESTS

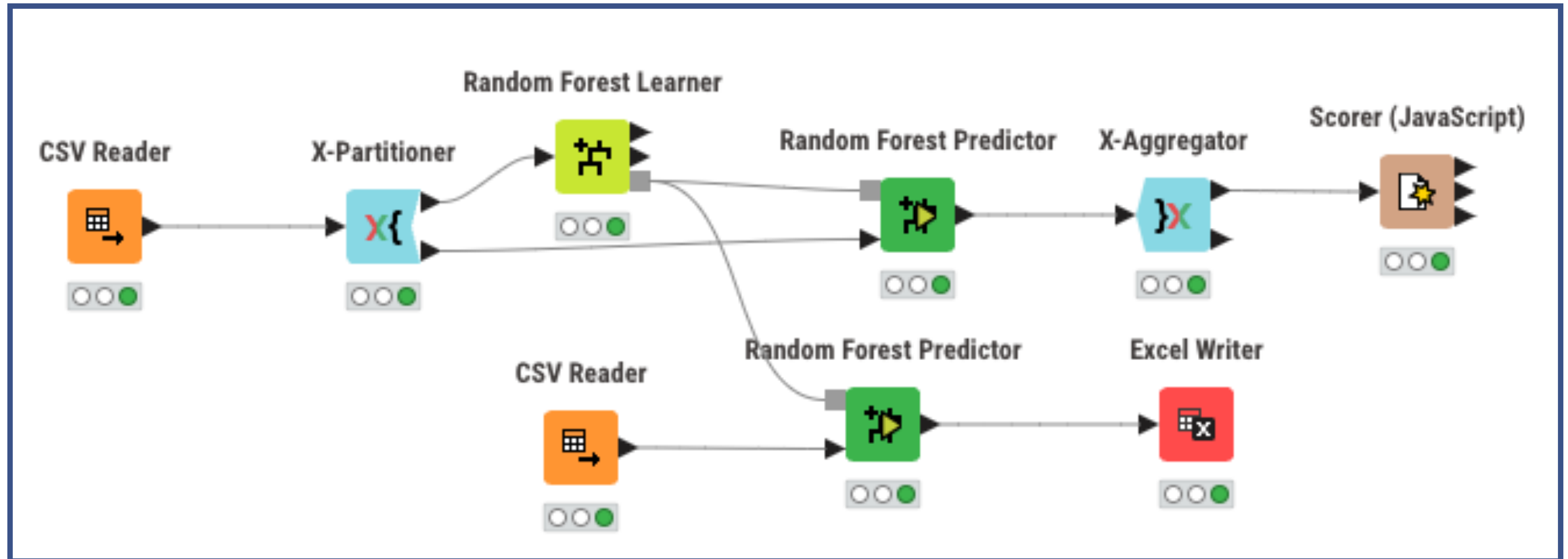


# Credit Risk Analysis

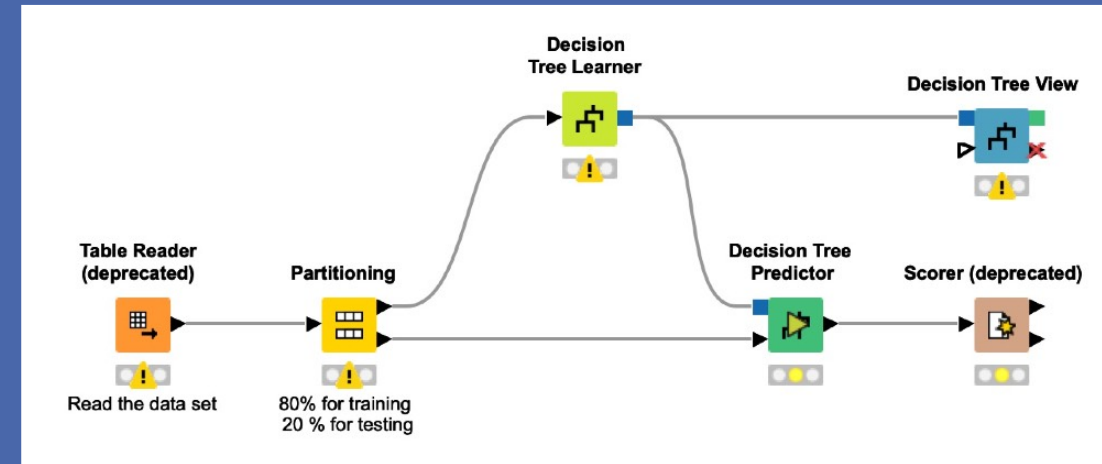
- We use the same dataset as Decision Trees, so predicting whether a customer is going to be defaulter or not

Data Attribute	Explanation
age	Age of Customer
ed	Eductation level of customer
employ	Tenure with current employer (in years)
address	Number of years in same address
income	Customer Income
debtinc	Debt to income ratio
creddebt	Credit to Debt ratio
othdebt	Other debts
Default	Customer defaulted in the past (1= defaulted, 0=Never defaulted) ( <b>dependent variable</b> ). Note that this is a categorical value even though the values are 0 and 1.

