Taller #3

Business Intelligence

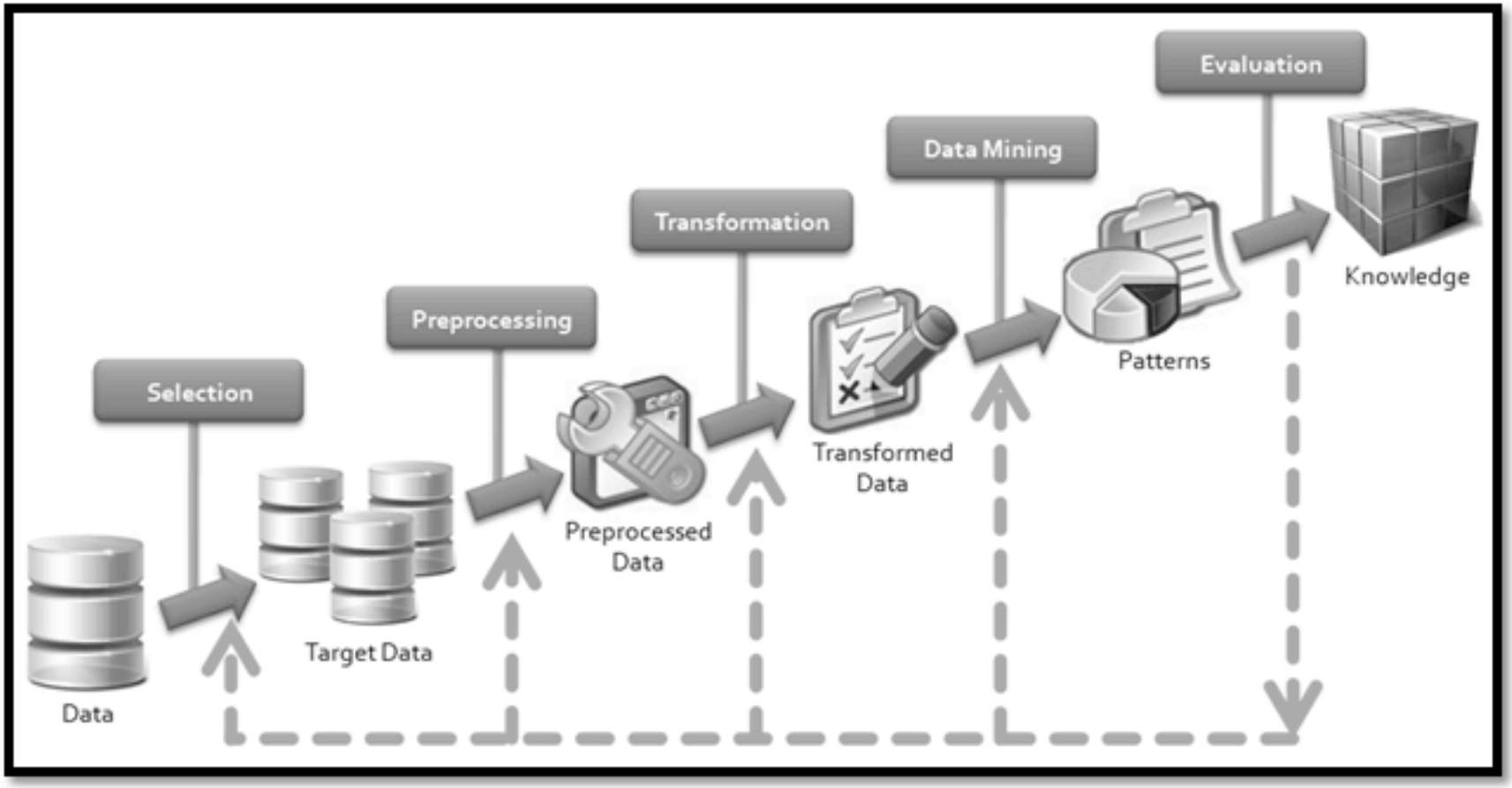
Carlos Reveco
creveco@dcc.uchile.cl

Cinthya Vergara
cvergarasilv@ing.uchile.cl

Agenda

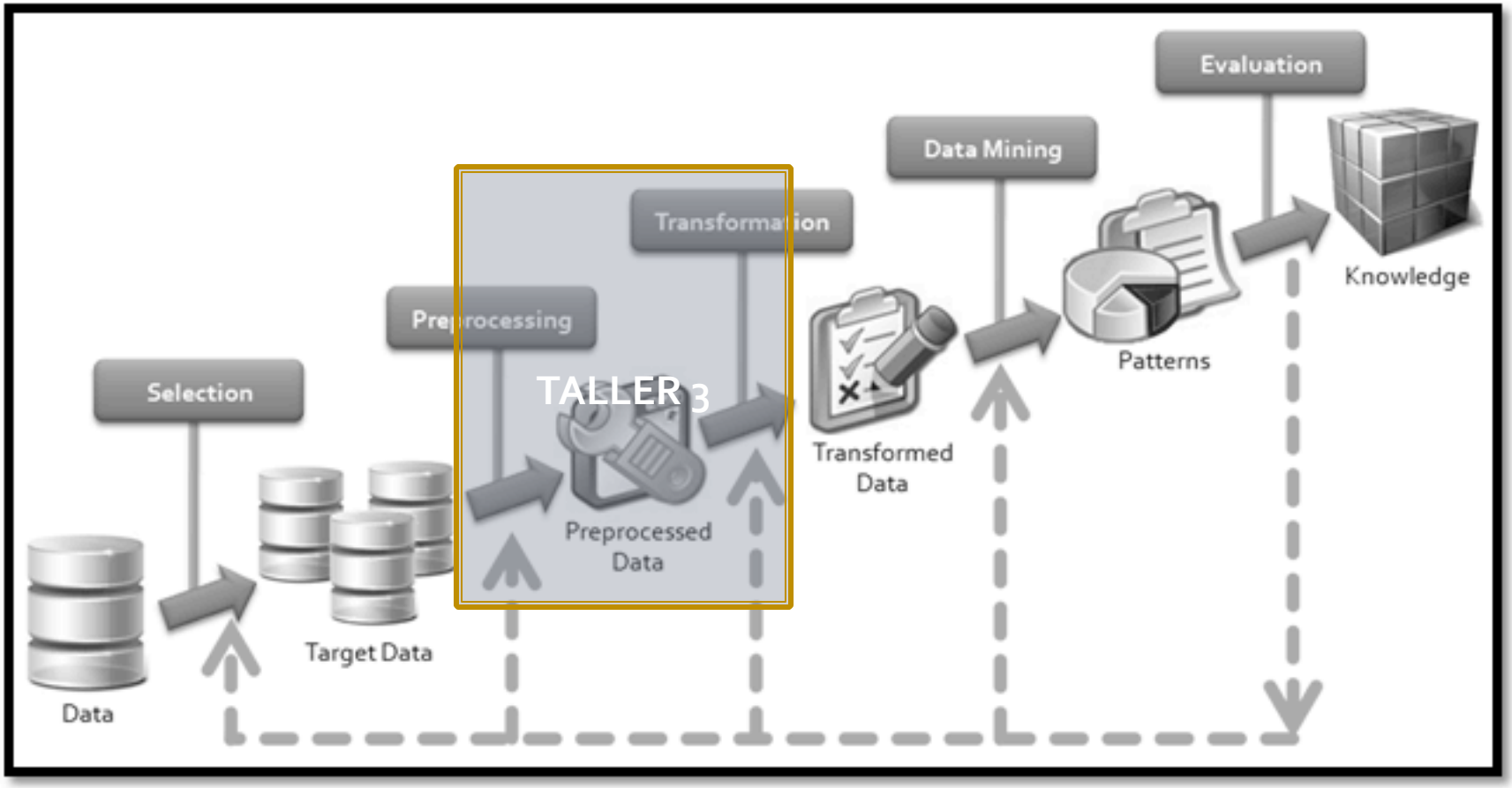
- Taller#3 - Uso de RapidMiner 5.0
  - Limpieza y selección de datos
    - Data Cleansing: Eliminación de valores nulos o valores fuera de rango (outliers)
    - Muestreo
  - Transformación y Selección de atributos
    - Generación de atributos y re-codificación de atributos
    - Discretización de atributos
    - Selección por correlación y Selección por pesos
    - Backward (Forward) Selection

Proceso KDD



Knowledge **D**iscovery in **D**atabases → **KDD**

Proceso KDD



Knowledge **D**iscovery in **D**atabases → **KDD**

Limpieza y selección de datos

Repaso de Conceptos

Limpieza de Datos

- Es necesario generar una base de datos limpia y en la que sea posible aplicar modelos de minería de datos.
- Acciones:
  - Limpieza de datos nulos.
    - La gran mayoría de modelos no son capaces de manejar datos nulos.
    - Excepción importante: Árboles de decisión.
  - Limpieza de variables irrelevantes.
    - Variables concentradas.
  - Eliminación de valores fuera de rango.

