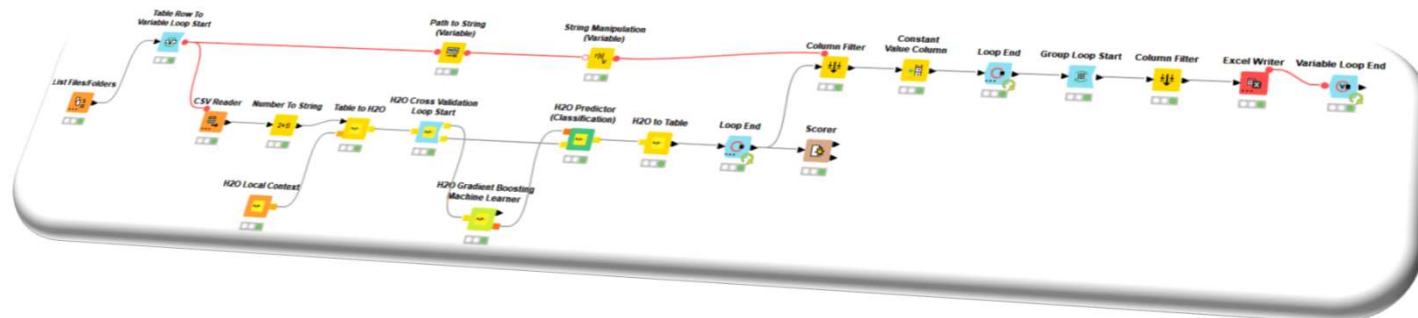


Gépi tanulás KNIME környezetben: példák a gyógyszer- és anyagtudomány területéről

Bajusz-Rác Anita

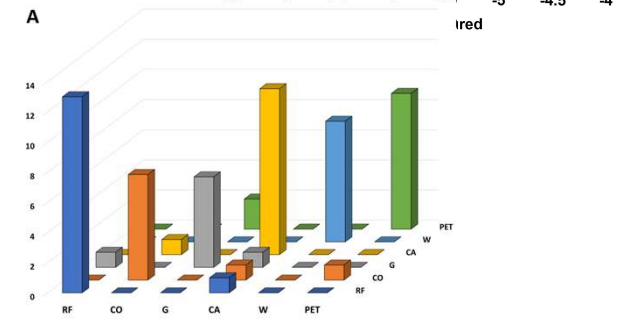
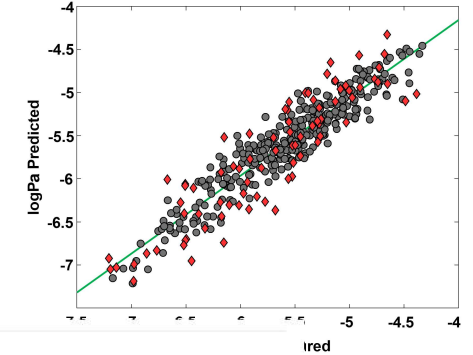
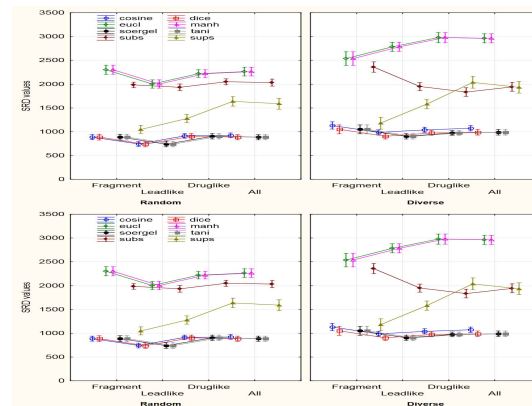
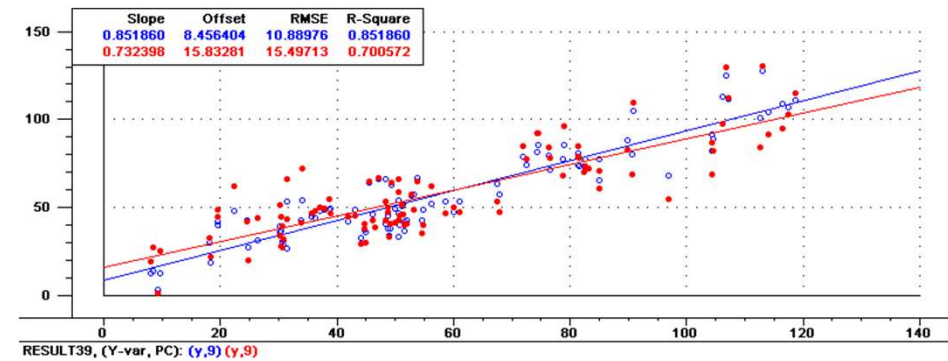
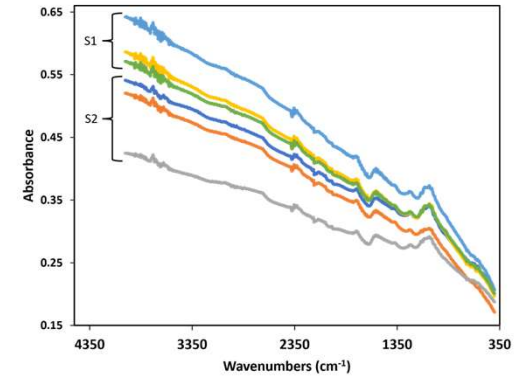
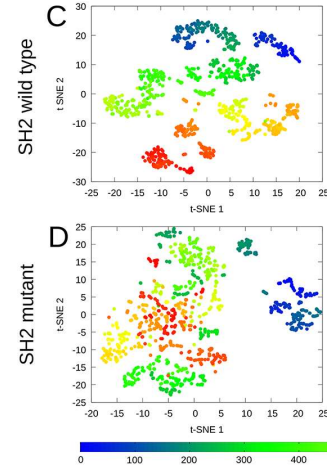
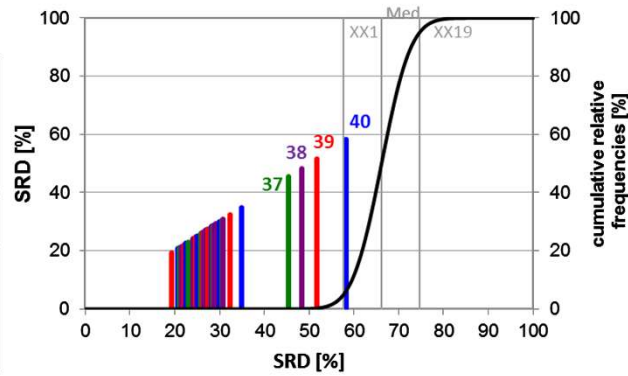
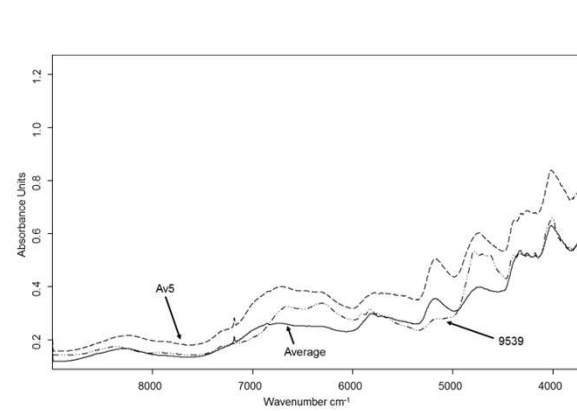
HUN-REN TTK Anyag- és Környezetkémiai Intézet



Occam borotvája vagy gépi tanulás Workshop, Szeged, 2024. március 8.

Bevezetés

- Honnan indultam? Hova érkeztem?



Miért éppen KNIME?

