CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Business Analytics and Data-Driven Decision Making
S15: Hands-on-sessionPredictive Analytics-2

Decision Trees, Random Forests, Cross Validation with KNIME

Raghava Mukkamala

Associate Professor & Director,
Centre for Business Data Science
Copenhagen Business School, Denmark
Email: rrm.digi@cbs.dk, Centre: https://cbsbda.github.io/

Slides based on KNIME Analytics Platform

Source: https://www.knime.com/sites/default/files/2021-07/slides-l1-ds.pdf

EQUIS ACCREDITED AACSB ACCREDITED ASSOCIATION AMBA ACCREDITED CEMS PIM

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

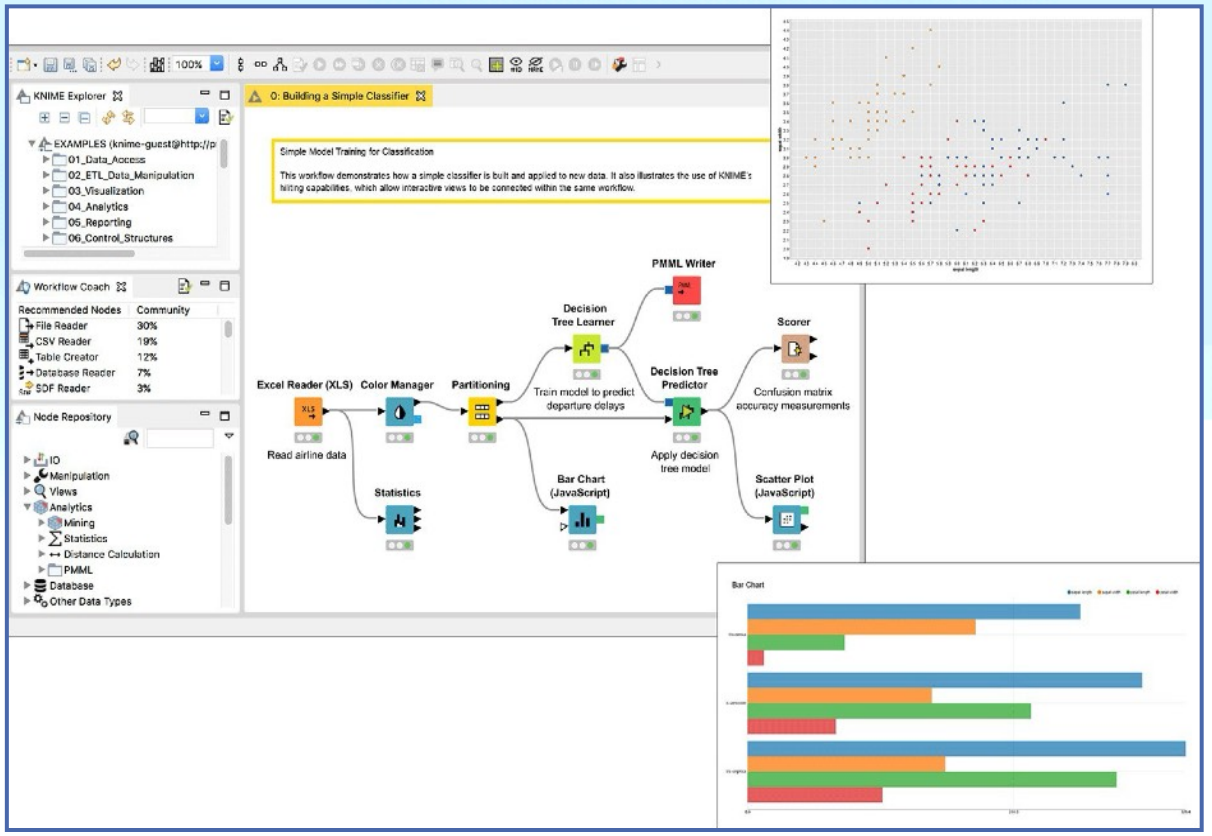
Overview: KNIME Analytics Platform

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

CBS

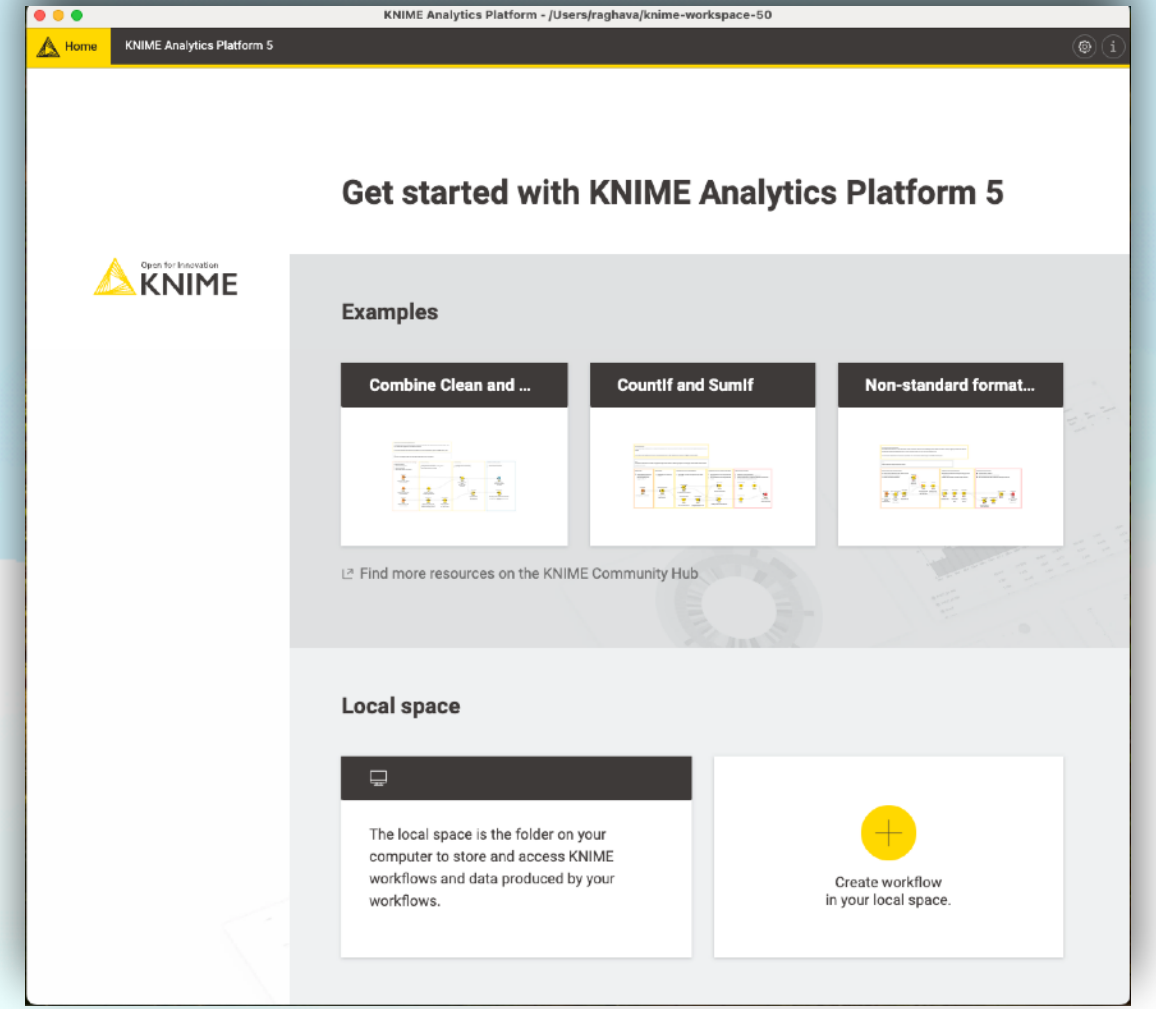
What is KNIME Analytics Platform?

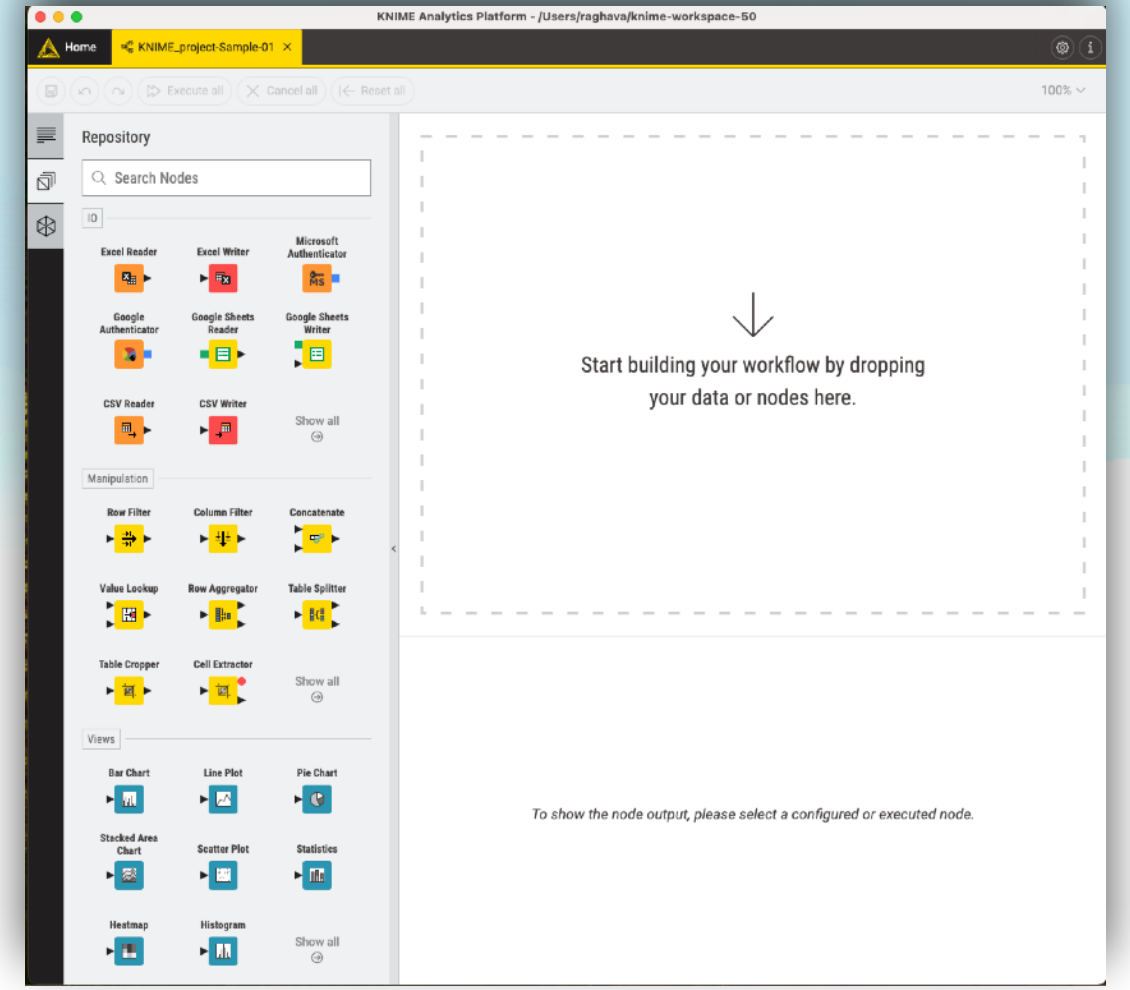
- A tool for data analysis, manipulation, visualization, and reporting
- Based on the graphical programming paradigm
- Provides a diverse array of extensions:
  - Machine learning,
  - Text Mining,
  - Network Mining



KNIME Classic look (before 5.0 version)

KNIME Analytics Platform 5.0 - Modern look

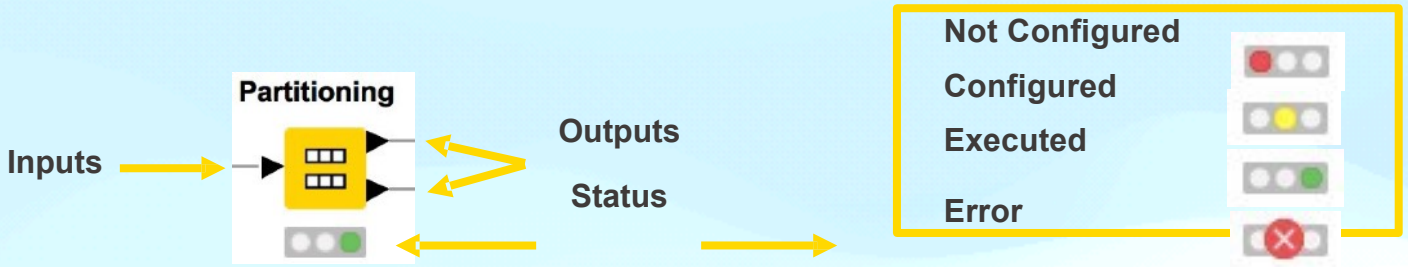




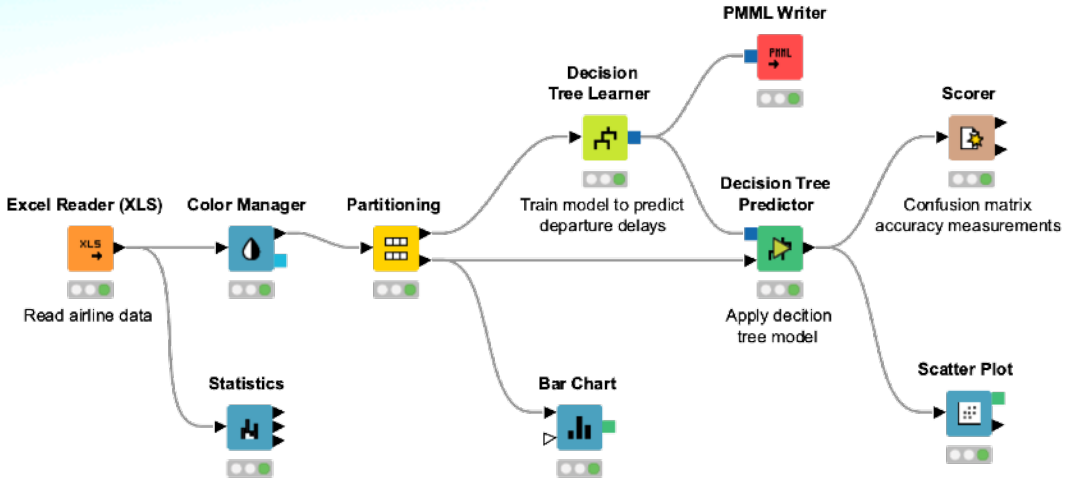
Open for Innovation KNIME

Visual KNIME Workflows

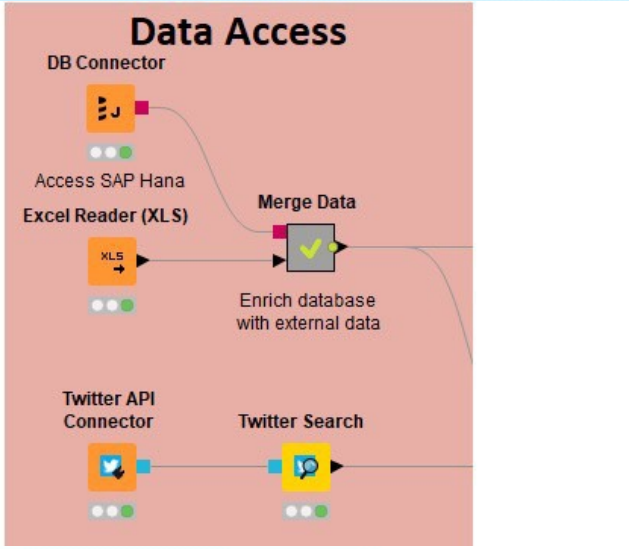
- **NODES** perform individual tasks on data
- Nodes has ports for data in and out



