NovaMechanics | Chemoinformatics & Bioinformatics Solutions

KNIME Enalos+ Molecular Descriptor nodes

A Brief Tutorial

Novamechanics Ltd
Contact: info@novamechanics.com
Version 1, June 2017

Table of Contents

Introduction

Rapid development of information and communication technologies during the last few decades has dramatically changed our capabilities of collecting, analyzing, storing and disseminating all types of data. This process has had a profound influence on the scientific research in many disciplines, including the development of new generations of effective and selective medicines. Large databases containing millions of chemical compounds tested in various biological assays such as PubChem are increasingly available as online collections. In order to find new drug leads, there is a need for efficient and robust procedures that can be used to screen chemical databases and virtual libraries against molecules with known activities or properties.

Molecular descriptors play a fundamental role in chemistry, pharmaceutical sciences, environmental protection policy, and health researches, as well as in quality control, being the way molecules, thought of as real bodies, are transformed into numbers, allowing some mathematical treatment of the chemical information contained in the molecule.

By this definition, the molecular descriptors are divided into two main categories: experimental measurements, such as log P, molar refractivity, dipole moment, polarizability, and, in general, physico-chemical properties, and theoretical molecular descriptors, which are derived from a symbolic representation of the molecule and can be further classified according to the different types of molecular representation.

Enalos+ nodes built upon the existing KNIME infrastructure are divided in five main categories (Modelling, Molecular Descriptors, NCI, PubChem and UniChem) and significantly increase the number of the available nodes, the data handling tools and bridge different chemoinformatics and modelling tools upon the same interface.

The current tutorial is designed to help the user in going step-by-step through the process of building a KNIME workflow, using the Molecular Descriptor Enalos+ nodes of Novamechanics Ltd. This case study deals with a Luminescence Cell-Based Counter screen to Identify Inhibitors of A1 Apoptosis (AID 449761).

Step 1-Workbench overview

The KNIME workbench is organized as follows:

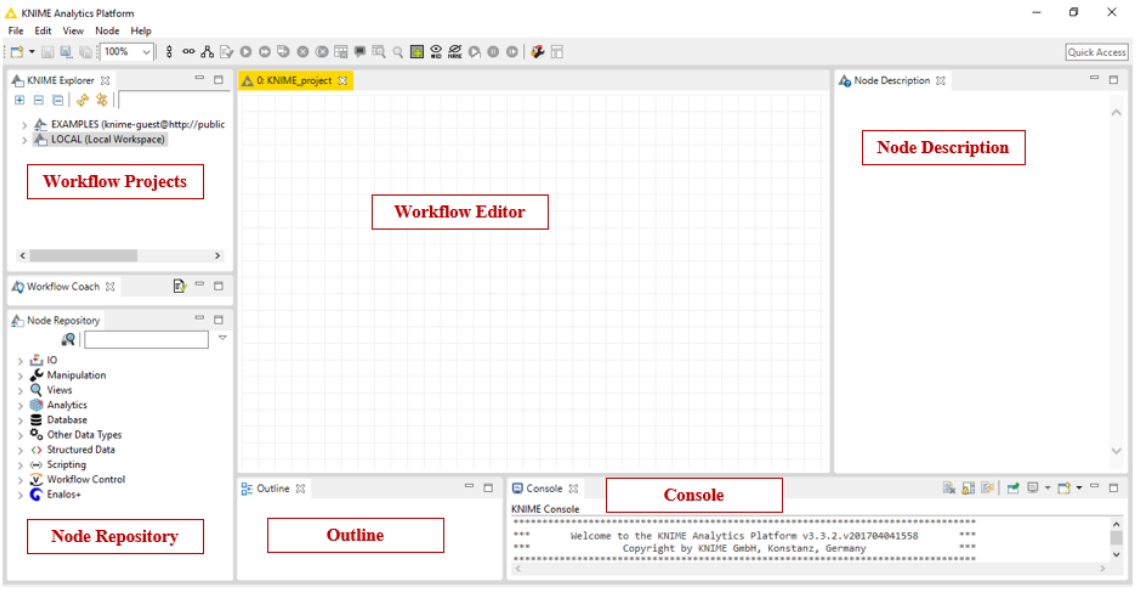


Fig. 1: KNIME workbench

It is composed of 6 main "windows": The Workflow Projects, the Workflow Editor, the Node Description, the Node Repository, the Outline and the Console. A short description of the KNIME's interface windows follows in Table 1:

Table 1: Description of KNIME interface

Workflow Projects	**Workflow Editor**	**Node Description**
Each workflow refers to a workflow project. All projects are displayed here. Import and export of workflows is supported. Status (closed, idle, executing and executed) is indicated by an icon.	Here the workflows are assembled by dragging nodes onto this editor, connecting, configuring and executing them.	Provides help about the selected node, its dialog options, views, expected input data and resulting output.
Node Repository	**Outline**	**Console**
Find all KNIME nodes here, ordered by categories. Help for selected nodes is displayed in the Node Description. Drag them onto the editor in order to add them to the workflow.	Overview over the workflow and navigation help for large workflows.	Status information, warnings and error messages are logged here. This information is also written to a log file.

Step 2-Building a workflow

The nodes are the basic processing units of a KNIME workflow. A workflow is built by dragging nodes from the Node Repository onto the Workflow Editor and connecting them, creating pipelines: Each node has a number of input-and/or output ports. Data (or a model according to each particular case) is transferred over a connection from an out-port to the in-port of another node.

1. Node status

When a node is dragged onto the workflow editor the status light shows red, which means that the node has to be configured in order to be able to be executed. A node is configured by right clicking it, choosing "Configure", and adjusting the necessary settings in the node's dialog. When the dialog is closed by pressing the "OK" button, the node is configured and the status light changes to yellow: the node is ready to be executed. Right-click on the node again shows an enabled "Execute" option; pressing it will execute the node and the result of this node will be available at the out-port (Fig. 2). After a successful execution the status light of the node is green. The result(s) can be inspected by exploring the out-port view(s): the last entries in the context menu open them. The above options "Configure", "Execute" and "View" are also available in the top ribbon of the KNIME interface window.

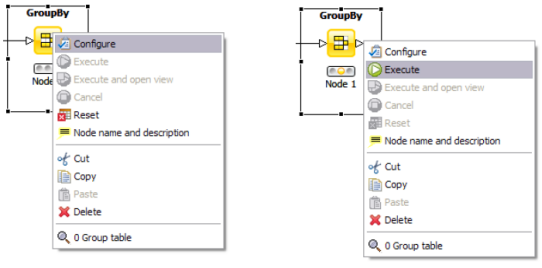


Fig. 2: Configuring and executing nodes

2. Ports

Ports on the left are input ports, where the data from the out-port of the predecessor node are provided. Ports on the right side of the node are called out-ports. The result of the node's operation on the data is provided at the out-port to successor nodes.

Step 3-Activate the Enalos+ nodes

In order to activate the Enalos+ nodes, the user has to copy the .jar file in the plugins folder and the .lic file in the license folder in the KNIME file location.

Step 4-A Brief Introduction

The Enalos+ nodes are divided into 5 main categories: Modelling, Molecular Descriptors, NCI, PubChem and UniChem.

1. Modelling

Modelling contains 11 nodes specified for data handling, preprocessing, testing modeling robustness and testing the accuracy of the predictions:

Table 2: Modelling nodes

Create New Molecules	**Domain APD**	**Domain Leverage**
Create New Molecules enables the user to create a list of molecules by combining a series of substituents with a core molecule.	*Domain APD* enables the user to define the domain of applicability of the model using a method based on the Euclidean distances.	*Domain Leverage* enables the user to define the domain of applicability of the model using a method based on the extent of extrapolation.
Int 2 Double	**Kennard and Stone**	**MLR**
Int 2 Double converts integer values of all columns to doubles.	*Kennard-Stone* node allows the selection of two representative subsets (as training and test sets) with a uniform distribution over an initial dataset.	*MLR* node performs Multiple Linear Regression in order to model the relationships between a scalar dependent variable y and two or more independent variables denoted as X.
Model Acceptability Criteria	**Remove Column**	**Remove Duplicates**
Model Acceptability Criteria gives information about the Quality of Fit and Predictive Ability of a continuous QSAR Model.	*Remove Column* node removes the selected input columns of the table that contain the same values at a percentage equal or higher than a specified cutoff limit.	*Remove Duplicates* enables the user to remove the rows of the input table that contain the same values in selected columns. The filtered table contains all rows that are unique and the first one of each repeated row.
Sphere Exclusion	**Y Randomization**	
Sphere Exclusion node allows the selection of two representative subsets (such as training and test sets). This method attempts to specify compounds which most effectively cover the available data space.	*Y Randomization* (or Y-scrambling) is a technique, applied to ensure a QSAR model's robustness.	