- Egyre nagyobb igény az ML módszerek használatára
- KNIME: híd a klasszikus statisztikai szoftverek és a programozás világa között
- Egyszerű használat – GUI és drag-and-drop mechanizmus
- Kiterjeszthetőség: Integrálható eszközök és algoritmusok más szoftverkörnyezetből (R, Python, Java, Weka stb.)
- Ingyenesen elérhető
- Közösségi támogatás

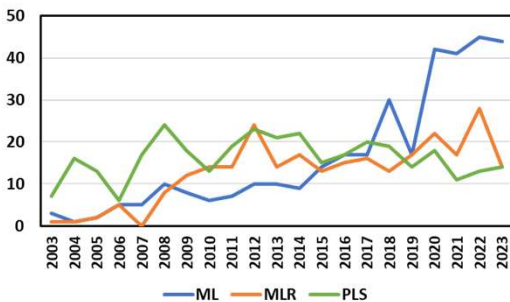


Miért a gyógyszerkutatás és miért az anyagtudomány?

- „Témát lehet adni és kapni. A téma adódik. Kialakulhat. Létrejöhet és beindulhat. A témával kapcsolatban minden mozgásforma lehetséges, csak éppen a választás lehetősége kizárt.”

(Ezésez Géza karrierje)

- A gyógyszerkutatás – választás
- Az anyagtudomány – létrejött
- Gyógyszerkutatás: QSAR, reakció tervezés stb. – több évtizedes hagyomány
- Anyagtudomány: feltörekvő trend



nature materials

nature reviews drug discovery

Explore content | About the journal | Publish with us

nature > nature materials > perspectives > article

Perspective | Published: 18 April 2019

Exploiting machine learning for e discovery and development

Sean Ekins, Ana C. Puhl, Kimberley M. Zorn, Thomas R. Lane, J Hickey & Alex M. Clark

Review Article | Published: 11 April 2019

Applications of machine learning for drug discovery

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer & Shanrong Zhao

nature > npj computational materials > review articles > article

Review Article | Open access | Published: 08 August 2019

Recent advances and applications of machine learning in solid-state materials science

Review Article | Open access | Published: 05 April 2022

Recent advances and applications of deep learning methods in materials science

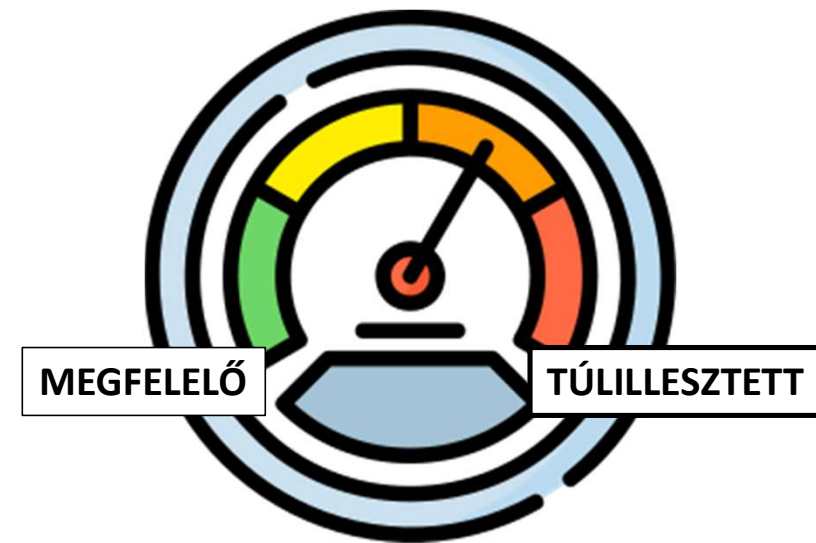
Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon J. L. Billinge, Elizabeth Holm, Shyue Ping Ong & Chris Wolverton

npj Computational Materials 8, Article number: 59 (2022) | Cite this article

„machine learning”/”pls”/”mlr” AND „qsar” AND „drug design”

Machine learning – Gépi tanulás osztályozási és regressziós feladatokra

- A méret a lényeg?
 - Regresszió: igen, Klasszifikáció: nem feltétlen
- Linearitás vs. nemlinearitás
- Üres cellák?
- Dimenziók száma, dimenzió redukció
- Modell validálás és teljesítmény értékelés
 - Túlillesztett modell kiszűrése
- Leggyakrabban használt algoritmusok: ANN, Fa-alapú (XGBoost, Random forest), SVM



A KNIME szoftverkörnyezet

- KNIME: drag-and-drop alapú programozás
- Definíciók: Workspace, workflow, node
- A vonalak jelentik a kapcsolatot az egységek között (adatáramlás)
- Node repository
- Explorer
- Node description
- Console

```
dropdownRender={this.renderDropdown}  
notFoundContent="No Matches"  
onDropdownVisibleChange={this.handleDropdownVisibleChange}  
onInputKeyDown={this.handleEnterKeyDown}  
onSearch={debounce(this.handleSearch, 3000, {leading: true})}  
onSelect={this.handleSelect}  
open={isDropdownVisible}  
placeholder="Search or enter new field name"  
ref={this.selectRef}  
showArrow={false}  
value={undefined}
```

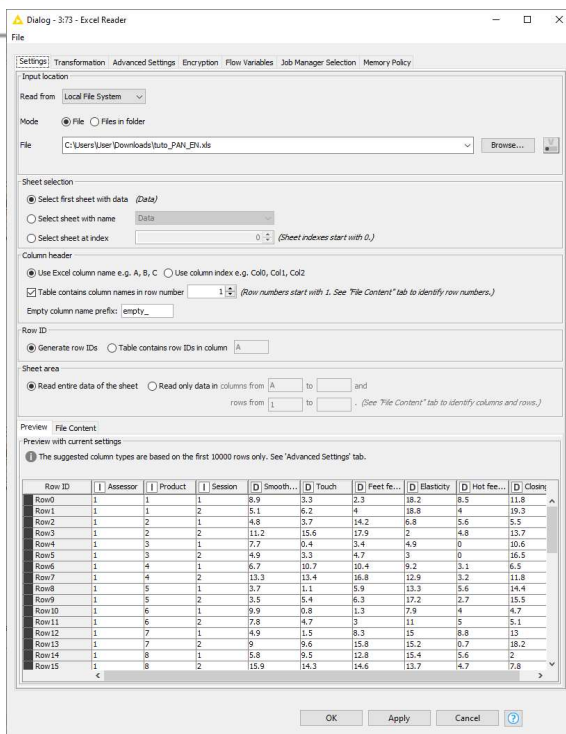
VS.

The screenshot displays the KNIME software interface. On the left, the 'Node Repository' is visible, showing a tree structure of nodes categorized by IO, Manipulation, Views, Analytics, Mining, and Statistics. The main workspace shows a workflow with nodes connected by lines, representing data flow. The nodes include 'List Files/folders', 'Table Row To Variable Loop Start', 'CSV Reader', 'Number To String', 'Table to H2O', 'H2O Cross Validation Loop Start', 'H2O Gradient Boosting Machine Learner', 'H2O Predictor (Classification)', 'H2O to Table', 'Loop End', 'Scorer', 'Column Filter', 'Constant Value Column', and 'Loop'. On the right, the 'Description' panel is open for the 'H2O Gradient Boosting Machine Learner' node, showing 'General Settings' and 'Algorithm Settings'.