Nodes are combined to create
WORKFLOWS

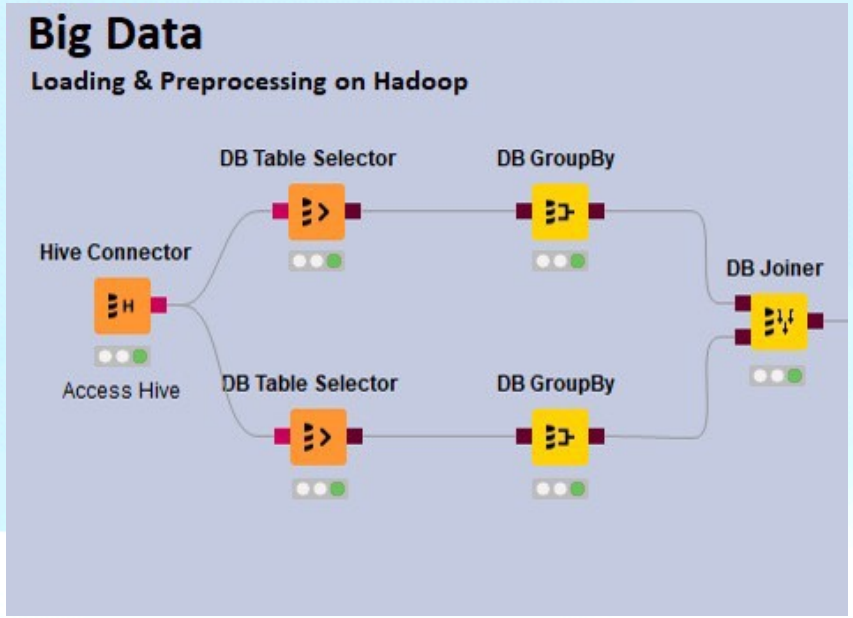


Data Access



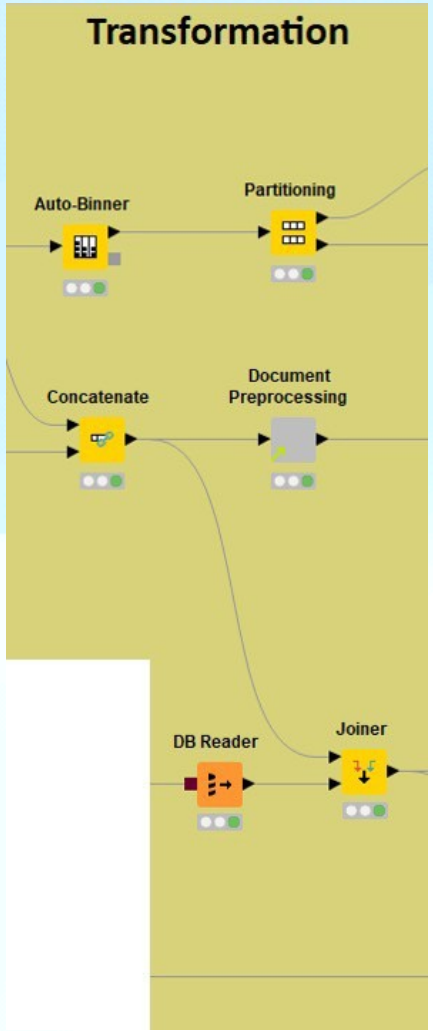
- Databases
  - MySQL, PostgreSQL, Oracle
  - Theobald
  - any JDBC (DB2, MS SQL Server)
  - Amazon DynamoDB
- Files
  - CSV, txt, Excel, Word, PDF
  - SAS, SPSS
  - XML, JSON, PMML
  - Images, texts, networks
- Other
  - Twitter, Google
  - Amazon S3, Azure Blob Store
  - Sharepoint, Salesforce
  -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

Big Data



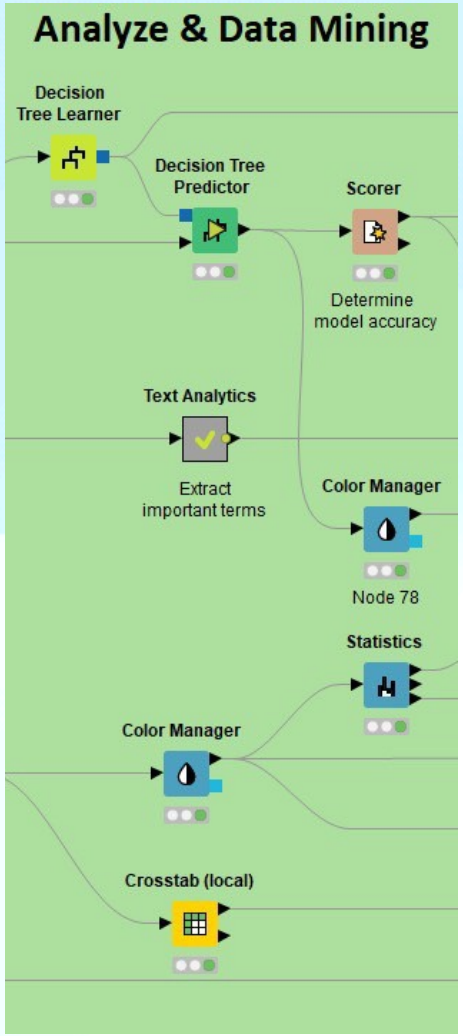
- Spark & Databricks
- HDFS support
- Hive
- Impala
- In-database processing

Data Transformation



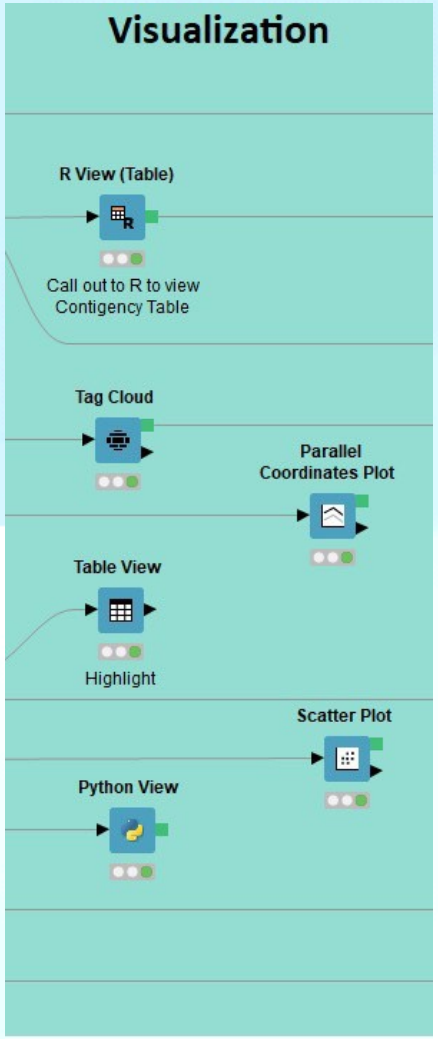
- Preprocessing
  - Row, column, matrix based
- Data blending
  - Join, concatenate, append
- Aggregation
  - Grouping, pivoting, binning
- Feature Creation and Selection

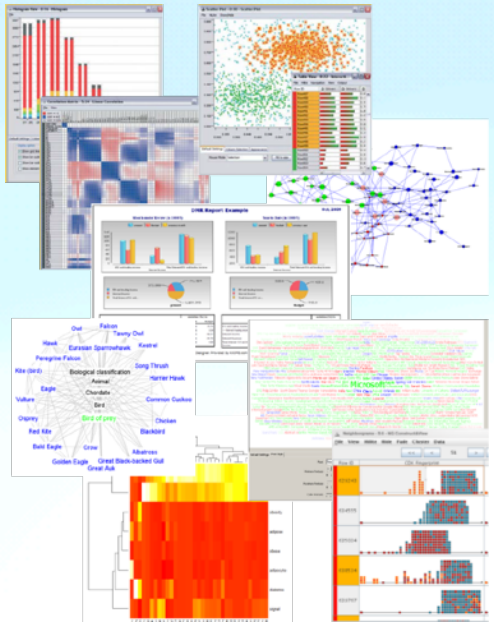
Analysis & Data Mining



- 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

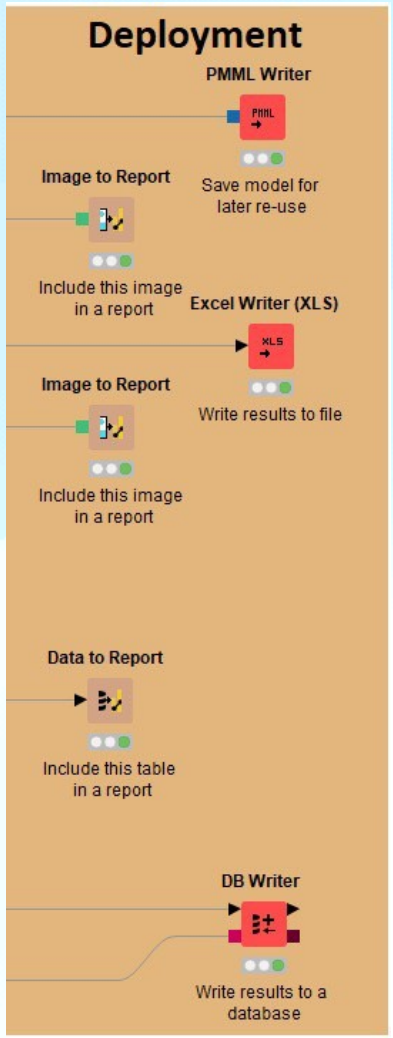
Visualization





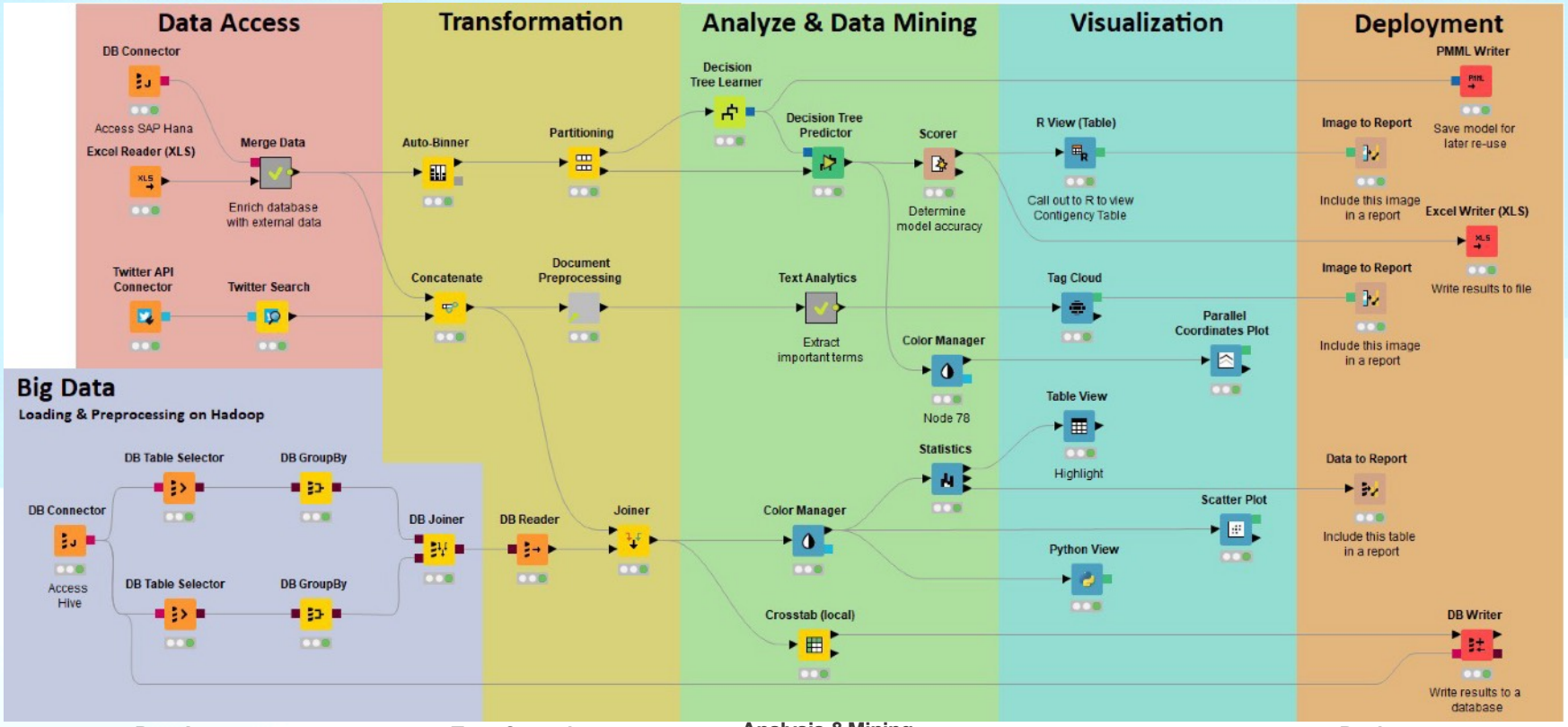
- Interactive Visualizations
- JavaScript-based nodes
  - Scatter Plot, Box Plot, Line Plot
  - Networks, ROC Curve, Decision Tree
  - Plotly Integration
  - Adding more with each release!
- Misc
  - Tag cloud, open street map, molecules
- Script-based visualizations
  - R, Python

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

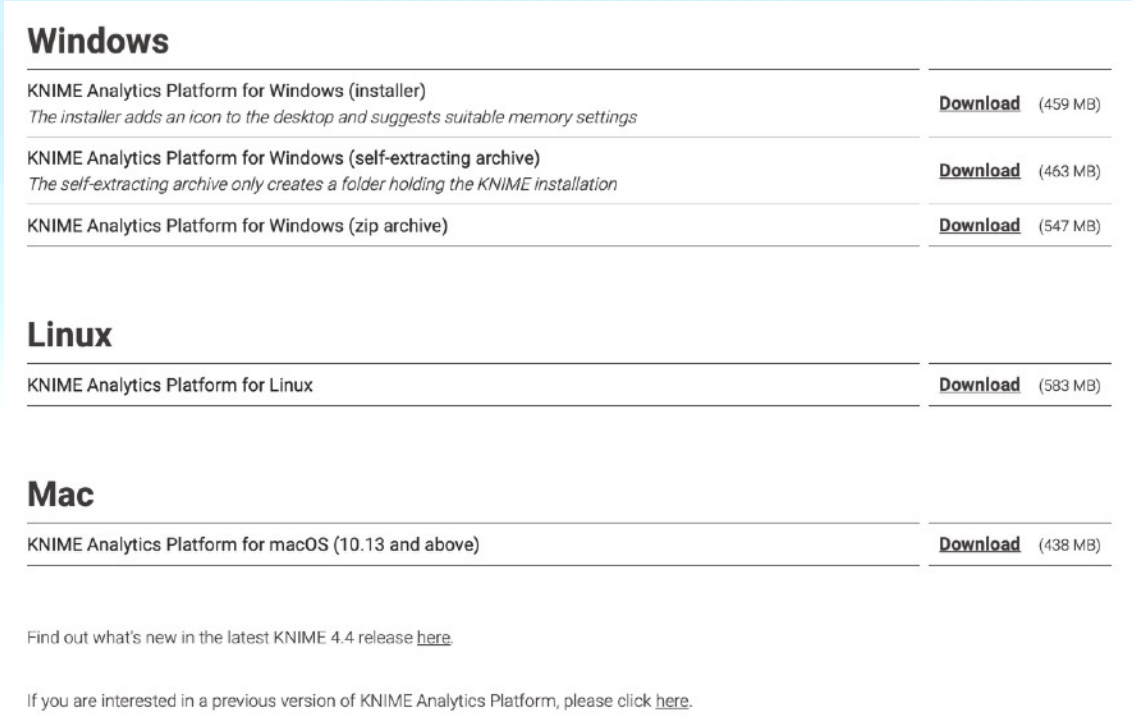
Visualization
R
JFreeChart
JavaScript Plotly
Community / 3rd

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

Linux

KNIME Analytics Platform for Linux	Download (583 MB)

Mac

KNIME Analytics Platform for macOS (10.13 and above)	Download (438 MB)

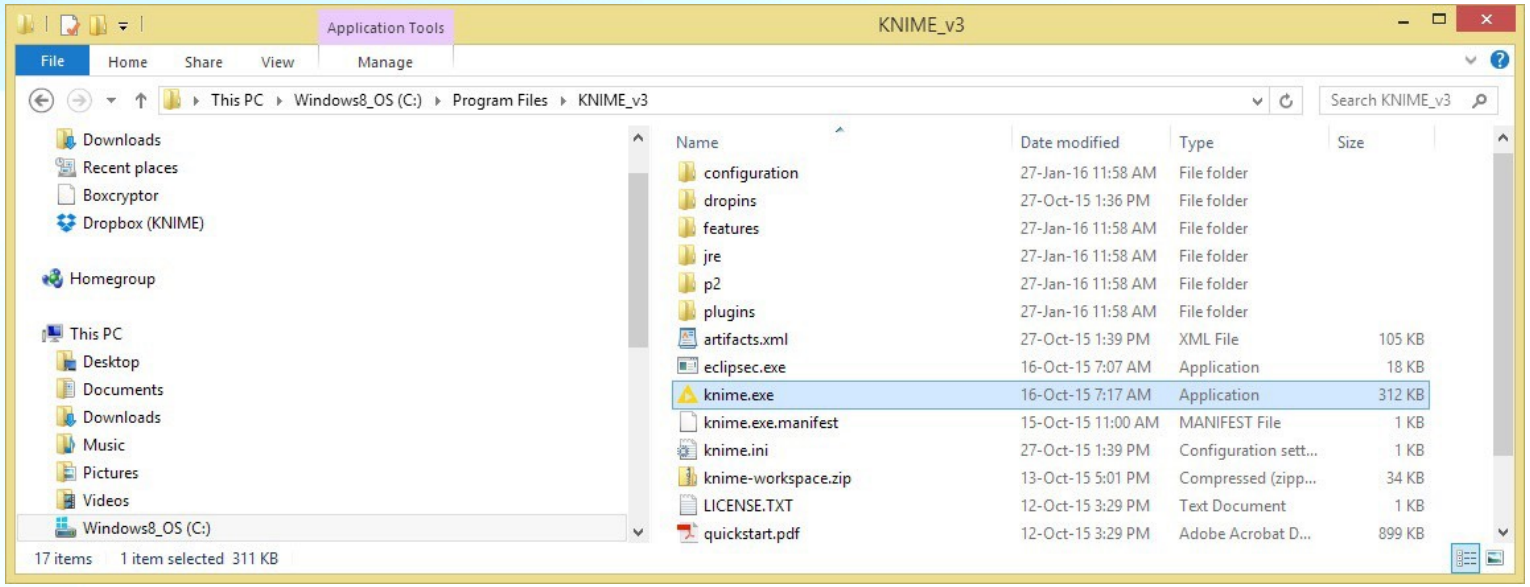
Find out what's new in the latest KNIME 4.4 release here.

