Módulo 8

8.4 Aprendizaje No Supervisado: Clustering y Reglas de Asociación en KNIME

Por **María Martínez Rojas**
Profesora Titular en CA, Universidad de Granada

Por **José Manuel Soto Hidalgo**
Profesor Titular en ICAR, Universidad de Granada

1. INTRODUCCIÓN

Esta cápsula se centra en la implementación en KNIME de los distintos algoritmos de aprendizaje no supervisado que se han introducido en el módulo 6. Se realizarán flujos de datos representativos del ciclo de vida de Ciencia de Datos para resolver problemas de clustering y reglas de asociación.

Al igual que en la cápsula anterior, se utilizarán como ejemplo los conjuntos de datos ya utilizados en el módulo 2, 3 y 6. En concreto, los datos de expresión genética comentados en el módulo 2 y utilizados como datos de ejemplo en el módulo 6.

2. CLUSTERING

En este apartado se van a crear flujos de datos para resolver un problema de clustering centrado en identificar grupos en los datos sin utilizar ninguna información a priori sobre categorías, tipos, clases o grupos conocidos en los datos. El flujo representado en la Figura 1 muestra una implementación de los dos tipos de clustering comentados en el módulo 6: clustering jerárquico y k-medias. El clustering jerárquico se implementa de dos formas distintas, incluyendo un modelo de distancias externo y con un nodo propio que incluye la matriz de distancias. Finalmente se visualizan los resultados a través de un heatmap, y se muestran distintas opciones de visualización: continua y discreta.

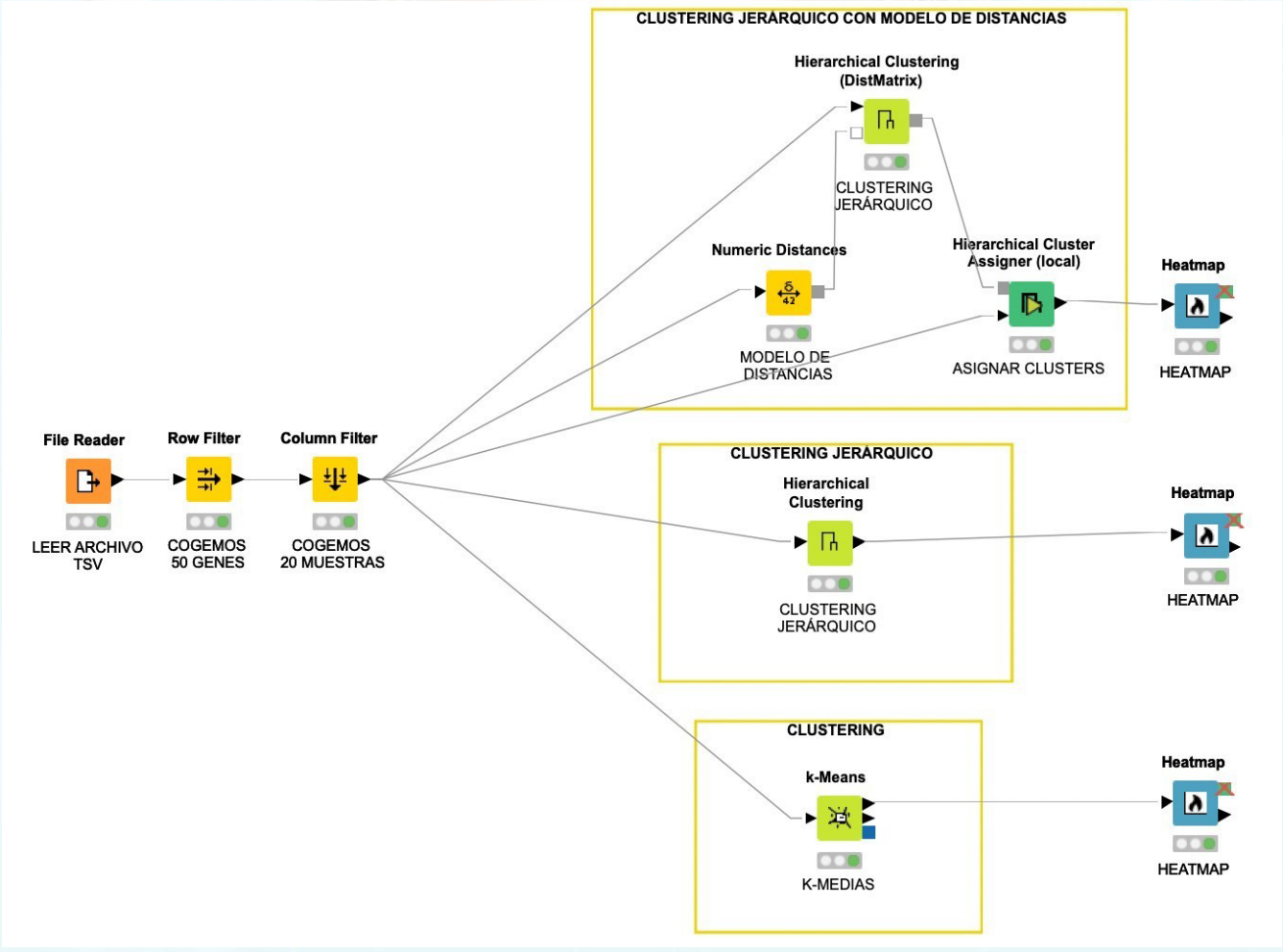


Figura 1: Flujo de datos con clustering jerárquico y k-medias.

Primero leemos los datos (matriz de expresión) con el nodo *File Reader* y a modo de ejemplo para facilitar el entendimiento de los dendogramas, se seleccionará sólo una porción de la matriz de expresión (50 genes y 20 muestras) con los nodos *Row Filter* y *Column Filter*, respectivamente. Una vez seleccionados los 50 genes y 20 muestras mostramos varias formas de hacer clustering jerárquico con KNIME. Por un lado, se puede crear un modelo de distancias (nodo *Numeric Distances*) que es el que utilizará el nodo *Hierarchical Clustering (DistMatrix)* junto con los datos para hacer el clustering jerárquico. La Figura 2 muestra las distintas opciones de configuración del nodo *Numeric Distances*, donde se pueden utilizar distintas distancias: Euclídea, Manhattan, Máximo, etc. las variables (muestras) a considerar en el clustering así como tratamiento de valores perdidos.