Legfontosabb Node egységek - adatelőkezelés

- Filterelés, szelektálás, vágás, módosítás

- IO
- Read
 - Excel Reader
 - File Reader
 - File Reader (Complex Format)
 - ARFF Reader
 - CSV Reader
 - Line Reader
 - Table Reader
 - PMML Reader
 - Fixed Width File Reader
 - Model Reader
 - Read Excel Sheet Names
 - Read Images
 - Explorer Browser
- Write
- Connectors
- File Folder Utility
- Other



Excel Reader

Dialog - 3/73 - Excel Reader

Settings | Transformation | Advanced Settings | Encryption | Flow Variables | Job Manager Selection | Memory Policy

Input location

Read from: Local File System

Mode: File Files in folder

File: C:\Users\User\Downloads\tuto_PAN_EN.xls

Sheet selection

Select first sheet with data (Data)

Select sheet with name: Data

Select sheet at index: 0 (Sheet indexes start with 0.)

Column header

Use Excel column name e.g. A, B, C Use column index e.g. Col0, Col1, Col2

Table contains column names in row number: 1 (Row numbers start with 1. See 'File Contents' tab to identify row numbers.)

Empty column name prefix: empty_

Row ID

Generate row IDs Table contains row IDs in column: A

Sheet area

Read entire data of the sheet Read only data in columns from: A to and rows from: 1 to

Preview | File Content


The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	Assessor	Product	Session	Smooth...	Touch	Feet fe...	Elasticity	Hot fee...	Closin
Row0	1	1	1	8.9	3.3	2.3	18.2	8.5	11.8
Row1	1	1	2	5.1	6.2	4	18.8	4	19.3
Row2	1	2	1	4.8	3.7	14.2	6.8	5.6	5.5
Row3	1	2	2	11.2	15.6	17.9	2	4.8	15.7
Row4	1	3	1	7.7	9.4	3.4	4.9	0	10.6
Row5	1	3	2	4.9	3.3	4.7	3	0	16.5
Row6	1	4	1	6.7	10.7	10.4	9.2	3.1	16.5
Row7	1	4	2	13.3	13.4	16.8	12.9	3.2	11.8
Row8	1	5	1	3.7	1.1	5.9	13.3	5.6	14.4
Row9	1	5	2	3.5	5.4	6.3	17.2	2.7	15.5
Row10	1	6	1	9.9	9.8	1.3	7.9	4	6.7
Row11	1	6	2	7.8	4.7	3	11	5	5.1
Row12	1	7	1	4.9	1.5	8.3	15	8.8	13
Row13	1	7	2	9	9.6	15.3	15.2	0.7	18.2
Row14	1	8	1	5.8	9.5	12.8	15.4	5.6	2
Row15	1	8	2	15.9	14.3	14.6	13.7	4.7	7.8

- Row
 - Filter
 - Duplicate Row Filter
 - Filter Apply
 - Filter Apply Row Splitter
 - Filter Definition Merger
 - HiLite Row Splitter
 - Nominal Value Row Filter
 - Nominal Value Row Splitter
 - Numeric Row Splitter
 - Reference Row Filter
 - Reference Row Splitter
 - Row Filter
 - Row Splitter
 - Rule-based Row Filter
 - Rule-based Row Filter (Dictionary)
 - Rule-based Row Splitter
 - Rule-based Row Splitter (Dictionary)


Oszlop

Column Auto Type Cast




Node 1

Column Rename




Node 3

String To Number




Node 4

Column Filter




Node 5

Joiner




Node 6

Target Shuffling




Node 15

Transpose




Node 14

Math Formula



Node 2


Rule Engine



Node 13


Sor

Row Filter




Node 8

Equal Size Sampling




Node 10

Partitioning




Node 7

Sorter



Node 12

GroupBy

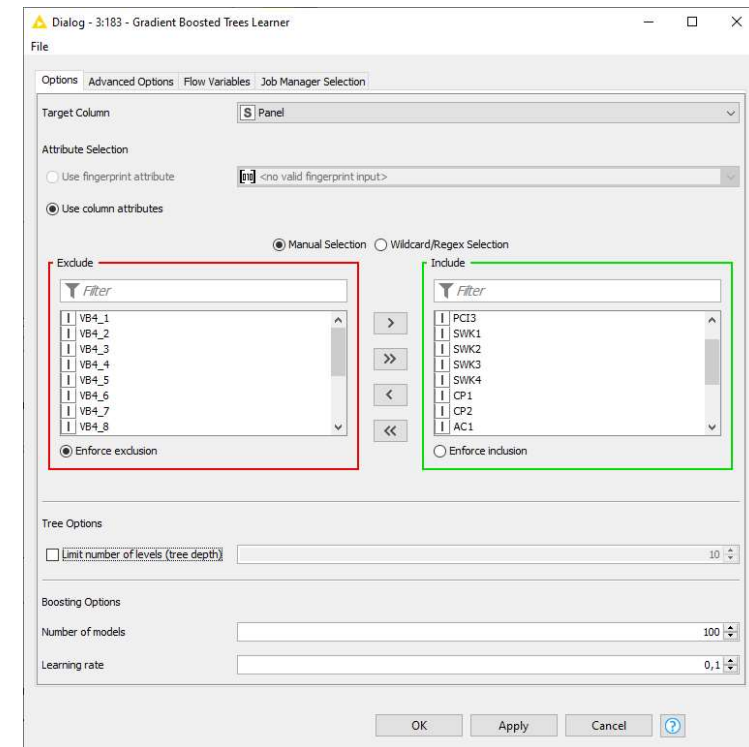
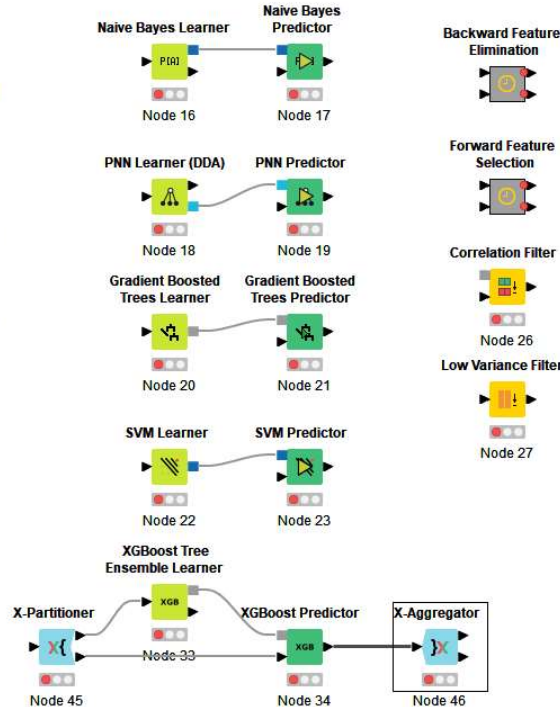
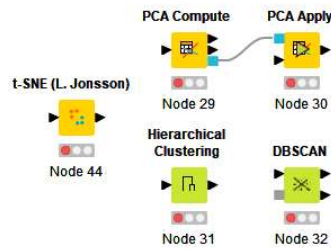
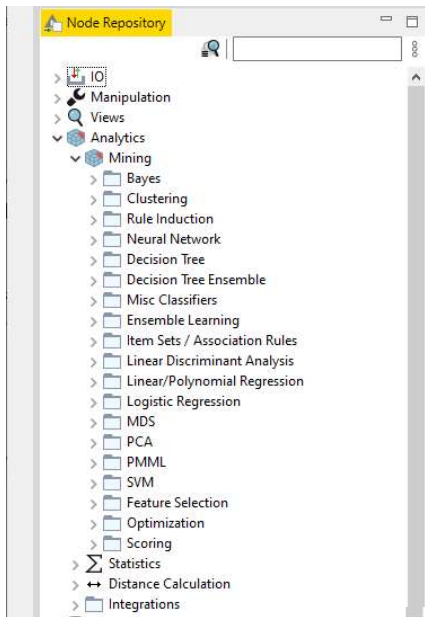