If you are interested in a previous version of KNIME Analytics Platform, please click here.

Start KNIME Analytics Platform

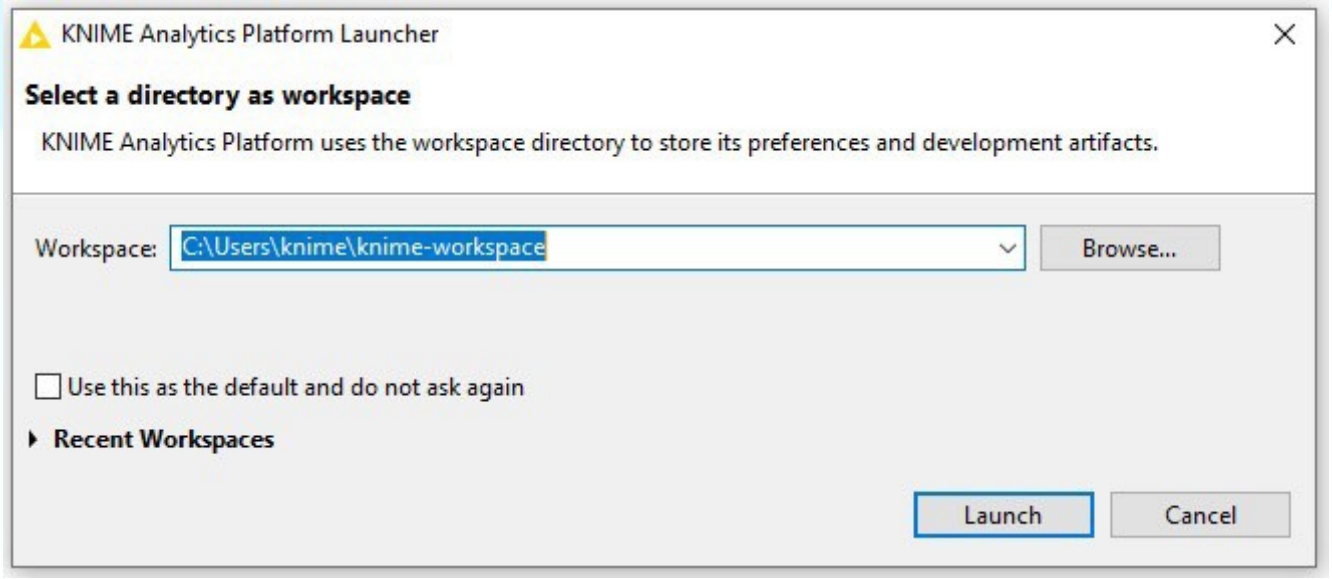
- Use the shortcut created by the installer

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

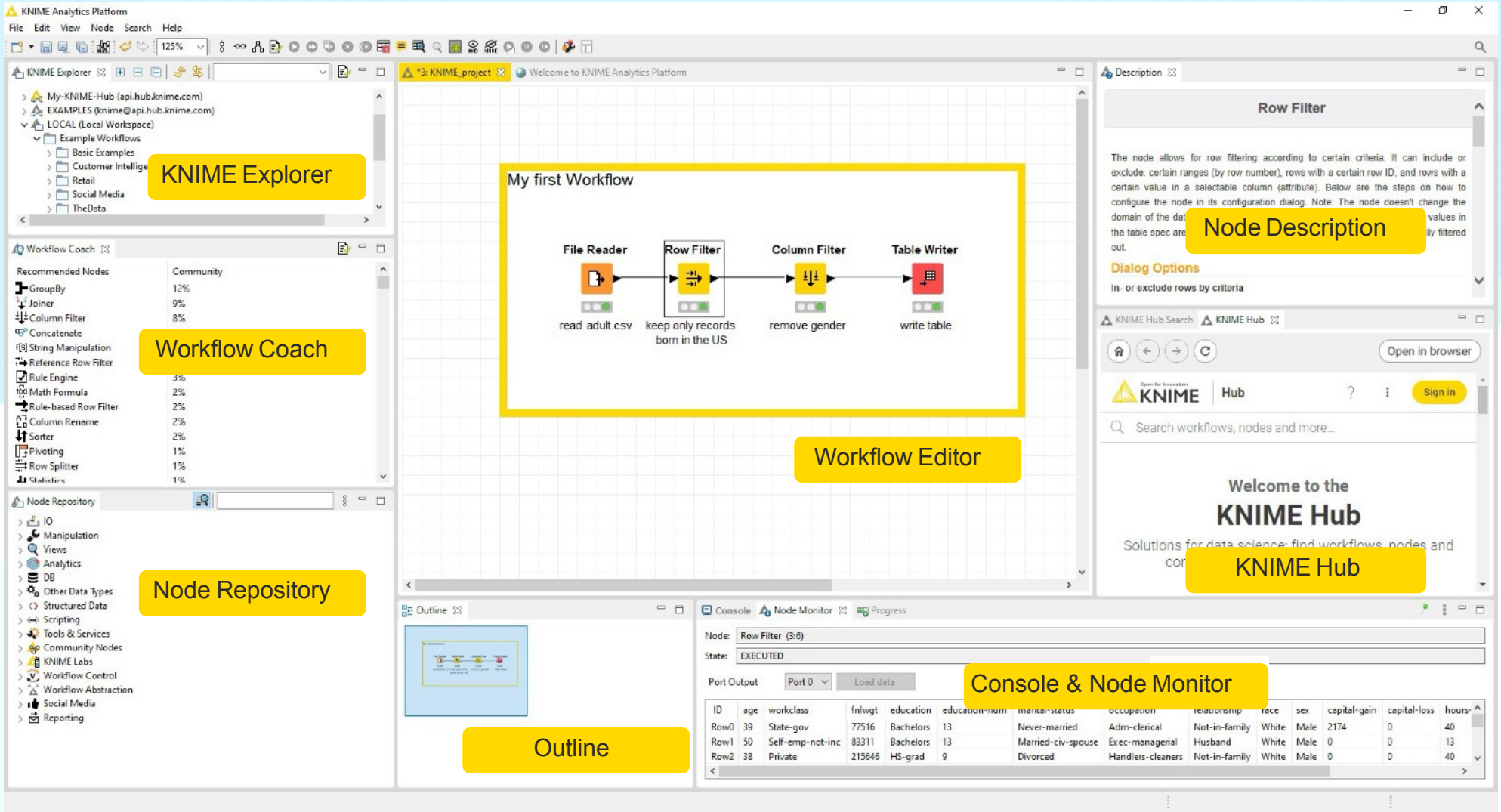


The KNIME Workspace

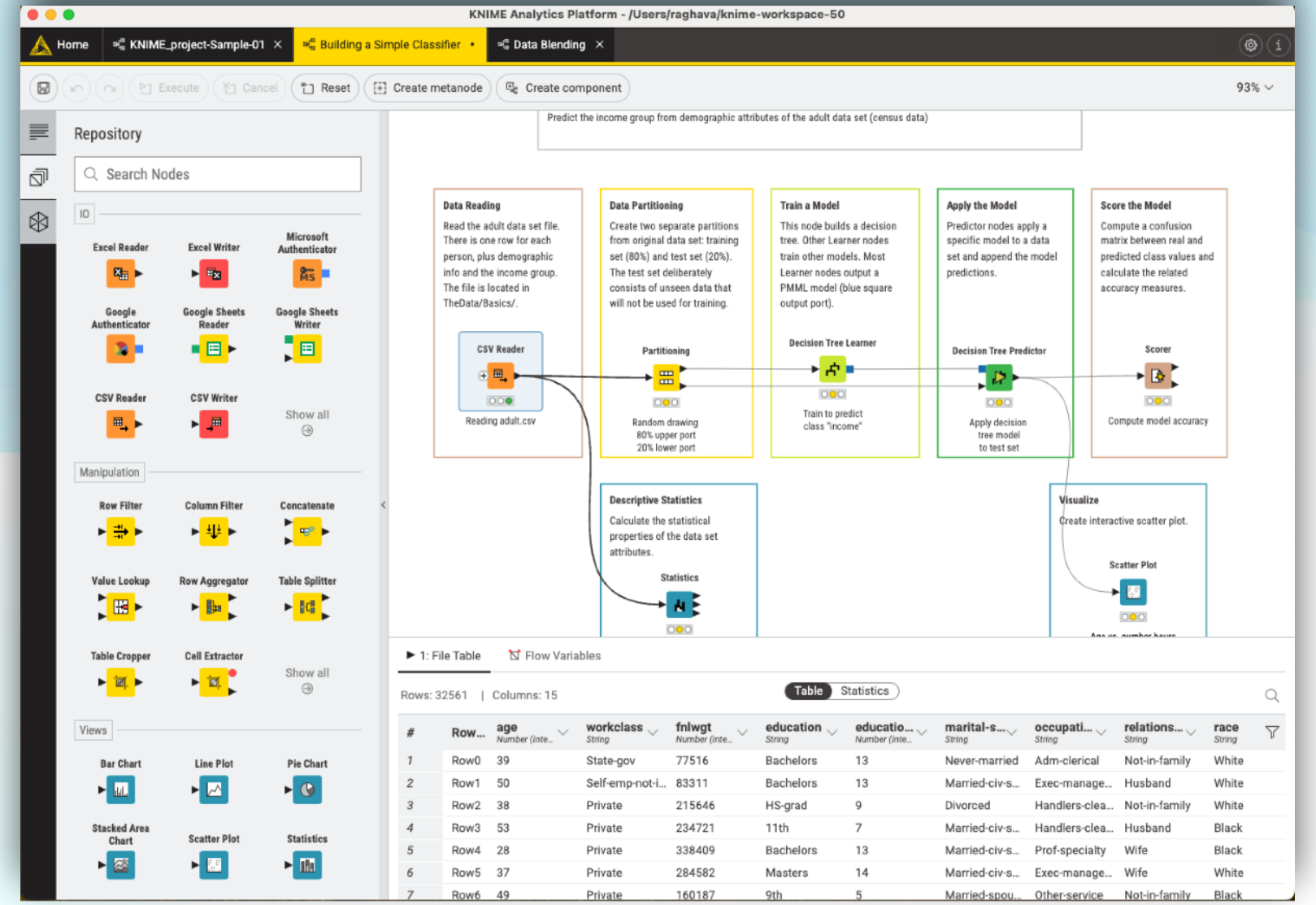
- The workspace is the **folder/directory** in which workflows (and potentially data files) are stored for the current KNIME session.
- Workspaces are portable (just like KNIME)



The KNIME Analytics Platform Workbench (classic UI)



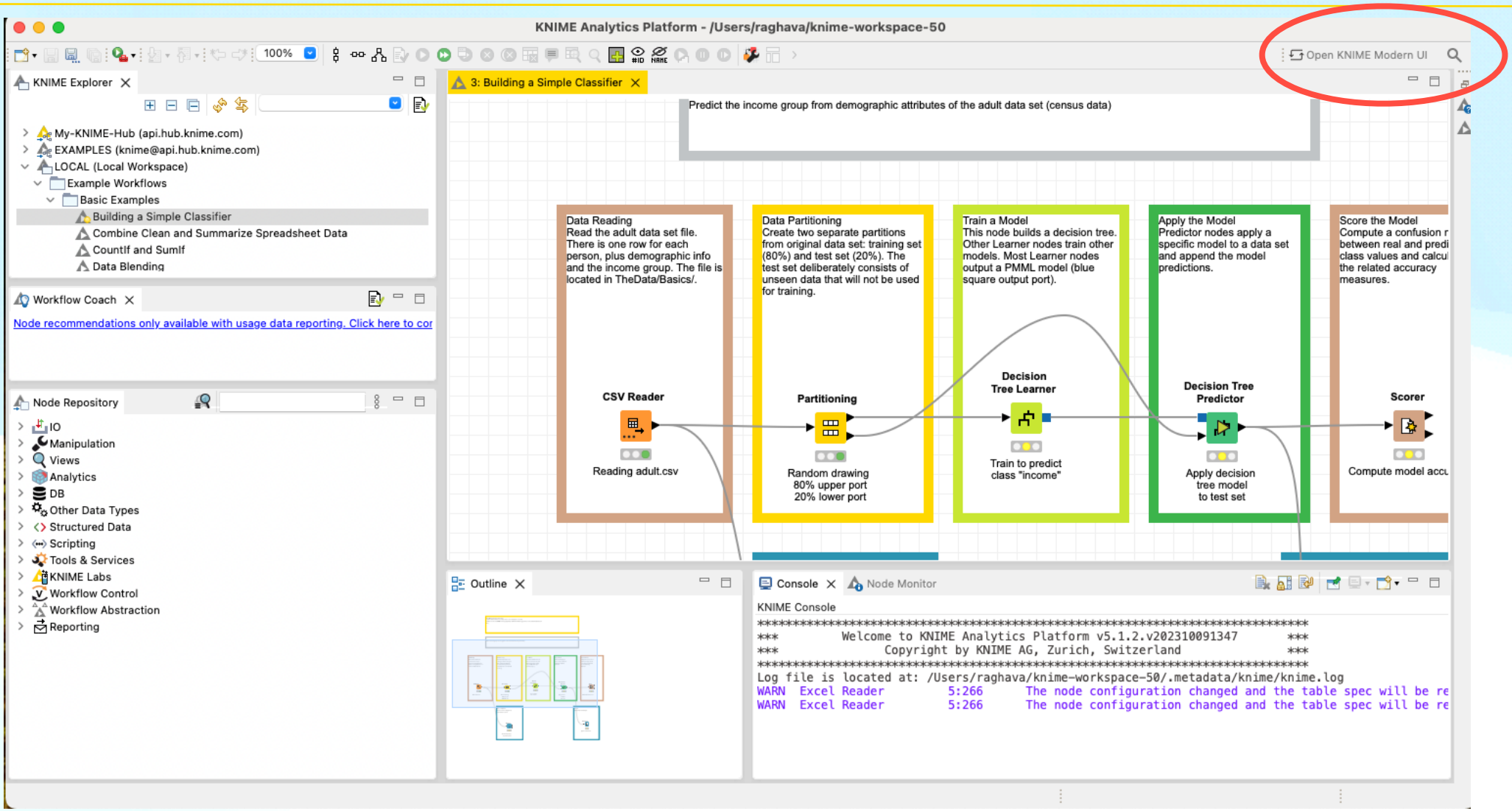
The KNIME Analytics Platform Workbench (Modern UI)