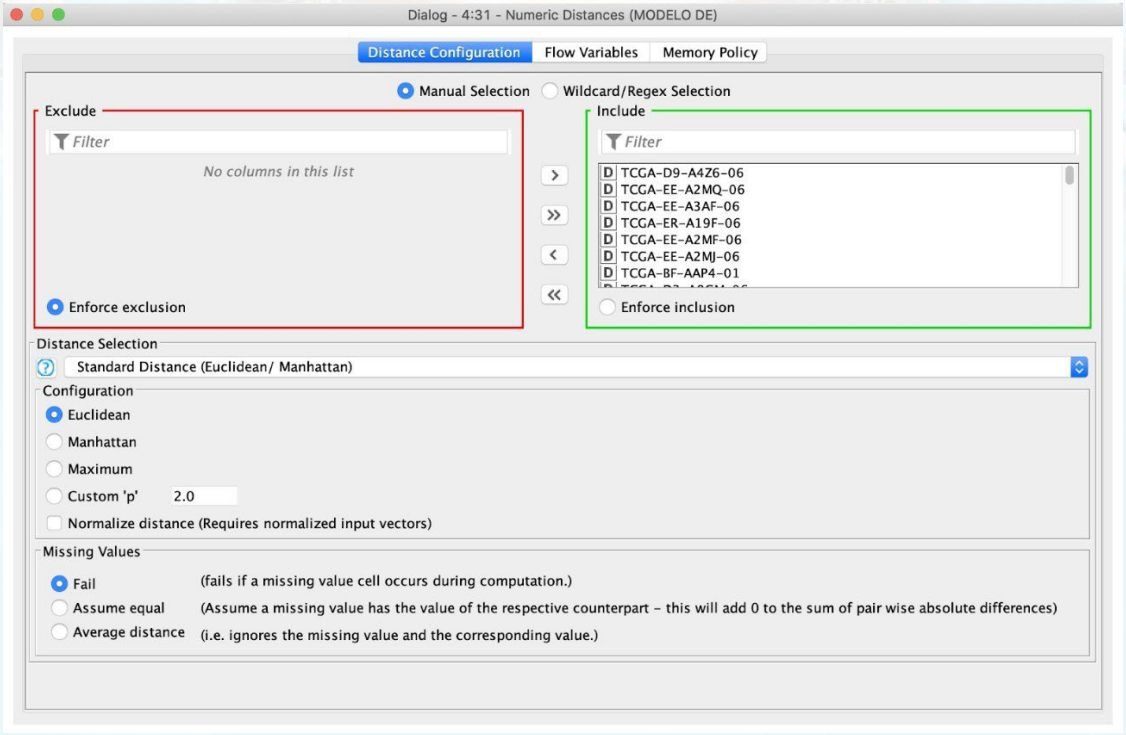


Figura 2: Opciones y parámetros del nodo Numeric Distances.

El nodo *Hierarchical Clustering (DistMatrix)* genera un modelo de datos con los cluster que junto con el nodo *Hierarchical Cluster Assigner* y los datos, asigna cada entrada a un número de cluster. Finalmente, podemos dibujar el heatmap con el nodo *Heatmap* e interactuar con el. Por ejemplo, podemos visualizar el heatmap con representación de colores continua (Figura 3) o discreta (Figura 4).

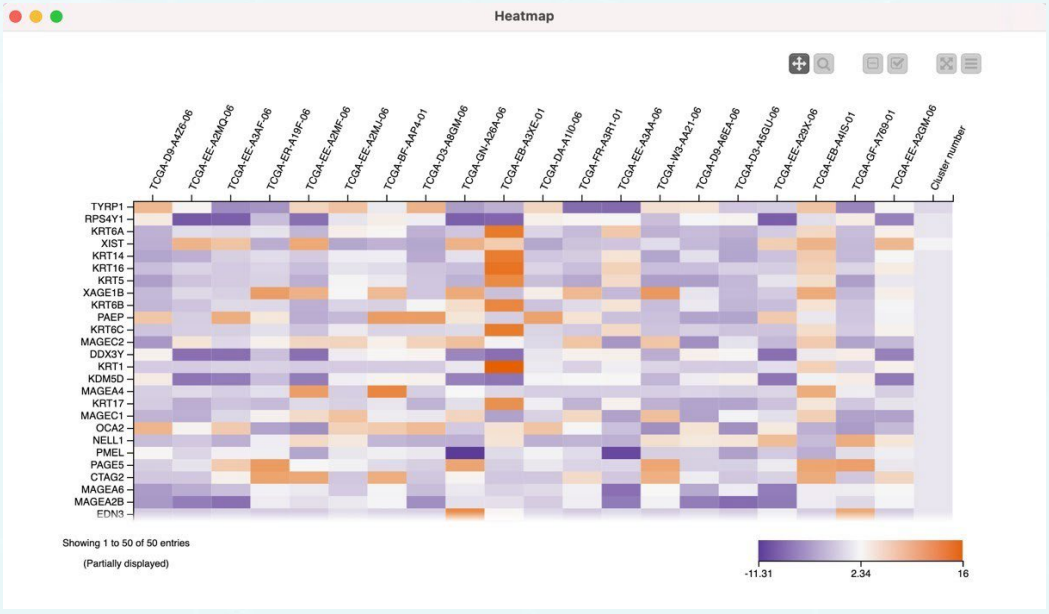


Figura 3: Visualización de resultados a través de heatmap con representación continua.

Figura 4: Visualización de resultados a través de heatmap con representación discreta.

También se puede realizar clustering jerárquico sin utilizar un modelo de distancias a través del nodo *Hierarchical Clustering*. Este nodo presenta opciones como el tipo de distancia a utilizar, número de cluster de salida y tipo de unión, así como las variables a considerar (Figura 5). Con este nodo, se puede visualizar el dendograma e interactuar con él. Para ello, una vez ejecutado el nodo, basta con hacer clic con el botón derecho y seleccionar la opción "View: Dendogram/distance view", y el resultado es el que se muestra en la Figura 6.