2. Molecular Descriptors

Molecular Descriptors contains *EnalosMold2* node.

3. NCI

NCI contains *CIR* node.

Table 3: Molecular Descriptors and NCI

EnalosMold2	**CIR**
Molecular Descriptors by *EnalosMold2* calculates a large and diverse set of molecular descriptors (777) encoding two-dimensional chemical structure information.	Enalos+ *CIR* node enables the user to get direct access to CIR (Chemical Identifier Resolver) through KNIME. The user has the option to select several output formats through a GUI menu.

4. PubChem

PubChem contains 8 nodes that give direct access to PubChem database through KNIME in order to extract useful information:

Table 4: PubChem nodes

Assay	**Assay Class**
Assay node gives the user access to PubChem database via substance or compound IDs (SID and CID), in order to find the Assays where a particular compound is tested. Using this node the user can download in KNIME information about the Assay and the Assay outcome.	*Assay Class* node searches in PubChem database according to one or more given AIDs (BioAssay identification numbers) and displays only the active or inactive compounds.
Main PubChem	**Patent**
Main PubChem node enables the user to search the PubChem database and obtain the following information for thousands of compounds with one request: PubChem CID (Compound ID), IUPAC Name, InChI, InChI-Key Molecular Formula, Molecular Weight, Canonical SMILES and the direct PubChem URL.	*Patent* node gives the user straight access to the PubChem database in order to obtain information about the patent coverage information for thousands of compounds with one request.
Patent to Sid	**Sid**
Patent to Sid node helps the user to search the PubChem database and obtain the SIDs (Substance IDs) of the compounds covered by the patents in request.	*Sid* node exports the CIDs (Compound IDs) of a given list of SIDs (Substance IDs), searching the PubChem database. The user can search the PubChem database and obtain information about the CIDs for thousands of compounds with one request.
Similarity	**Vendor**
Via *Similarity* node, the user can search the whole PubChem database for similar compounds (Tanimoto Similarity) and obtain the following information for thousands of compounds with one request: PubChem CID (Compound ID), Molecular Formula, Molecular Weight and Number of Rotatable Bonds.	*Vendor* node enables the user to search the PubChem database and obtain information about the commercial availability for thousands of compounds with one request.

5. UniChem

UniChem contains 2 nodes for accessing UniChem databases:

Table 5: UniChem nodes

UniChem	UniChem Connectivity
Enalos *UniChem* gives the user direct access to UniChem databases through KNIME. UniChem is a superset of all 27 available databases, separated in 5 friendly and easily recognizable categories.	*UniChem Connectivity* is an expanded version of the standard UniChem tool that allows you to find related molecules. Connectivity Search allows molecules to be first matched on the basis of complete identity between the connectivity layer of their corresponding Standard InChIs, and the remaining layers then compared to highlight stereo-chemical and isotopic differences

Step 5-Adding Nodes

In the Node Depository, expand the *Enalos+* and the contained *Molecular Descriptors* category and choose *EnalosMold2* node (Fig. 3). Then, drag & drop the *EnalosMold2* icon into the Workflow Editor window. Do it twice, in order to have 2 *EnalosMold2* in the Workflow editor.

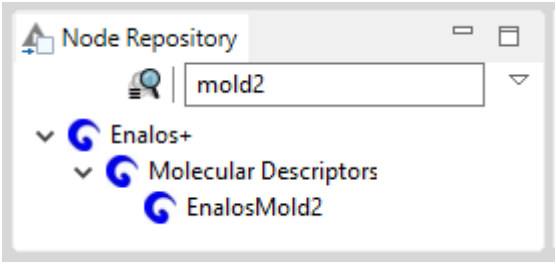


Fig. 3: Node Depository interface

Then, expand the *IO* category followed by the *Read* and drag into the Workflow Editor 2 *Excel Reader (XLS)* nodes. To complete the workflow, you will also need to add 2 more *Column Splitter* nodes, 2 *Joiner* nodes and the *Concatenate*, *Kennard and Stone*, *IBk (3.7)*, *Weka Predictor (3.7)* and *Scorer* nodes. You can search all these nodes by name in the Node Repository.

You can rename the nodes as shown below (Fig. 4).

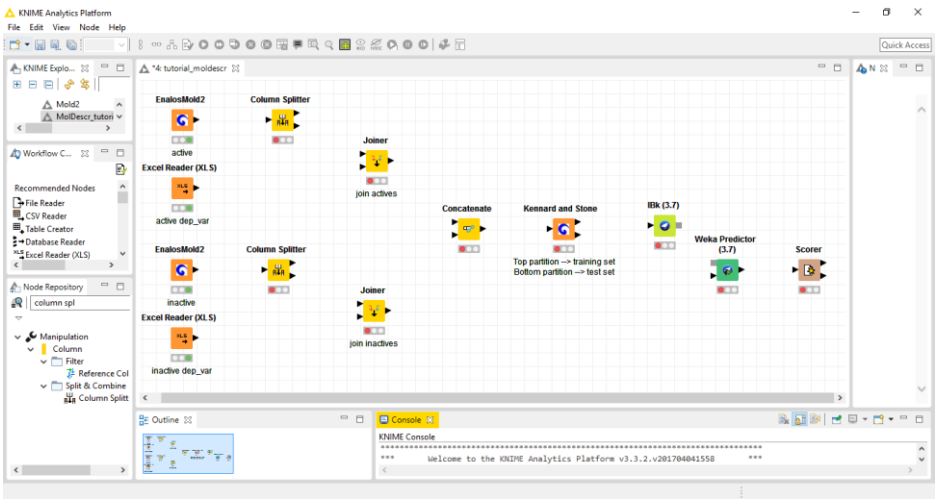


Fig. 4: Workflow editor

Step 6-Connecting Nodes

Now, you need to connect the nodes, in order to get the data flowing. Click an output port and drag the connection to an appropriate input port. Complete the flow as pictured below (Fig. 5). The nodes will not show a green status as long as they are not yet configured and executed.

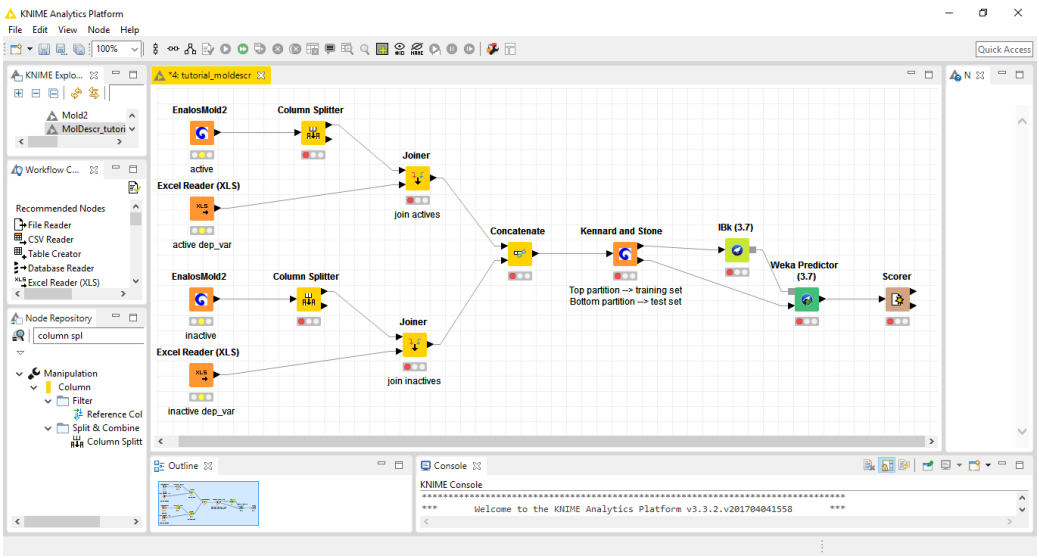


Fig. 5: Connecting nodes

Step 7-Configuring nodes

Fully connected nodes showing a red status icon need to be configured. Start with the *EnalosMold2* "active" node, right click it and select "Configure" from the menu. Press "Browse" button, select the "Mold2" .exe file to read and an .sdf file downloaded from PubChem referring to the active compounds of the AID 449761 (Fig. 6). Press "Apply" and "OK" to close the dialog of the *EnalosMold2* node. Once the node has been configured correctly, it switches to yellow (meaning ready for execution). Follow the same steps for the "inactive" *EnalosMold2* node. In this case select the corresponding .sdf file is referring to the inactive compounds of the AID 449761 (Fig. 7). Both active and inactive .sdf files are downloaded from PubChem database.

