University of Minho
School of Engineering

Machine Learning with Knime

Similarity Based Systems

Perfil ML:FA@MiEI/4º ano - 1º Semestre
@MES/2º ano - 1º Semestre

Bruno Fernandes, Paulo Novais

12/12/2019

Contents

- Association Rules Learning
- External tool and If Switch nodes
- Deployment Examples
  - PMML
  - KNIME workflows from the command line
- Hands On

Association Rules Learning

Association Rule Learning is a ML method for discovering relations between variables! It aims at finding frequent patterns, associations or correlations among sets of data.

- Extremely useful for Recommender Systems, market basket analysis and fraud detection
- Typically does not consider the order of items
- Rules do not extract an individual's preference but, instead, look for relationships among data

Association Rules Learning

Rules in the form:

Antecedent -> Consequent

{Pulp Fiction, Silence of the Lambs} -> {The Shawshank Redemption}

Itemset: {Pulp Fiction, Silence of the Lambs, The Shawshank Redemption}

Association Rules Learning

{Pulp Fiction, Silence of the Lambs} -> {The Shawshank Redemption}

- How to boost the number of views of *The Shawshank Redemption*?

…

Association Rules Learning

{Pulp Fiction, Silence of the Lambs} -> {The Shawshank Redemption}

- What happens if we remove *Pulp Fiction* from the movie catalogue?

…

Association Rules Learning

Properties and Metrics:

- Support
  - Gives an idea of how frequent an itemset is in all existing transactions
  - Helps identifying rules worth considering. For example, to consider itemsets that occur, at least, 100 times out of a total of 10000 transactions, support = 0.01

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

- Confidence
  - An indication of how often a rule has been found to be true, i.e., the proportion of transactions containing X which also contain also Y

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

Association Rules Learning

Properties and Metrics:

- Support
- Confidence

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

User Id	Seen Movies
1	A, C
2	A, D
3	A, B, E, D
4	B, C, E
5	A, D, E, B

$$s(\{A\} \rightarrow \{B\}) = \frac{2}{5} \qquad c(\{A\} \rightarrow \{B\}) = \frac{1}{2}$$

$$s(\{A\} \rightarrow \{D\}) = \frac{3}{5} \qquad c(\{A\} \rightarrow \{D\}) = \frac{3}{4}$$

$$s(\{B, E\} \rightarrow \{D\}) = \frac{2}{5} \qquad c(\{B, E\} \rightarrow \{D\}) = \frac{2}{3}$$

Association Rules Learning

Properties and Metrics:

- Lift
  - Measures how much better the rule is at predicting the presence of Y compared to just relying on the raw probability of Y in the dataset
  - If lift < 1 than items are negatively correlated, i.e., the items are substitute to each other. Items have negative effect on each other!
  - If lift > 1 than items are positively correlated, i.e., tells us the degree to which those two occurrences are dependent on one another (useful for prediction!)
  - If lift = 1 than items are independent (no rule can be drawn involving those two items)

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y}$$

Association Rules Learning

Properties and Metrics:

- Lift $$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y}$$

User Id	Seen Movies
1	A, C
2	A, D
3	A, B, E, D
4	B, C, E
5	A, D, E, B

$$Lift(\{A\} \rightarrow \{B\}) = \frac{2}{4 * (3/5)} = 0.83$$

$$Lift(\{A\} \rightarrow \{D\}) = \frac{3}{4 * (3/5)} = 1.25$$

$$Lift(\{B, E\} \rightarrow \{D\}) = \frac{2}{3 * (3/5)} = 1.11$$

Association Rules Learning

Given a dataset D, the goal is to find all rules that have:

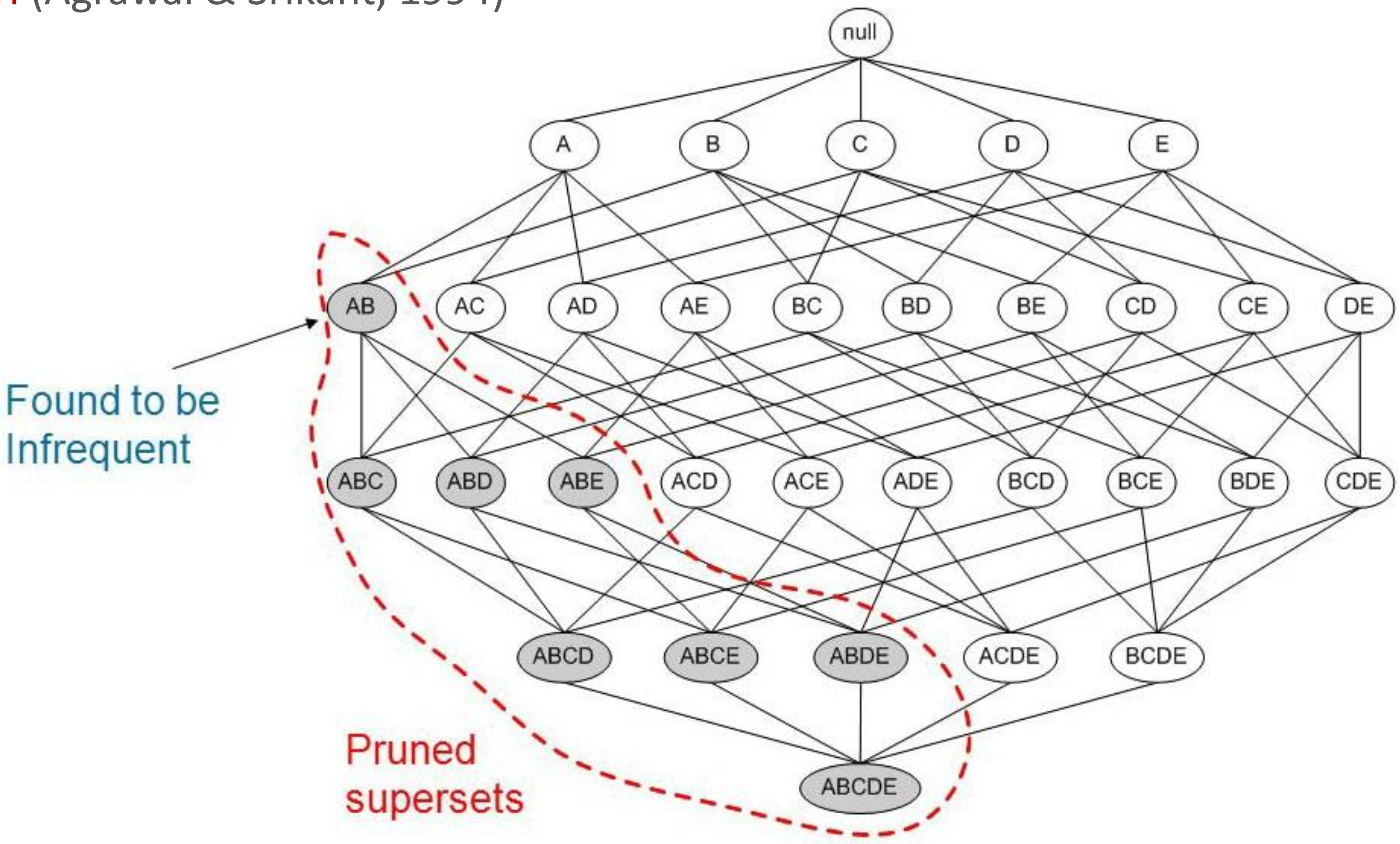
- Support ≥ minsup (threshold) <- Usually the first step!
- Confidence ≥ minconf (threshold) <- Usually, the second!

Most common algorithm - Apriori (Agrawal & Srikant, 1994):

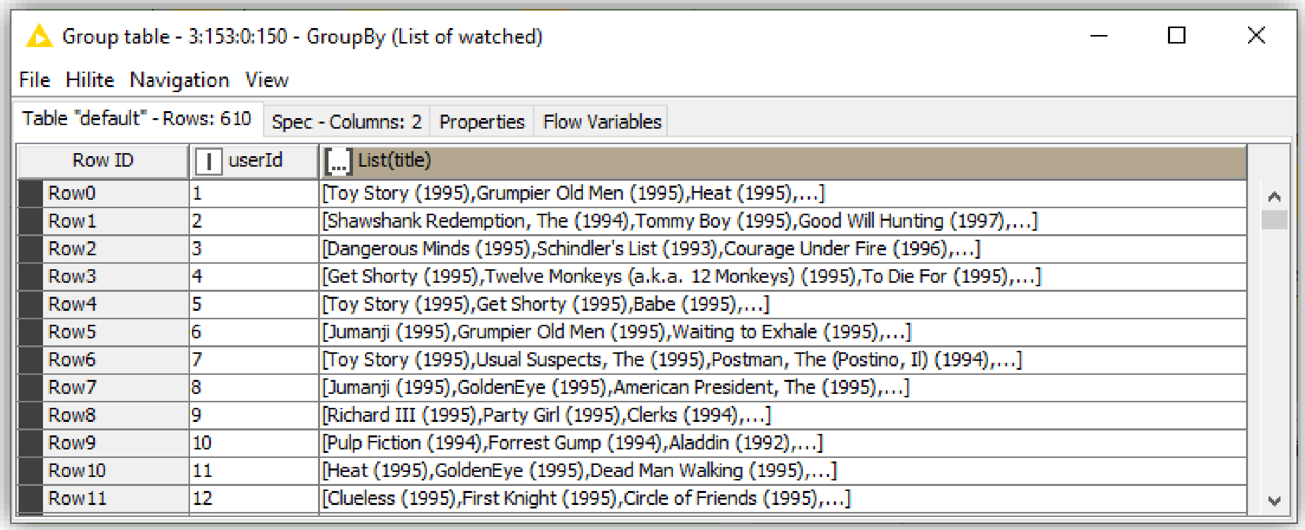
- If a itemset is frequent then all its sub-itemsets should be frequent as well
- If {Pulp Fiction, Silence of the Lambs, The Shawshank Redemption} is frequent than {Pulp Fiction, Silence of the Lambs} must also be frequent!
- If an itemset is infrequent, then all of its supersets must also be infrequent