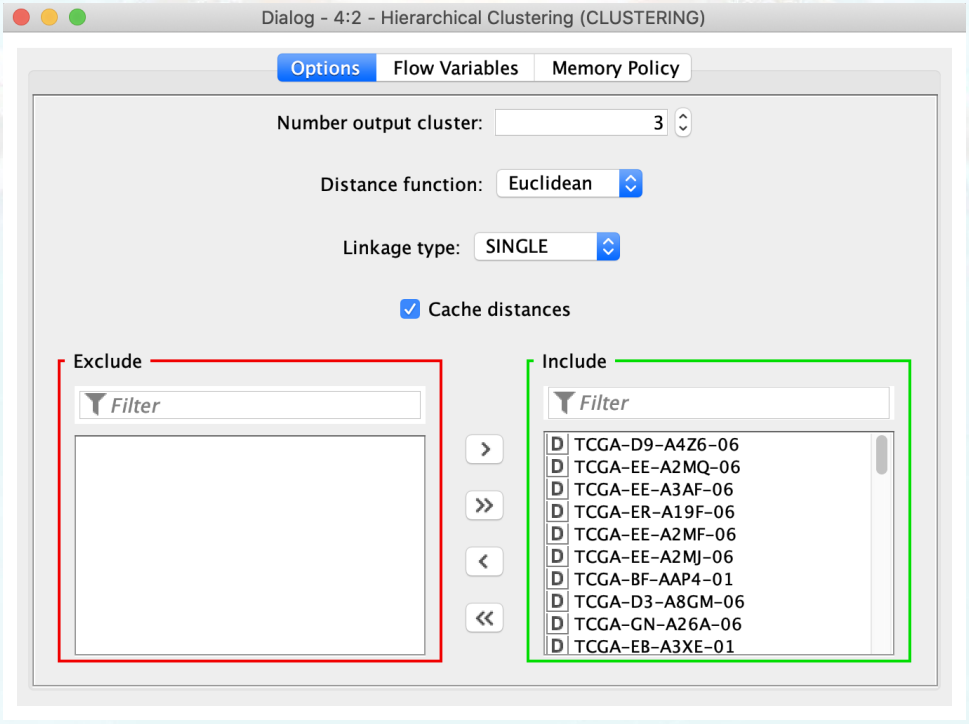


Figura 5: Opciones y parámetros del nodo Hierarchical Clustering

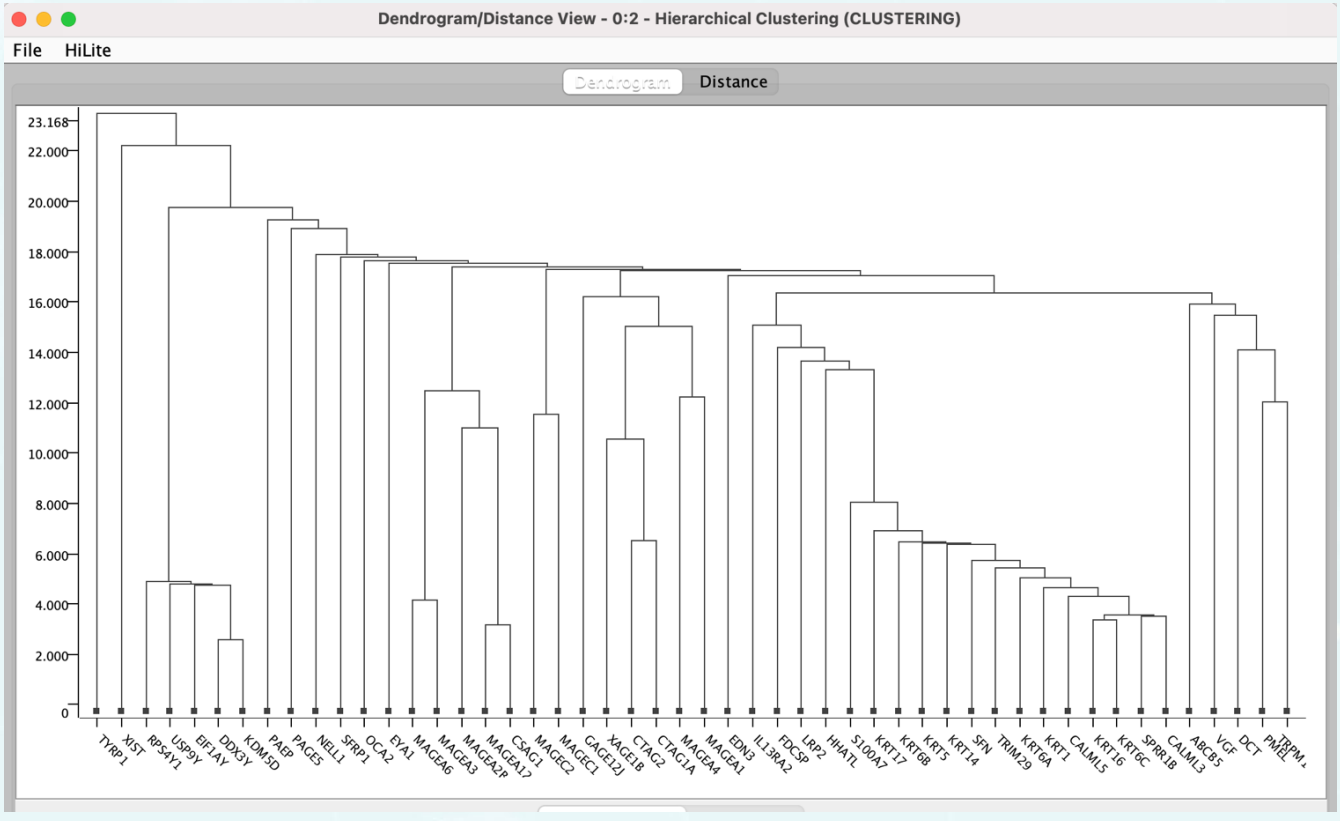


Figura 6: Visualización de resultados a través de un dendograma

Por último, en el flujo de la Figura 1 también se muestra la opción de clustering k-medias. El nodo *k-Means* permite realizar el clustering de k-medias de manera muy sencilla. Se pueden fijar el número de clusters a obtener, iniciación aleatoria, número máximo de iteraciones, etc. La Figura 7 muestra dichas opciones.