In KNIME 5.0 or later versions

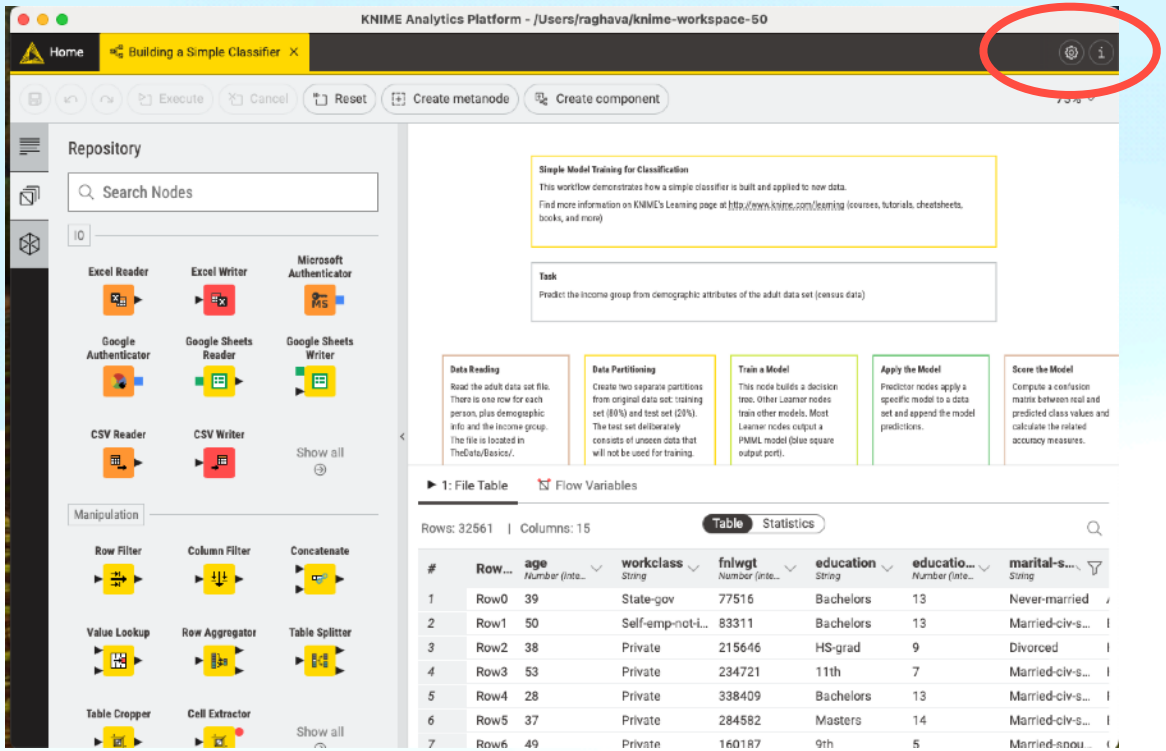
Open for Innovation KNIME

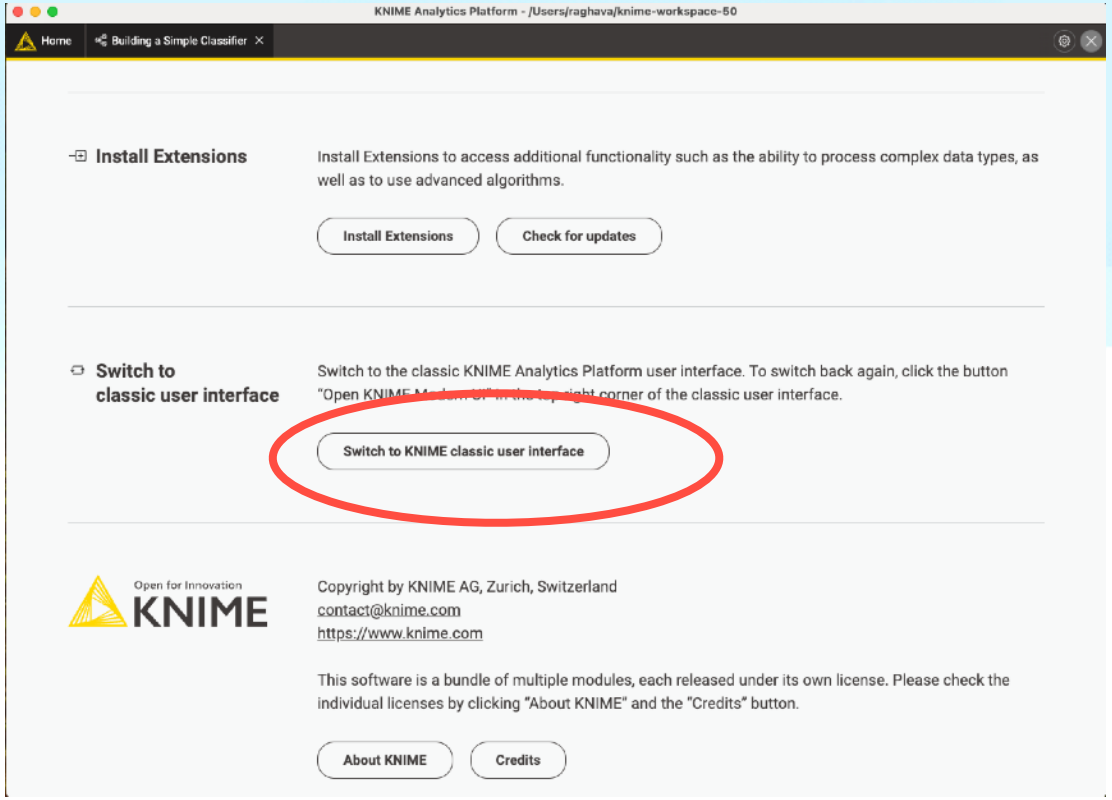
Switching from Classical UI to Modern UI (in KNIME >5.0)



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

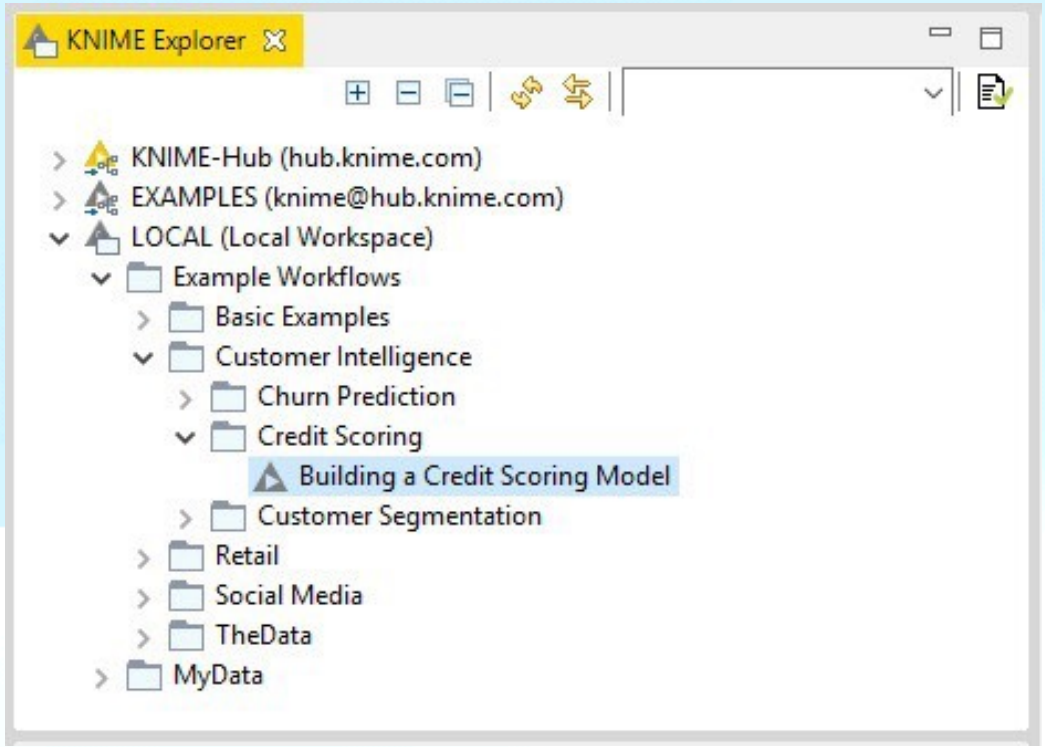
Click on icon i





Open for Innovation KNIME

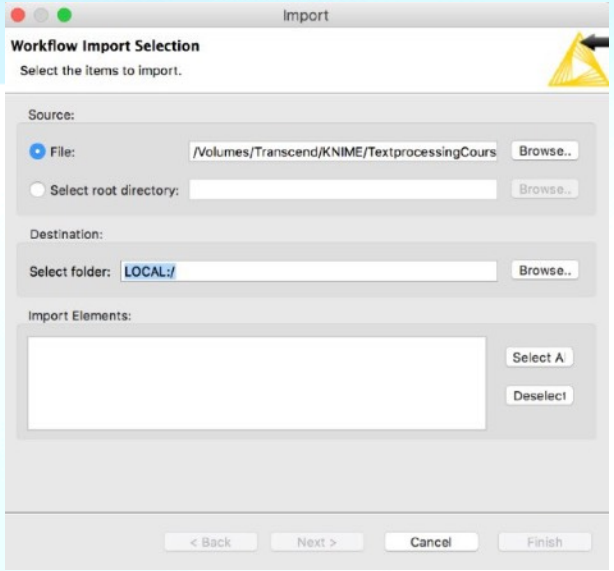
KNIME Explorer

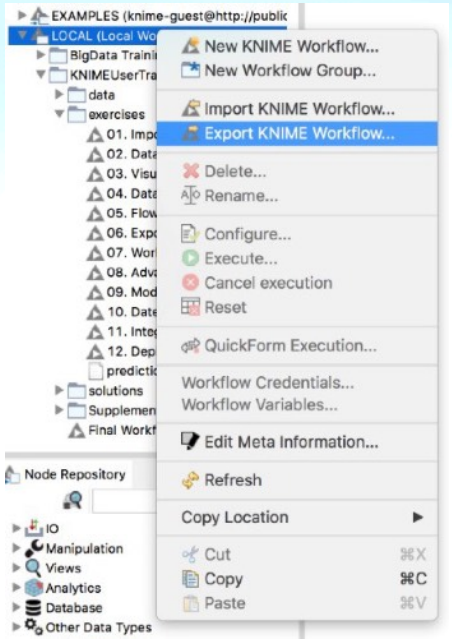


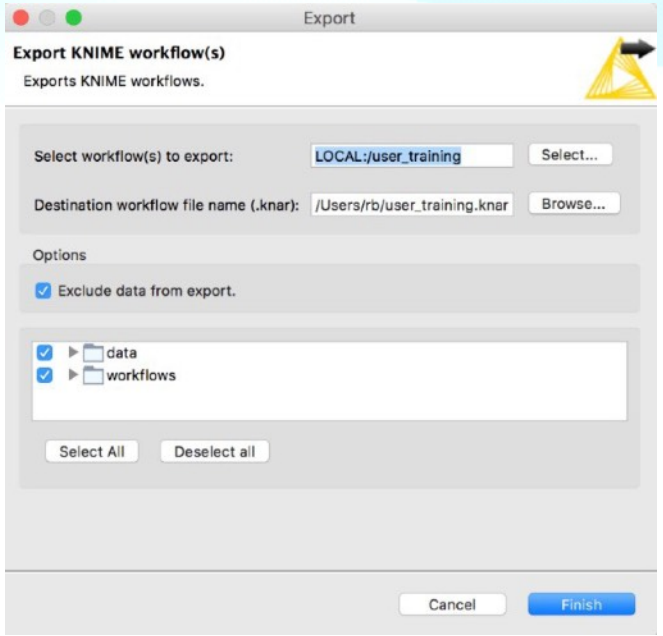
- In LOCAL you can access your own workflow projects.
- Other mountpoints allow you to connect to
  - EXAMPLE Server
  - KNIME Hub
  - KNIME Server
- The Explorer toolbar on the top has a search box and buttons to
  - select the workflow displayed in the active editor
  - refresh the view
- The KNIME Explorer can contain 4 types of content:
  - Workflows
  - Workflow groups
  - Data files
  - Shared Components

Creating New Workflows, Importing and Exporting

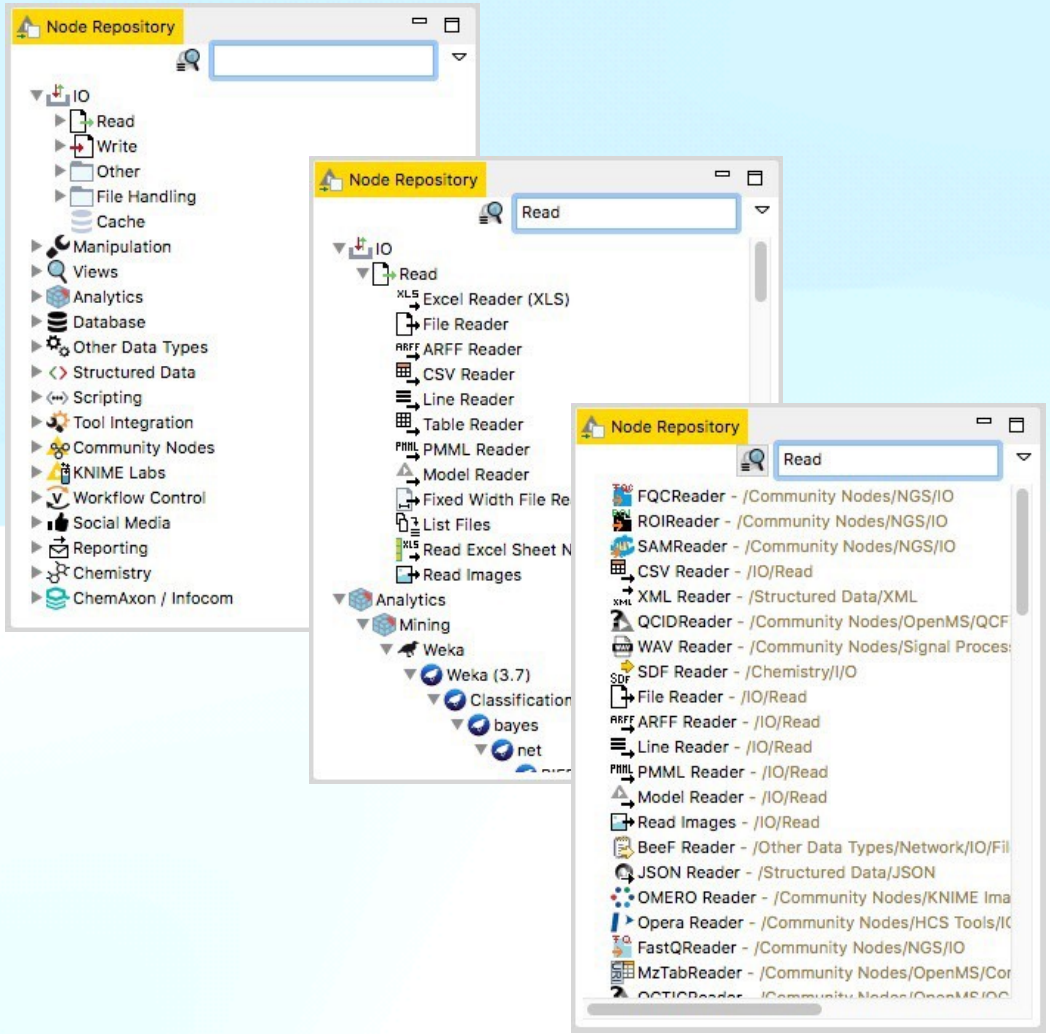
- Right-click inside the KNIME Explorer to create a new workflow or a workflow group, or to import a workflow
- Right-click the workflow or workflow group to export







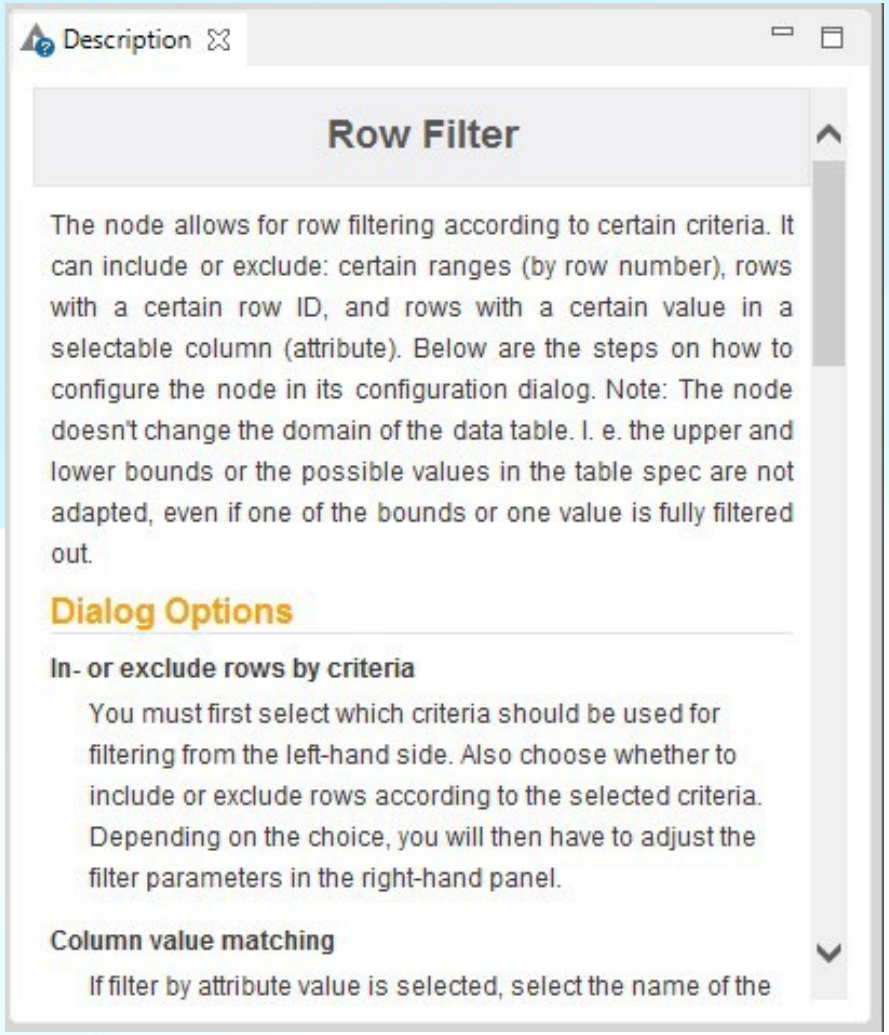
Node Repository



- The Node Repository lists all KNIME nodes

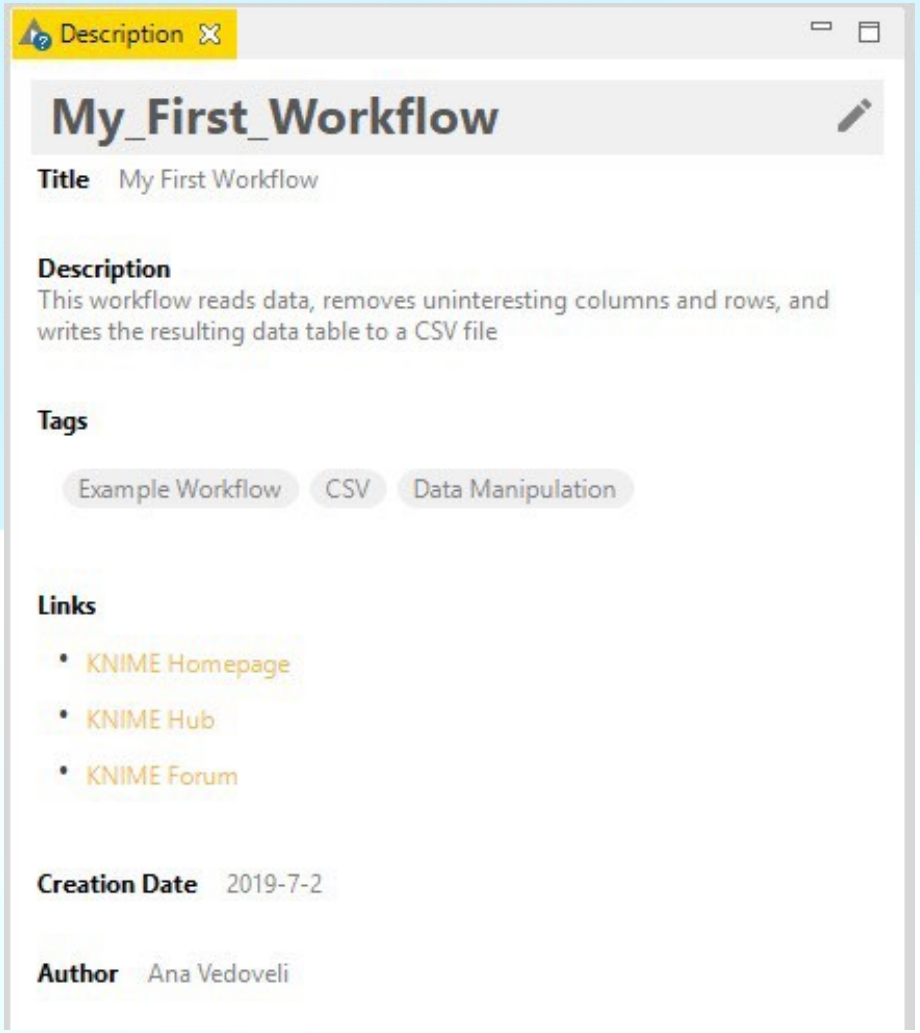
- The search box has 2 modes
  - Standard Search – exact match of node name
  - Fuzzy Search – finds the most similar node name