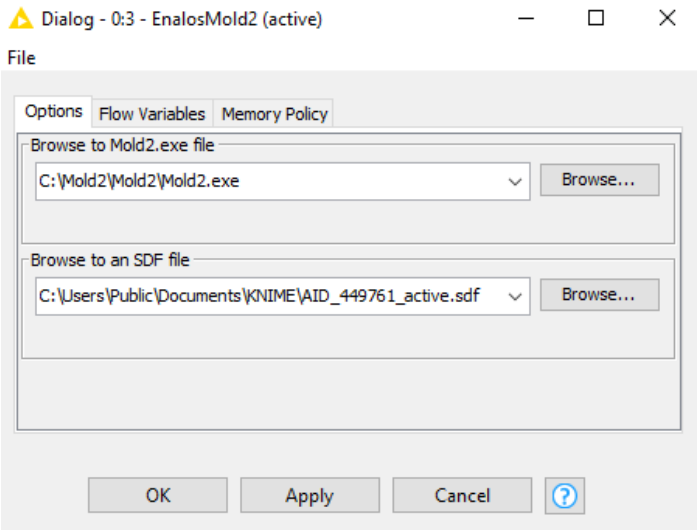


Fig. 6: Configuring the EnalosMold2 "active" node

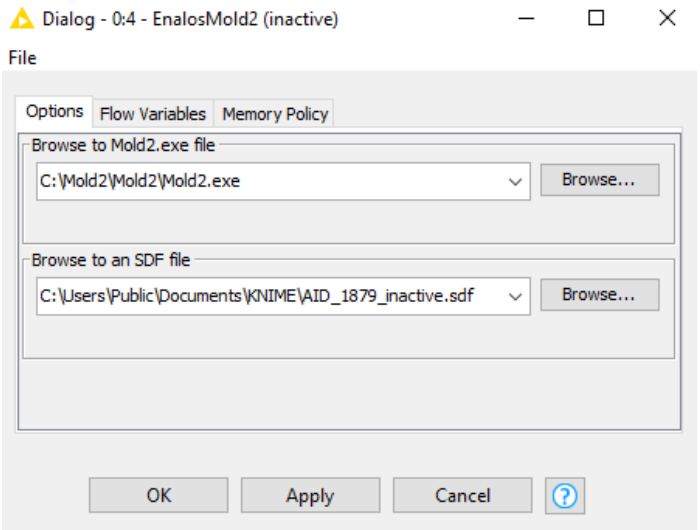


Fig. 7: Configuring the EnalosMold2 "inactive" node

Then, configure *Excel Reader (XLS)* nodes. For the "active dep_var" *Excel Reader (XLS)* node browse for an .xls file containing the values of the 1st set's depended variable (active) (Fig. 8). Repeat the process for the "inactive dep-var" node and select the 2nd set of the depended variable (inactive) (Fig. 9).

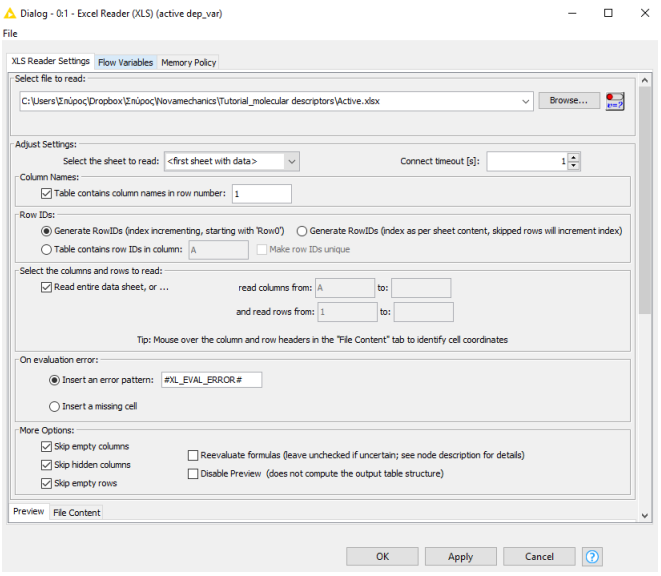


Fig. 8: Configuring the "active dep_var" Excel Reader (XLS) node

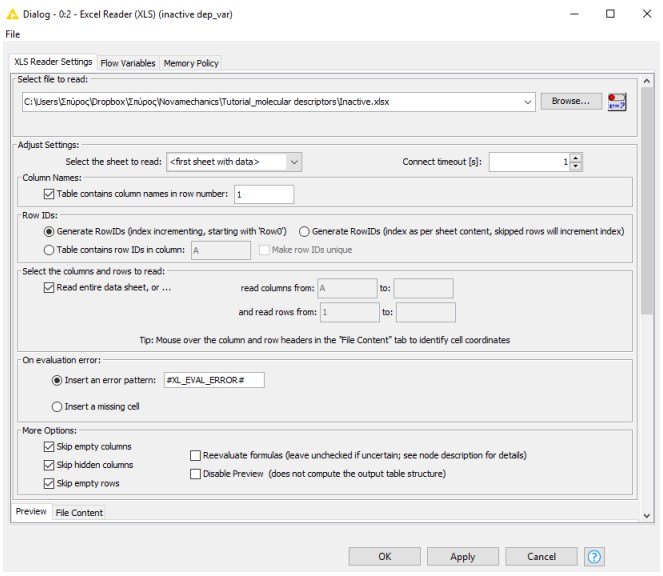


Fig. 9: Configuring the "inactive dep_var" Excel Reader (XLS) node

Subsequently, you will need to configure the *Column Splitter* nodes in order to exclude from the *EnalosMold2* output the column referring to the row numbers (see Fig. 10).

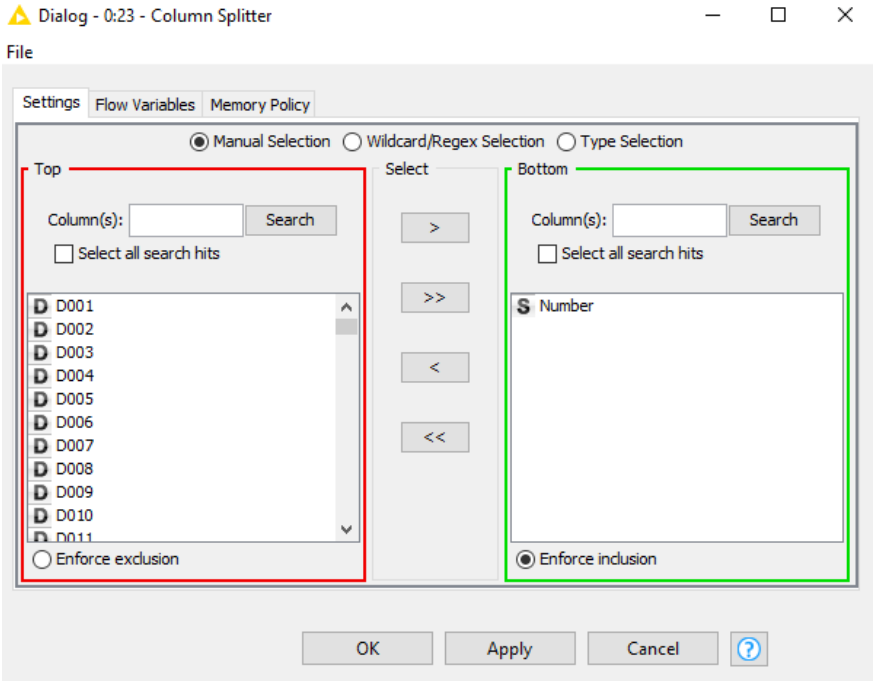


Fig. 10: Configuring Column Splitter nodes

The *Joiner* nodes join two tables in a database-like way. The join is based on the joining columns of both tables. The Joiner nodes' configuring menu is shown below and is exactly the same for both "join actives" and "join inactives" *Joiner* nodes (Fig. 11).