Limpieza de Datos
Imputación de datos nulos

- Tipos de Datos perdidos (Taxonomía Clásica) [Little and Rubin, 1987]:
  - **Missing Completely at Random** (MCAR):
    - Los valores perdidos no se relacionan con las variables en la base de datos
  - **Missing at Random** (MAR):
    - Los valores perdidos se relacionan con los valores de las otras variables dentro de la base de datos.
  - **Not Missing at Random or Nonignorable** (NMAR):
    - Los valores perdidos dependen del valor de la variable.

Limpieza de Datos
Imputación de datos nulos

- MCAR:
  - Si no existe conocimiento de por qué se generaron estos casos, no se pueden generar imputaciones correctas.
    - En general, se pueden eliminar estos casos porque no suele ser muchos.
    - Mejor idea: Si son pocos casos nulos en la variable, mantenerlo y llegar a la fase de selección de atributos.
    - Si el atributo sirve, intentar rellenar casos nulos, si no, simplemente eliminarlo.
  - En general, es necesario utilizar el criterio.

Limpieza de Datos
Imputación de datos nulos

- Si una variable posee muchos valores concentrados en un solo valor (e.g. cero), es recomendable eliminarla.
- Variable sin varianza, no posee información relevante, por lo que puede ser seguramente descartada.
- Criterios:
  - Nominales: Si se concentra en más de un cierto porcentaje (95% de repetidos).
  - Numéricas: Si desviación estándar es menor a un cierto valor

Limpieza de Datos
Imputación de datos nulos

- Si existe conocimiento, es posible generar "modelos" para rellenar estos datos.
  - MAR: Es posible utilizar árboles de decisión, regresiones u otro modelo para poder generar los valores faltantes.
  - NMAR: Aquí los datos nulos son un valor en sí mismos, por lo que es buena idea intentar rescatarlo.
    - Raramente se tiene este tipo de caso.