Node 9

Matematikai műveletek

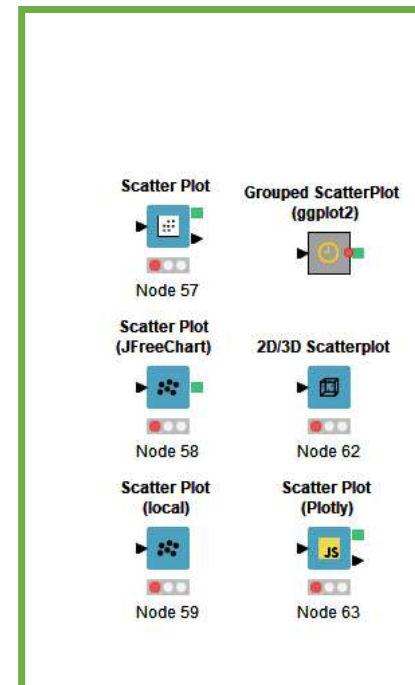
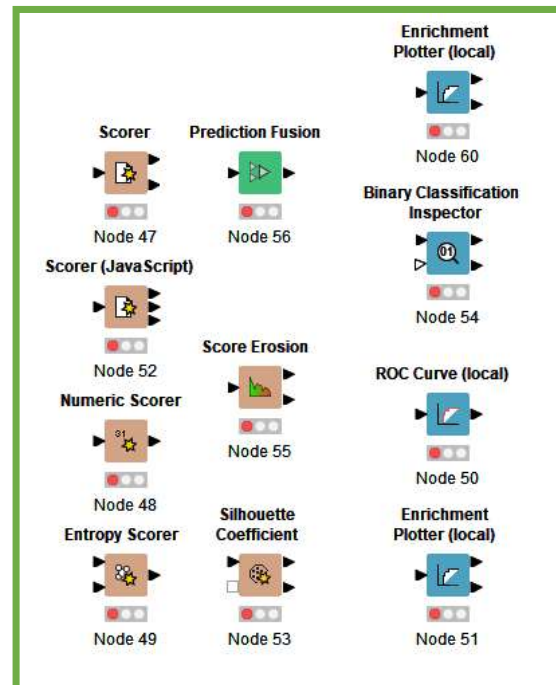
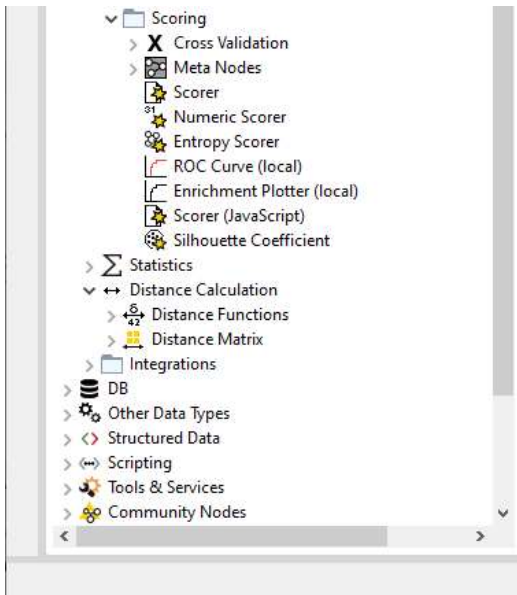
Legfontosabb Node egységek - Modellezés

- Beépített egység csomagok



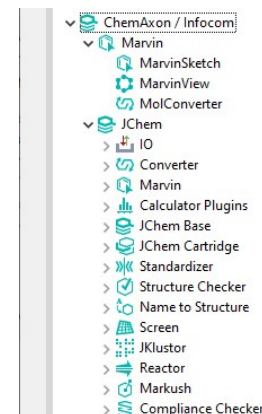
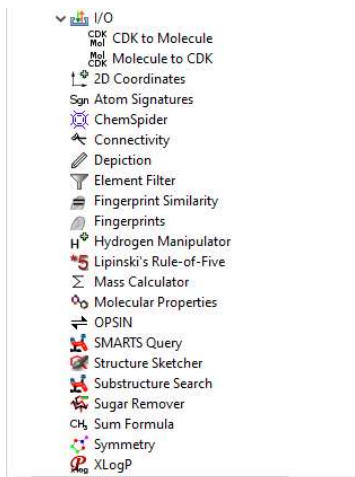
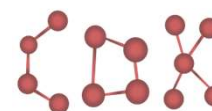
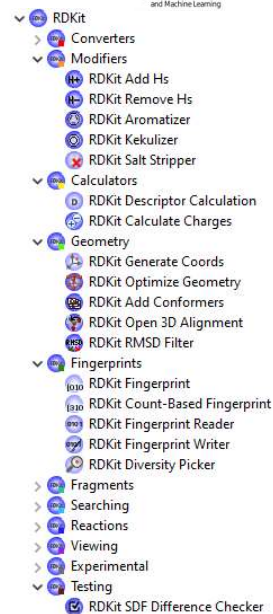
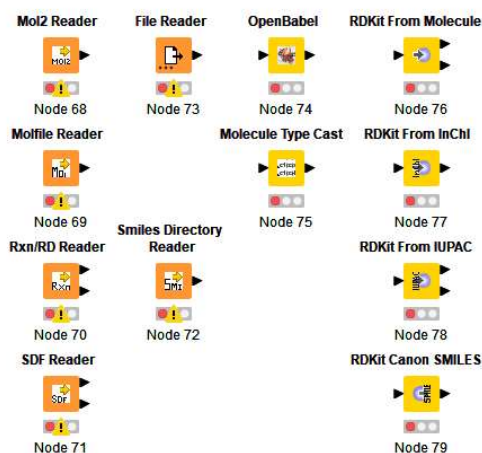
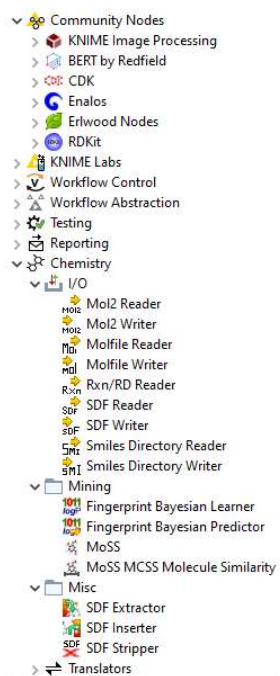
Legfontosabb Node egységek - Kiértékelés

- Teljesítmény paraméterek, valószínűségi értékek transzformálása, vizualizáció, exportálás