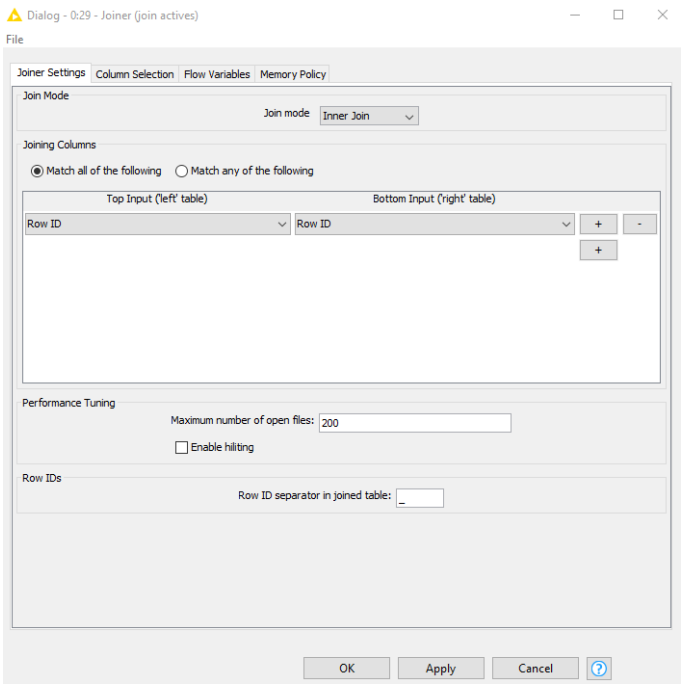


Fig. 11: Configuring Joiner nodes

The *Concatenate* node concatenates the "active" and the "inactive" tables and takes as input, the output of the 2 *Joiner* nodes. Configure this node as depicted in Fig. 12:

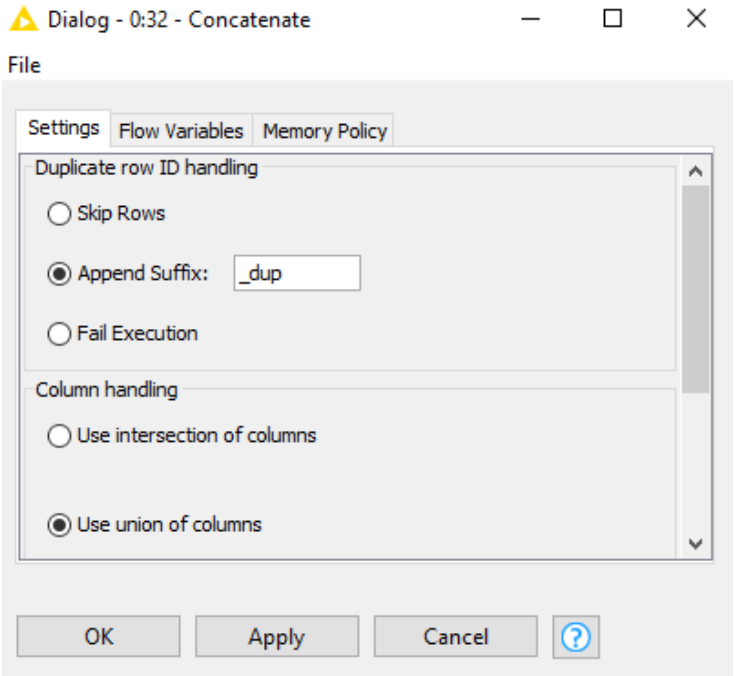


Fig. 12: Configuring Concatenate node

The Kennard and Stone method allows the selection of two representative subsets (as training and test sets) with a uniform distribution over an initial dataset. *Kennard and Stone* node takes as input the table that is going to be partitioned and outputs the 2 partitions (training and test set). Configure *Kennard and Stone* by defining the "Target column" which refers to the depended variable and the "Model percentage". In general, the test set should be about 15-20% of the entire dataset. So, you can select, say, 80% as a "Model percentage" (Fig. 13).

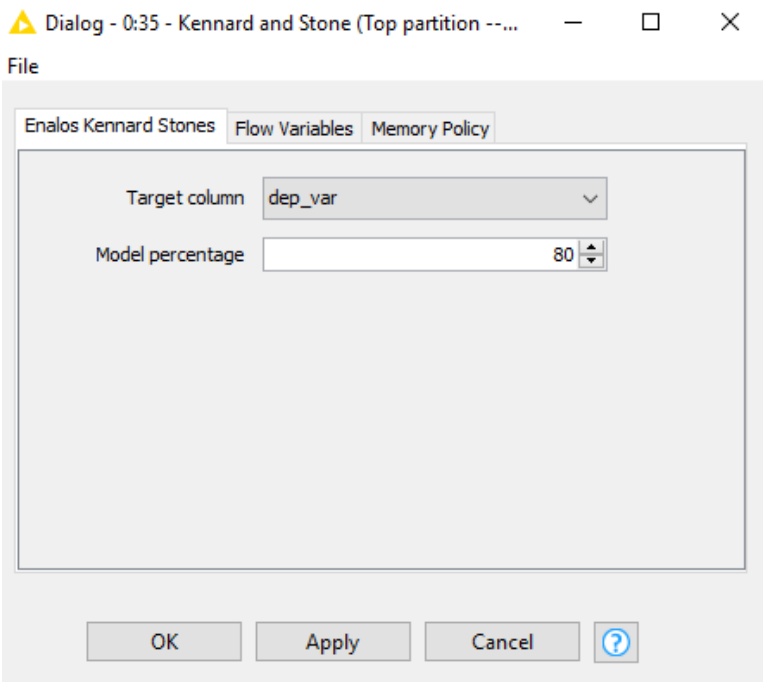


Fig. 13: Configuring Kennard and Stone node

IBk (3.7) node is a k-Nearest-Neighbor (kNN) classifier. In the configuring menu, you can select an appropriate value of K based on cross-validation. You can also do distance weighting. Select, say, 3 in the kNN menu (Fig. 14).

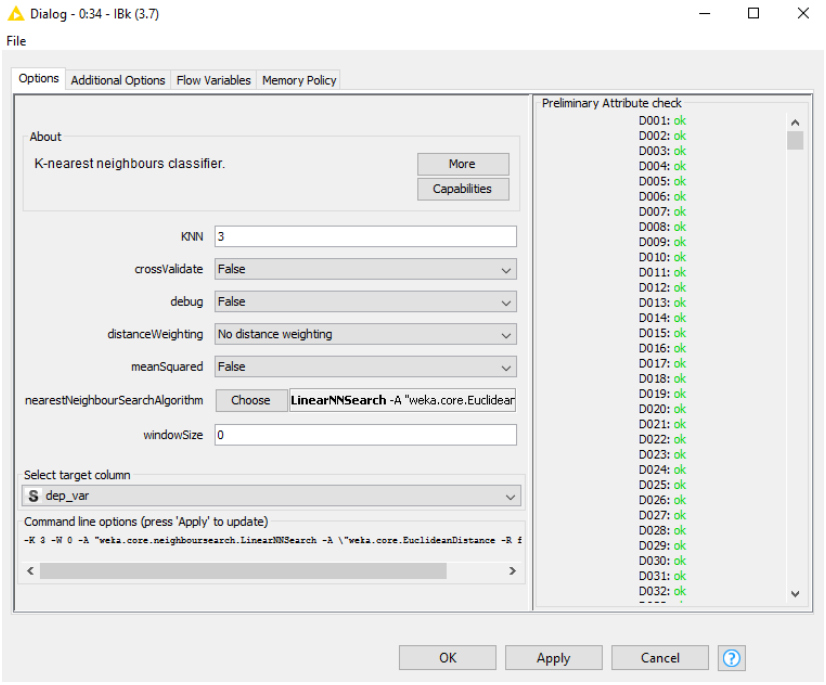


Fig. 14: Configuring IBk (3.7) node

The *Weka Predictor* node takes a model generated in a weka node and classifies the test data at the import (Fig. 15).

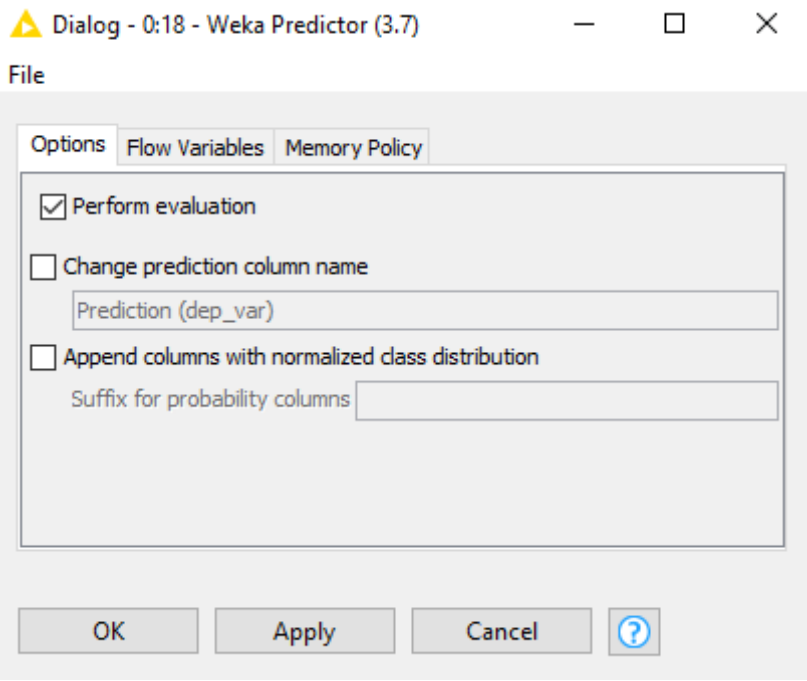


Fig. 15: Configuring Weka Predictor node

Scorer node compares two columns by their attribute value pairs and shows the confusion matrix, i.e. how many rows of which attribute and their classification match. In the configuring menu select the depended variable in the "First Column" and the prediction of the depended variable in the "Second Column" (Fig. 16).