Description



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

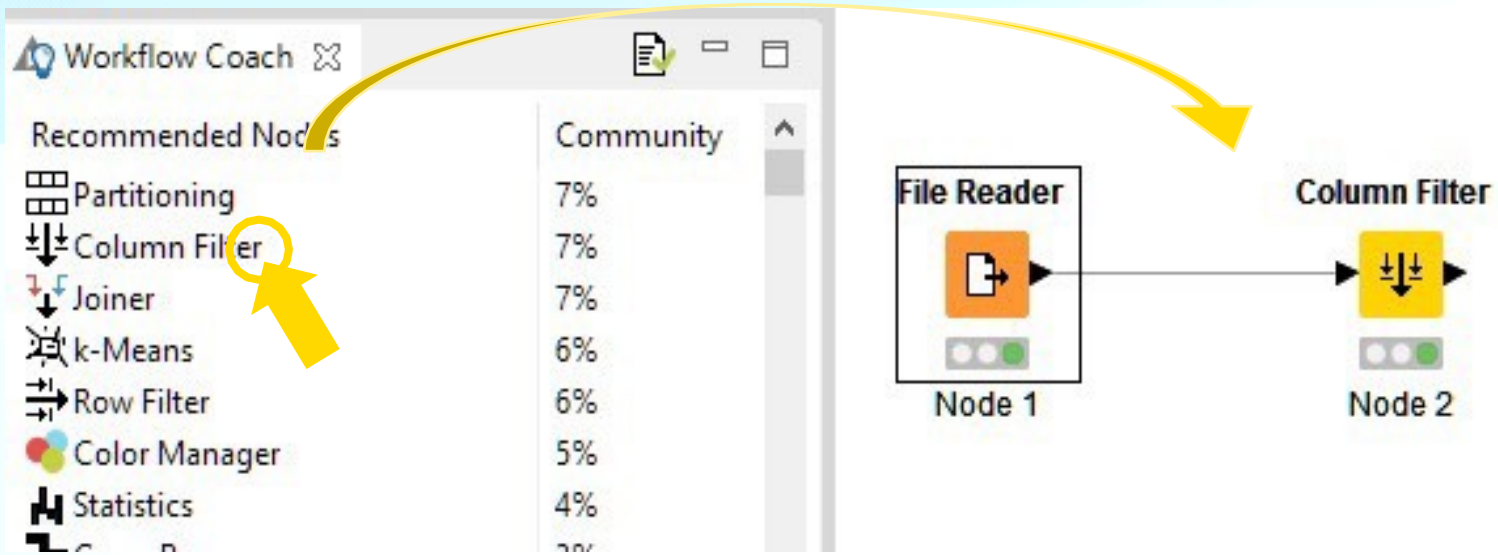
Workflow Description



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

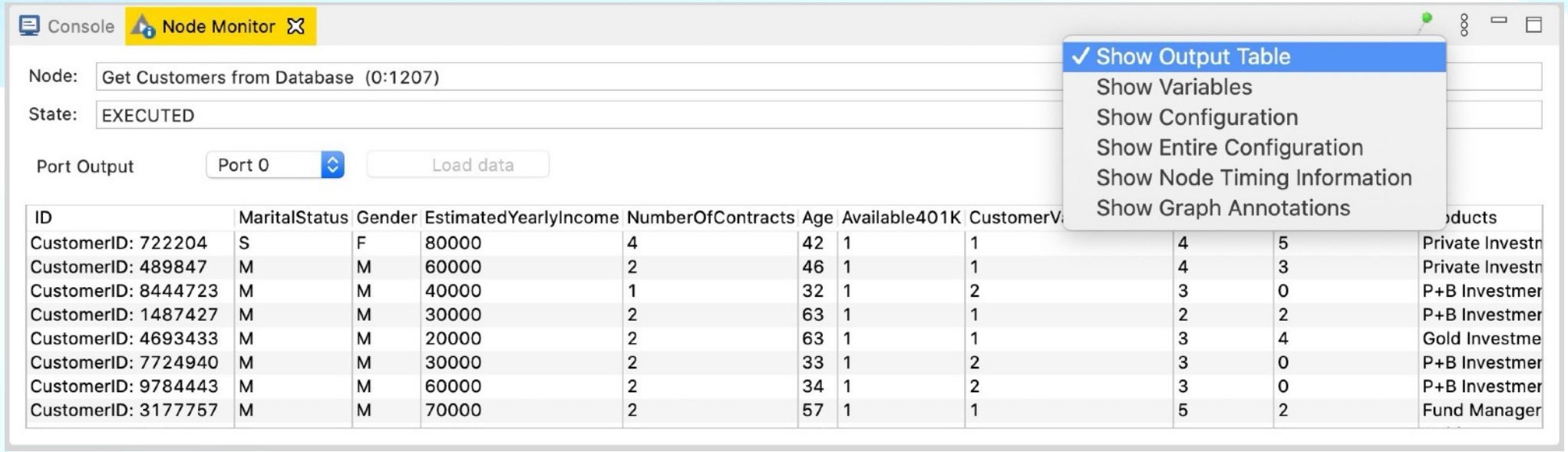
Workflow Coach

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



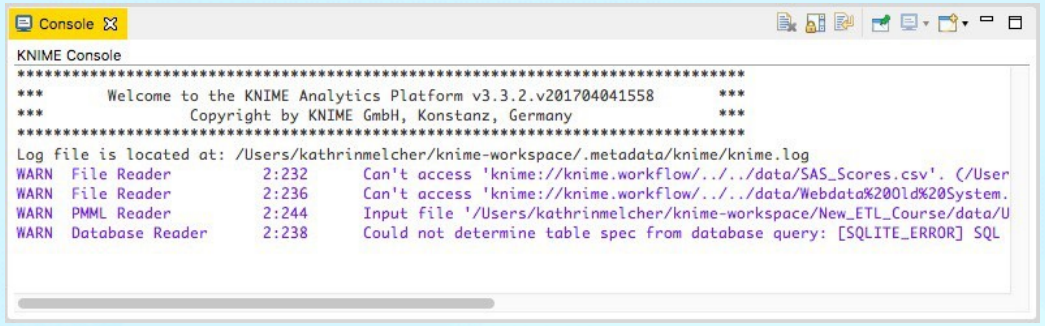
Node Monitor

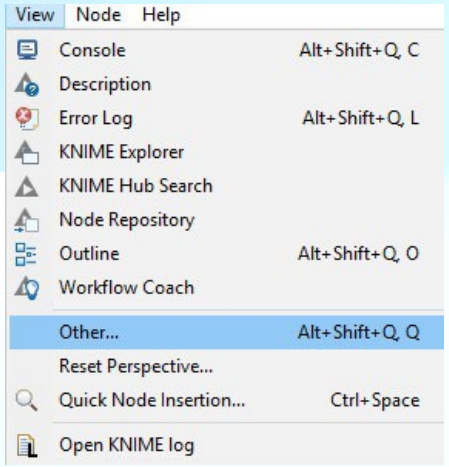
- By default the Node Monitor shows you the output table of the node selected in the workflow editor
- Click on the three dots on the upper right to show the flow variables, configuration, etc.

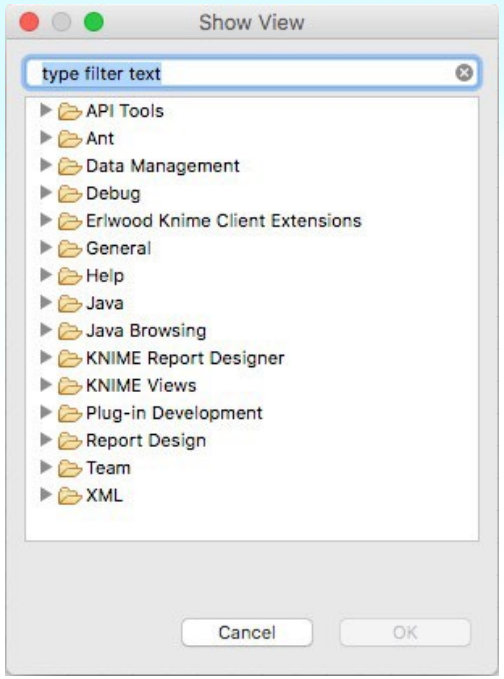


ID	MaritalStatus	Gender	EstimatedYearlyIncome	NumberOfContracts	Age	Available401K	CustomerV			oducts
CustomerID: 722204	S	F	80000	4	42	1	1	4	5	Private Investn
CustomerID: 489847	M	M	60000	2	46	1	1	4	3	Private Investn
CustomerID: 8444723	M	M	40000	1	32	1	2	3	0	P+B Investmer
CustomerID: 1487427	M	M	30000	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



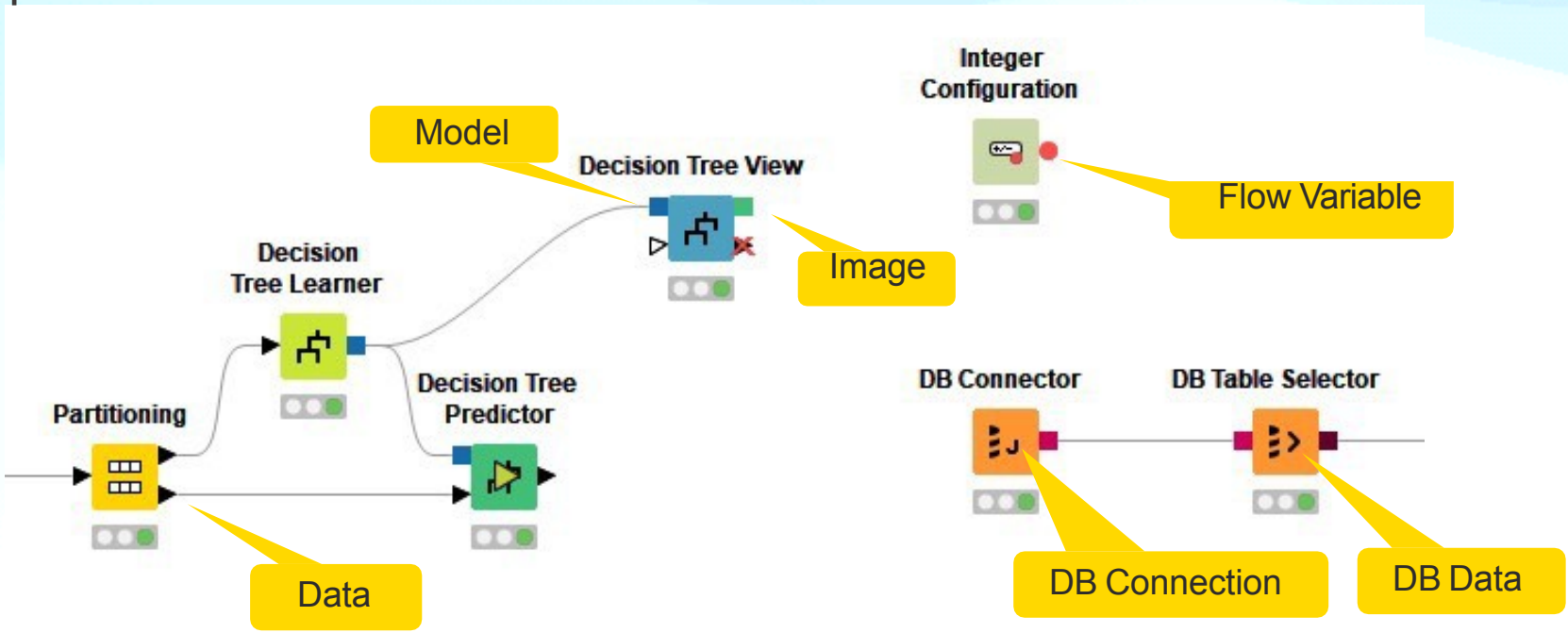




- Console view prints out error and warning messages about what is going on under the hood
- Click on View and select Other… to add different views
  - Node Monitor, Licenses, etc.

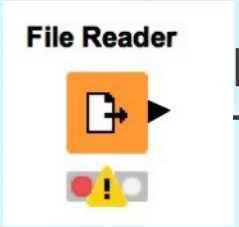
Inserting and Connecting Nodes

- 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:

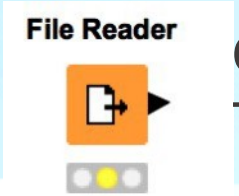


More on Nodes…

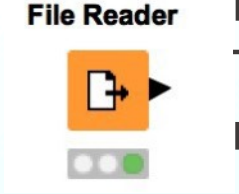
- A node can have 4 states:



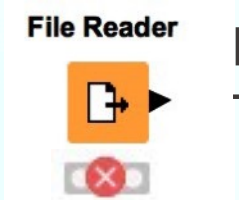
Not Configured:
The node is waiting for configuration or incoming data.



Configured:
The node has been configured correctly and can be executed.



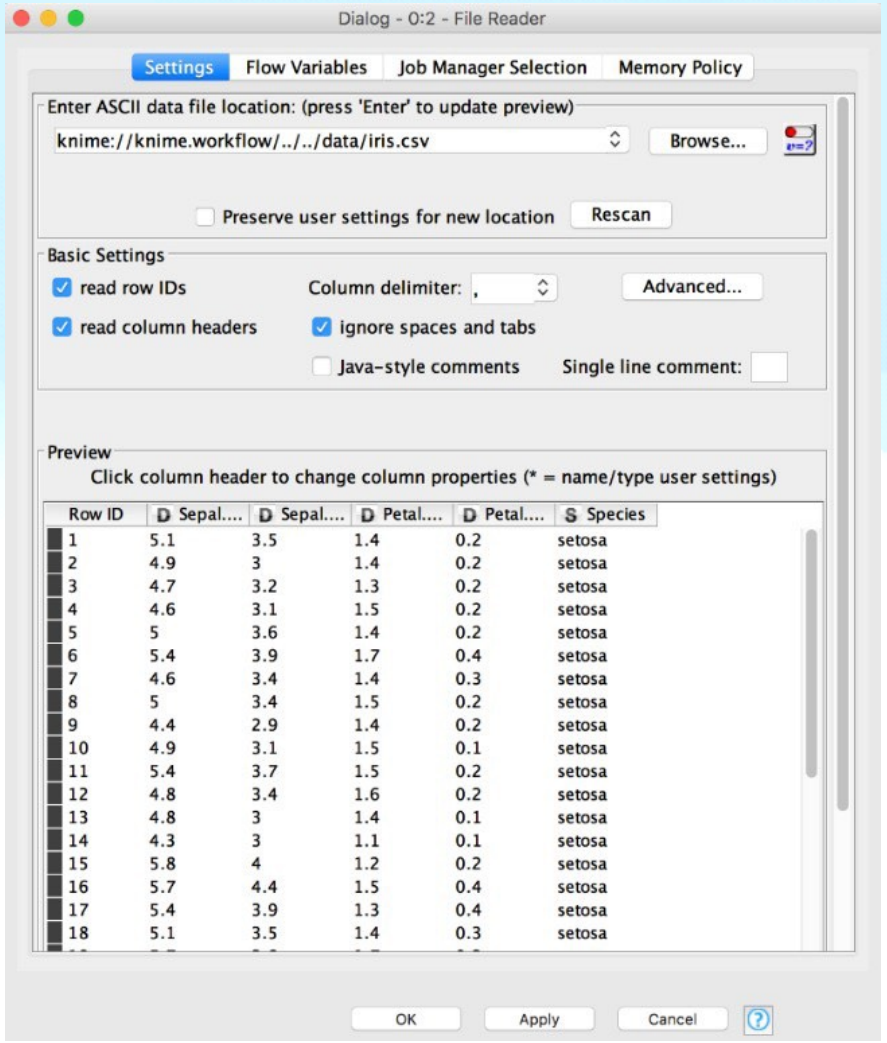
Executed:
The node has been successfully executed. Results may be viewed and used in downstream nodes.



Error:
The node has encountered an error during execution.