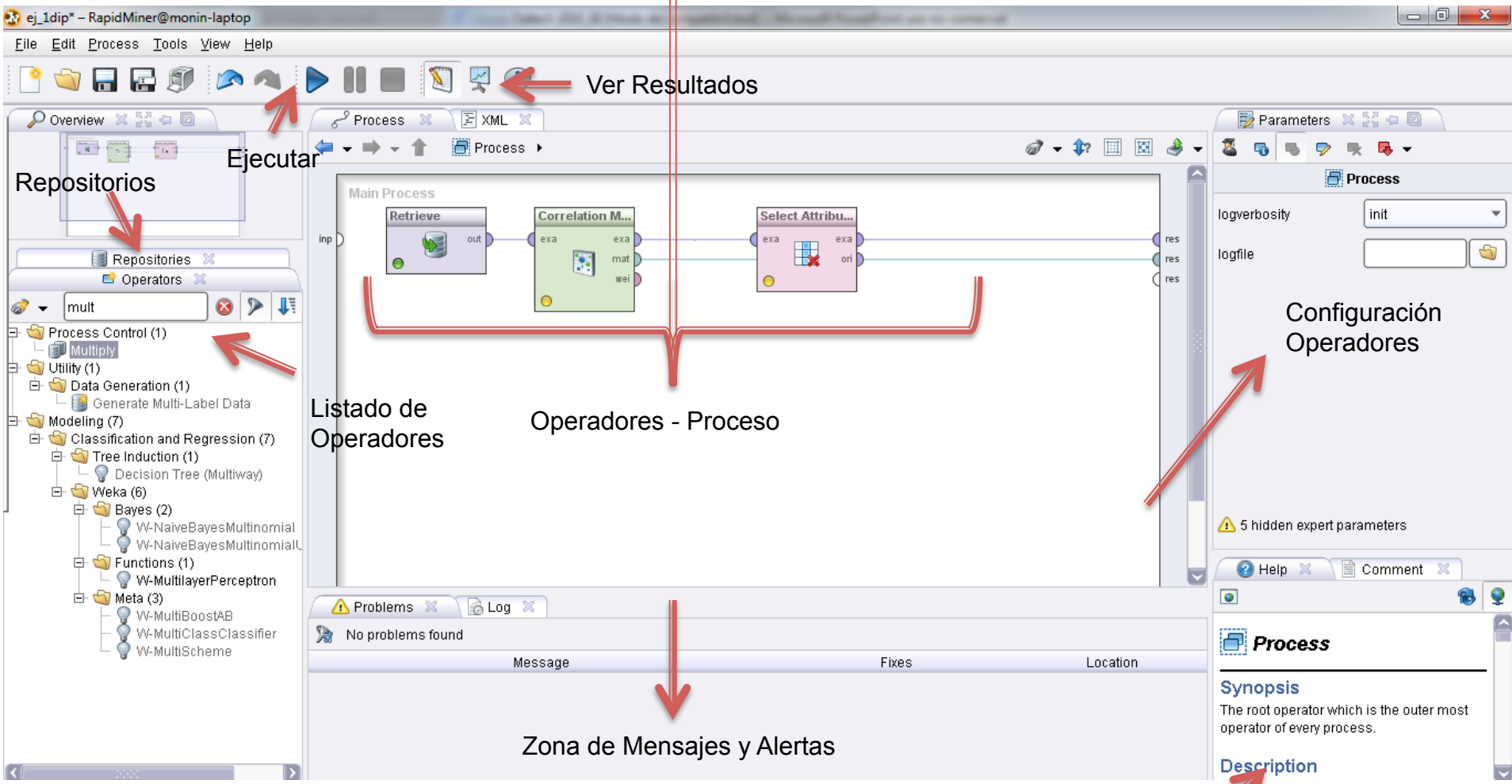
Base de datos Bankloan
Descripción

- **Customer**: ID (e.g. RUT) del cliente. (Etiqueta del caso).
- **Age**: Edad del cliente en años.
- **Education**: Nivel de educación del cliente.
  - 1: Básica completa. 2: Media completa. 3: Superior incompleta. 4: Superior completa. 5: Con posgrado.
- **Employ**: Antigüedad laboral.
- **Address**: Antigüedad en la vivienda actual.
- **Income**: Ingreso en MU$.
- **DebtInc**: Ratio Deuda/Ingreso.
- **CredDebt**: Deuda en tarjetas de crédito.
- **OthDebt**: Otra deuda en miles.
- **Default**: 1 Si cliente dejó de pagar el crédito, o si no (clase del objeto, variable objetivo).