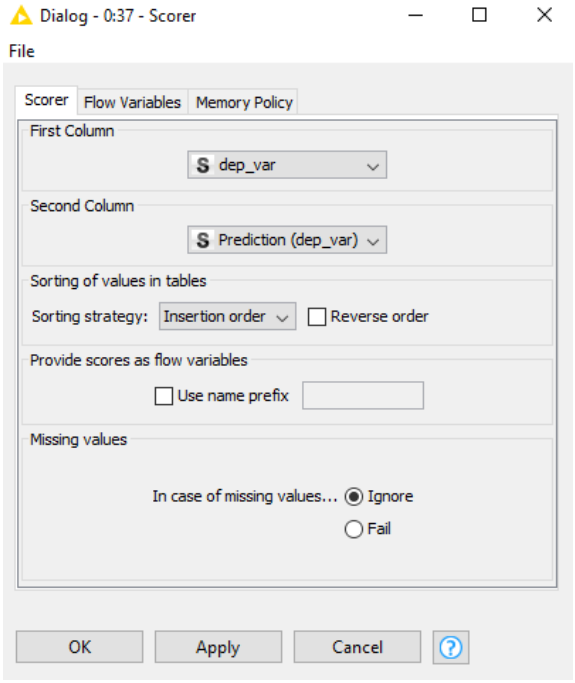


Fig. 16: Configuring Scorer node

Step 8-Executing nodes

Now, right click on the *EnalosMold2* nodes and execute them. Then execute the *Scorer* node. The workbench will execute all predecessor nodes for you. In a larger, more complex flow, you could

select multiple nodes and trigger execution for all of them. The workflow manager will execute the nodes as needed, if possible in parallel. To execute all executable nodes press (Shift+F7).

Step 9-Inspecting the results

In order to examine the data and the results, open the nodes' views. From *EnalosMold2* output ports the table read in is extracted (Fig. 17, Fig. 18).

Descriptors - 0:3 - EnalosMold2 (active)

File

Table "default" - Rows: 132 | Spec - Columns: 778 | Properties | Flow Variables

Row ID	Number	D001	D002	D003	D004
Row0	1	3	0	0	2
Row1	2	3	0	0	1
Row2	3	3	0	0	0
Row3	4	2	0	0	2
Row4	5	4	0	0	0
Row5	6	2	0	0	1
Row6	7	1	0	0	0
Row7	8	1	0	0	0
Row8	9	2	0	0	0
Row9	10	2	0	0	2
Row10	11	3	0	0	1
Row11	12	3	0	0	1
Row12	13	1	0	0	1
Row13	14	1	0	0	2
Row14	15	1	0	0	1
Row15	16	1	0	0	0
Row16	17	3	0	0	1

Fig. 17: "active" EnalosMold2 results

Descriptors - 0:4 - EnalosMold2 (inactive)

File

Table "default" - Rows: 244 | Spec - Columns: 778 | Properties | Flow Variables

Row ID	Number	D001	D002	D003	D004
Row0	1	2	0	0	1
Row1	2	1	0	0	1
Row2	3	2	0	0	1
Row3	4	2	0	0	0
Row4	5	2	0	0	1
Row5	6	2	0	0	0
Row6	7	2	0	0	2
Row7	8	2	0	0	1
Row8	9	1	0	0	1
Row9	10	2	0	0	1
Row10	11	1	0	0	2
Row11	12	2	0	0	0
Row12	13	1	0	0	1
Row13	14	2	0	0	1
Row14	15	2	0	0	1
Row15	16	2	0	0	0
Row16	17	1	0	0	0

Fig. 18: "inactive" EnalosMold2 results

The *Excel Reader (XLS)* and the *Joiner* nodes output is depicted below (Fig. 19, Fig. 20) and (Fig. 21, Fig. 22).

Output table - 0:1 - Excel Reader (XLS) (active dep_var)

File

Table "Active.xlsx [Sheet1]" - Rows: 131 | Spec - Column: 1 | Properties | Flow Variables

Row ID	dep_var
Row0	active
Row1	active
Row2	active
Row3	active
Row4	active
Row5	active
Row6	active
Row7	active
Row8	active
Row9	active
Row10	active
Row11	active
Row12	active
Row13	active
Row14	active
Row15	active
Row16	active
Row17	active

Fig. 19: "active dep_var" Excel Reader (XLS) results

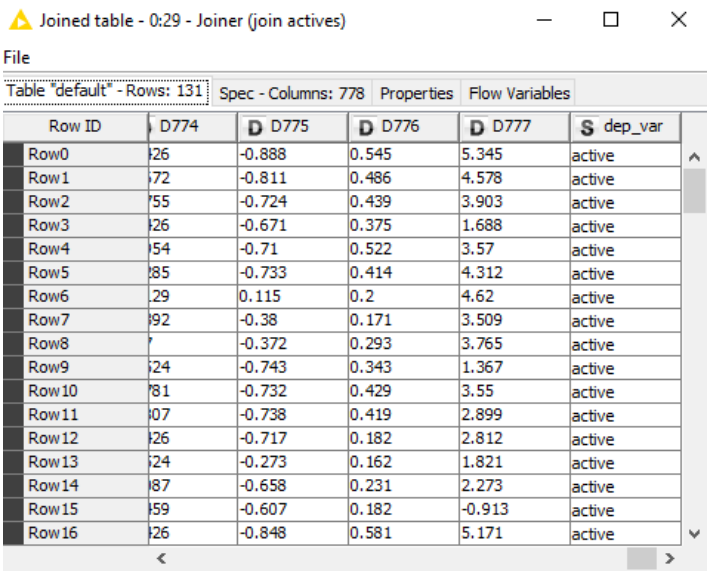
Output table - 0:2 - Excel Reader (XLS) (inactive dep_var)

File

Table "Inactive.xlsx [Sheet1]" - Rows: 114 | Spec - Column: 1 | Properties | Flow Variables

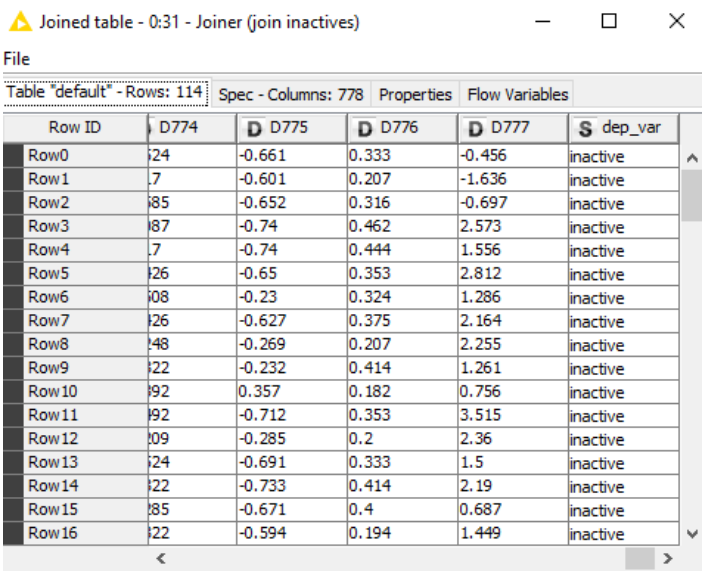
Row ID	dep_var
Row0	inactive
Row1	inactive
Row2	inactive
Row3	inactive
Row4	inactive
Row5	inactive
Row6	inactive
Row7	inactive
Row8	inactive
Row9	inactive
Row10	inactive
Row11	inactive
Row12	inactive
Row13	inactive
Row14	inactive
Row15	inactive
Row16	inactive
Row17	inactive