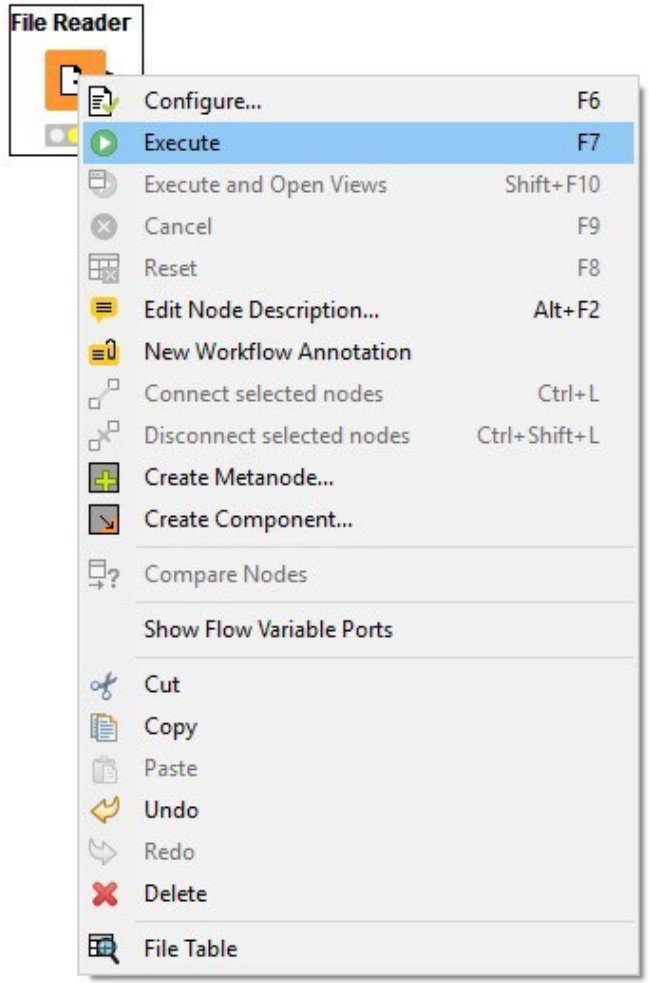
Node Configuration

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

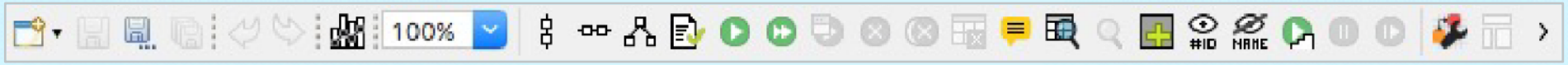


Node Execution

- 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

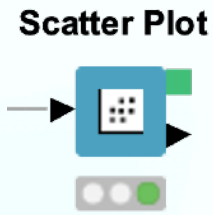


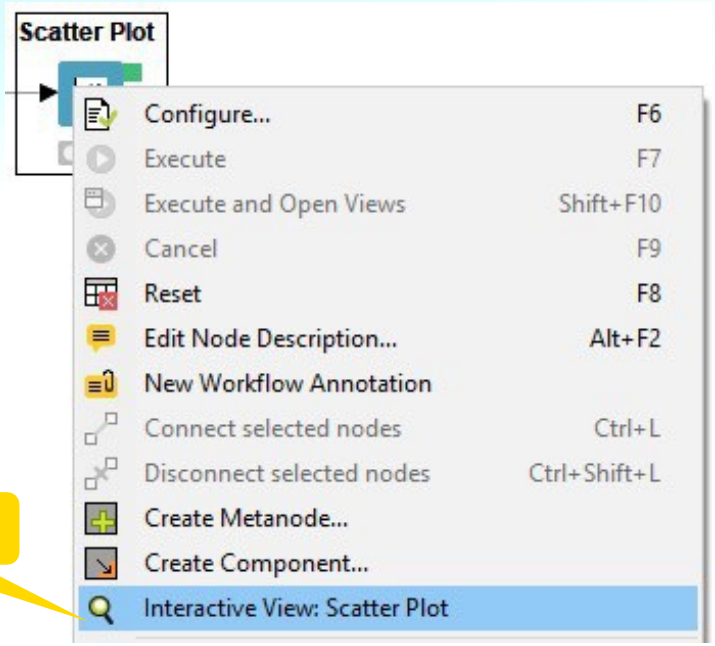
The buttons in the toolbar can be used for the active workflow. The most important buttons are:

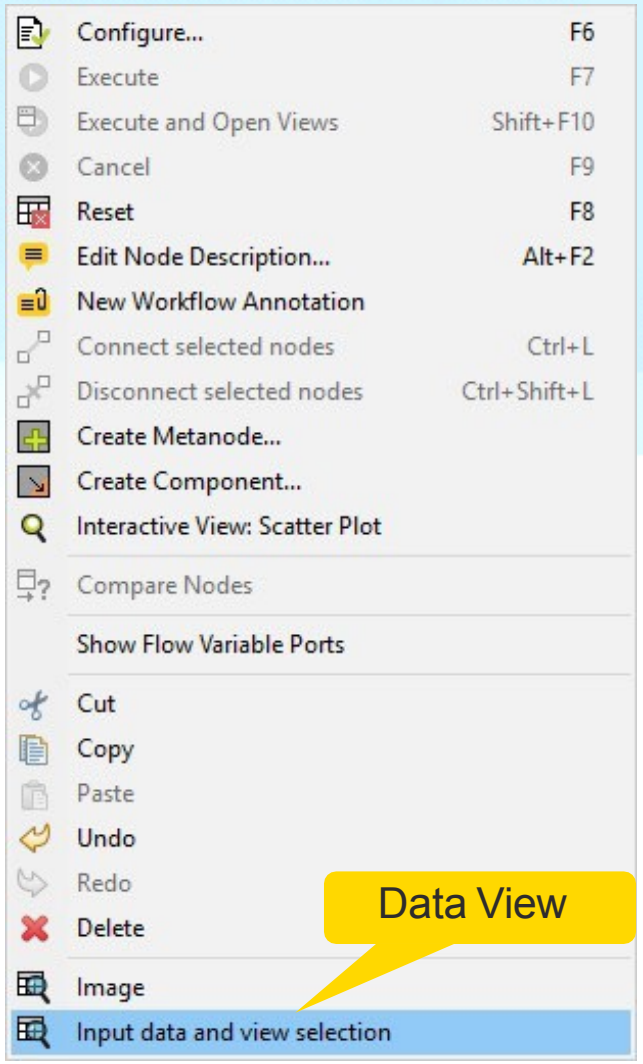
- Execute selected and executable nodes (F7)
- Execute all executable nodes
- Execute selected nodes and open first view
- Cancel all selected, running nodes (F9)
- Cancel all running nodes

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

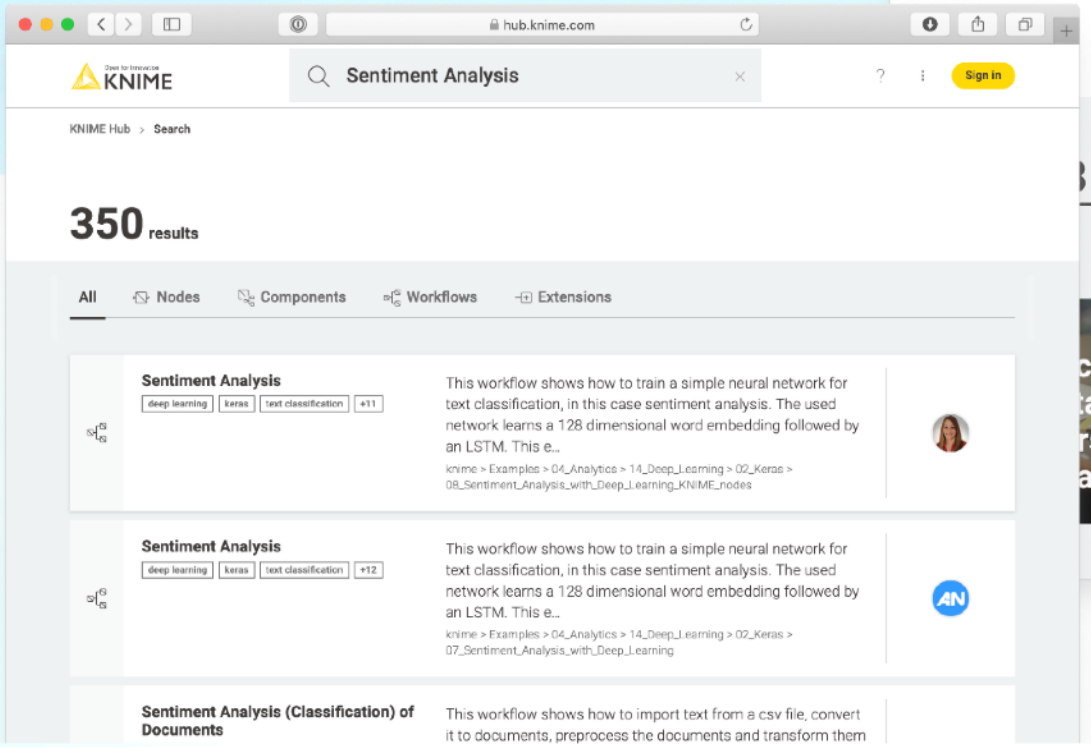
Dedicated file extensions for workflows and workflow groups associated with KNIME Analytics Platform

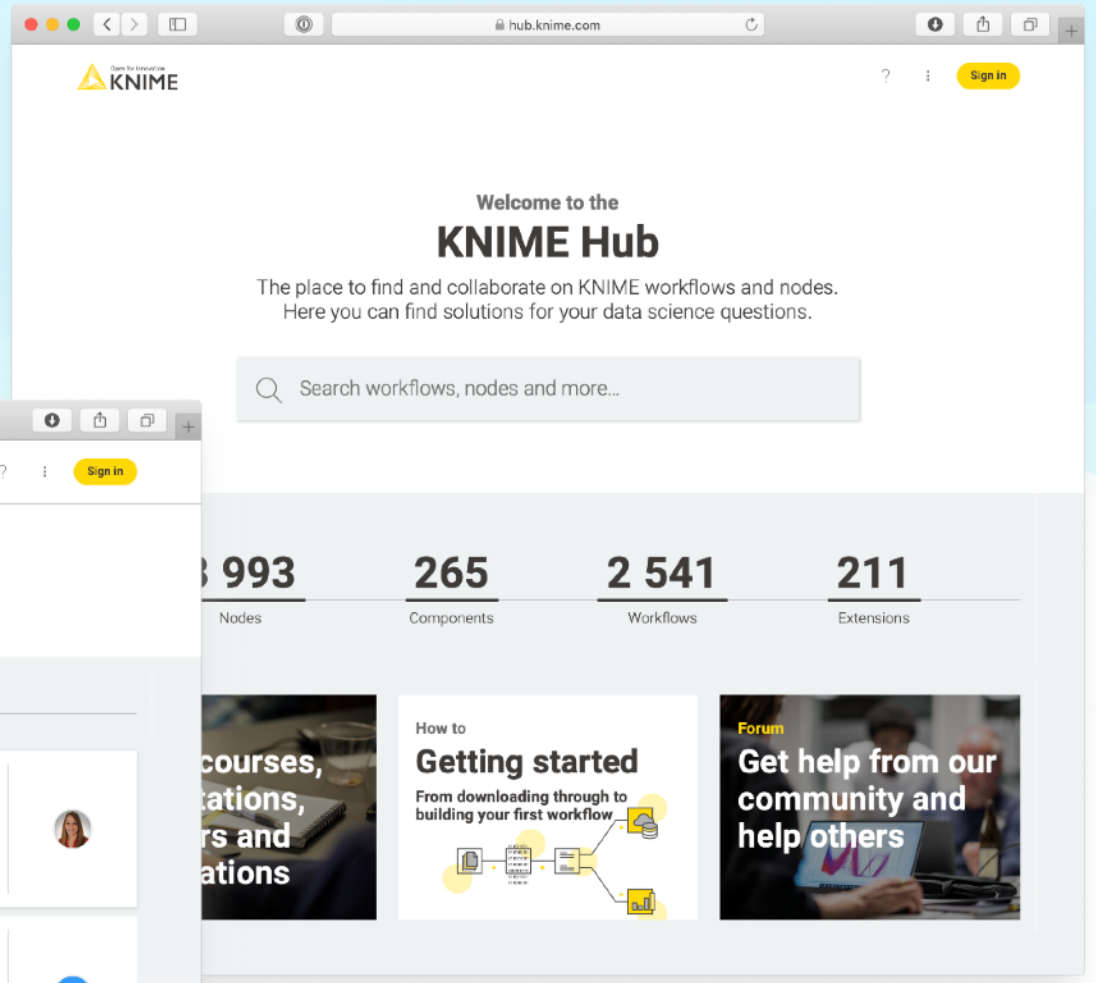
- ***.knwf** for KNIME Workflow Files

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

Getting Started: KNIME Hub

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

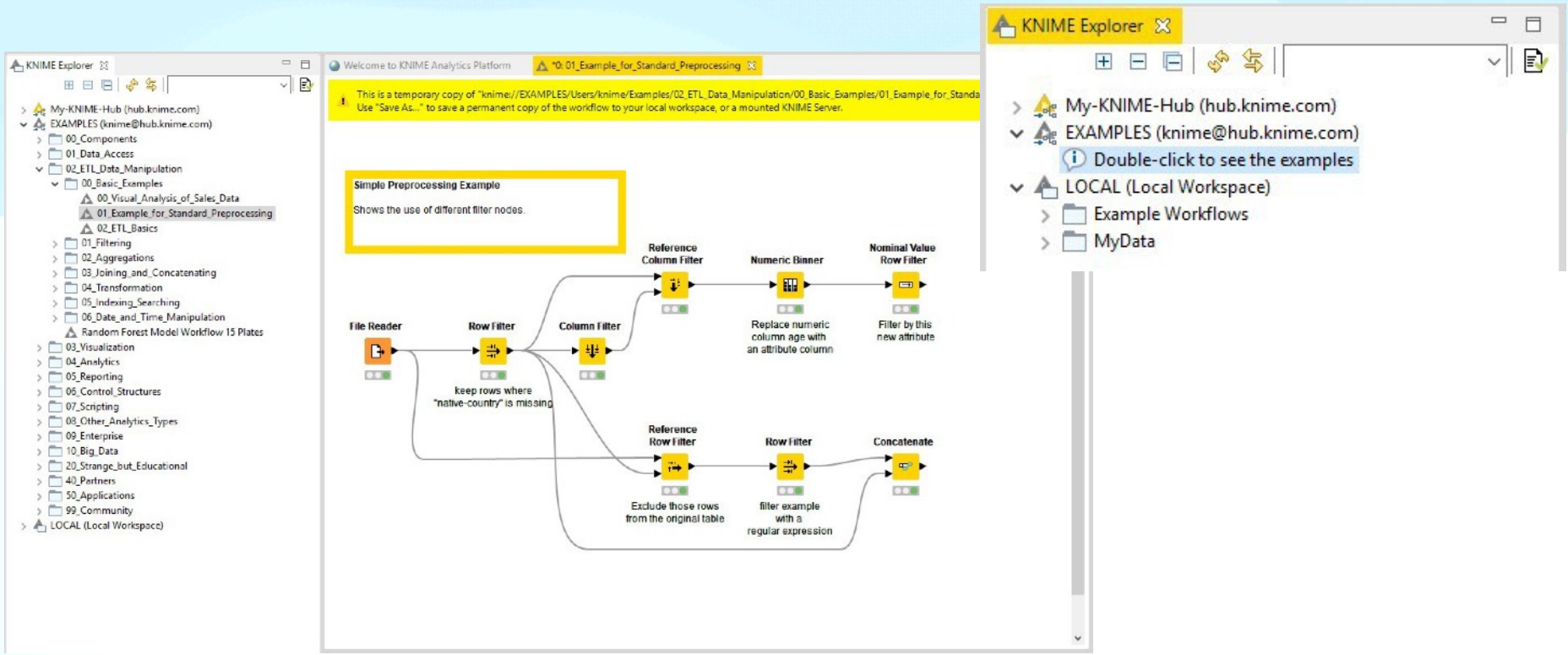




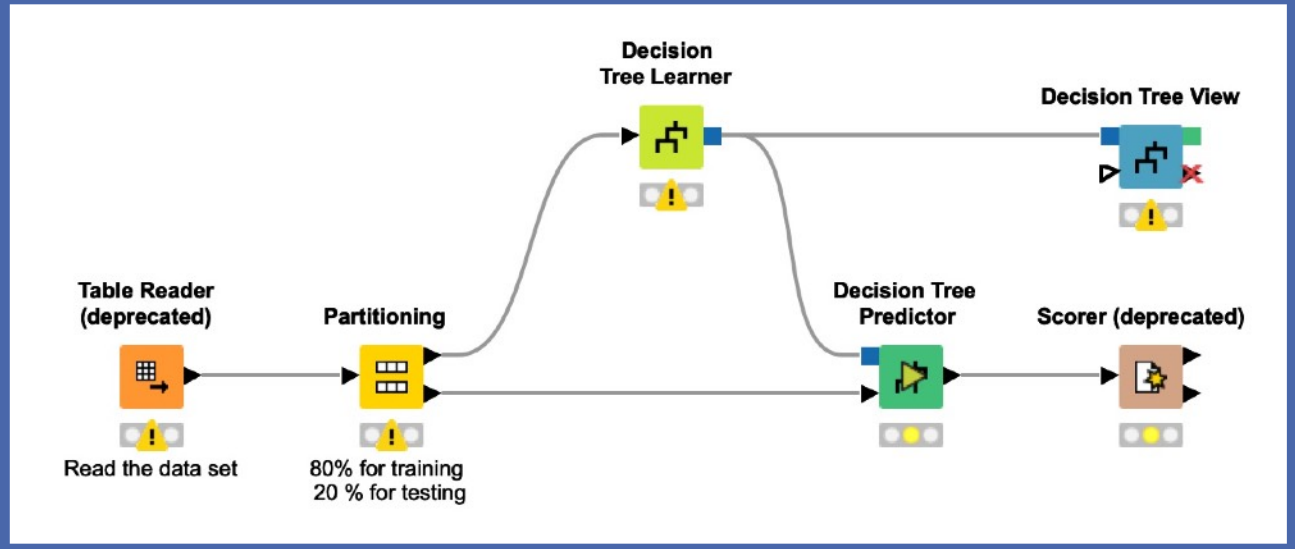
https://hub.knime.com

Getting Started: KNIME Example Server

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



Use Case I: Credit Risk Analysis Using Decision Trees



CBS

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).