Association Rules Learning

Apriori (Agrawal & Srikant, 1994)



Association Rules Learning



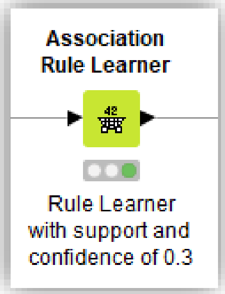
Group table - 3:153:0:150 - GroupBy (List of watched)

File Hilite Navigation View

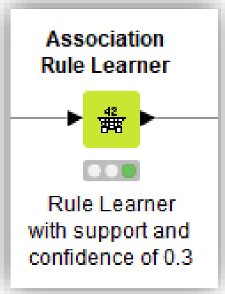
Table "default" - Rows: 610 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	userId	List(title)
Row0	1	[Toy Story (1995),Grumpier Old Men (1995),Heat (1995),...]
Row1	2	[Shawshank Redemption, The (1994),Tommy Boy (1995),Good Will Hunting (1997),...]
Row2	3	[Dangerous Minds (1995),Schindler's List (1993),Courage Under Fire (1996),...]
Row3	4	[Get Shorty (1995),Twelve Monkeys (a.k.a. 12 Monkeys) (1995),To Die For (1995),...]
Row4	5	[Toy Story (1995),Get Shorty (1995),Babe (1995),...]
Row5	6	[Jumanji (1995),Grumpier Old Men (1995),Waiting to Exhale (1995),...]
Row6	7	[Toy Story (1995),Usual Suspects, The (1995),Postman, The (Postino, Il) (1994),...]
Row7	8	[Jumanji (1995),GoldenEye (1995),American President, The (1995),...]
Row8	9	[Richard III (1995),Party Girl (1995),Clerks (1994),...]
Row9	10	[Pulp Fiction (1994),Forrest Gump (1994),Aladdin (1992),...]
Row10	11	[Heat (1995),GoldenEye (1995),Dead Man Walking (1995),...]
Row11	12	[Clueless (1995),First Knight (1995),Circle of Friends (1995),...]

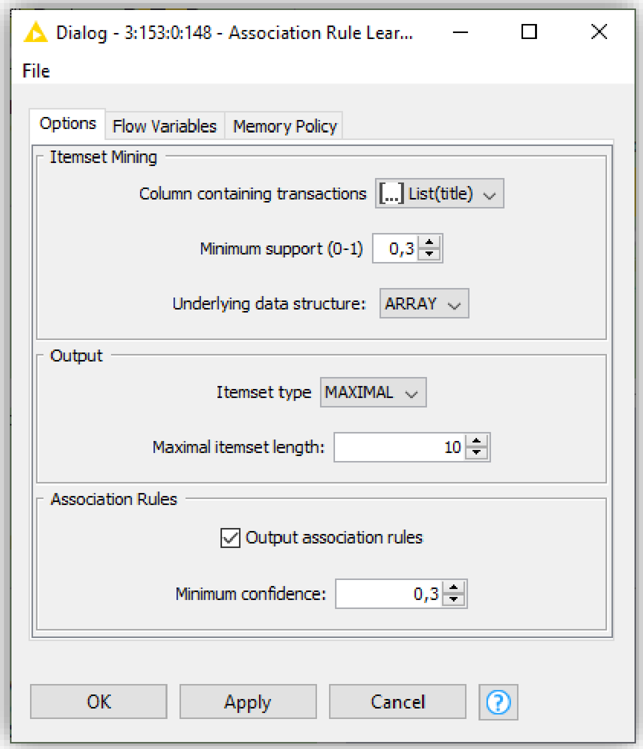
Pass a list of *something* as input to the Association Rule Learner node!



Association Rules Learning



Rule Learner with support and confidence of 0.3



Itemset Mining

Column containing transactions [...] List(title)

Minimum support (0-1) 0,3

Underlying data structure: ARRAY

Output

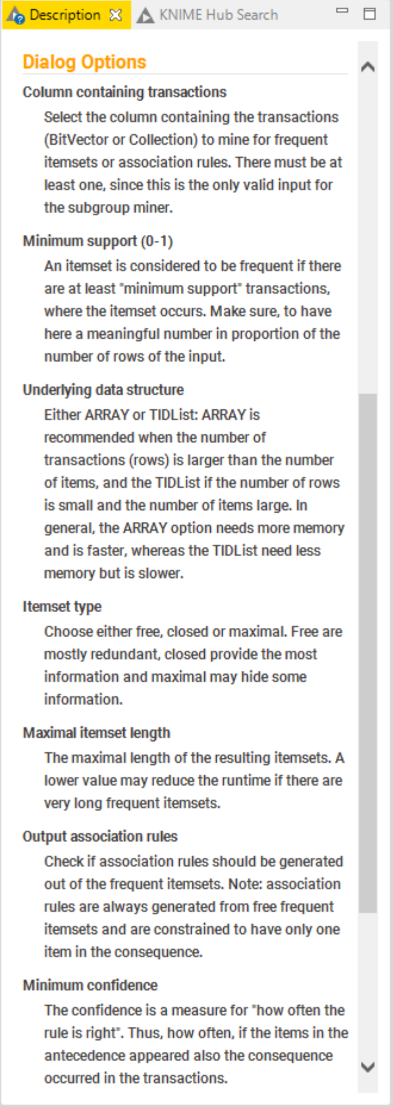
Itemset type MAXIMAL

Maximal itemset length: 10

Association Rules

☑ Output association rules

Minimum confidence: 0,3



Dialog Options

Column containing transactions

Select the column containing the transactions (BitVector or Collection) to mine for frequent itemsets or association rules. There must be at least one, since this is the only valid input for the subgroup miner.

Minimum support (0-1)

An itemset is considered to be frequent if there are at least "minimum support" transactions, where the itemset occurs. Make sure, to have here a meaningful number in proportion of the number of rows of the input.

Underlying data structure

Either ARRAY or TIDList: ARRAY is recommended when the number of transactions (rows) is larger than the number of items, and the TIDList if the number of rows is small and the number of items large. In general, the ARRAY option needs more memory and is faster, whereas the TIDList need less memory but is slower.

Itemset type

Choose either free, closed or maximal. Free are mostly redundant, closed provide the most information and maximal may hide some information.

Maximal itemset length

The maximal length of the resulting itemsets. A lower value may reduce the runtime if there are very long frequent itemsets.

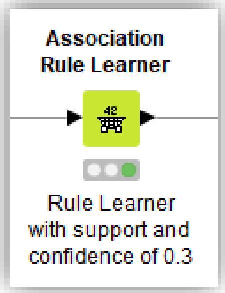
Output association rules

Check if association rules should be generated out of the frequent itemsets. Note: association rules are always generated from free frequent itemsets and are constrained to have only one item in the consequence.

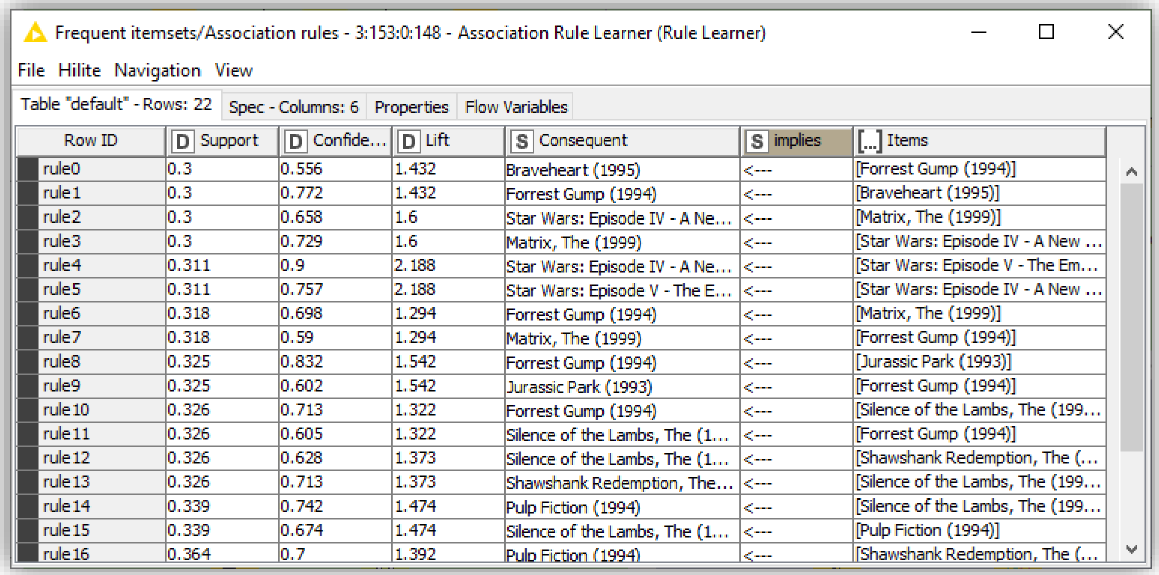
Minimum confidence

The confidence is a measure for "how often the rule is right". Thus, how often, if the items in the antecedence appeared also the consequence occurred in the transactions.