Fig. 20: "inactive dep_var" Excel Reader (XLS) results



Row ID	D774	D775	D776	D777	dep_var
Row0	426	-0.888	0.545	5.345	active
Row1	672	-0.811	0.486	4.578	active
Row2	755	-0.724	0.439	3.903	active
Row3	426	-0.671	0.375	1.688	active
Row4	954	-0.71	0.522	3.57	active
Row5	885	-0.733	0.414	4.312	active
Row6	1.29	0.115	0.2	4.62	active
Row7	992	-0.38	0.171	3.509	active
Row8	?	-0.372	0.293	3.765	active
Row9	624	-0.743	0.343	1.367	active
Row10	781	-0.732	0.429	3.55	active
Row11	807	-0.738	0.419	2.899	active
Row12	426	-0.717	0.182	2.812	active
Row13	624	-0.273	0.162	1.821	active
Row14	987	-0.658	0.231	2.273	active
Row15	959	-0.607	0.182	-0.913	active
Row16	426	-0.848	0.581	5.171	active

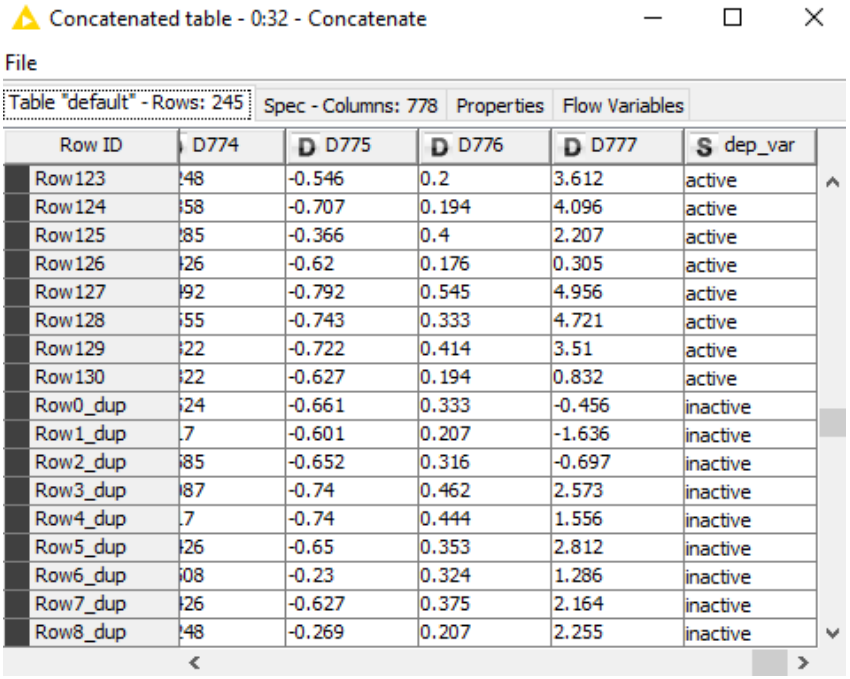
Fig. 21: "join actives" Joiner results



Row ID	D774	D775	D776	D777	dep_var
Row0	624	-0.661	0.333	-0.456	inactive
Row1	1.7	-0.601	0.207	-1.636	inactive
Row2	885	-0.652	0.316	-0.697	inactive
Row3	987	-0.74	0.462	2.573	inactive
Row4	1.7	-0.74	0.444	1.556	inactive
Row5	426	-0.65	0.353	2.812	inactive
Row6	608	-0.23	0.324	1.286	inactive
Row7	426	-0.627	0.375	2.164	inactive
Row8	748	-0.269	0.207	2.255	inactive
Row9	822	-0.232	0.414	1.261	inactive
Row10	992	0.357	0.182	0.756	inactive
Row11	992	-0.712	0.353	3.515	inactive
Row12	209	-0.285	0.2	2.36	inactive
Row13	624	-0.691	0.333	1.5	inactive
Row14	822	-0.733	0.414	2.19	inactive
Row15	885	-0.671	0.4	0.687	inactive
Row16	822	-0.594	0.194	1.449	inactive

Fig. 22: "join inactives" Joiner results

The Concatenate node exports a table with rows from both the input tables (Fig. 23).



Row ID	D774	D775	D776	D777	dep_var
Row123	748	-0.546	0.2	3.612	active
Row124	958	-0.707	0.194	4.096	active
Row125	885	-0.366	0.4	2.207	active
Row126	426	-0.62	0.176	0.305	active
Row127	992	-0.792	0.545	4.956	active
Row128	755	-0.743	0.333	4.721	active
Row129	822	-0.722	0.414	3.51	active
Row130	822	-0.627	0.194	0.832	active
Row0_dup	624	-0.661	0.333	-0.456	inactive
Row1_dup	1.7	-0.601	0.207	-1.636	inactive
Row2_dup	885	-0.652	0.316	-0.697	inactive
Row3_dup	987	-0.74	0.462	2.573	inactive
Row4_dup	1.7	-0.74	0.444	1.556	inactive
Row5_dup	426	-0.65	0.353	2.812	inactive
Row6_dup	608	-0.23	0.324	1.286	inactive
Row7_dup	426	-0.627	0.375	2.164	inactive
Row8_dup	748	-0.269	0.207	2.255	inactive

Fig. 23: Concatenate results

The *Kennard and Stone* node extracts a top partition with 196 rows (training set) and a bottom partition with 49 rows (test set).

First Partition - 0:35 - Kennard and Stone (Top partition --...

File

Table "default" - Rows: 196 | Spec - Columns: 778 | Properties | Flow Variables

Row ID	D001	D002	D003	D004	D005
Row0	3	0	0	2	3
Row29	3	0	0	0	4
Row78	0	1	0	1	4
Row115	4	0	0	0	6
Row79_dup	2	0	0	1	2
Row42_dup	4	0	0	0	4
Row45_dup	2	0	0	1	2
Row35	1	1	0	2	2
Row2	3	0	0	0	5
Row77_dup	3	0	0	1	3
Row4	4	0	0	0	5
Row16_dup	1	0	0	0	2
Row37_dup	3	0	0	1	4
Row32	1	0	0	1	1
Row61_dup	2	0	0	1	4
Row74_dup	4	0	0	0	4
Row73	3	0	0	1	4

Fig. 24: Kennard and Stone results (Top partition)

Second Partition - 0:35 - Kennard and Stone (Top partition...

File

Table "default" - Rows: 49 | Spec - Columns: 778 | Properties | Flow Variables

Row ID	D001	D002	D003	D004	D005
Row14	1	0	0	1	2
Row21	1	0	0	1	2
Row42	1	0	0	3	1
Row43	1	0	0	1	3
Row45	1	0	0	1	2
Row47	0	0	0	0	3
Row48	2	0	0	2	2
Row62	1	0	0	2	1
Row63	1	0	0	2	1
Row69	2	0	0	1	2
Row76	2	0	0	0	2
Row79	2	0	0	2	2
Row84	2	0	0	1	2
Row89	2	0	0	1	2
Row91	2	0	0	1	2
Row94	2	0	0	0	2
Row100	3	0	0	1	3

Fig. 25: Kennard and Stone results (Bottom partition)

The Weka Predictor (3.7) node outputs the classified test data (Fig. 26).

Classified Test data - 0:18 - Weka Predictor (3.7)

File

Table "default" - Rows: 49 | Spec - Columns: 779 | Properties | Flow Variables

Row ID	D775	D776	D777	dep_var	Predicti...
Row14	658	0.231	2.273	active	active
Row21	836	0.273	3.994	active	inactive
Row42	646	0.207	2.535	active	active
Row43	742	0.194	3.823	active	active
Row45	338	0.176	2.367	active	inactive
Row47	869	0	2.941	active	active
Row48	79	0.429	3.323	active	inactive
Row62	684	0.214	3.423	active	active
Row63	658	0.231	2.366	active	active
Row69	735	0.364	3.521	active	active
Row76	776	0.6	3.059	active	active
Row79	788	0.324	1.914	active	active
Row84	762	0.5	3.321	active	active
Row89	699	0.462	3.428	active	active
Row91	696	0.414	2.748	active	active
Row94	696	0.429	2.633	active	active
Row100	743	0.514	3.227	active	active

Fig. 26: Weka Prediction (3.7) results

Finally, the Scorer node exports the confusion matrix (1st output-Fig. 27) and the accuracy statistics table (2nd output-Fig. 28).