(source: https://www.kaggle.com/karanagarwal/bankloans)

Credit Risk Analysis

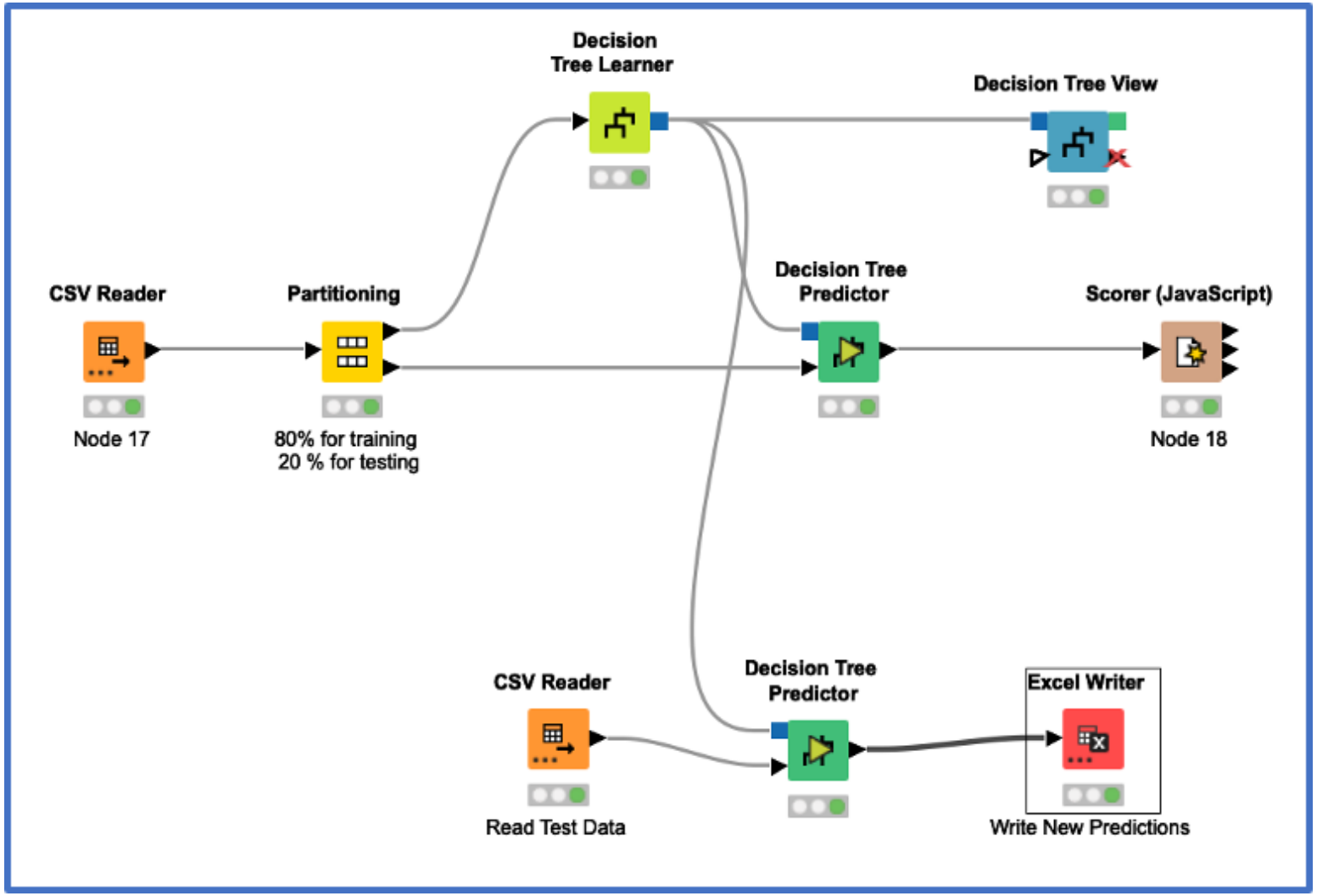
- A lender commonly makes two types of decisions: first, whether to grant credit to a new applicant and second, whether to increase credit limits for existing customers.
- In both cases, we look into a large sample of previous customers with their application details, behavioral patterns, and subsequent 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.

(source: https://www.kaggle.com/karanagarwal/bankloans)

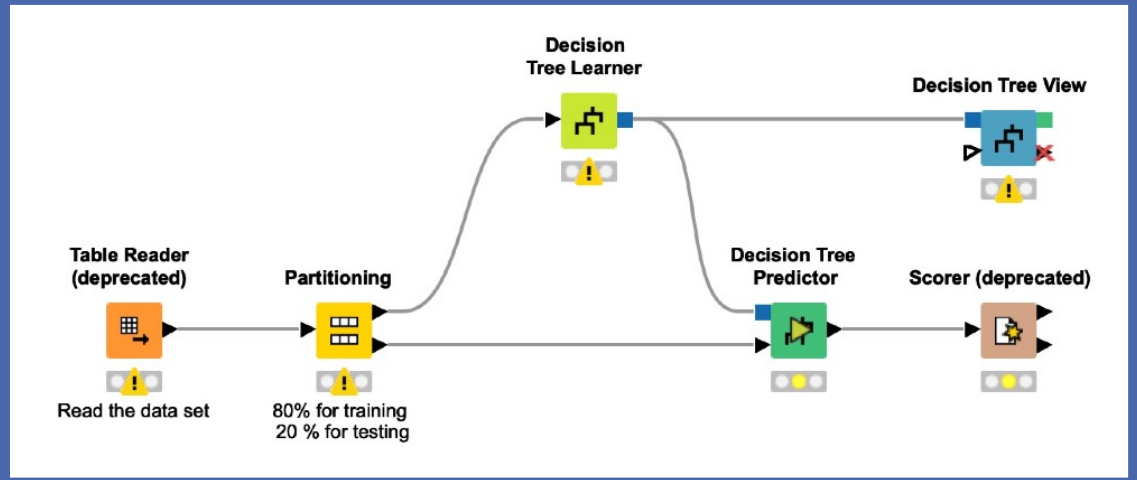
Dataset

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) (**dependent variable**). 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



CBS

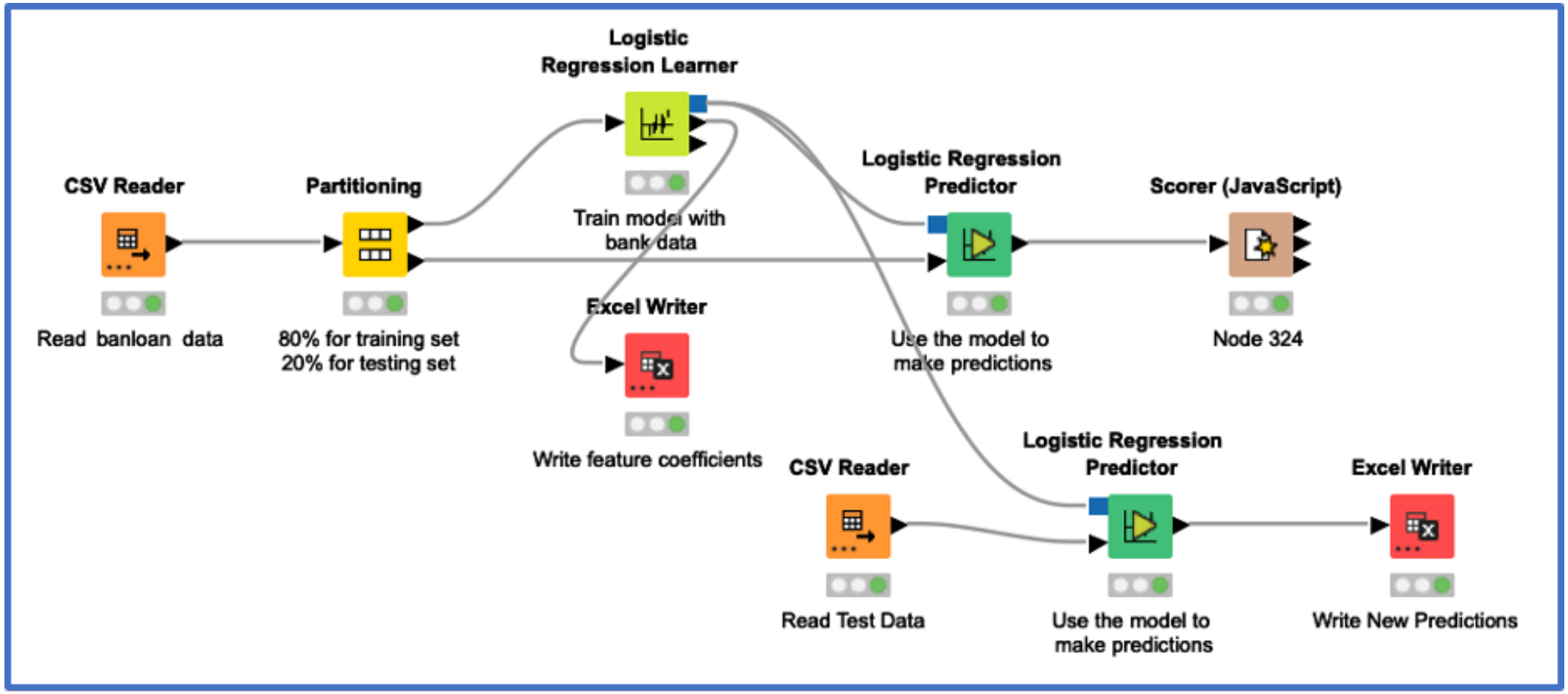
Credit Risk Analysis

- We use the same dataset as Decision Trees

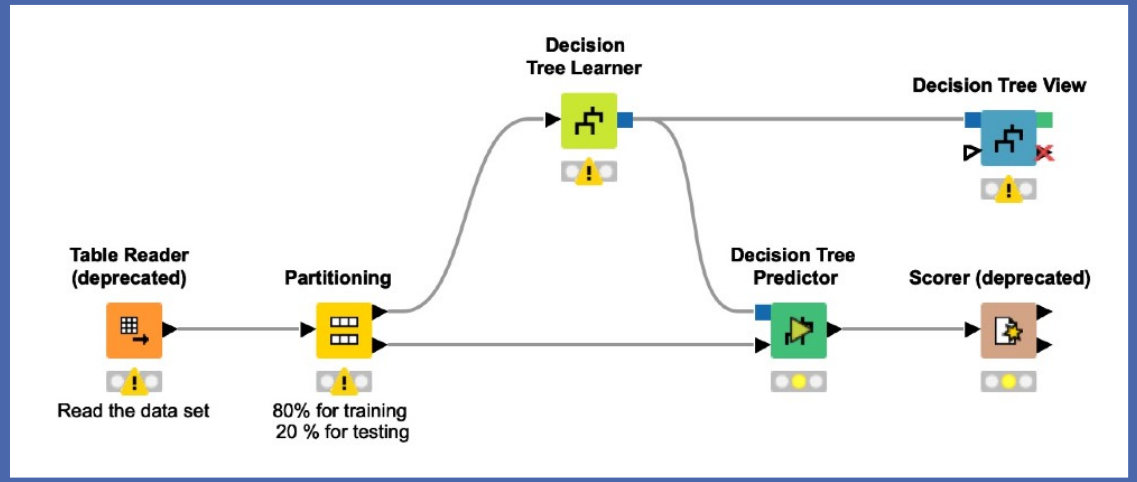
Data Attribute	Explanation
age	Age of Customer
ed	Eductation level of customer
employ	Tenure with current employer (in years)
address	Number of years in same address
income	Customer Income
debtinc	Debt to income ratio
creddebt	Credit to Debt ratio
othdebt	Other debts
Default	Customer defaulted in the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

(source: https://www.kaggle.com/karanagarwal/bankloans)