Association Rules Learning

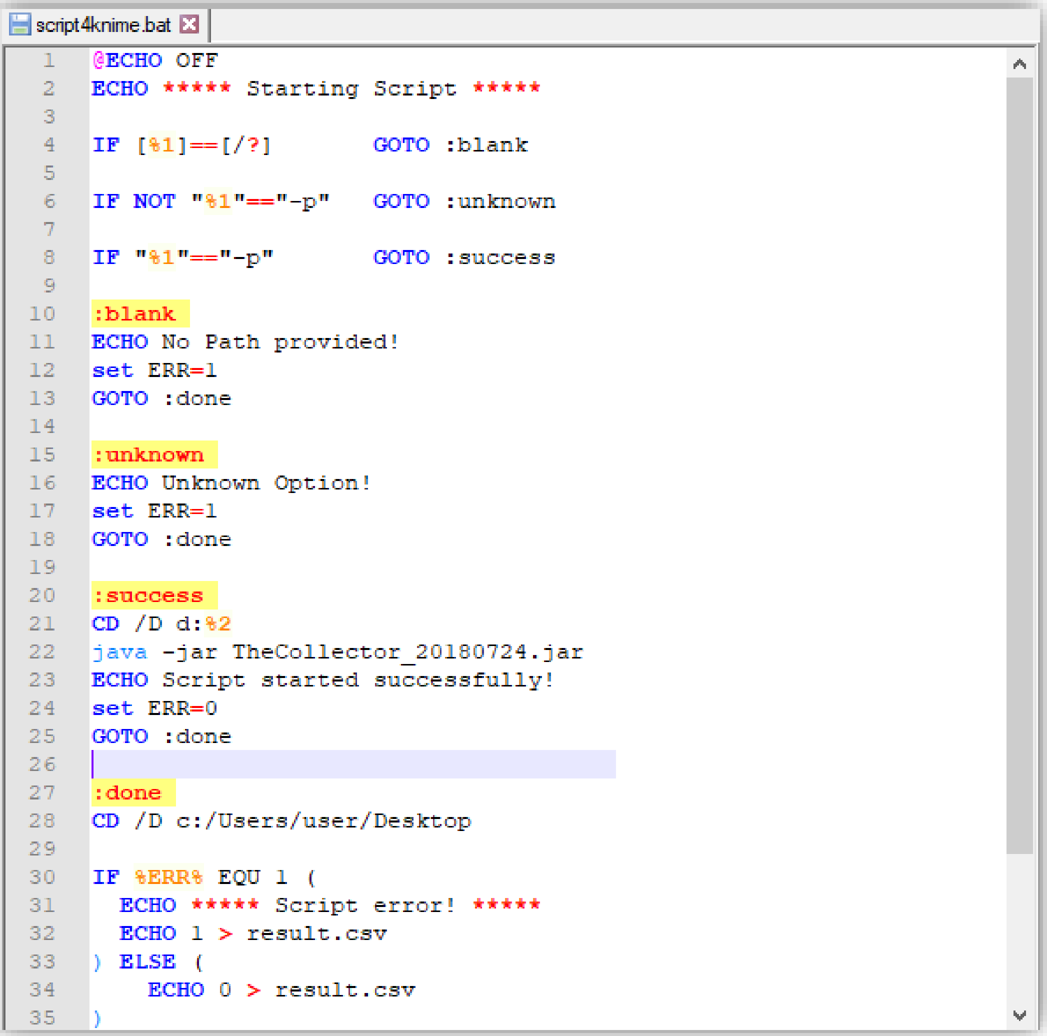


Rule Learner with support and confidence of 0.3


Frequent itemsets/Association rules - 3:153:0:148 - Association Rule Learner (Rule Learner)
File Hilite Navigation View
Table "default" - Rows: 22
Spec - Columns: 6
Properties
Flow Variables

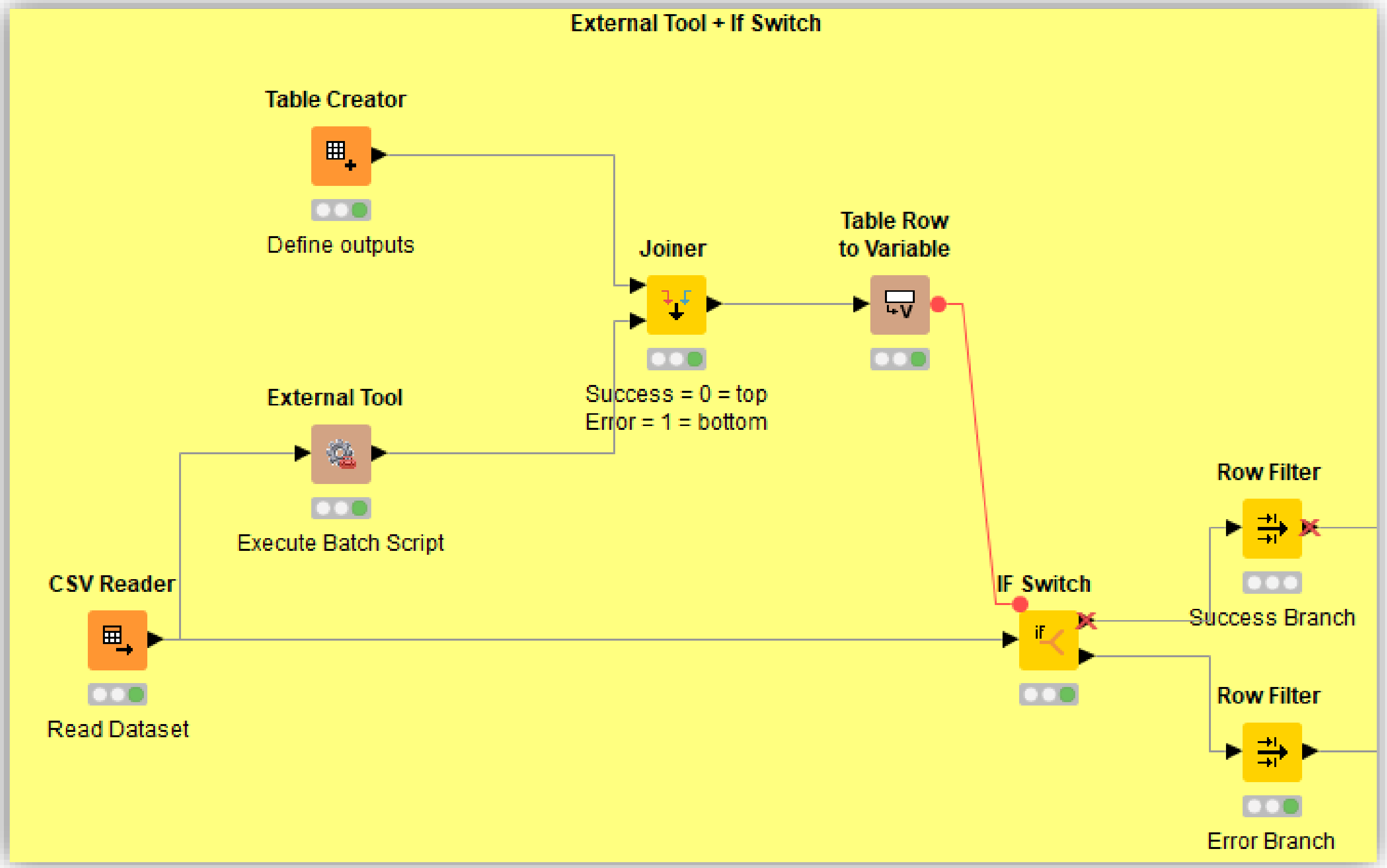
Row ID	D Support	D Confide...	D Lift	S Consequent	S implies	[...] Items
rule0	0.3	0.556	1.432	Braveheart (1995)	<---	[Forrest Gump (1994)]
rule1	0.3	0.772	1.432	Forrest Gump (1994)	<---	[Braveheart (1995)]
rule2	0.3	0.658	1.6	Star Wars: Episode IV - A Ne...	<---	[Matrix, The (1999)]
rule3	0.3	0.729	1.6	Matrix, The (1999)	<---	[Star Wars: Episode IV - A New ...
rule4	0.311	0.9	2.188	Star Wars: Episode IV - A Ne...	<---	[Star Wars: Episode V - The Em...
rule5	0.311	0.757	2.188	Star Wars: Episode V - The E...	<---	[Star Wars: Episode IV - A New ...
rule6	0.318	0.698	1.294	Forrest Gump (1994)	<---	[Matrix, The (1999)]
rule7	0.318	0.59	1.294	Matrix, The (1999)	<---	[Forrest Gump (1994)]
rule8	0.325	0.832	1.542	Forrest Gump (1994)	<---	[Jurassic Park (1993)]
rule9	0.325	0.602	1.542	Jurassic Park (1993)	<---	[Forrest Gump (1994)]
rule10	0.326	0.713	1.322	Forrest Gump (1994)	<---	[Silence of the Lambs, The (199...
rule11	0.326	0.605	1.322	Silence of the Lambs, The (1...	<---	[Forrest Gump (1994)]
rule12	0.326	0.628	1.373	Silence of the Lambs, The (1...	<---	[Shawshank Redemption, The (...
rule13	0.326	0.713	1.373	Shawshank Redemption, The...	<---	[Silence of the Lambs, The (199...
rule14	0.339	0.742	1.474	Pulp Fiction (1994)	<---	[Silence of the Lambs, The (199...
rule15	0.339	0.674	1.474	Silence of the Lambs, The (1...	<---	[Pulp Fiction (1994)]
rule16	0.364	0.7	1.392	Pulp Fiction (1994)	<---	[Shawshank Redemption, The (...