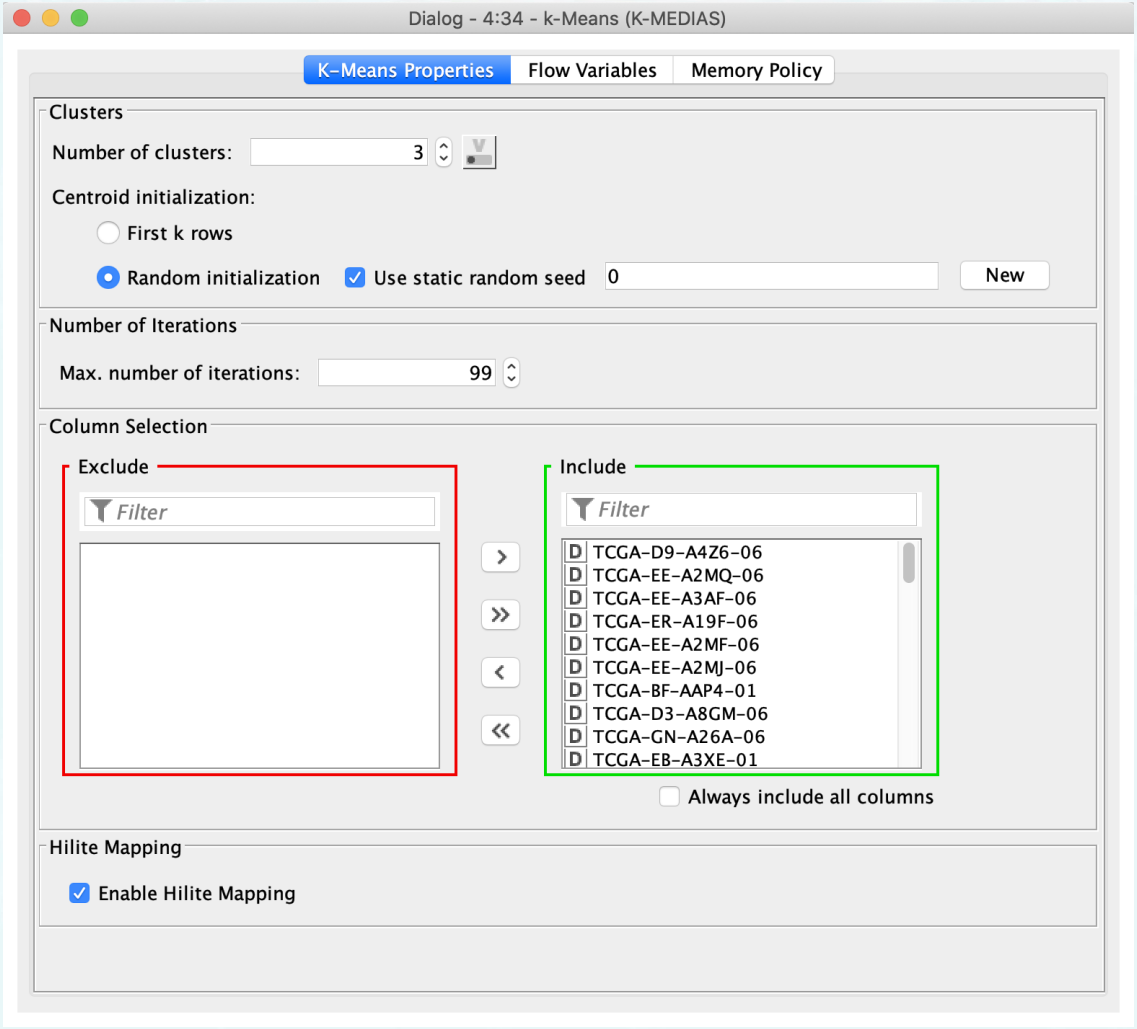


Figura 7: Opciones y parámetros del nodo k-Means

3. REGLAS DE ASOCIACIÓN

En este apartado se va a crear un flujo de datos (Figura 8) para resolver un problema de reglas de asociación donde se ilustrará cómo obtener ítem set frecuentes y varias formas de obtener reglas de asociación. • Para ello, se va a utilizar el mismo conjunto de datos utilizado en el módulo 6 (cápsula 3), que consta de seis variables: MUTATIONSUBTYPES, UV-signature, RNASEQ-CLUSTER_CONSENHIER, MethTypes.201408, MIRCluster, LYMPHOCYTE.SCORE.