KNIME a gyógyszerkutatásban (QSAR/QSPR) - adatelőkészítés, adat generálás

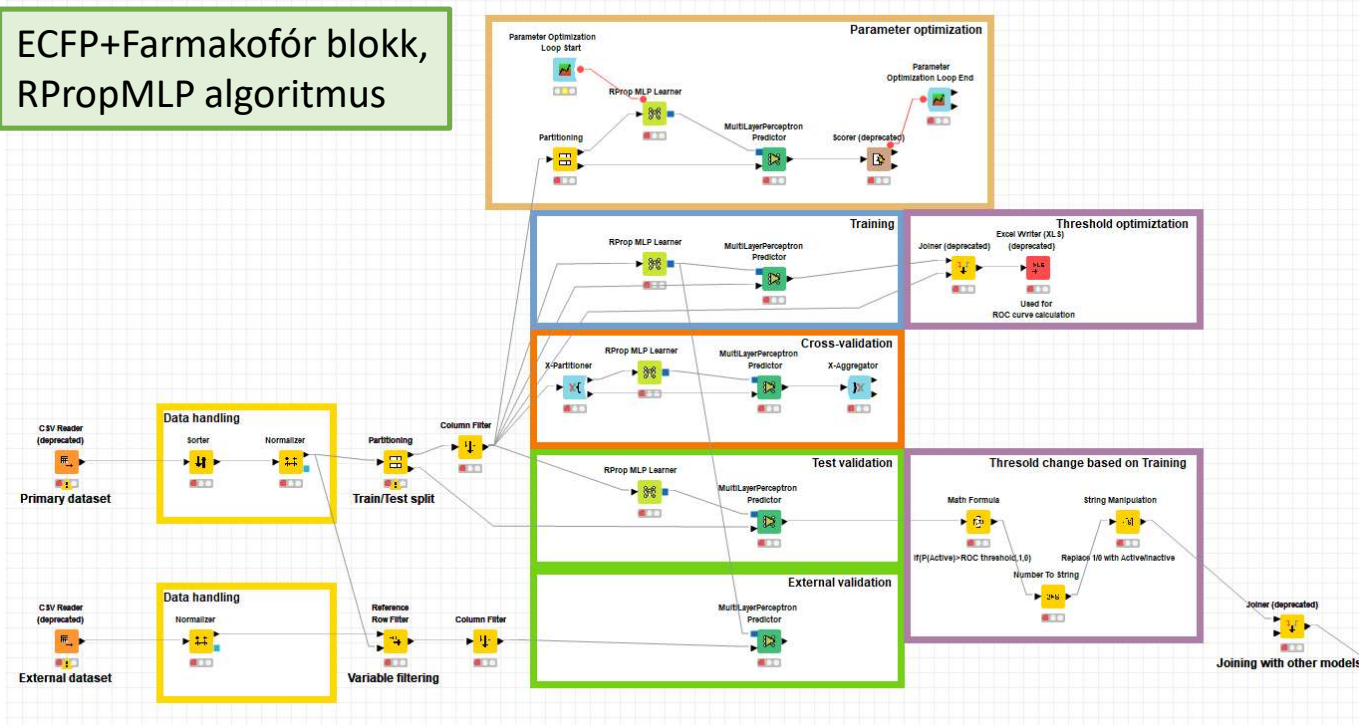
- Gyógyszerkutatásban, keminformatikában használható node-ok
- Közösségi (Community) node csomagok



ADME modellezés KNIME környezetben

- Cytochrome P450 (CYP) 2C9 izoenzimjén való aktivitás
- 45000 különböző molekula
- XGBoost, RPropMLP algoritmus
- 1D,2D,3D deskriptorok, ECFP + Farmakofór ujjlenyomatok, Kölcsönhatás ujjlenyomat + Dokkolás

ECFP+Farmakofór blokk,
RPropMLP algoritmus

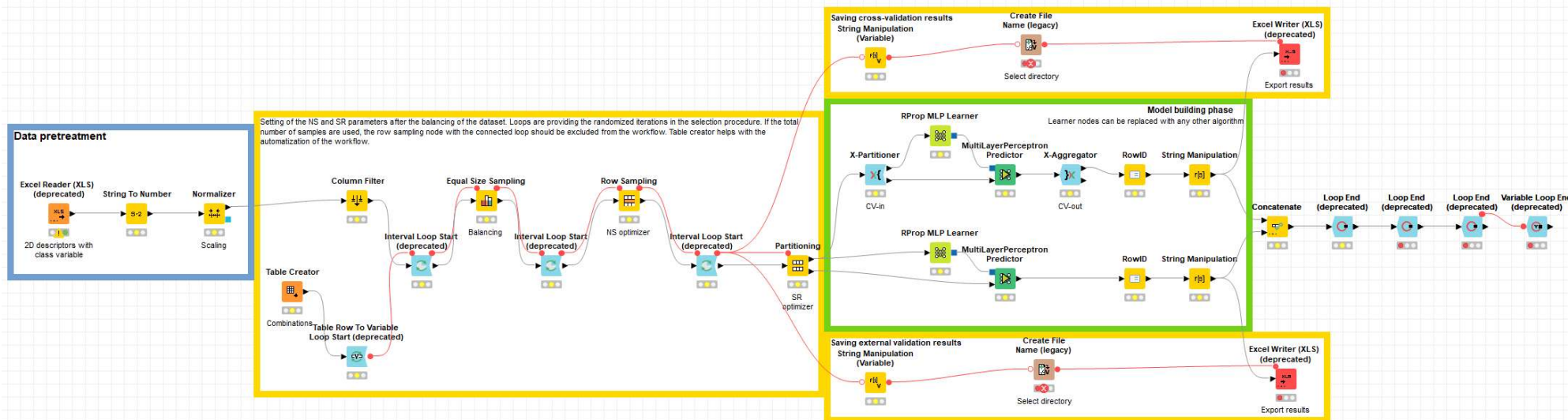
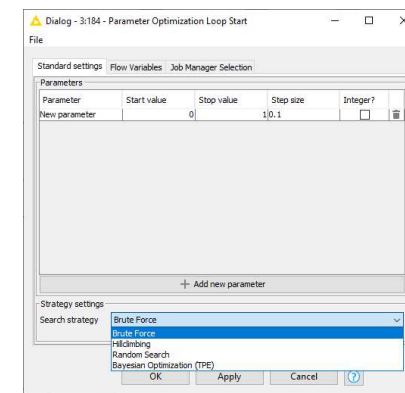
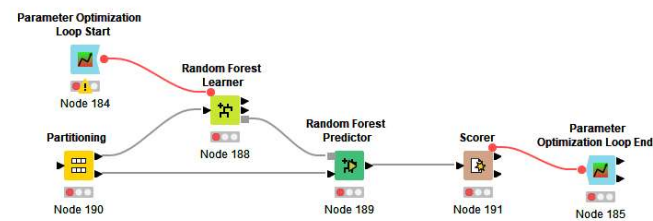


- Valószínűségi limitértékek optimalálása
- Konszenzus voksolás a hat modell alapján
- AUC: 0.85 (Belső teszt), 0.84 (Külső teszt)

Rácz, A., Keserű, G.M. Large-scale evaluation of cytochrome P450 2C9 mediated drug interaction potential with machine learning-based consensus modeling. *J Comput Aided Mol Des* **34**, 831–839 (2020). <https://doi.org/10.1007/s10822-020-00308-y>

Automatizálhatósági lehetőségek

- Probléma specifikus loop egységek
- A loop start – loop end node-ok
- Paraméter optimálás →
- Iteratív modell futtatások



Rácz, A.; Bajusz, D.; Héberger, K. Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. *Molecules* **2021**, *26*, 1111. <https://doi.org/10.3390/molecules26041111>

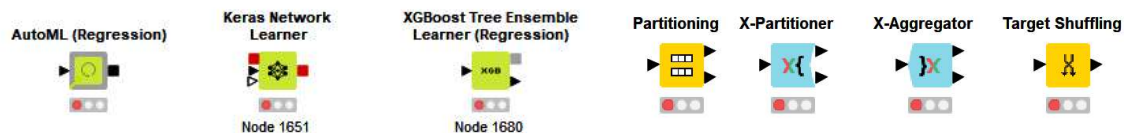
Anyagtudomány és spektrumok KNIME környezetben

- **Osztályozás**

- Polietilén minták osztályozása adalékanyag szerint
- Szén nanostruktúrák osztályozása a kiindulási alapanyag szempontjából

- **Regresszió**

- Polimer minták merevségének meghatározása
- Spektrum adatelőkezelés: előzetesen kivitelezhető
- Outlier szelekció csak numerikus szinten elérhető
- Modellézés: beépített egységek mellett – Keras, Sklearn, R stb. csomagok