External Tool Node & If Switch
The Script to Execute



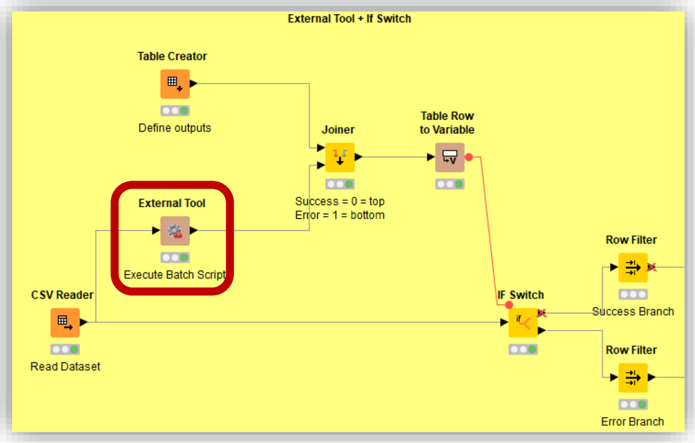
```
@ECHO OFF
ECHO ***** Starting Script *****

IF [%1]==[/?]        GOTO :blank

IF NOT "%1"=="-p"    GOTO :unknown

IF "%1"=="-p"        GOTO :success

:blank
ECHO No Path provided!
set ERR=1
GOTO :done

:unknown
ECHO Unknown Option!
set ERR=1
GOTO :done

:success
CD /D d:%2
java -jar TheCollector_20180724.jar
ECHO Script started successfully!
set ERR=0
GOTO :done

:done
CD /D c:/Users/user/Desktop

IF %ERR% EQU 1 (
  ECHO ***** Script error! *****
  ECHO 1 > result.csv
) ELSE (
    ECHO 0 > result.csv
)
```

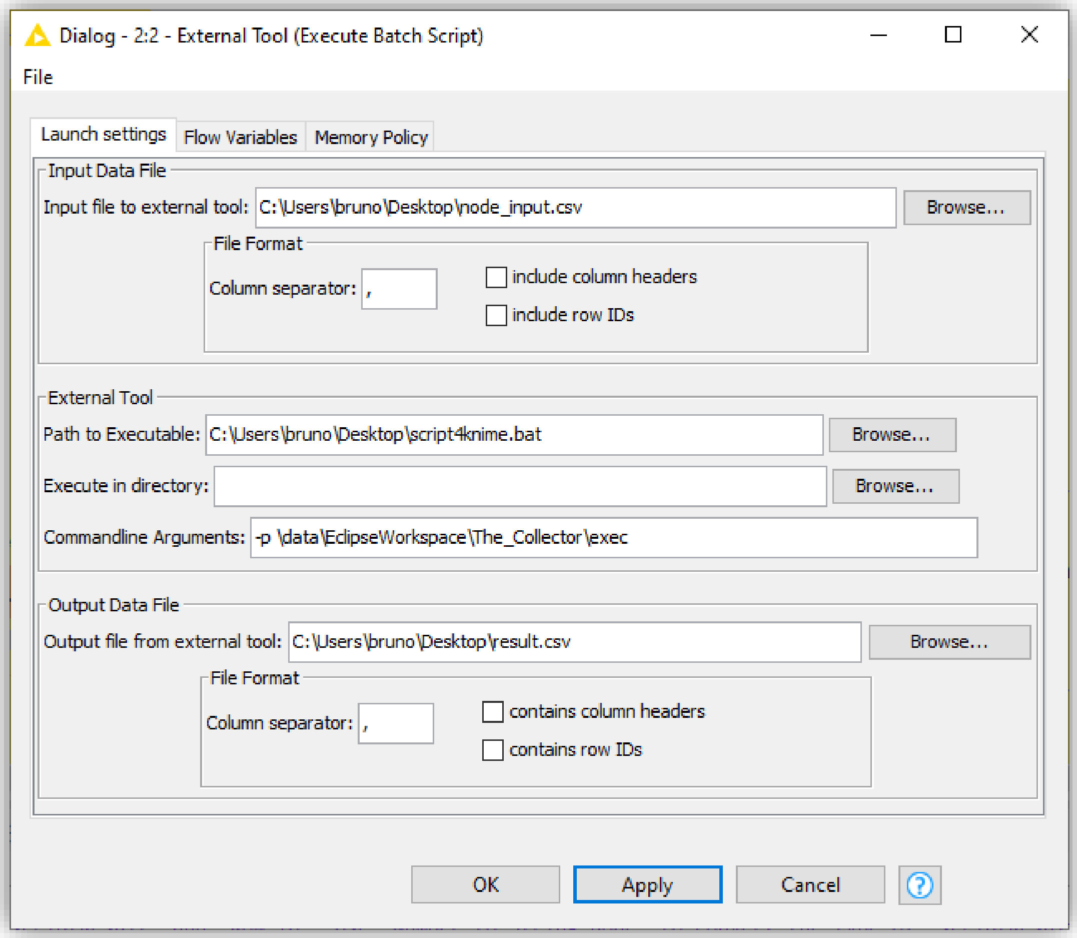
External Tool Node & If Switch
A Workflow



Install KNIME External Tool Support

External Tool Node & If Switch
External Tool Node Configuration





Input Data File

Input file to external tool: C:\Users\bruno\Desktop\node_input.csv [Browse...]

File Format

Column separator: ,

□ include column headers

□ include row IDs

External Tool

Path to Executable: C:\Users\bruno\Desktop\script4knime.bat [Browse...]

Execute in directory: [Browse...]

Commandline Arguments: -p \data\EclipseWorkspace\The_Collector\exec

Output Data File

Output file from external tool: C:\Users\bruno\Desktop\result.csv [Browse...]

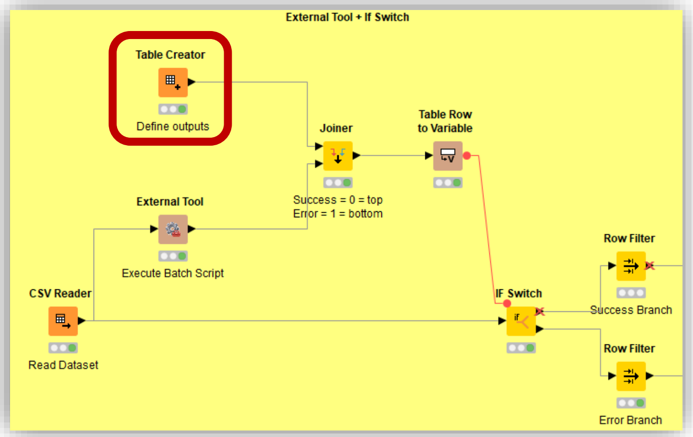
File Format

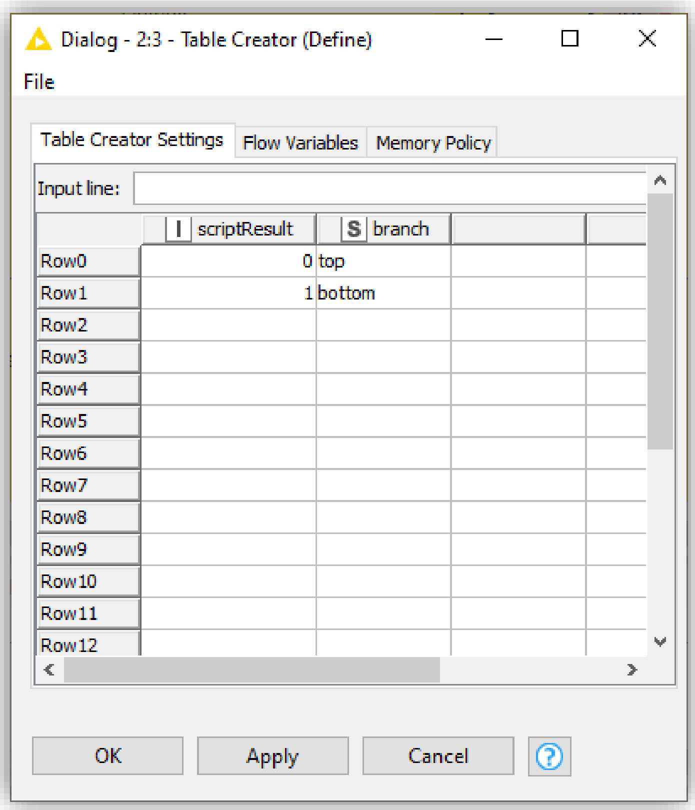
Column separator: ,

□ contains column headers

□ contains row IDs

External Tool Node & If Switch Branch Definition

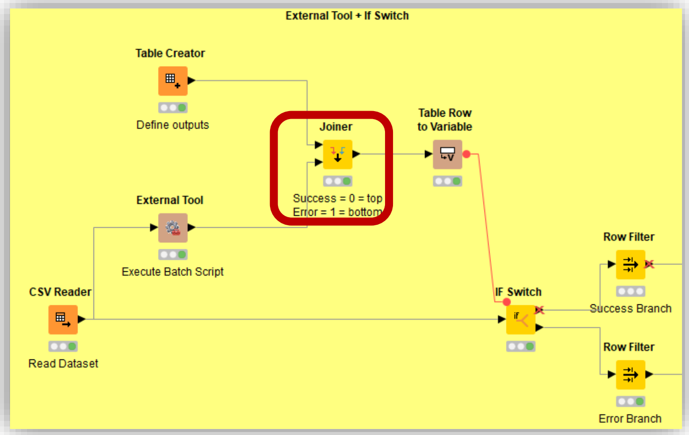




Input line:

	I scriptResult	S branch		
Row0	0	top		
Row1	1	bottom		
Row2				
Row3				
Row4				
Row5				
Row6				
Row7				
Row8				
Row9				
Row10				
Row11				
Row12				

External Tool Node & If Switch
Joiner Configuration



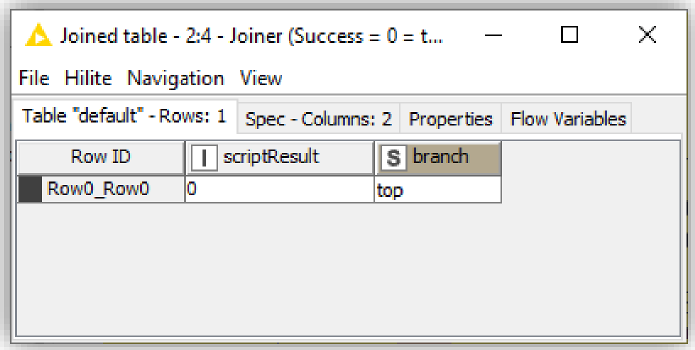
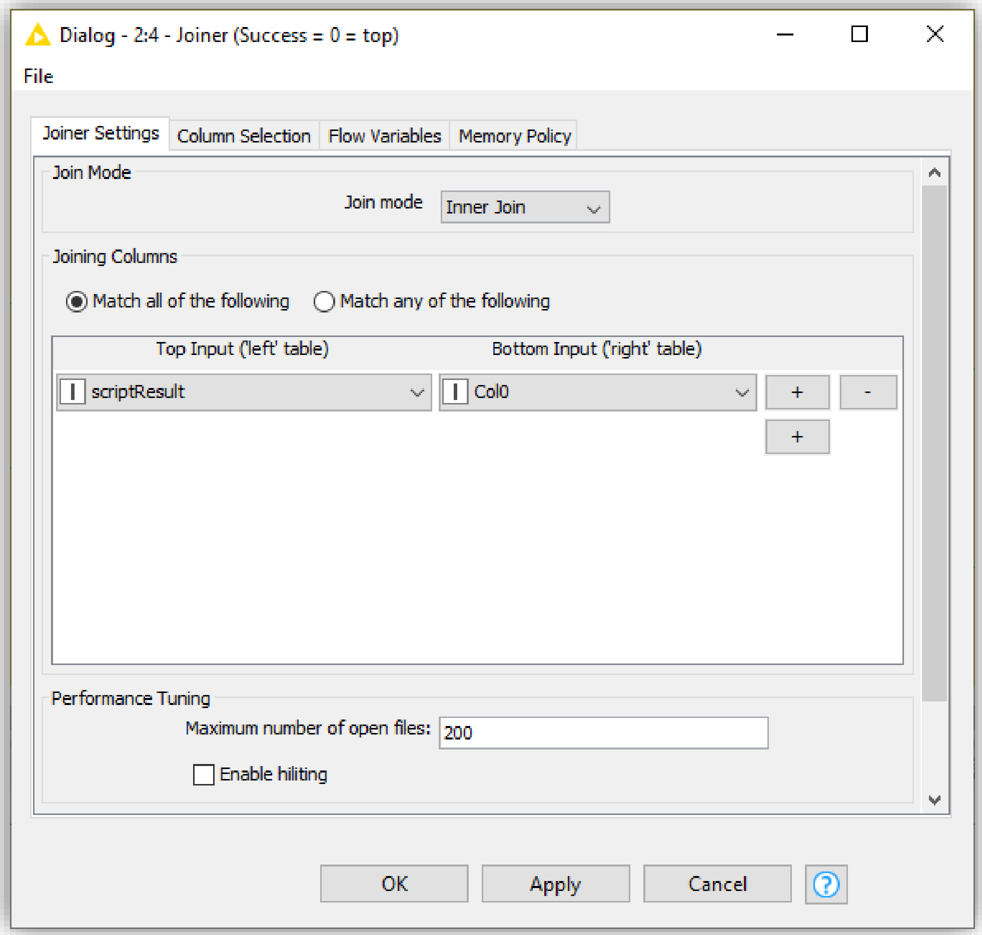

Joined table - 2:4 - Joiner (Success = 0 = t...

Table "default" - Rows: 1 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	I scriptResult	S branch
Row0_Row0	0	top


Dialog - 2:4 - Joiner (Success = 0 = top)

Joiner Settings | Column Selection | Flow Variables | Memory Policy

Join Mode

Join mode: Inner Join

Joining Columns

● Match all of the following ○ Match any of the following

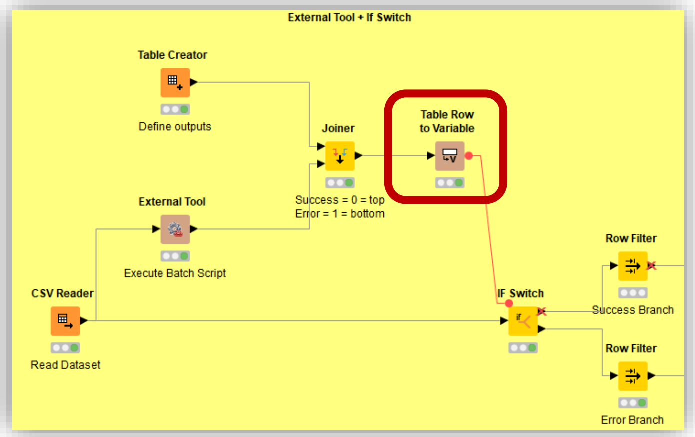
Top Input ('left' table)	Bottom Input ('right' table)
I scriptResult	I Col0

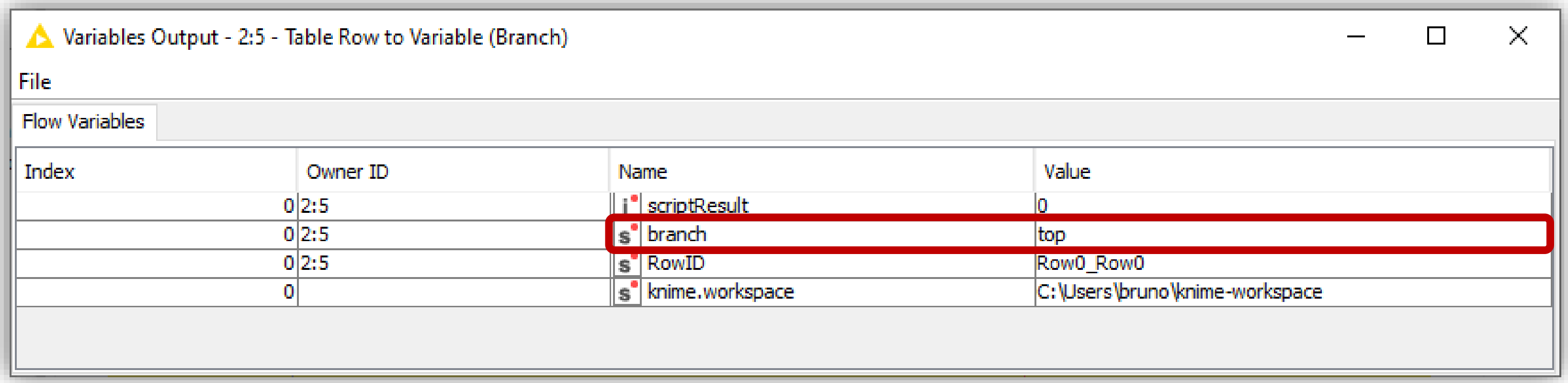
Performance Tuning

Maximum number of open files: 200

□ Enable hiliting

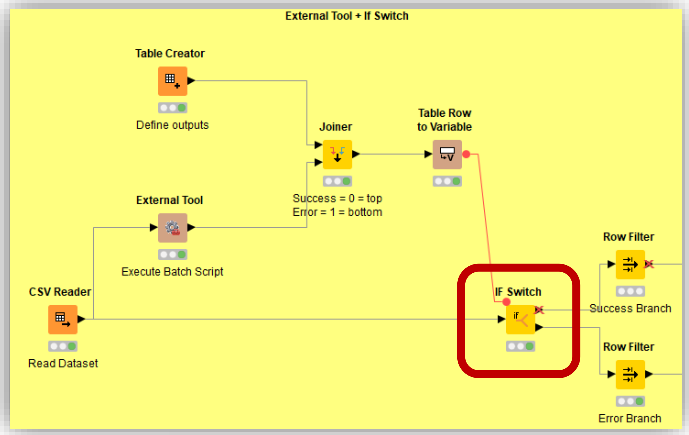
External Tool Node & If Switch
Table Row to Variable Configuration