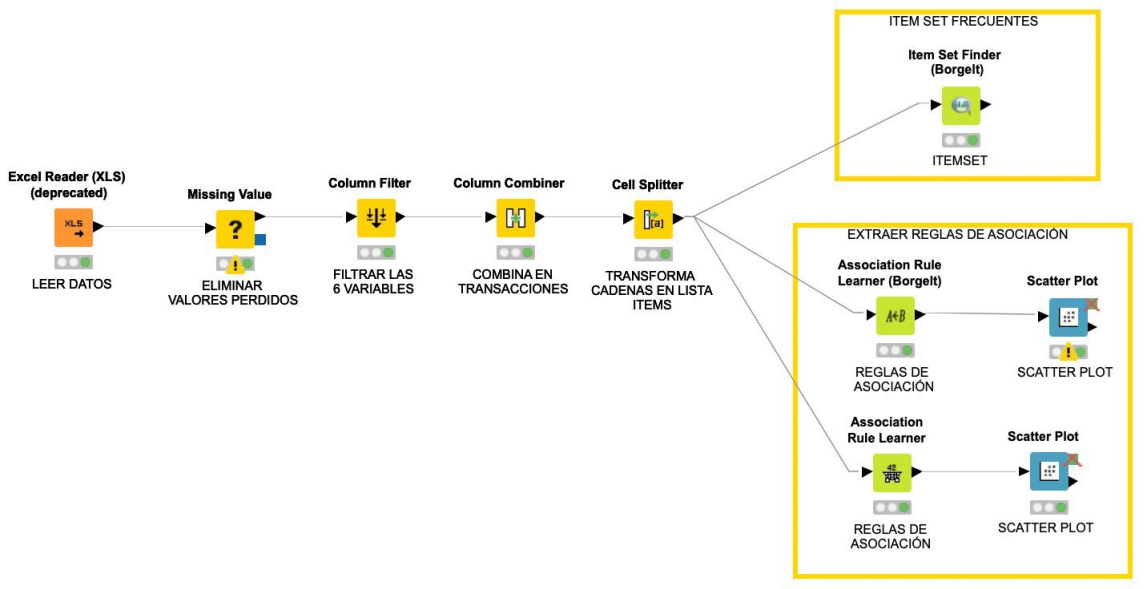


Figura 8: Flujo de datos Reglas de Asociación

En primer lugar, se leen los datos del conjunto de datos con el nodo *Excel Reader (XLS)*. Este fichero contiene todas las variables del conjunto, por lo que es necesario el nodo *Column Filter* para seleccionar las 6 variables. En las opciones de configuración de dicho nodo se incluyen en la parte de la derecha las variables que se desean incluir, en concreto, las 6 variables mencionadas anteriormente (Figura 9).

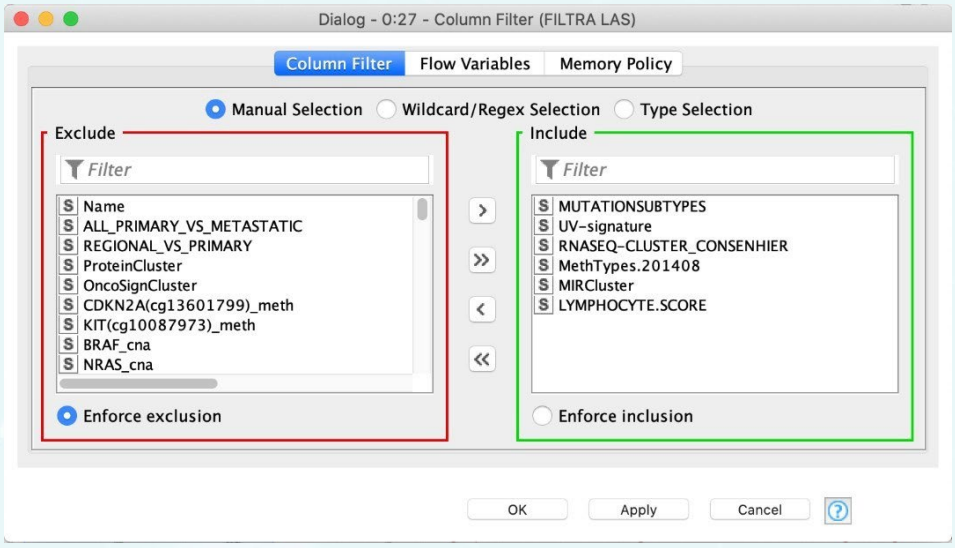
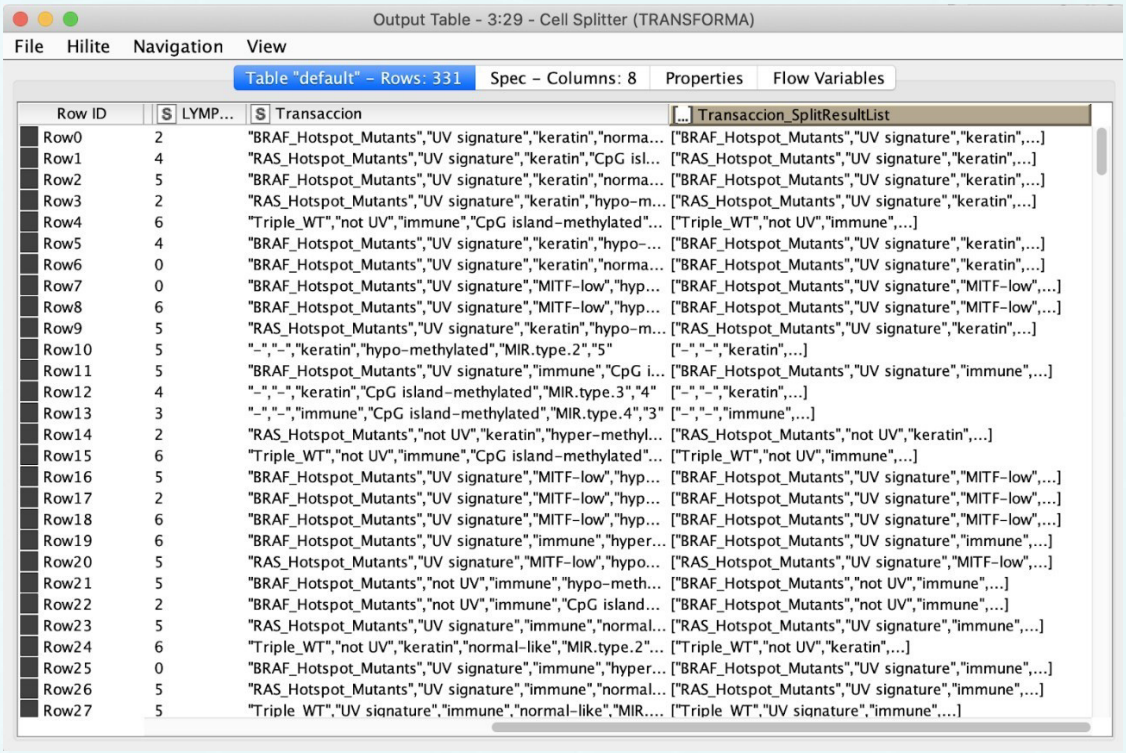


Figura 9: Menú de configuración del nodo Column Filter

Tras estos dos pasos ya tenemos preprocesados los datos con los que realizar las transacciones y calcular las reglas de asociación. Como se explicaba en el módulo 6, el primer paso para poder obtener reglas de asociación es identificar qué define los ítems y las transacciones de los datos. Para ello, se utilizarán los nodos *Column Combiner* y el *Cell Splitte*r que unen los valores que toman cada una de las variables en una única columna y las agrupa en un array, respectivamente. La Figura 10 muestra la salida del nodo *Cell Splitte*r donde se puede observar como la columna nueva que se ha creado llamada Transacción contiene todos los valores para cada una de las variables en un array.