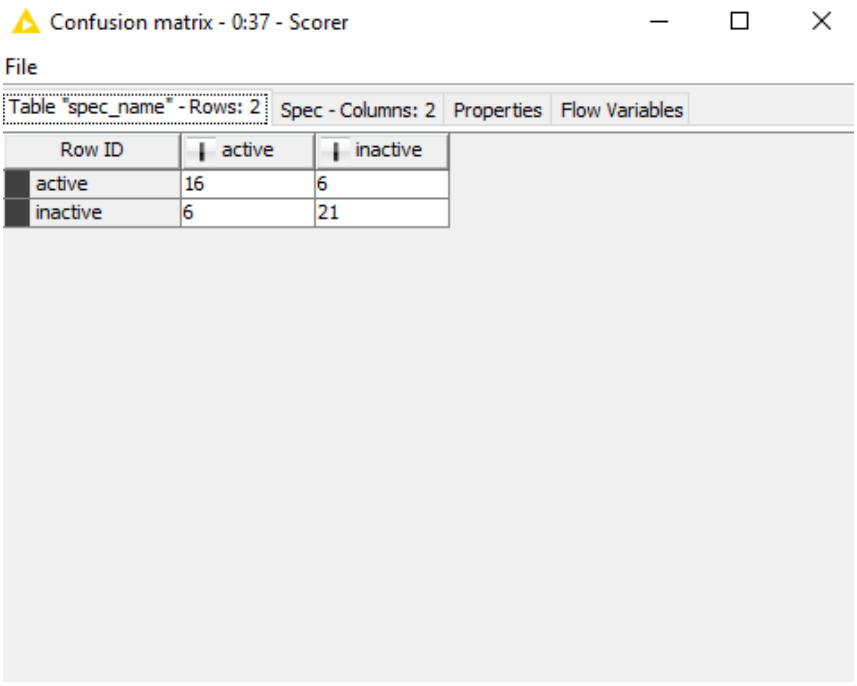


Fig. 27: Scorer results (1)

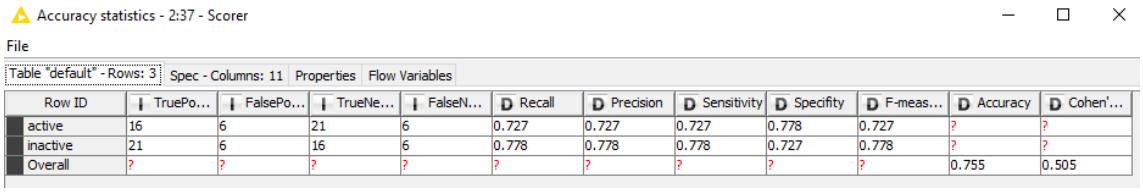


Fig. 28: Scorer results (2)

The executed workflow is depicted in Fig. 29:

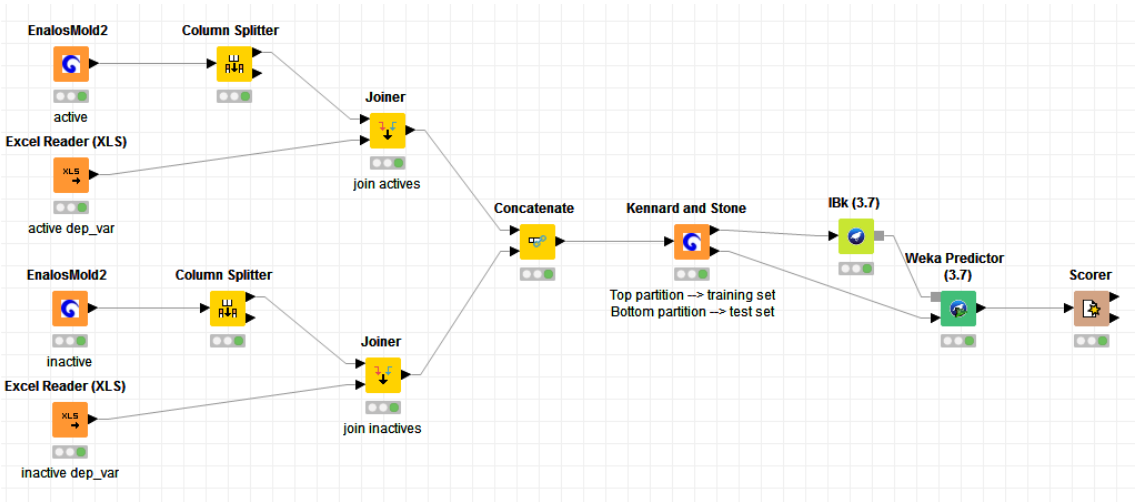


Fig. 29: Workflow using Enalos+ Molecular Descriptor nodes

Step 10-Extending the main Workflow

In order to develop a more consistent and robust model, you can extend the main workflow.

- First of all, add the Enalos+ *Remove Column* node. This node removes the selected input columns of the table that contain the same values at a percentage equal or higher than a specific cutoff limit. In the configuring menu remove the column containing the depended variable and set 10% as Threshold (Fig. 30)
- Then, drag and drop *Shuffle* node in the Workflow editor. This node shuffles the rows of the input tables such that they are in random order (Fig. 31)
- You can also add *Normalizer* node, which normalizes the values of all (numeric) columns. In the dialog, you can choose the columns you want to work on and set a method of normalization, for example Z-Score Normalization (Fig. 32).

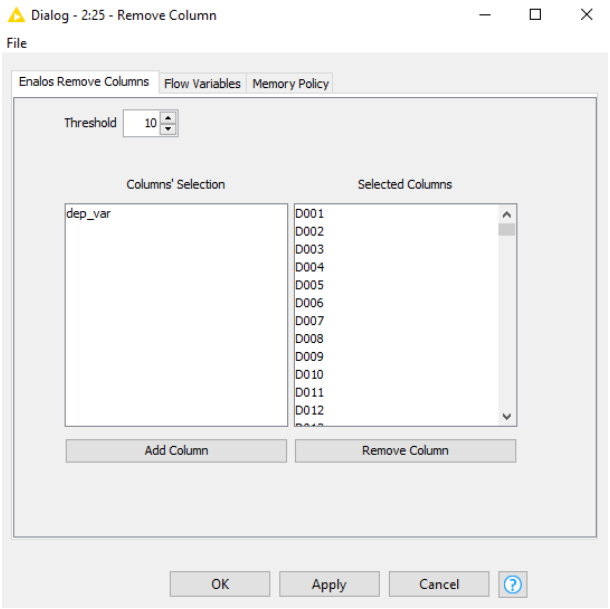


Fig. 30: Configuring Remove Column node

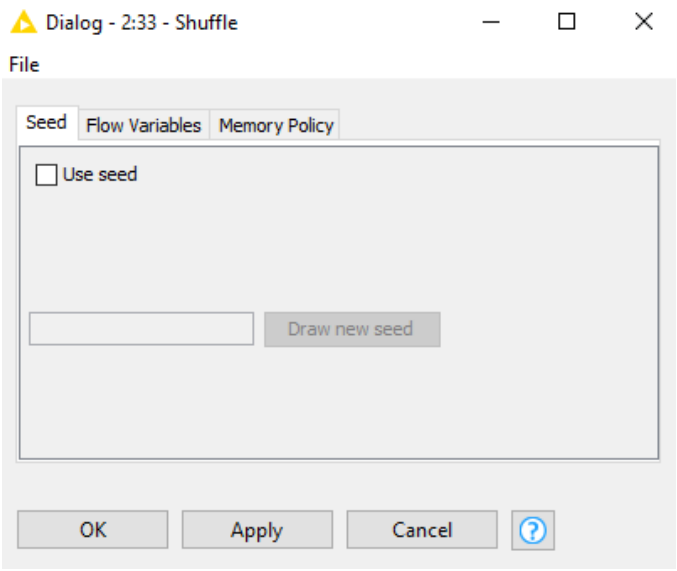


Fig. 31: Configuring Shuffle node

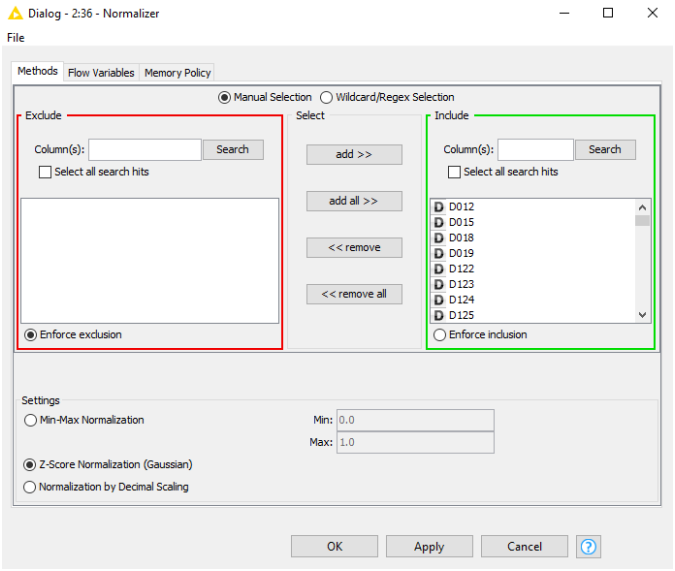


Fig. 32: Configuring Normalizer node

The updated workflow is depicted in Fig. 33.