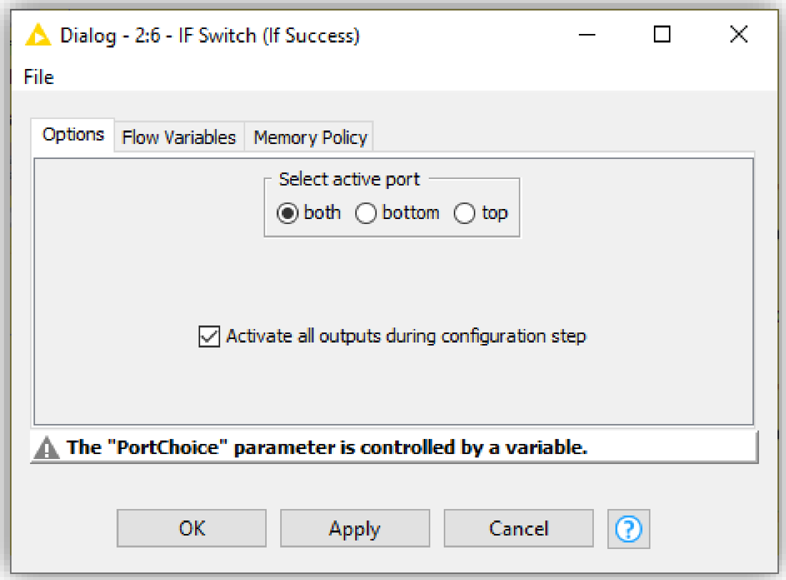




Index	Owner ID	Name	Value
0	2:5	scriptResult	0
0	2:5	branch	top
0	2:5	RowID	Row0_Row0
0		knime.workspace	C:\Users\bruno\knime-workspace

External Tool Node & If Switch
If Switch Node Configuration




Dialog - 2:6 - IF Switch (If Success)

File

Options | Flow Variables | Memory Policy

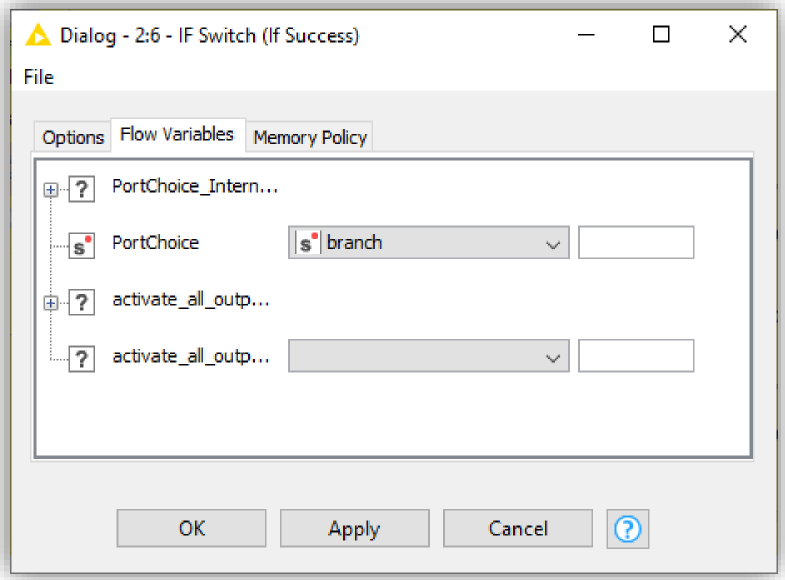
Select active port

● both ○ bottom ○ top

☑ Activate all outputs during configuration step

⚠ **The "PortChoice" parameter is controlled by a variable.**

OK | Apply | Cancel | ?


Dialog - 2:6 - IF Switch (If Success)

File

Options | Flow Variables | Memory Policy

PortChoice_Intern...

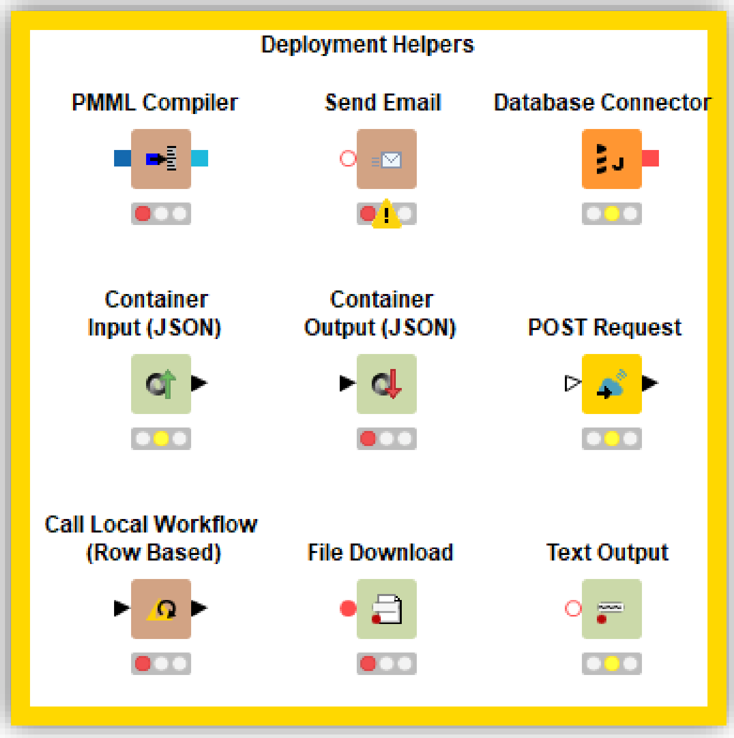
PortChoice: branch

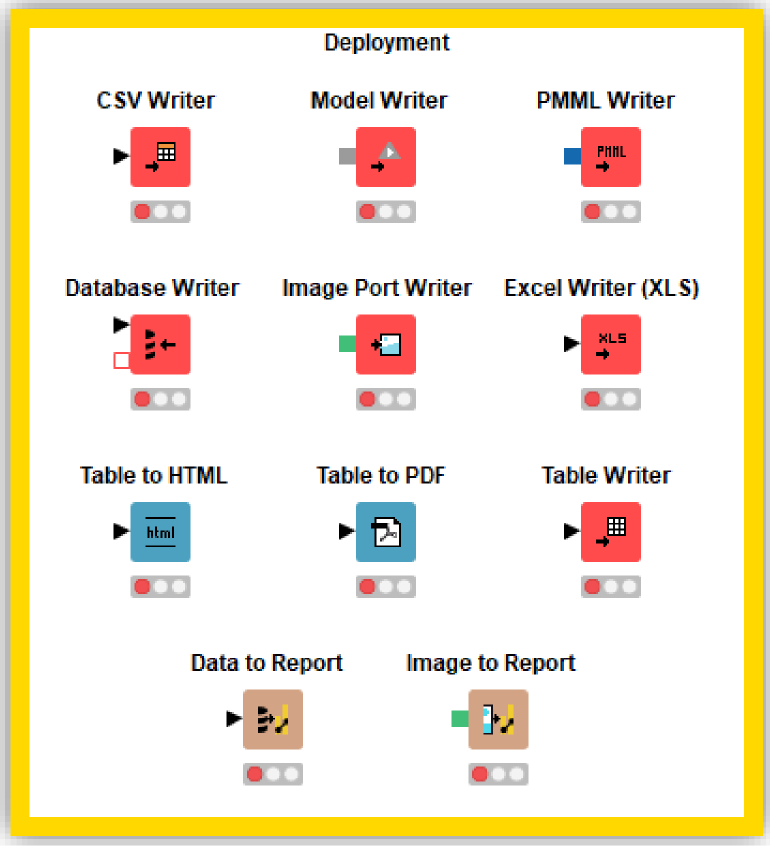
activate_all_outp...

activate_all_outp...

OK | Apply | Cancel | ?