Row ID	LYMP...	Transaccion	Transaccion_SplitResultList
Row0	2	"BRAF_Hotspot_Mutants","UV signature","keratin","norma...	["BRAF_Hotspot_Mutants","UV signature","keratin",...]
Row1	4	"RAS_Hotspot_Mutants","UV signature","keratin","CpG isl...	["RAS_Hotspot_Mutants","UV signature","keratin",...]
Row2	5	"BRAF_Hotspot_Mutants","UV signature","keratin","norma...	["BRAF_Hotspot_Mutants","UV signature","keratin",...]
Row3	2	"RAS_Hotspot_Mutants","UV signature","keratin","hypo-m...	["RAS_Hotspot_Mutants","UV signature","keratin",...]
Row4	6	"Triple_WT","not UV","immune","CpG island-methylated"...	["Triple_WT","not UV","immune",...]
Row5	4	"BRAF_Hotspot_Mutants","UV signature","keratin","hypo-...	["BRAF_Hotspot_Mutants","UV signature","keratin",...]
Row6	0	"BRAF_Hotspot_Mutants","UV signature","keratin","norma...	["BRAF_Hotspot_Mutants","UV signature","keratin",...]
Row7	0	"BRAF_Hotspot_Mutants","UV signature","MITF-low","hyp...	["BRAF_Hotspot_Mutants","UV signature","MITF-low",...]
Row8	6	"BRAF_Hotspot_Mutants","UV signature","MITF-low","hyp...	["BRAF_Hotspot_Mutants","UV signature","MITF-low",...]
Row9	5	"RAS_Hotspot_Mutants","UV signature","keratin","hypo-m...	["RAS_Hotspot_Mutants","UV signature","keratin",...]
Row10	5	"-","-","keratin","hypo-methylated","MIR.type.2","5"	["-","-","keratin",...]
Row11	5	"BRAF_Hotspot_Mutants","UV signature","immune","CpG i...	["BRAF_Hotspot_Mutants","UV signature","immune",...]
Row12	4	"-","-","keratin","CpG island-methylated","MIR.type.3","4"	["-","-","keratin",...]
Row13	3	"-","-","immune","CpG island-methylated","MIR.type.4","3"	["-","-","immune",...]
Row14	2	"RAS_Hotspot_Mutants","not UV","keratin","hyper-methyl...	["RAS_Hotspot_Mutants","not UV","keratin",...]
Row15	6	"Triple_WT","not UV","immune","CpG island-methylated"...	["Triple_WT","not UV","immune",...]
Row16	5	"BRAF_Hotspot_Mutants","UV signature","MITF-low","hyp...	["BRAF_Hotspot_Mutants","UV signature","MITF-low",...]
Row17	2	"BRAF_Hotspot_Mutants","UV signature","MITF-low","hyp...	["BRAF_Hotspot_Mutants","UV signature","MITF-low",...]
Row18	6	"BRAF_Hotspot_Mutants","UV signature","MITF-low","hyp...	["BRAF_Hotspot_Mutants","UV signature","MITF-low",...]
Row19	6	"BRAF_Hotspot_Mutants","UV signature","immune","hyper...	["BRAF_Hotspot_Mutants","UV signature","immune",...]
Row20	5	"RAS_Hotspot_Mutants","UV signature","MITF-low","hypo...	["RAS_Hotspot_Mutants","UV signature","MITF-low",...]
Row21	5	"BRAF_Hotspot_Mutants","not UV","immune","hypo-meth...	["BRAF_Hotspot_Mutants","not UV","immune",...]
Row22	2	"BRAF_Hotspot_Mutants","not UV","immune","CpG island...	["BRAF_Hotspot_Mutants","not UV","immune",...]
Row23	5	"RAS_Hotspot_Mutants","UV signature","immune","normal...	["RAS_Hotspot_Mutants","UV signature","immune",...]
Row24	6	"Triple_WT","not UV","keratin","normal-like","MIR.type.2"...	["Triple_WT","not UV","keratin",...]
Row25	0	"BRAF_Hotspot_Mutants","UV signature","immune","hyper...	["BRAF_Hotspot_Mutants","UV signature","immune",...]
Row26	5	"RAS_Hotspot_Mutants","UV signature","immune","normal...	["RAS_Hotspot_Mutants","UV signature","immune",...]
Row27	5	"Triple_WT","UV signature","immune","normal-like","MIR....	["Triple_WT","UV signature","immune",...]

Figura 10: Salida del nodo Cell Spliter con las transacciones en forma de array

A partir de las transacciones podemos determinar un conjunto de ítem set frecuentes y extraer reglas de asociación. A continuación, ilustramos cómo representar estos dos ejemplos en KNIME:

1) Determinar conjunto de ítem set frecuentes:

Con el nodo nodo *Item Set Finder* se pueden buscar elementos frecuentes en una lista de conjuntos de elementos. Este nodo proporciona diferentes algoritmos para esta tarea, como se puede ver en la Figura 11: A priori, FP-growth, RElim, Sam, JIM, DICE, TANIMOTO. Asimismo, tiene opción de determinar el objetivo: frecuente, cerrado y maximal, así como firmar el tamaño mínimo del ítem set y el soporte.

En este ejemplo, vamos a explorar el algoritmo A priori (similar al detallado en el módulo 6) con un valor de soporte de 0.015. La Figura 12 muestra los conjuntos de "item sets" obtenidos ordenados por porcentaje relativo del soporte.