Osztályozás

Szén nanostruktúrák kiindulási alapanyag szerinti osztályozása

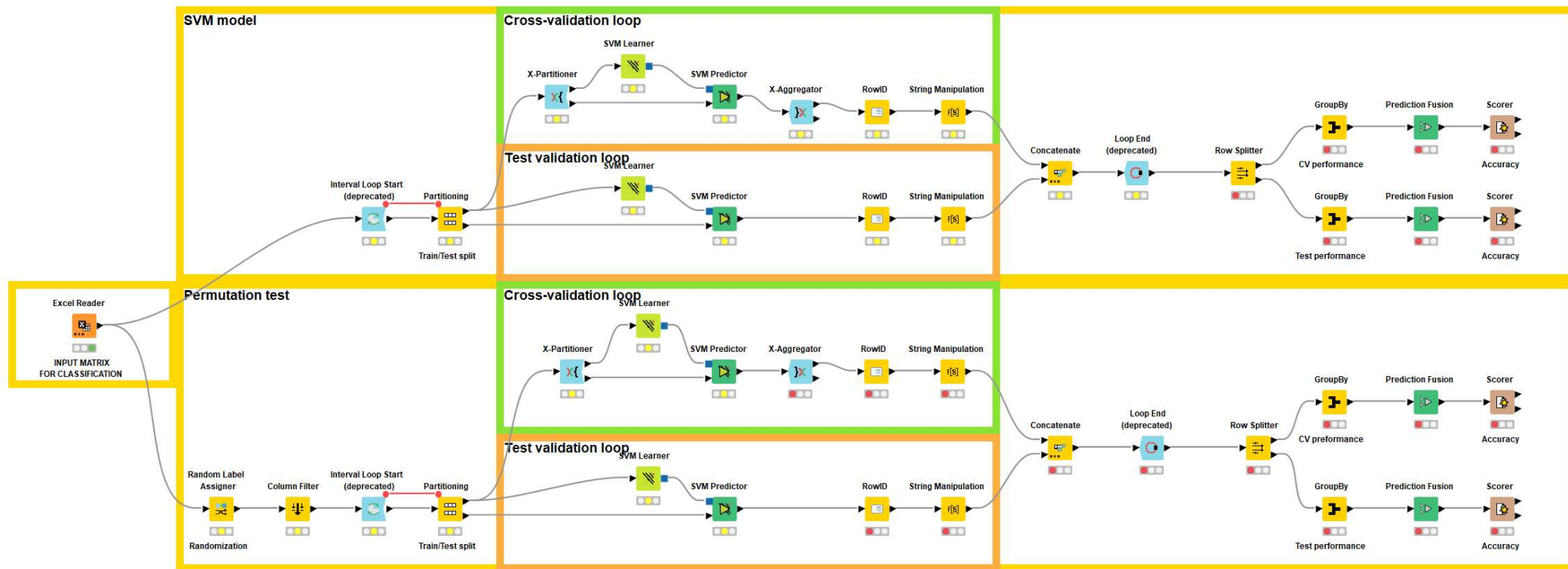
- 62 minta
- ATR-IR spektrum (1867 változó) – alapvonal korrekció + deriválás + SNV
- 6 csoport

Polietilén minták adalékanyag szerinti osztályozása

- 36 minta
- ATR-IR spektrum (3733 változó) – standardizálás
- 4 csoport

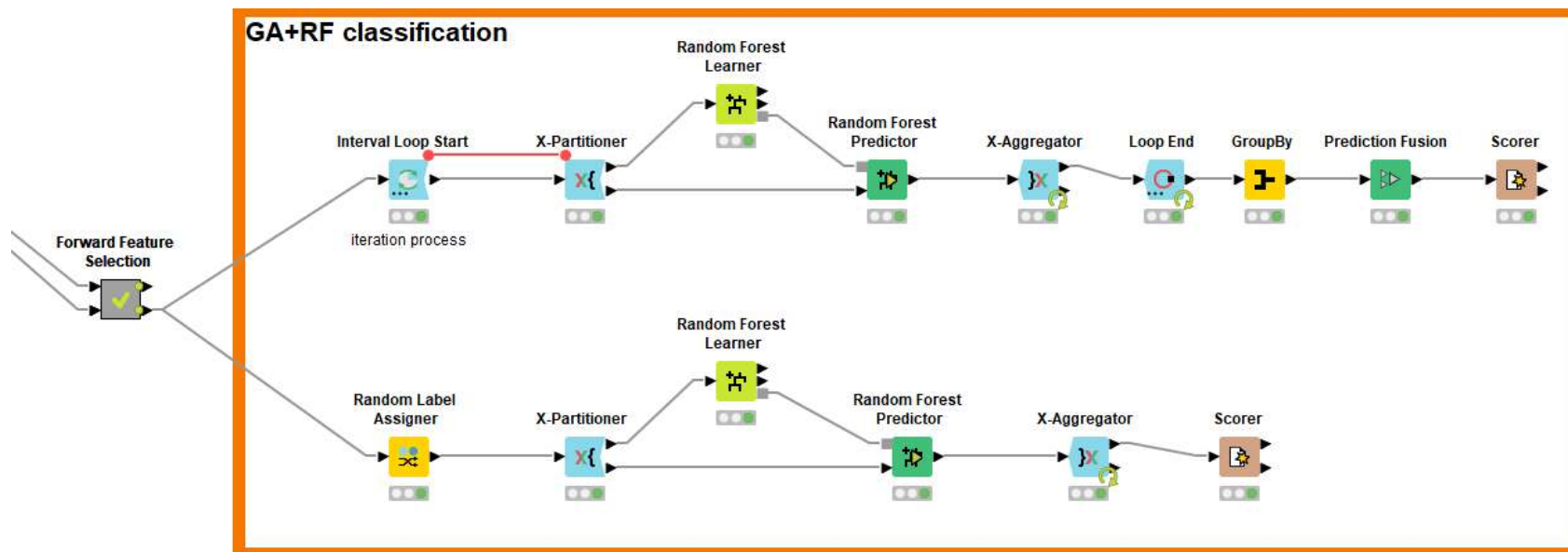
• Szén nanostruktúrák kiindulási alapanyag szerinti osztályozása

- SVM algoritmus
- Double-CV iteratív módon
- Permutációs teszt – túlílllesztés elkerülése
- BACC: 0.865 (CV)



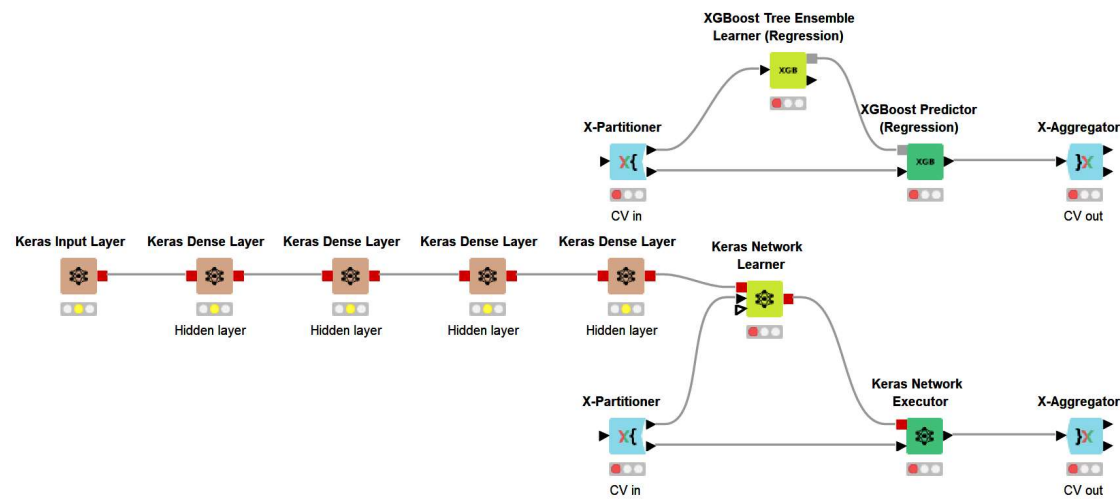
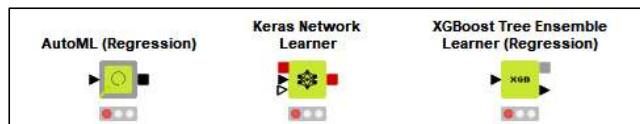
• Polietilén minták adalékanyag szerinti osztályozása