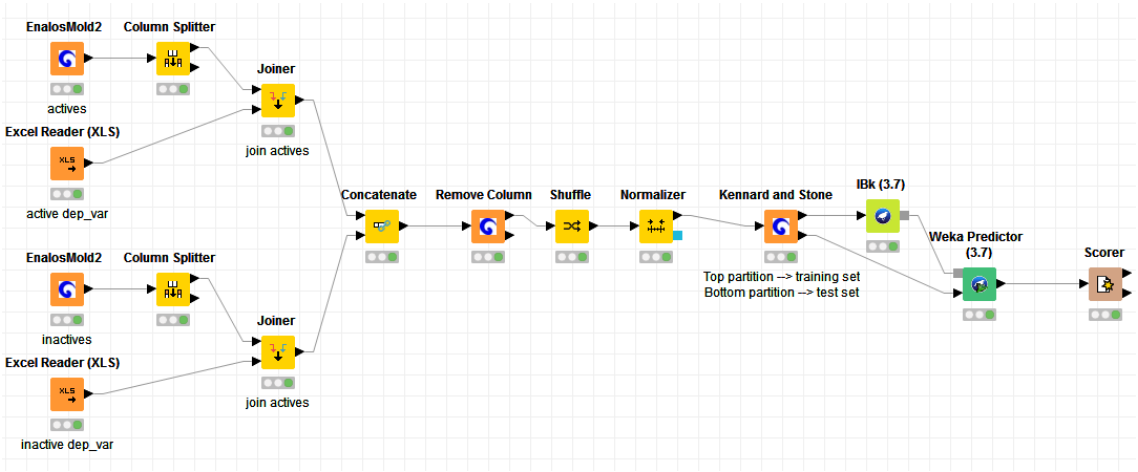


Fig. 33: Updated Workflow

The new, converted results of the workflow nodes are shown below:

- The *Remove Column* node's filtered table contains the 411 of the 778 initial columns (Fig. 34)
- The shuffled table (output of the *Shuffle* node) is depicted in Fig. 35
- The normalized table (output of the *Normalizer* node) is shown in Fig. 36
- The classified test data from the *Weka Predictor (3.7)* node are displayed in Fig. 37
- Concerning the *Scorer* node's output ports, the 1st one exports the confusion matrix (Fig. 38) and the 2nd the accuracy statistics (Fig. 39)

Filtered Table - 2:25 - Remove Column

File

Table "default" - Rows: 245 | Spec - Columns: 411 | Properties | Flow Variables

Row ID	D012	D015	D018	D019	D122
Row0	19	0.182	43	25	399.561
Row1	23	0.135	51	23	457.555
Row2	22	0.22	54	22	529.541
Row3	17	0.188	43	13	391.413
Row4	28	0.304	62	22	597.574
Row5	16	0.345	39	17	491.187
Row6	8	0.733	35	43	417.64
Row7	10	0.4	42	38	454.62
Row8	17	0.341	52	34	518.621
Row9	17	0.314	46	24	432.485
Row10	22	0.286	55	29	552.697
Row11	22	0.302	56	32	564.733
Row12	13	0.152	43	17	408.849
Row13	12	0.351	46	31	465.565
Row14	11	0.346	34	14	409.284
Row15	14	0.303	44	23	487.002
Row16	21	0.226	43	22	384.529

Fig. 34: Remove Column results

Shuffled - 2:33 - Shuffle

File

Table "default" - Rows: 245 | Spec - Columns: 411 | Properties | Flow Variables

Row ID	D774	D775	D776	D777	dep_var
Row64_dup	44	-0.654	0.316	2.634	inactive
Row16	26	-0.848	0.581	5.171	active
Row65	44	-0.729	0.24	2.555	active
Row114	09	-0.646	0.214	1.105	active
Row105_dup	85	-0.79	0.643	4.613	inactive
Row42	09	-0.646	0.207	2.535	active
Row130	22	-0.627	0.194	0.832	active
Row120	55	-0.35	0.333	2.323	active
Row59_dup	44	-0.737	0.308	2.989	inactive
Row15_dup	54	-0.775	0.522	4.237	inactive
Row76	07	-0.776	0.6	3.059	active
Row7	92	-0.38	0.171	3.509	active
Row113	26	-0.594	0.375	0.237	active
Row28_dup	54	-0.74	0.231	2.315	inactive
Row26	28	-0.679	0.3	0.214	active
Row74	22	-0.306	0.414	3.139	active
Row111_dup	48	-0.751	0.429	3.258	inactive

Fig. 35: Shuffle results

Normalized table - 2:36 - Normalizer

File

Table "default" - Rows: 245 | Spec - Columns: 411 | Properties | Flow Variables

Row ID	D012	D015	D018	D019	D122
Row64_dup	-1.251	0.117	-1.958	-1.39	-1.463
Row16	0.895	-0.895	0.029	0.115	-0.496
Row65	-0.894	-0.735	-1.075	-0.512	-0.972
Row114	-0.715	0.582	-0.633	-0.512	-0.648
Row105_dup	0.716	-1.426	-0.412	-1.014	-1.102
Row42	-0.894	-1.108	-0.633	-0.387	-0.473
Row130	-0.715	2.369	-0.302	0.115	0.212
Row120	0.18	0.002	0.471	0.491	0.422
Row59_dup	0.18	-0.263	0.802	0.491	0.748
Row15_dup	2.147	-1.234	2.127	0.616	1.678
Row76	-0.357	-0.061	-1.627	-0.888	-2.008
Row7	-1.072	1.064	-0.081	2.121	0.319
Row113	0.18	0.782	0.029	-0.763	-0.161
Row28_dup	-1.43	-1.272	-1.296	0.616	-0.642
Row26	0.537	1.626	1.133	0.742	1.096
Row74	0.001	0.443	-0.302	-0.387	-0.566
Row111_dup	0.001	-1.024	-0.523	-0.638	-0.899

Fig. 36: Normalizer results

Classified Test data - 2:18 - Weka Predictor (3.7)

File

Table "default" - Rows: 49 | Spec - Columns: 412 | Properties | Flow Variables

Row ID	D775	D776	D777	dep_var	Predicti...
Row120	8	-0.133	-0.368	active	active
Row113	101	0.137	-1.838	active	active
Row74	66	0.388	0.206	active	active
Row23_dup	362	0.299	-0.159	inactive	inactive
Row41_dup	92	-2.29	0.185	inactive	inactive
Row102	395	1.038	0.196	active	active
Row84	431	0.946	0.335	active	inactive
Row101	351	-0.133	0.091	active	active
Row7_dup	539	0.775	1.031	inactive	active
Row90	494	0.775	0.238	active	active
Row121	361	1.148	1.215	active	active
Row96	456	-0.071	0.154	active	active
Row21_dup	423	0.483	0.376	inactive	inactive
Row25	504	1.136	0.397	active	active
Row77	469	-0.133	-0.799	active	active
Row93	469	0.946	-0.043	active	active
Row107	334	-0.006	-0.138	active	inactive

Fig. 37: Weka Predictor (3.7) results

Confusion matrix - 2:37 - Scorer

File

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

Row ID	active	inactive
active	22	10
inactive	3	14

Fig. 38: Scorer results (1)

Accuracy statistics - 2:37 - Scorer

File

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specifity	F-meas...	Accuracy	Cohen'...
active	22	3	14	10	0.688	0.88	0.688	0.824	0.772	?	?
inactive	14	10	22	3	0.824	0.583	0.824	0.688	0.683	?	?
Overall	?	?	?	?	?	?	?	?	?	0.735	0.466

Fig. 39: Scorer results (2)

Embark your own voyage of discovery!

Now, you can compare the accuracy statistics of the simple example (Step 9) and the extended example (Step 10). This was just a simple example to get you started. There is a lot more to discover. Try to explore it! We tried to keep it simple and intuitive. We would love to receive your feedback and find out what you liked and what you did not like; things you find not functional or things that did not seem to work.