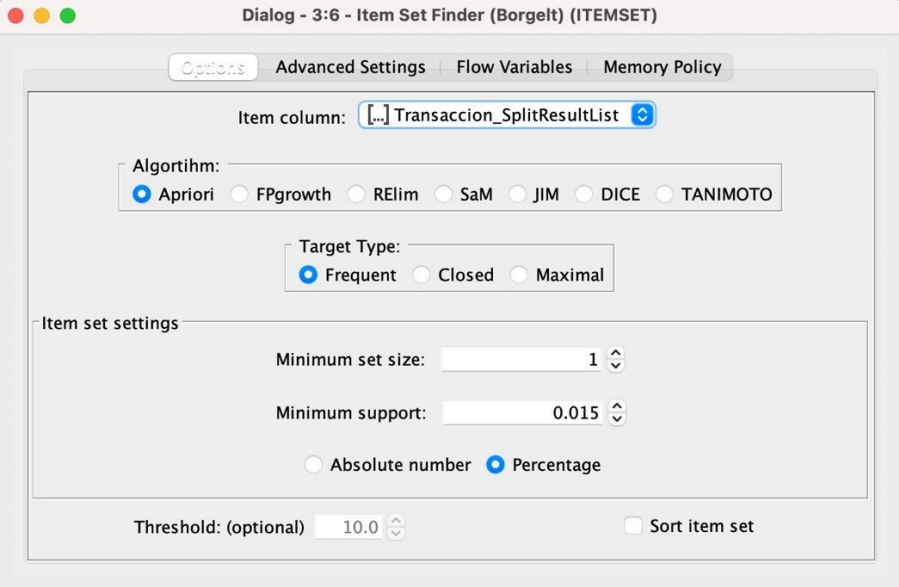
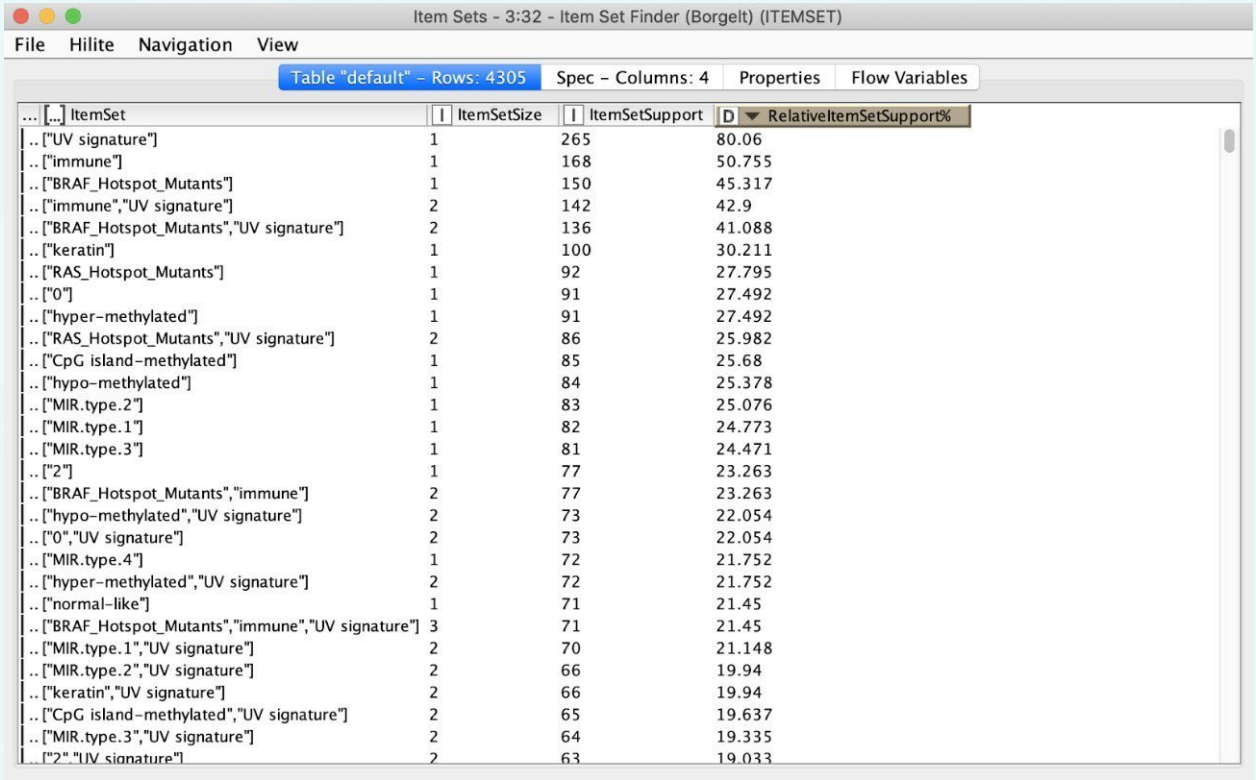