- Genetikus algoritmus
- Random Forest
- 3-fold CV 10x iterációban
- Permutációs teszt használata
- BACC: 0.917



Regressziós modell konszenzus értékeléssel

- Polimer minták merevségének predikciója
 - 195 minta ATR-IR spektruma
 - Y: merevség (modulus)
 - Spektrum előkezelés: Deriválás+SNV – 3451 változó

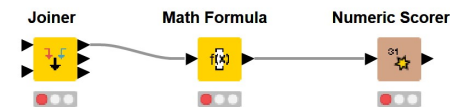


Konszenzus modell:

$$R^2_{CV} = 0.953$$

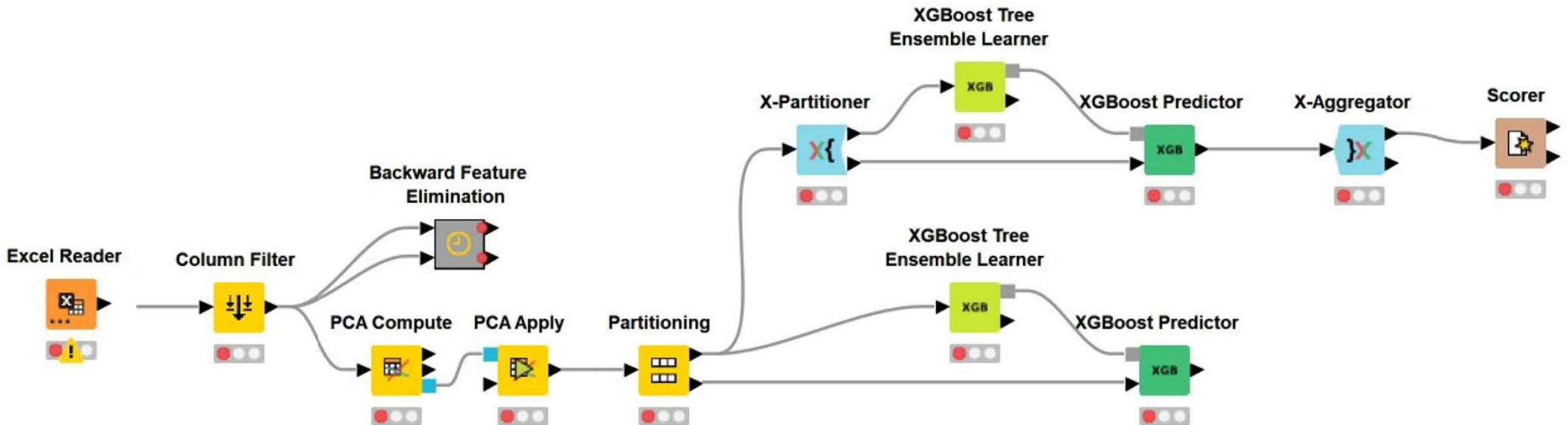
$$R^2_{Test} = 0.952$$

$$R^2_{Rnd} = -0.442$$



KNIME workflow építése lépésről lépésre

- Az adat importálástól a modellezésig

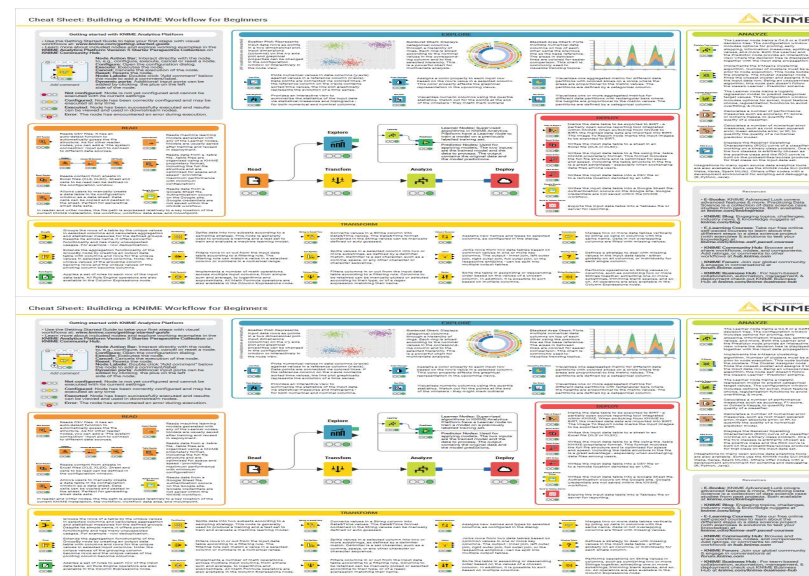
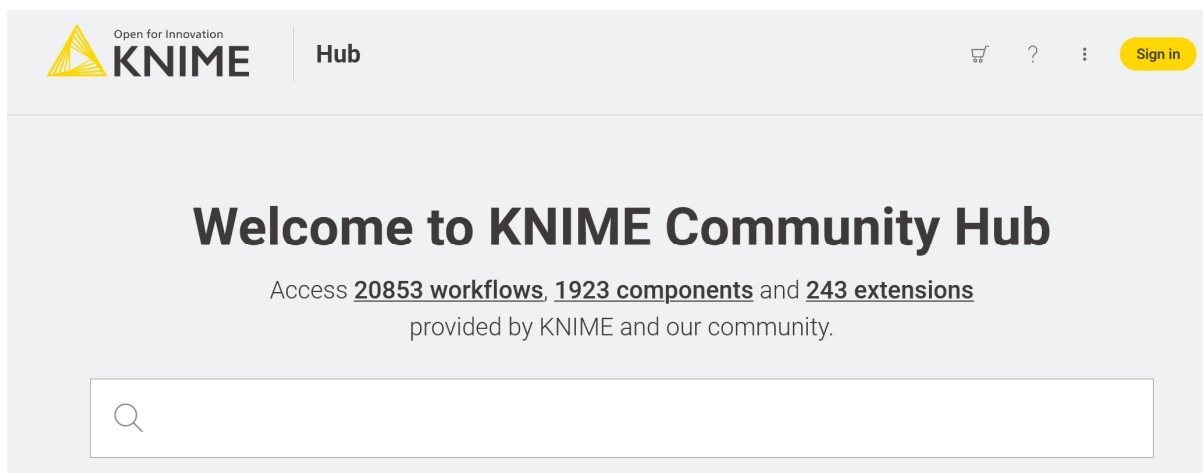


KNIME: kihívások és limitációk

- Bonyolult regressziós workflow-k
- Blokkonkénti változó szelekció hiánya
- Spektrum adatelőkezelés
- Memória/disc igényes
- Bonyolultan az egyszerűt
- Vizualizáció
- Batch processing vs. Real time analysis

Hasznos oldalak a KNIME használatához

- CheatSheets: <https://www.knime.com/knimepress#cheat-sheets>
- KNIME Hub, KNIME community forum, NodePit