KNIME Workflow for Logistics Regression



Cross Validation in ML Models

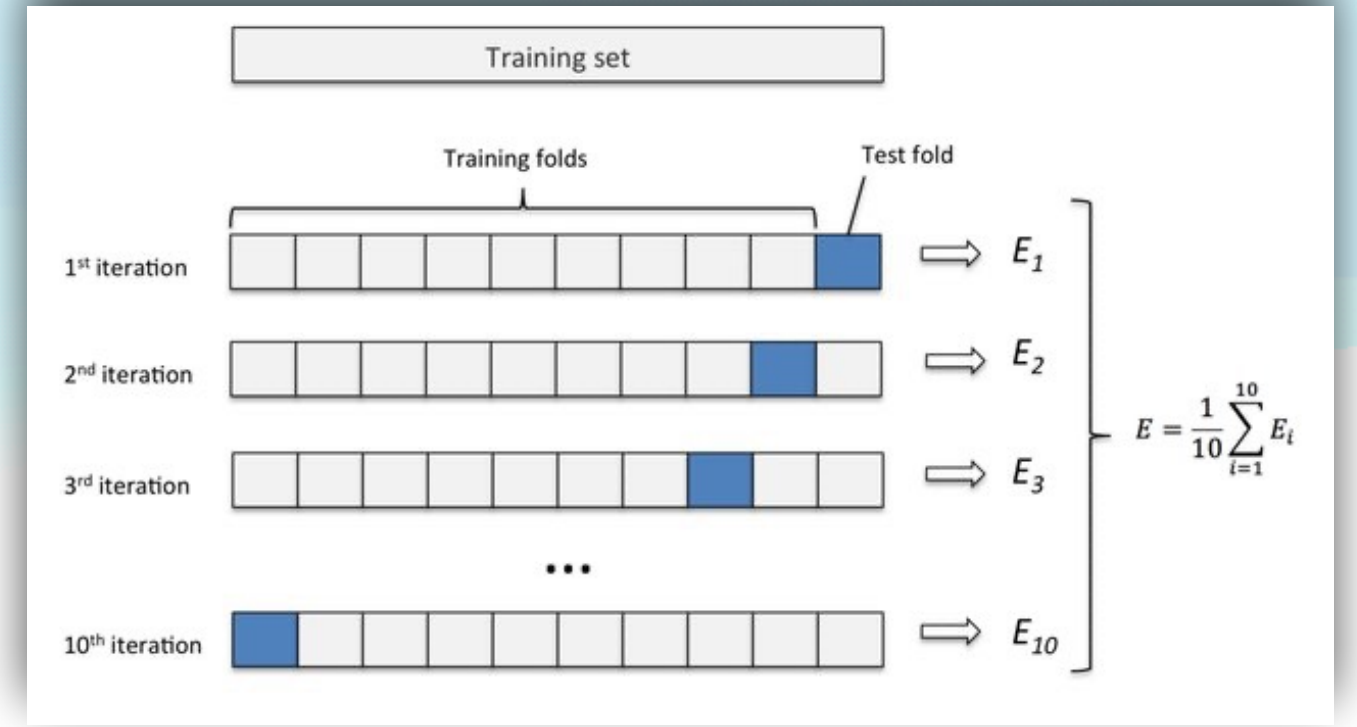


CBS

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)
```


Image from Introduction to Support Vector Machines and Kernel Methods, Ashfaque and Iqbal, 2019

Why Cross-validation?

Confusion Matrix

Scorer View

Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	88	15	85.44%
1 (Actual)	18	19	51.35%
	83.02%	55.88%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
76.43%	23.57%	0.378	107	33

Reset Apply Close

Confusion Matrix

Scorer View

Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	82	19	81.19%
1 (Actual)	23	16	41.03%
	78.10%	45.71%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
70.00%	30.00%	0.229	98	42

Reset Apply Close

Confusion Matrix

Scorer View

Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	75	20	78.95%
1 (Actual)	28	17	37.78%
	72.82%	45.95%	

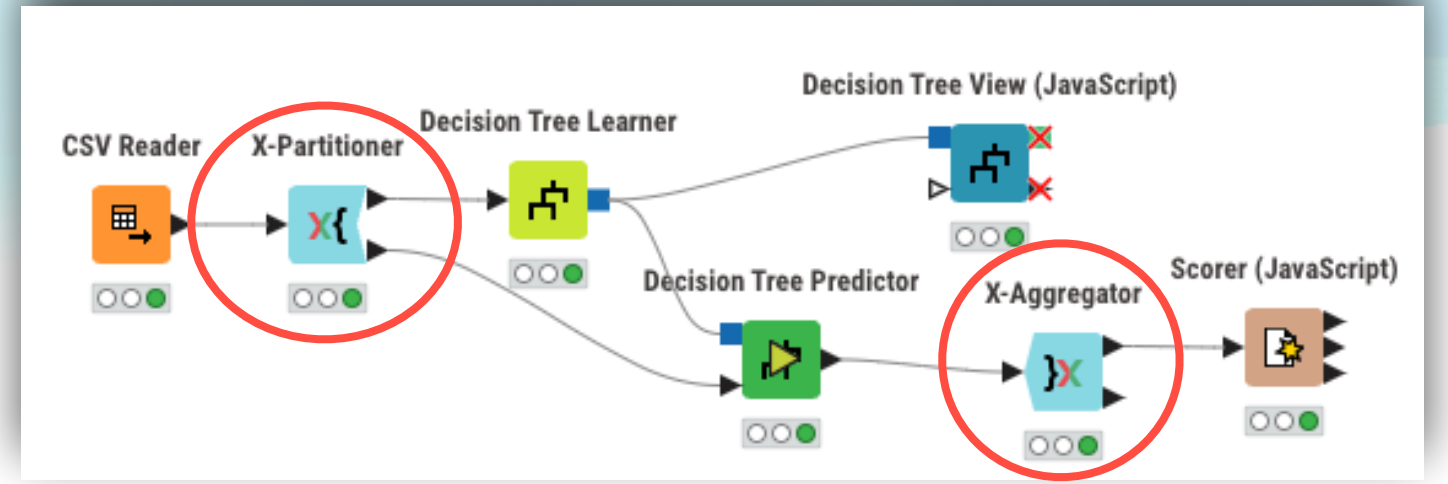
Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
65.71%	34.29%	0.175	92	48

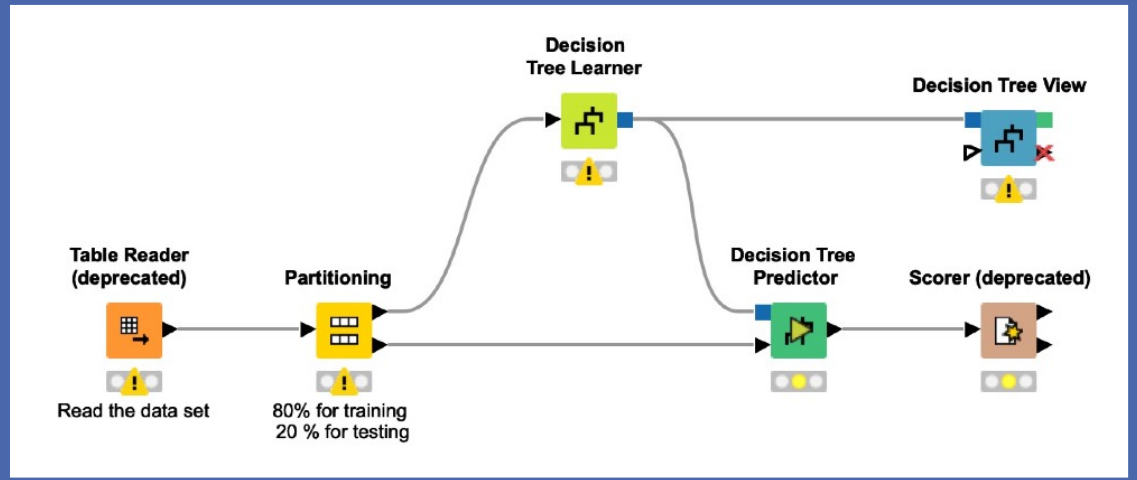
Reset Apply Close

How to do Cross-validation in KNIME?

- Replace the *Partitioning* node with the *x-Partitioner* node
- Set the 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



CBS

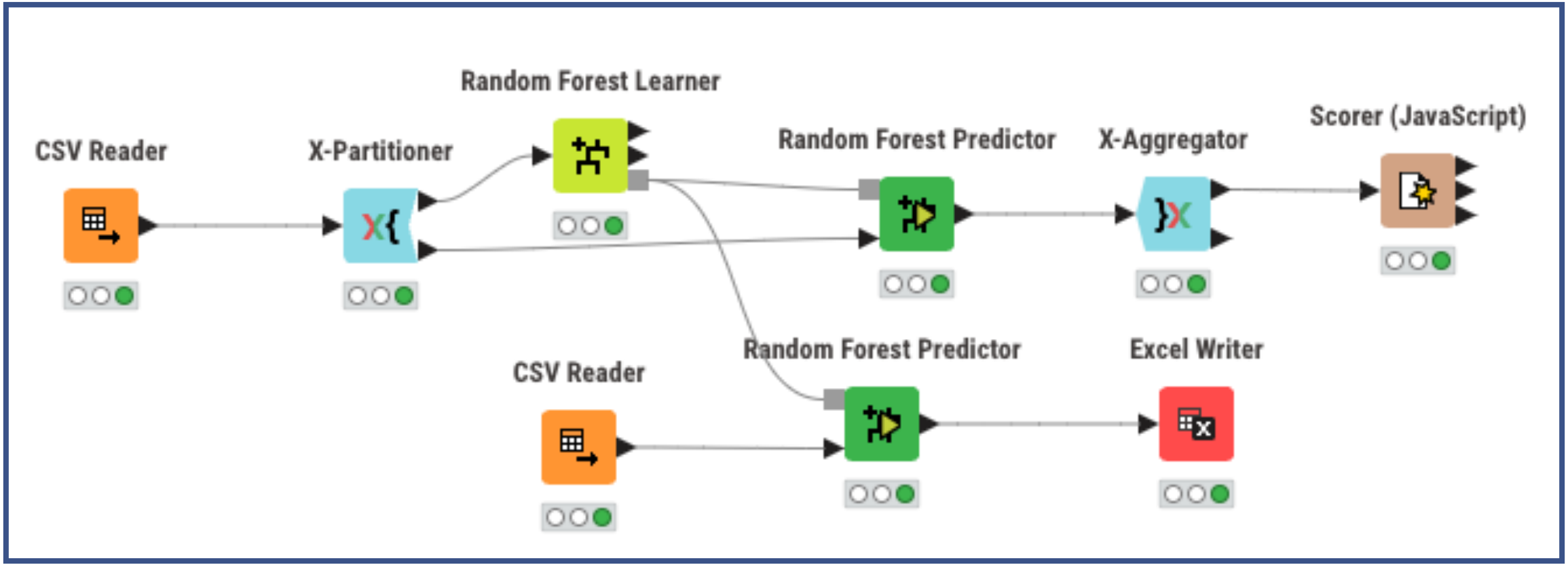
Credit Risk Analysis

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

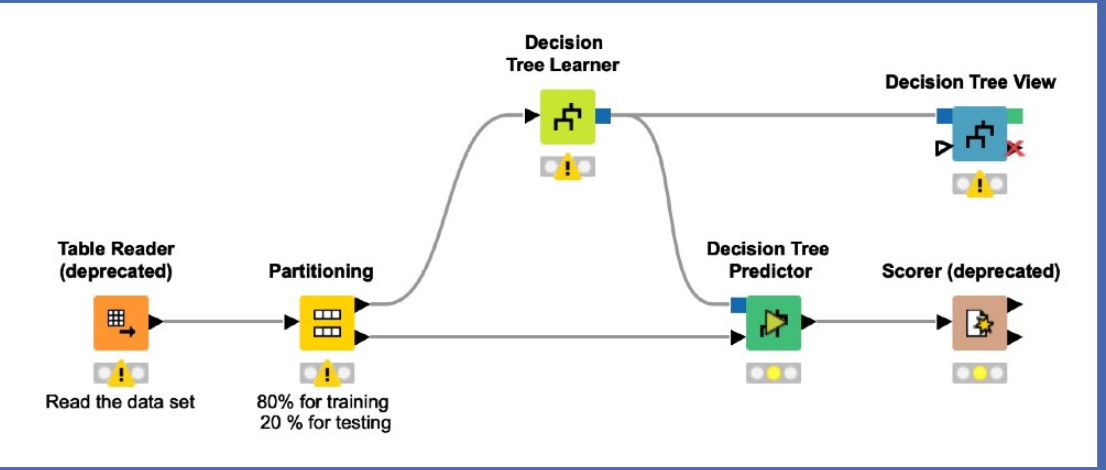
Data Attribute	Explanation
age	Age of Customer
ed	Eductation level of customer
employ	Tenure with current employer (in years)
address	Number of years in same address
income	Customer Income
debtinc	Debt to income ratio
creddebt	Credit to Debt ratio
othdebt	Other debts
Default	Customer defaulted in the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

(source: https://www.kaggle.com/karanagarwal/bankloans)

KNIME Workflow for Random Forests



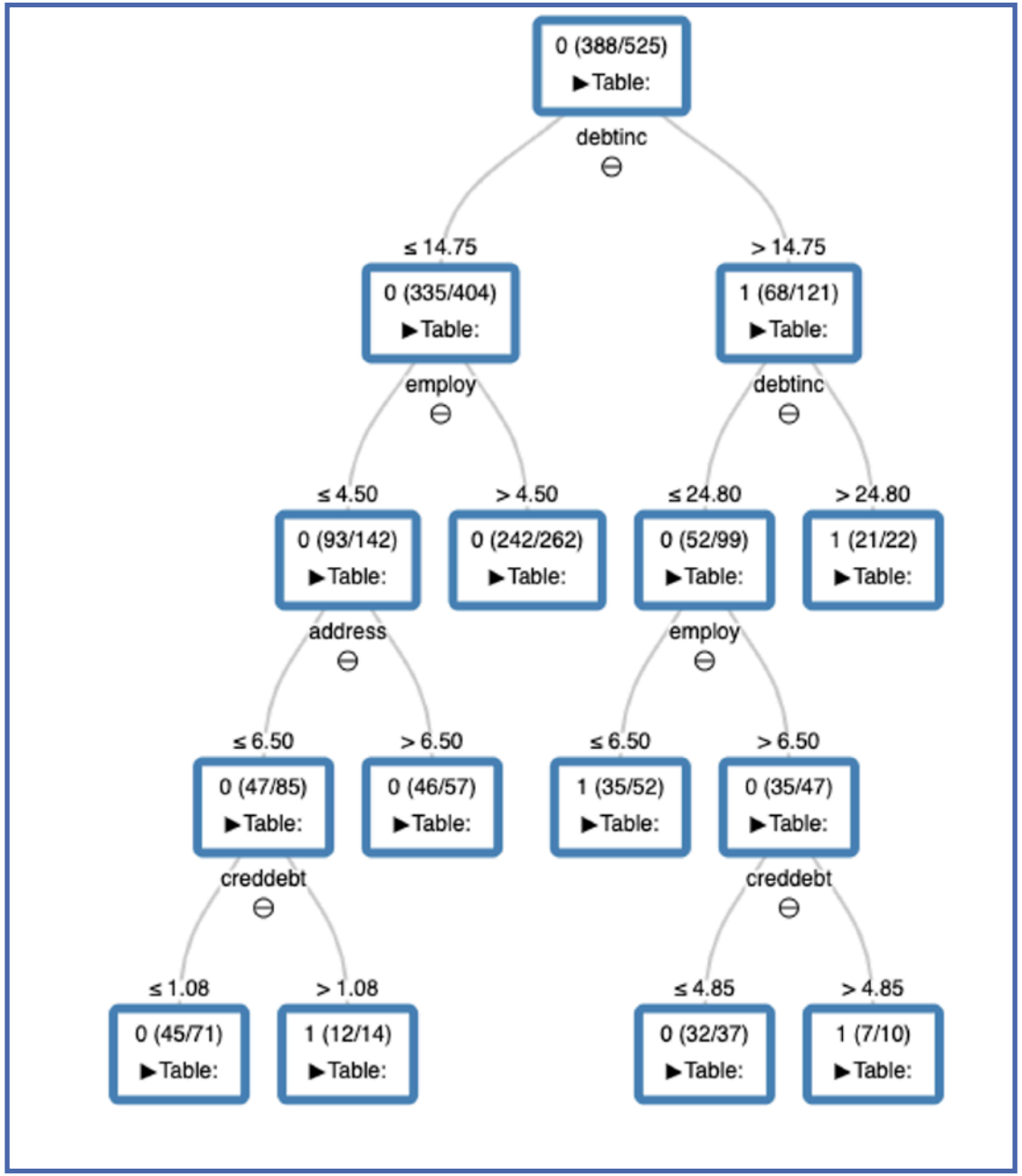
Comparing Results Across Different Models



CBS

Credit Risk Analysis - Decision Trees

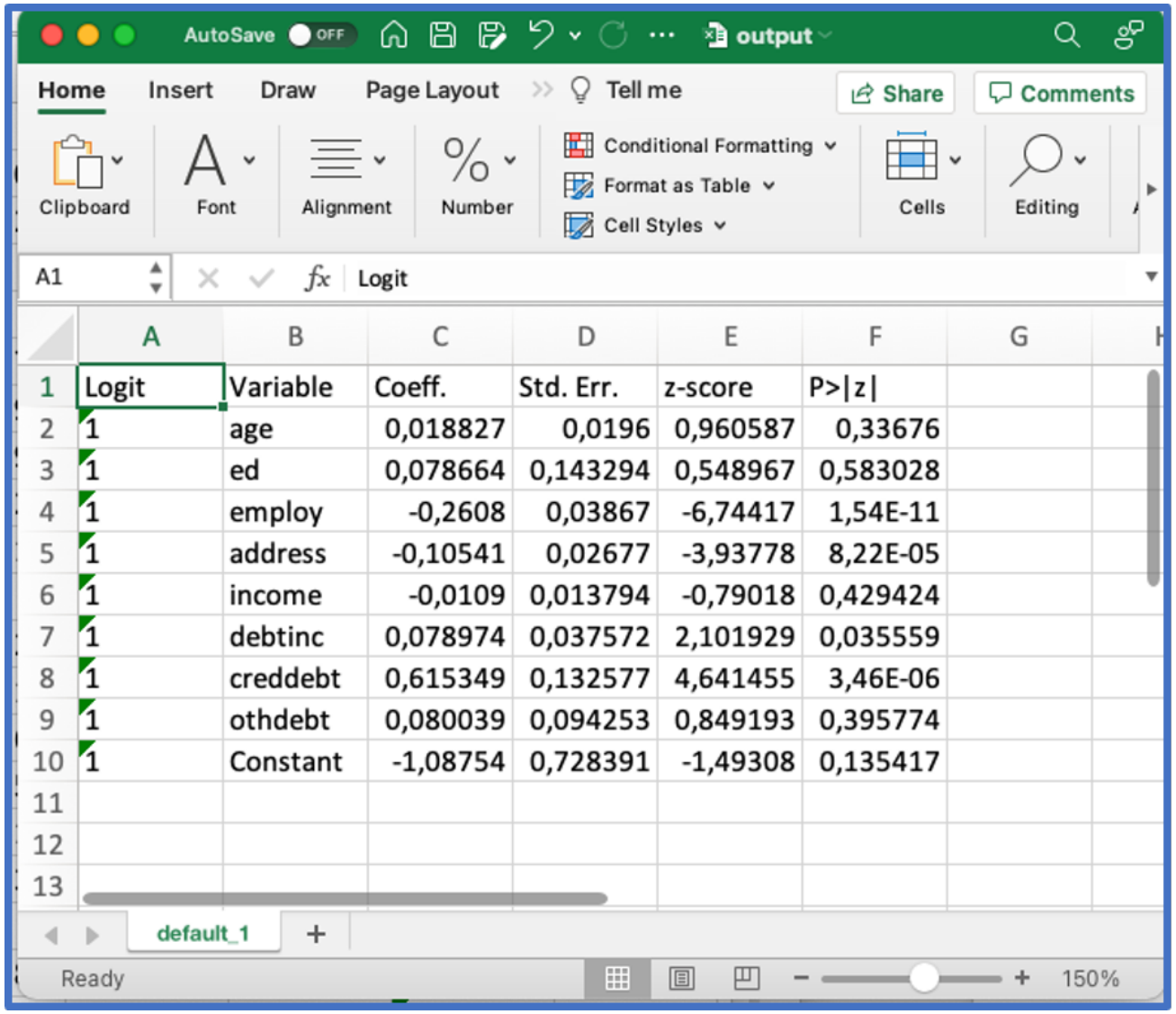
- According to Decision trees, the most important features that can influence whether a customer can become a defaulter or not are as follows.
  - debtinc
  - employ
  - address,
  - creddebt



CBS CENTRE FOR BUSINESS DATA ANALYTICS

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



	A	B	C	D	E	F
1	Logit	Variable	Coeff.	Std. Err.	z-score	P>|z|
2	1	age	0,018827	0,0196	0,960587	0,33676
3	1	ed	0,078664	0,143294	0,548967	0,583028
4	1	employ	-0,2608	0,03867	-6,74417	1,54E-11
5	1	address	-0,10541	0,02677	-3,93778	8,22E-05
6	1	income	-0,0109	0,013794	-0,79018	0,429424
7	1	debtinc	0,078974	0,037572	2,101929	0,035559
8	1	creddebt	0,615349	0,132577	4,641455	3,46E-06
9	1	othdebt	0,080039	0,094253	0,849193	0,395774
10	1	Constant	-1,08754	0,728391	-1,49308	0,135417
11						
12						
13						

https://quantifyinghealth.com/interpret-logistic-regression-coefficients/

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

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

https://raghavamukkamala.github.io/

https://cbsbda.github.io/