Figura 11: Opciones y parámetros del nodo Item Set Finder

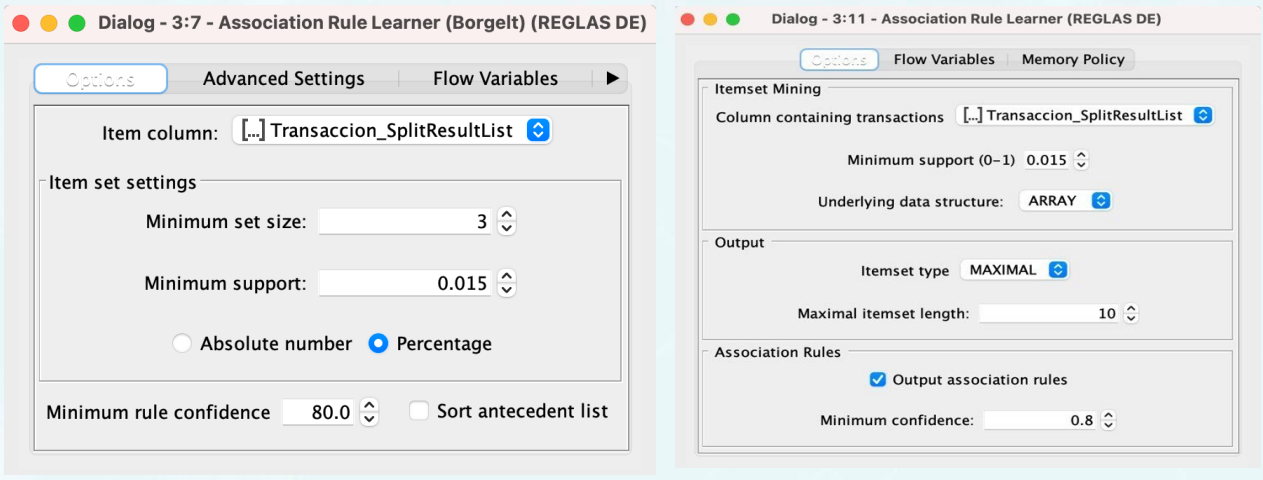


ItemSet	ItemSetSize	ItemSetSupport	RelativeItemSetSupport%
["UV signature"]	1	265	80.06
["immune"]	1	168	50.755
["BRAF_Hotspot_Mutants"]	1	150	45.317
["immune","UV signature"]	2	142	42.9
["BRAF_Hotspot_Mutants","UV signature"]	2	136	41.088
["keratin"]	1	100	30.211
["RAS_Hotspot_Mutants"]	1	92	27.795
["0"]	1	91	27.492
["hyper-methylated"]	1	91	27.492
["RAS_Hotspot_Mutants","UV signature"]	2	86	25.982
["CpG island-methylated"]	1	85	25.68
["hypo-methylated"]	1	84	25.378
["MIR.type.2"]	1	83	25.076
["MIR.type.1"]	1	82	24.773
["MIR.type.3"]	1	81	24.471
["2"]	1	77	23.263
["BRAF_Hotspot_Mutants","immune"]	2	77	23.263
["hypo-methylated","UV signature"]	2	73	22.054
["0","UV signature"]	2	73	22.054
["MIR.type.4"]	1	72	21.752
["hyper-methylated","UV signature"]	2	72	21.752
["normal-like"]	1	71	21.45
["BRAF_Hotspot_Mutants","immune","UV signature"]	3	71	21.45
["MIR.type.1","UV signature"]	2	70	21.148
["MIR.type.2","UV signature"]	2	66	19.94
["keratin","UV signature"]	2	66	19.94
["CpG island-methylated","UV signature"]	2	65	19.637
["MIR.type.3","UV signature"]	2	64	19.335
["2","UV signature"]	2	63	19.033

Figura 12: Conjuntos de ítem sets obtenidos ordenados por porcentaje relativo del soporte

2) Extraer y analizar reglas de asociación:

Con los nodos *Association Rule Learner (Borgelt)* y *Association Rule Learner* se pueden extraer reglas de asociación. En la Figura 13 se pueden ver los distintos parámetros y opciones de configuración que presentan los nodos *Association Rule Learner (Borgelt)* y *Association Rule Learner* entre las que destacar el número mínimo de ítems, el mínimo de soporte y el mínimo de confianza. En este ejemplo, se ha fijado 3 como número mínimo de ítems, un soporte de 0.015 y una confianza del 80%.



a) b)

Figura 13: Opciones y parámetros del nodo **a)** Association Rule Learner (Borgelt) y **b)** Association Rule Learner

La Figura 14 muestra un gráfico de dispersión donde se pueden ver cómo se distribuyen las reglas generadas en función de dos medidas. A modo de ejemplo se ha ilustrado con el porcentaje de confianza y soporte, aunque se podría utilizar de forma interactiva con otras medidas como "lift", etc. seleccionando las medidas que se deseen visualizar a través del menú superior derecho.

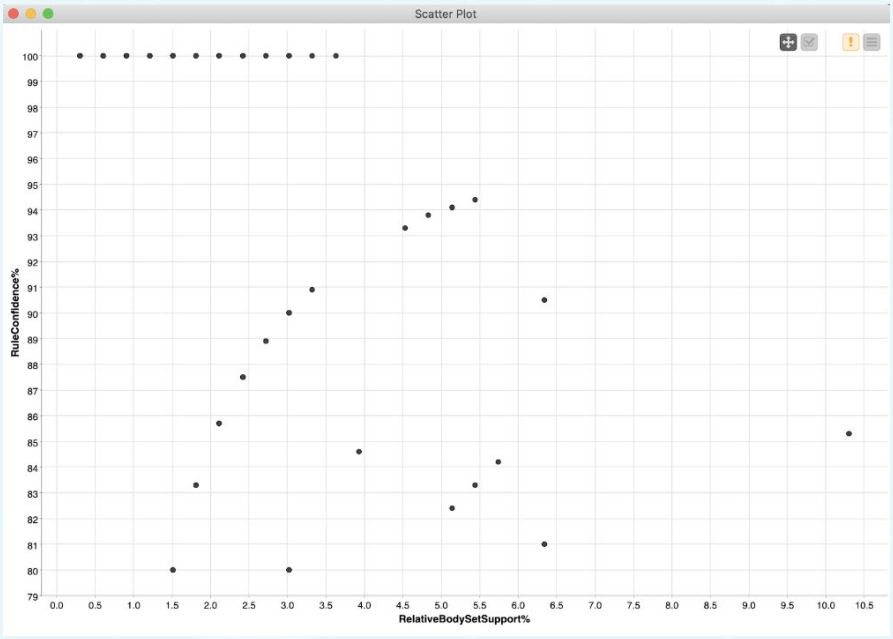


Figura 14: Gráfico de dispersión de las reglas generadas en función del soporte y confianza

4. REFERENCIAS BIBLIOGRÁFICAS

- Bakos, G. (2013). KNIME essentials. Packt Publishing Ltd.
- Blokdyk, G. (2019). KNIME a Complete Guide - 2019 Edition. Emereo Pty Limited, 2019.
- McCormick, K. (2019). Introduction to Machine Learning with KNIME. linkedin.com
- Silipo, R. (2016). Introduction to Data Analytics with KNIME: A Data Science Approach to Analytics. O'Reilly.
- Silipo, R., & Mazanetz, M. P. (2012). The KNIME cookbook. KNIME Press, Zürich, Switzerland.
- Strickland, J. (2016). Data Analytics Using Open-Source Tools. Lulu. com.