# KNIME Workflow for Random Forests



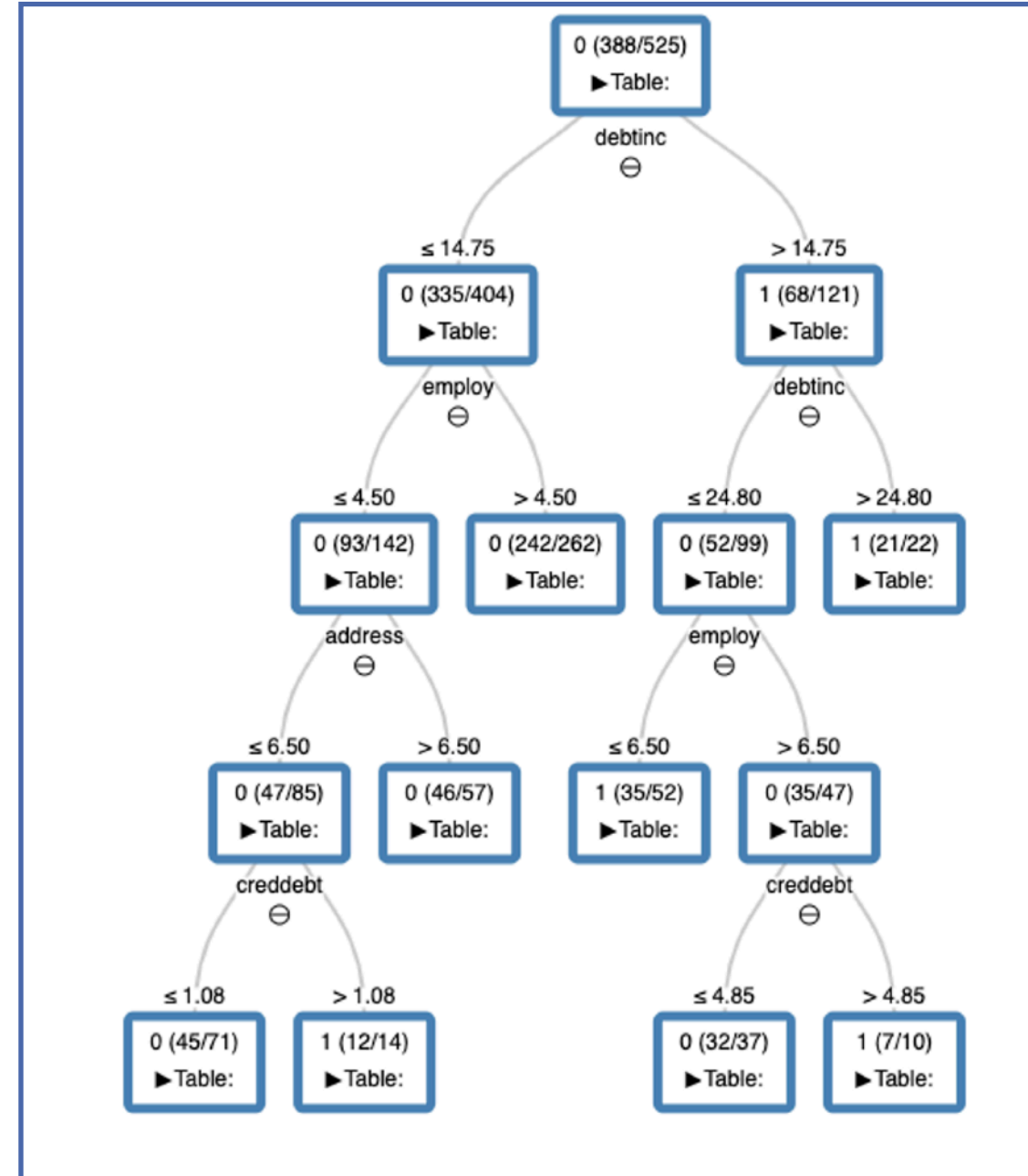
# COMPARING RESULTS ACROSS DIFFERENT MODELS



# Credit Risk Analysis - Decision Trees

• According to Decision trees, the most important features that can influence whether a customer can become a defaulter or not are as follows.

- debtinc
- employ
- address,
- creddebt



# Credit Risk Analysis - Logistic Regression

- According to Decision trees, the most important features that can influence whether a customer can become a defaulter or not are as follows.
  - debtinc
  - Address,
  - creddebt
  - employ

Logit	Variable	Coeff.	Std. Err.	z-score	P> z
1	age	0,018827	0,0196	0,960587	0,33676
1	ed	0,078664	0,143294	0,548967	0,583028
1	employ	-0,2608	0,03867	-6,74417	1,54E-11
1	address	-0,10541	0,02677	-3,93778	8,22E-05
1	income	-0,0109	0,013794	-0,79018	0,429424
1	debtinc	0,078974	0,037572	2,101929	0,035559
1	creddebt	0,615349	0,132577	4,641455	3,46E-06
1	othdebt	0,080039	0,094253	0,849193	0,395774
1	Constant	-1,08754	0,728391	-1,49308	0,135417

# References

- <https://quantifyinghealth.com/interpret-logistic-regression-coefficients/>
- <https://towardsdatascience.com/a-simple-interpretation-of-logistic-regression-coefficients-e3a40a62e8cf>
- <https://towardsdatascience.com/a-simple-interpretation-of-p-values-34db3777d907>
- <https://www.knime.com/example-workflows>

# Thank you!

Raghava Mukkamala

[rrm.digi@cbs.dk](mailto:rrm.digi@cbs.dk)

<https://www.cbs.dk/staff/rrmdigi>

<https://raghavamukkamala.github.io/>

<https://cbsbda.github.io/>