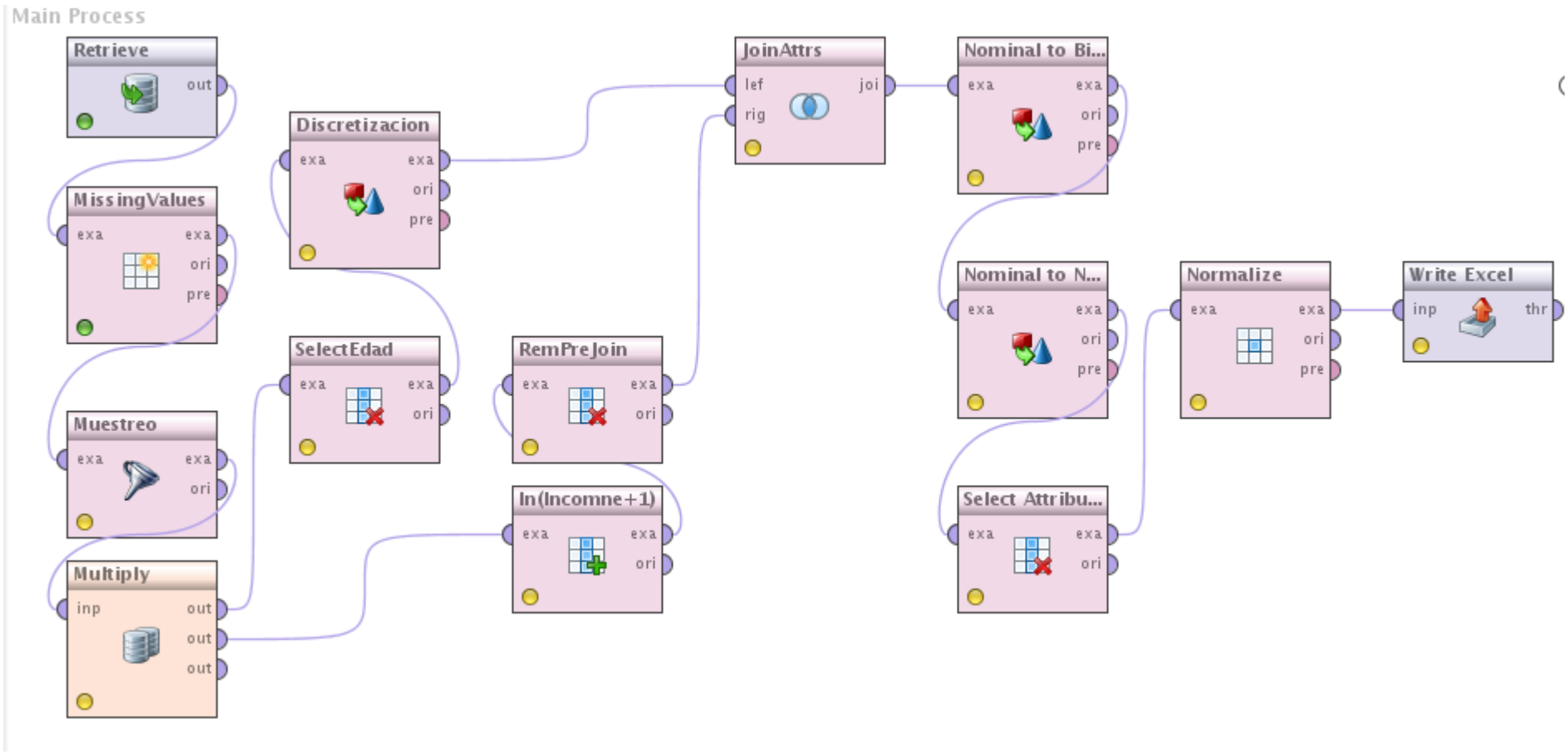
Taller #3

Uso de Herramienta de Minería de Datos y Ejercicio Práctico

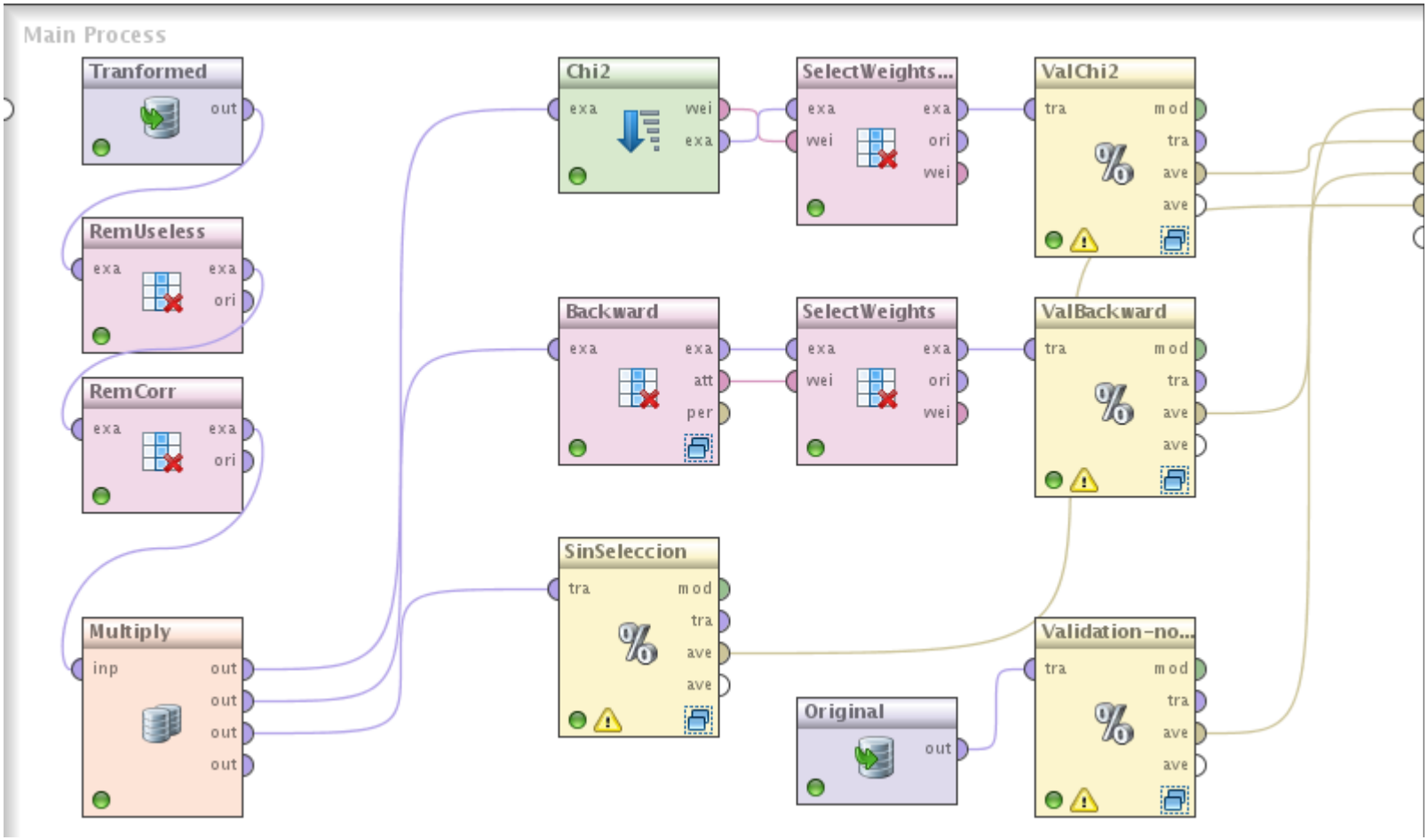
Herramienta RapidMiner



Objetivo 1:
Proceso Transformación

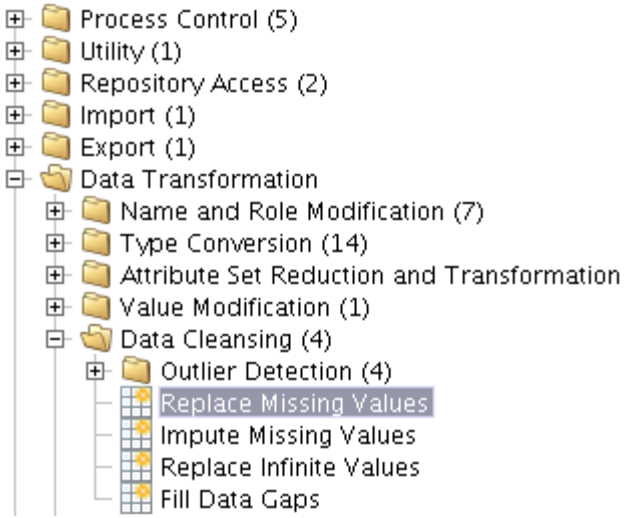