Workflow Deployment Nodes

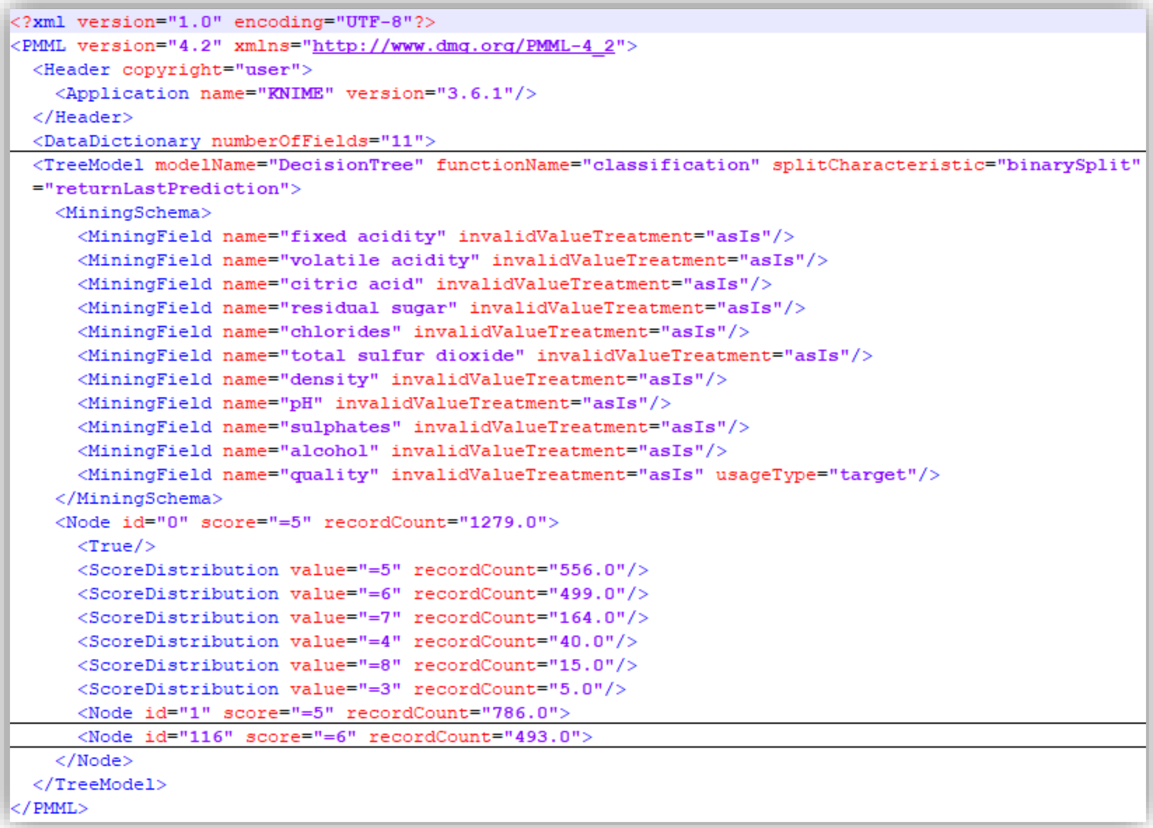




You may need to install extensions such as KNIME Compiled Model Export, KNIME PMML Translation and KNIME Report Designer/BIRT

Workflow Deployment
PMML

PMML, a XML-based format, is the leading standard on Data Mining and Machine Learning models representation, enabling the instant deployment of predictive solutions.



```xml
<?xml version="1.0" encoding="UTF-8"?>
<PMML version="4.2" xmlns="http://www.dmg.org/PMML-4_2">
  <Header copyright="user">
    <Application name="KNIME" version="3.6.1"/>
  </Header>
  <DataDictionary numberOfFields="11">
  <TreeModel modelName="DecisionTree" functionName="classification" splitCharacteristic="binarySplit"
  ="returnLastPrediction">
    <MiningSchema>
      <MiningField name="fixed acidity" invalidValueTreatment="asIs"/>
      <MiningField name="volatile acidity" invalidValueTreatment="asIs"/>
      <MiningField name="citric acid" invalidValueTreatment="asIs"/>
      <MiningField name="residual sugar" invalidValueTreatment="asIs"/>
      <MiningField name="chlorides" invalidValueTreatment="asIs"/>
      <MiningField name="total sulfur dioxide" invalidValueTreatment="asIs"/>
      <MiningField name="density" invalidValueTreatment="asIs"/>
      <MiningField name="pH" invalidValueTreatment="asIs"/>
      <MiningField name="sulphates" invalidValueTreatment="asIs"/>
      <MiningField name="alcohol" invalidValueTreatment="asIs"/>
      <MiningField name="quality" invalidValueTreatment="asIs" usageType="target"/>
    </MiningSchema>
    <Node id="0" score="=5" recordCount="1279.0">
      <True/>
      <ScoreDistribution value="=5" recordCount="556.0"/>
      <ScoreDistribution value="=6" recordCount="499.0"/>
      <ScoreDistribution value="=7" recordCount="164.0"/>
      <ScoreDistribution value="=4" recordCount="40.0"/>
      <ScoreDistribution value="=8" recordCount="15.0"/>
      <ScoreDistribution value="=3" recordCount="5.0"/>
      <Node id="1" score="=5" recordCount="786.0">
      <Node id="116" score="=6" recordCount="493.0">
    </Node>
  </TreeModel>
</PMML>
```

Workflows from the command line

There is a command line option allowing you to run KNIME workflows in batch mode!

But first you will need to add KNIME directory to the PATH environment variable:

- Windows

Use the Environment Variables GUI

- Linux & Mac

export PATH=$PATH:<*KNIME_DIRECTORY*>

- As Alternative

Execute the commands directly inside KNIME directory

Note:

On Mac, the executable is not directly located in the KNIME directory but, instead, inside a subfolder of the application bundle - knime.app/Contents/MacOS/knime

Workflows from the command line
List of Arguments

Windows

```
knime.exe -consoleLog -noexit -nosplash -application org.knime.product.KNIME_BATCH_APPLICATION
```

Linux

```
knime -nosplash -application org.knime.product.KNIME_BATCH_APPLICATION
```

Mac

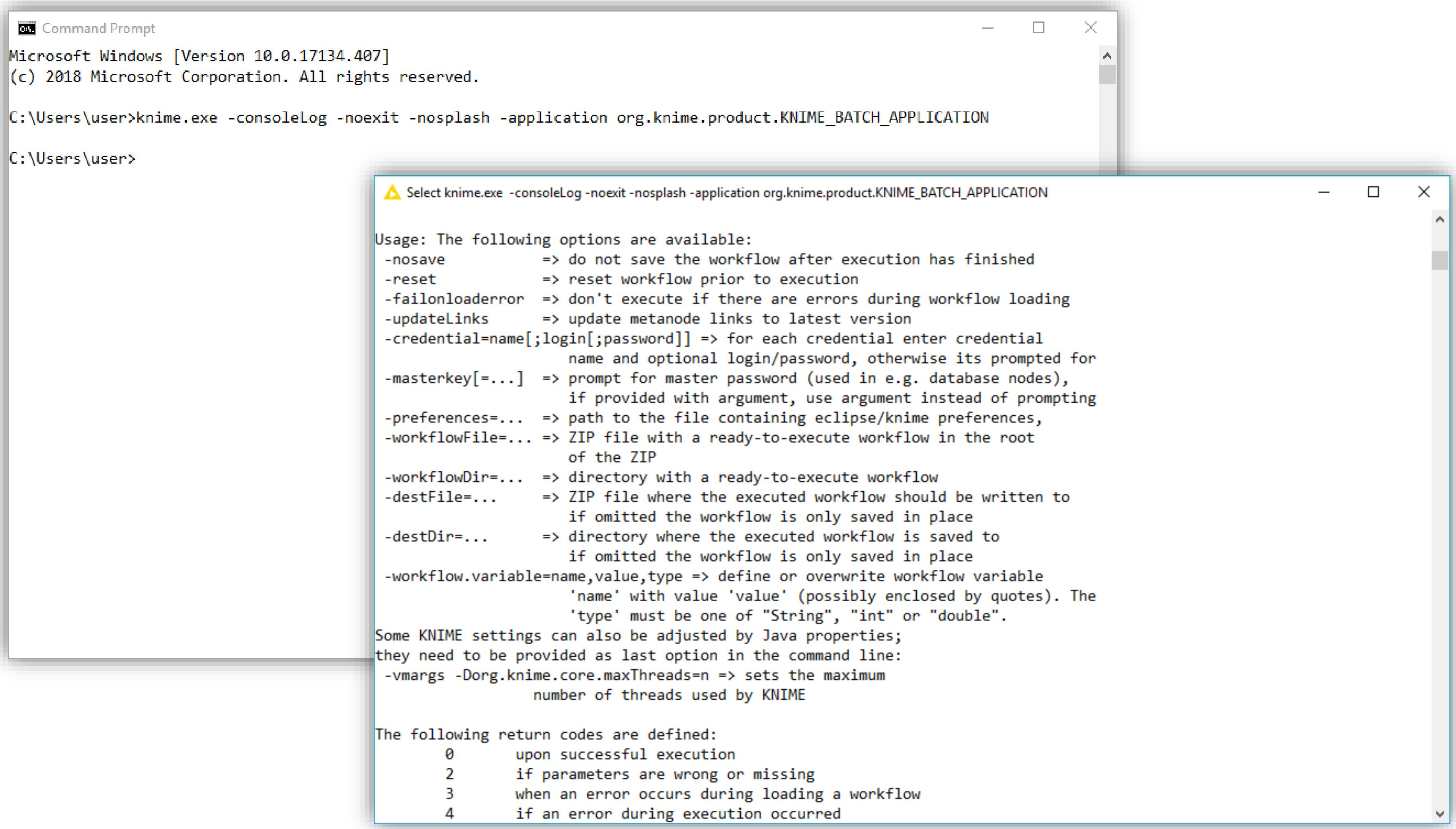
```
knime.app/Contents/MacOS/knime -nosplash -application org.knime.product.KNIME_BATCH_APPLICATION
```

Note:

In Windows, the arguments *-consoleLog -noexit* are required to redirect log messages to a new console window, which is automatically opened. *-nosplash* prevents the initial splash window with KNIME info from being shown.