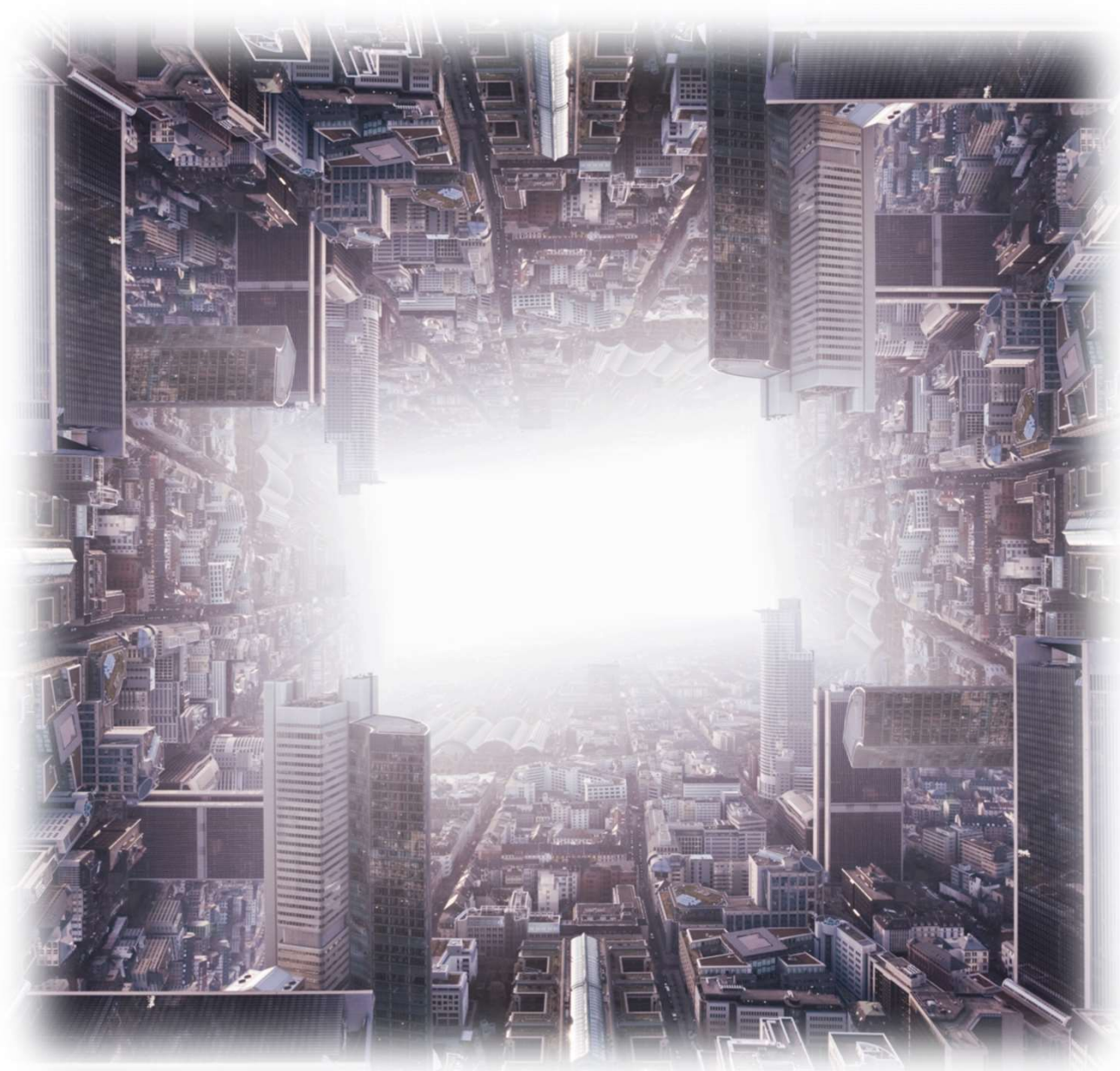
Welcome to the KNIME Community Forum

Advance your skills and connect with peers using KNIME.



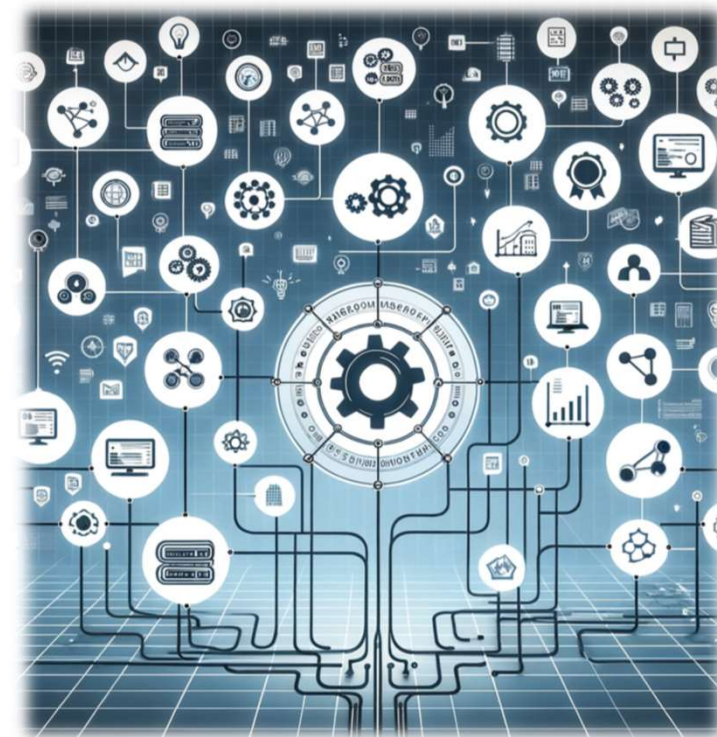
KNIME: jövőbeli lehetőségek

- Felhő alapú adatelemzés és modellezés
- Automatikus ML modellezés
- ML interpretációs lehetőségek
- Regressziós modellezés bővítése



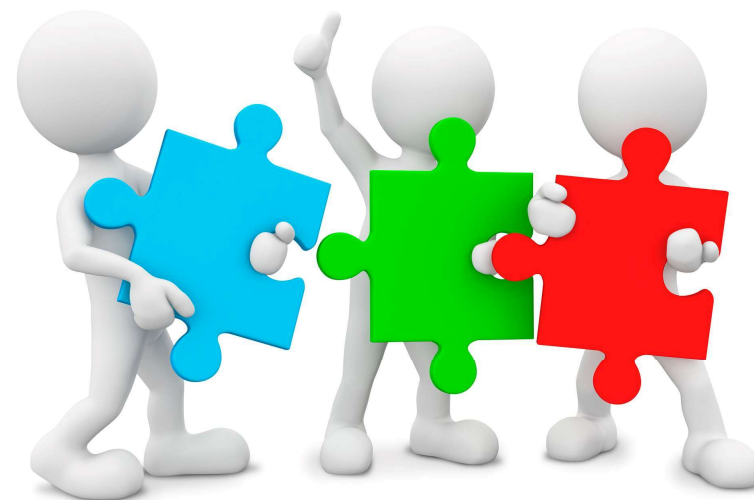
Összefoglalás

- Vizuális megjelenés = logikusabb, átláthatóbb a programot nem ismerő számára is
- Egyre több integráció, „community package”, ML algoritmus
- Nagyon jól használható keminformatikai és analitikai feladatokban
- Spektrum adatoknál „hybrid”-ként kezelhető
- Klasszifikációs algoritmusok száma végtelen
- A határait csak a számítógépes erőforrás jelenti



Köszönetnyilvánítás

- Héberger Károly
- Bajusz Dávid
- Keserű György Miklós
- Klébert Szilvia
- Tátraaljai Dóra
- László Krisztina
- Várdai Róbert
- Plazmakémiai Kutatócsoport



-
- OTKA K134260
 - A Bolyai János Kutatási Ösztöndíj támogatásával készült.
 - AZ INNOVÁCIÓS ÉS TECHNOLÓGIAI MINISZTERIUM ÚNKP-23-5 KÓDSZÁMÚ ÚJ NEMZETI KIVÁLÓSÁG PROGRAMJÁNAK A NEMZETI KUTATÁSI, FEJLESZTÉSI ÉS INNOVÁCIÓS ALAPBÓL FINANSZÍROZOTT SZAKMAI TÁMOGATÁSÁVAL KÉSZÜLT.

**Köszönöm a
figyelmet!**



„Draw me a half robot half human chemist in pen draw style“