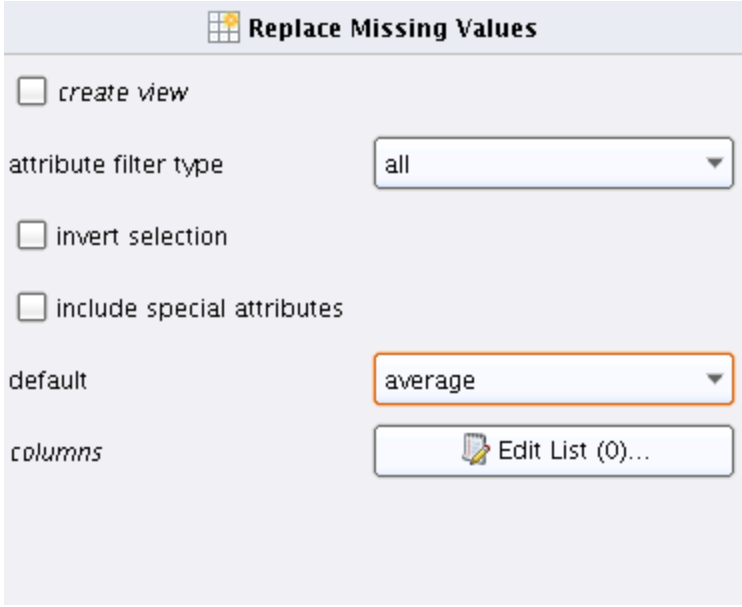
Objetivo 2:
Proceso Selección



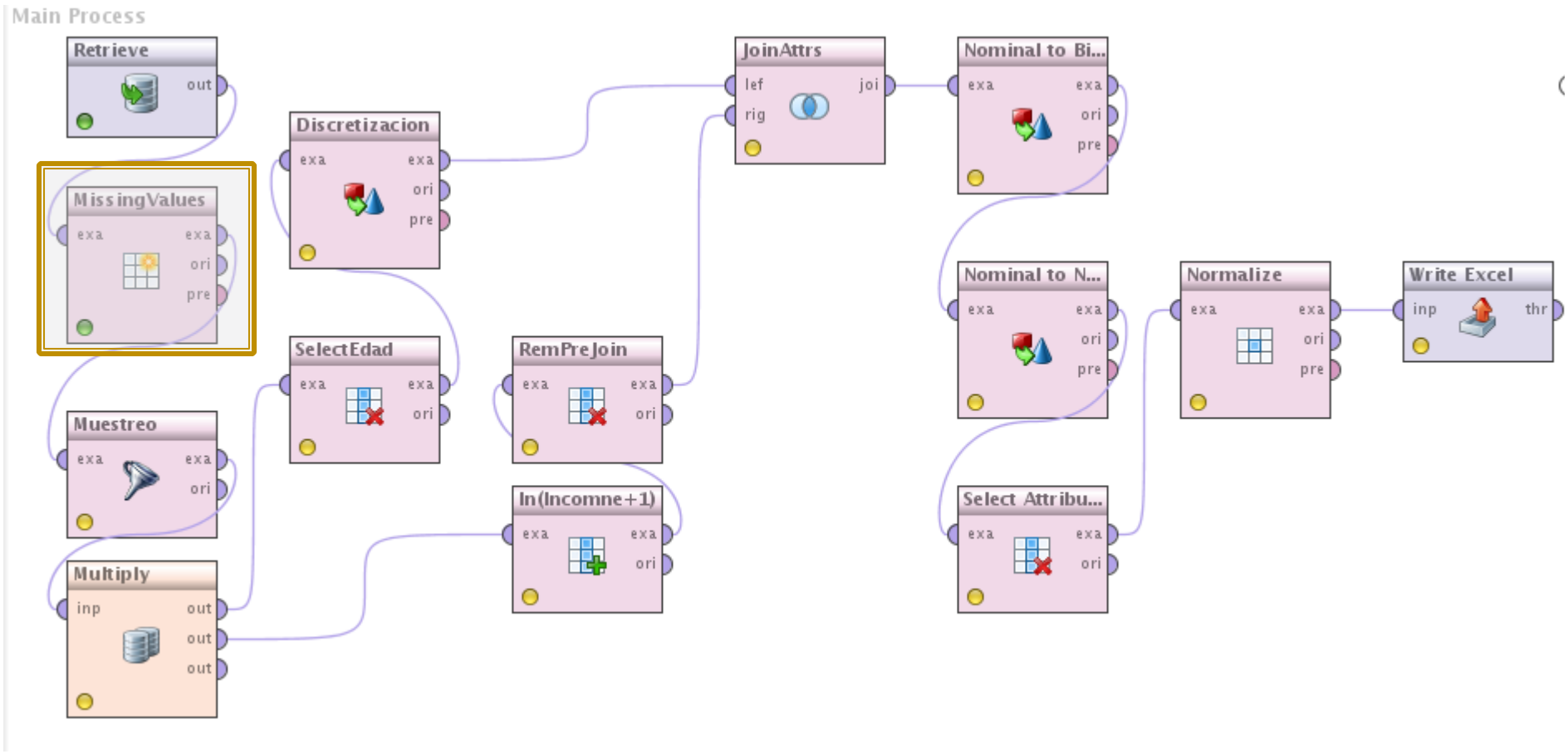
Limpieza de Datos
RapidMiner 5.0

- Existe un operador simple para reemplazar los valores faltantes.



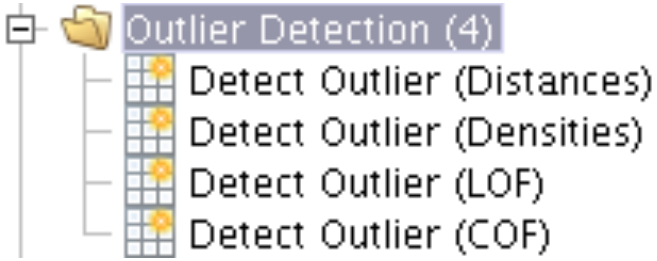


Proceso Transformación