Workflows from the command line

List of Arguments - Output



```
Microsoft Windows [Version 10.0.17134.407]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\user>knime.exe -consoleLog -noexit -nosplash -application org.knime.product.KNIME_BATCH_APPLICATION

C:\Users\user>
```

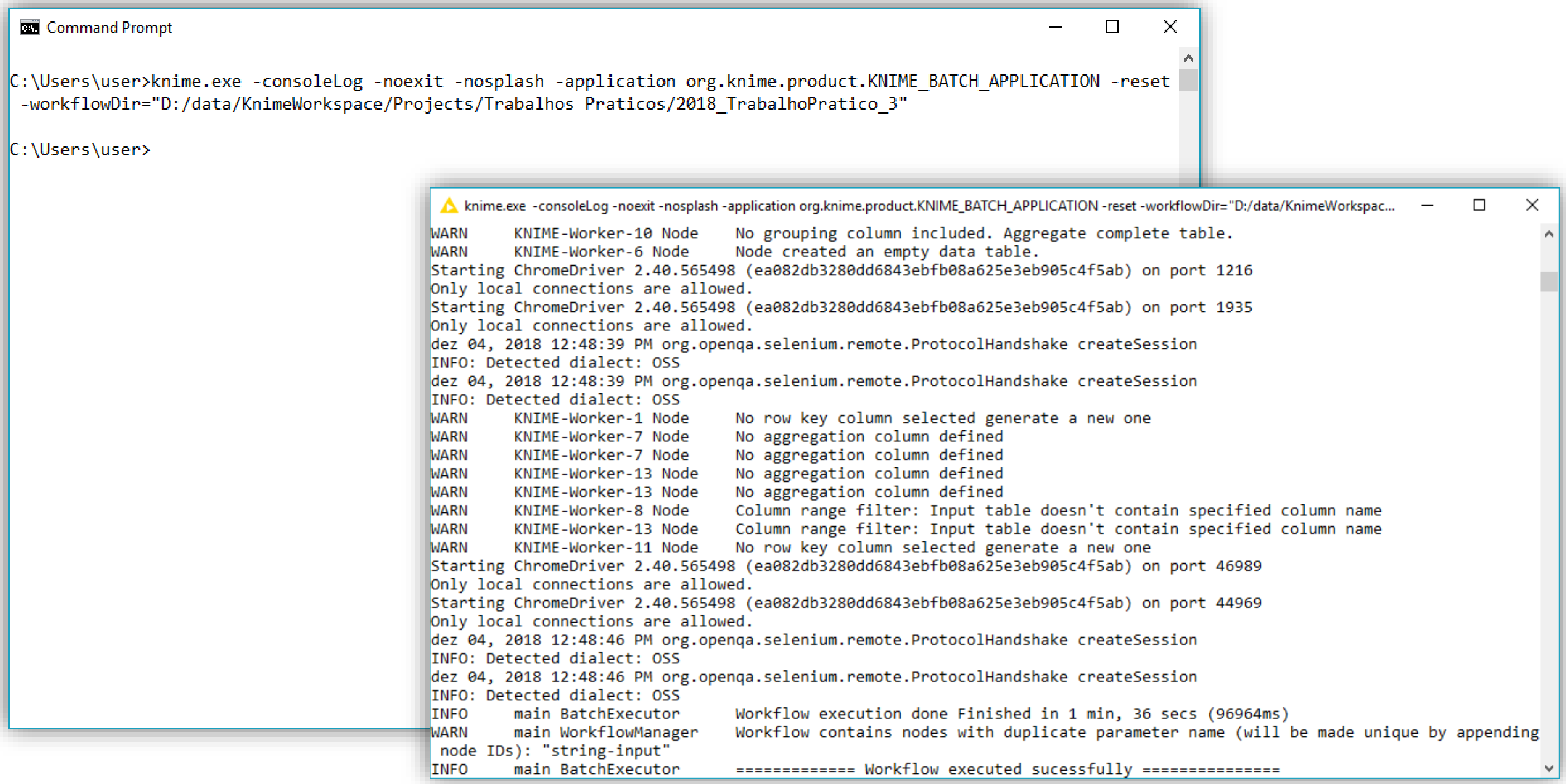
Select knime.exe -consoleLog -noexit -nosplash -application org.knime.product.KNIME_BATCH_APPLICATION

```
Usage: The following options are available:
 -nosave            => do not save the workflow after execution has finished
 -reset             => reset workflow prior to execution
 -failonloaderror   => don't execute if there are errors during workflow loading
 -updateLinks       => update metanode links to latest version
 -credential=name[;login[;password]] => for each credential enter credential
                       name and optional login/password, otherwise its prompted for
 -masterkey[=...]   => prompt for master password (used in e.g. database nodes),
                       if provided with argument, use argument instead of prompting
 -preferences=...   => path to the file containing eclipse/knime preferences,
 -workflowFile=... => ZIP file with a ready-to-execute workflow in the root
                       of the ZIP
 -workflowDir=...   => directory with a ready-to-execute workflow
 -destFile=...      => ZIP file where the executed workflow should be written to
                       if omitted the workflow is only saved in place
 -destDir=...       => directory where the executed workflow is saved to
                       if omitted the workflow is only saved in place
 -workflow.variable=name,value,type => define or overwrite workflow variable
                       'name' with value 'value' (possibly enclosed by quotes). The
                       'type' must be one of "String", "int" or "double".
Some KNIME settings can also be adjusted by Java properties;
they need to be provided as last option in the command line:
 -vmargs -Dorg.knime.core.maxThreads=n => sets the maximum
                   number of threads used by KNIME

The following return codes are defined:
        0       upon successful execution
        2       if parameters are wrong or missing
        3       when an error occurs during loading a workflow
        4       if an error during execution occurred
```

Workflows from the command line

Running a Workflow

Just add, to the previous command line, the argument -workflowDir (or -workflowFile - see the previous slide for diferences)



Command Prompt

```
C:\Users\user>knime.exe -consoleLog -noexit -nosplash -application org.knime.product.KNIME_BATCH_APPLICATION -reset
 -workflowDir="D:/data/KnimeWorkspace/Projects/Trabalhos Praticos/2018_TrabalhoPratico_3"

C:\Users\user>
```

knime.exe -consoleLog -noexit -nosplash -application org.knime.product.KNIME_BATCH_APPLICATION -reset -workflowDir="D:/data/KnimeWorkspac...

```
WARN     KNIME-Worker-10 Node    No grouping column included. Aggregate complete table.
WARN     KNIME-Worker-6 Node     Node created an empty data table.
Starting ChromeDriver 2.40.565498 (ea082db3280dd6843ebfb08a625e3eb905c4f5ab) on port 1216
Only local connections are allowed.
Starting ChromeDriver 2.40.565498 (ea082db3280dd6843ebfb08a625e3eb905c4f5ab) on port 1935
Only local connections are allowed.
dez 04, 2018 12:48:39 PM org.openqa.selenium.remote.ProtocolHandshake createSession
INFO: Detected dialect: OSS
dez 04, 2018 12:48:39 PM org.openqa.selenium.remote.ProtocolHandshake createSession
INFO: Detected dialect: OSS
WARN     KNIME-Worker-1 Node     No row key column selected generate a new one
WARN     KNIME-Worker-7 Node     No aggregation column defined
WARN     KNIME-Worker-7 Node     No aggregation column defined
WARN     KNIME-Worker-13 Node    No aggregation column defined
WARN     KNIME-Worker-13 Node    No aggregation column defined
WARN     KNIME-Worker-8 Node     Column range filter: Input table doesn't contain specified column name
WARN     KNIME-Worker-13 Node    Column range filter: Input table doesn't contain specified column name
WARN     KNIME-Worker-11 Node    No row key column selected generate a new one
Starting ChromeDriver 2.40.565498 (ea082db3280dd6843ebfb08a625e3eb905c4f5ab) on port 46989
Only local connections are allowed.
Starting ChromeDriver 2.40.565498 (ea082db3280dd6843ebfb08a625e3eb905c4f5ab) on port 44969
Only local connections are allowed.
dez 04, 2018 12:48:46 PM org.openqa.selenium.remote.ProtocolHandshake createSession
INFO: Detected dialect: OSS
dez 04, 2018 12:48:46 PM org.openqa.selenium.remote.ProtocolHandshake createSession
INFO: Detected dialect: OSS
INFO     main BatchExecutor      Workflow execution done Finished in 1 min, 36 secs (96964ms)
WARN     main WorkflowManager    Workflow contains nodes with duplicate parameter name (will be made unique by appending
 node IDs): "string-input"
INFO     main BatchExecutor      ============= Workflow executed sucessfully ===============
```

Hands On

KNIME Analytics Platform
File Edit View Node Help
100%
Quick Access
KNIME Explorer
My-KNIME-Hub (hub.knime.com)
EXAMPLES (knime@hub.knime.com)
LOCAL (Local Workspace)
Aulas Exemplos
Aulas Exercicios
Example Workflows
ML Projects
KNIME_Aula_Exercicio
Workflow Coach
Recommended Nodes
File Reader
CSV Reader
Excel Reader (XLS)
Table Creator
Database Reader (legacy)
Table Reader
List Files
Database Connection T...
Node Repository
IO
Manipulation
Views
Analytics
DB
Other Data Types
Structured Data
Scripting
Tools & Services
Workflow Control
Workflow Abstraction
Reporting
0: KNIME_Aula_Exercicio
HANDS ON
Description
KNIME Hub Search
Search workflows, nodes, and more...
Console
Outline