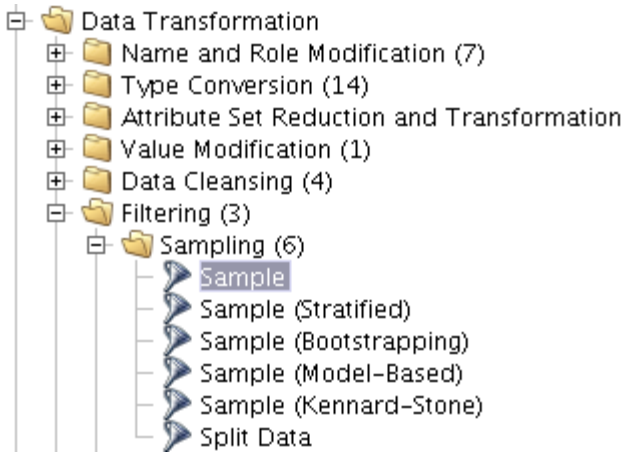
Valores Fuera de rango
RapidMiner 5.0

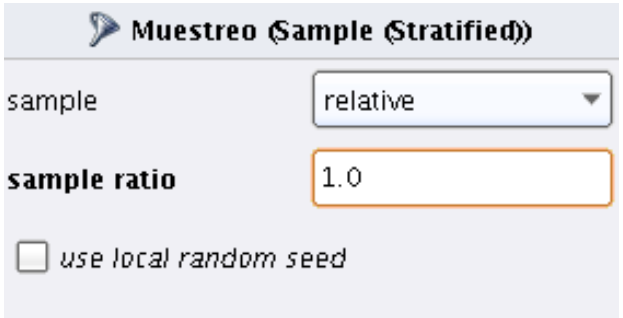
- Existen 4 métodos implementados, basados en publicaciones recientes.
  - Altamente costosos (computacional)
  - Difíciles de usar (usarlos con cuidado)



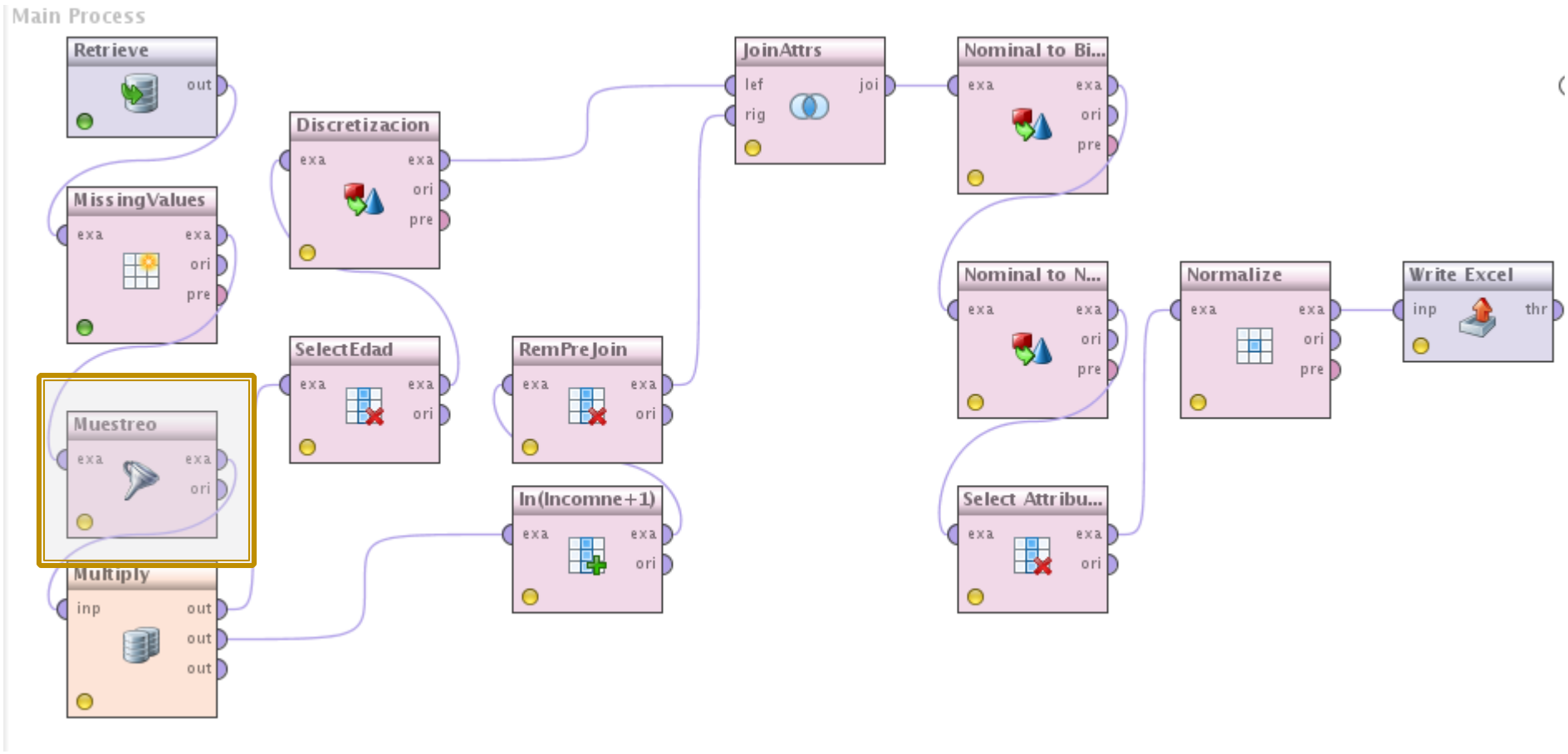
Muestreo
RapidMiner 5.0

- Existen varios métodos de muestreo que permiten seleccionar particiones de los datos.
  - Recomendable cuando el volumen de datos es difícil de manejar.





Proceso Transformación

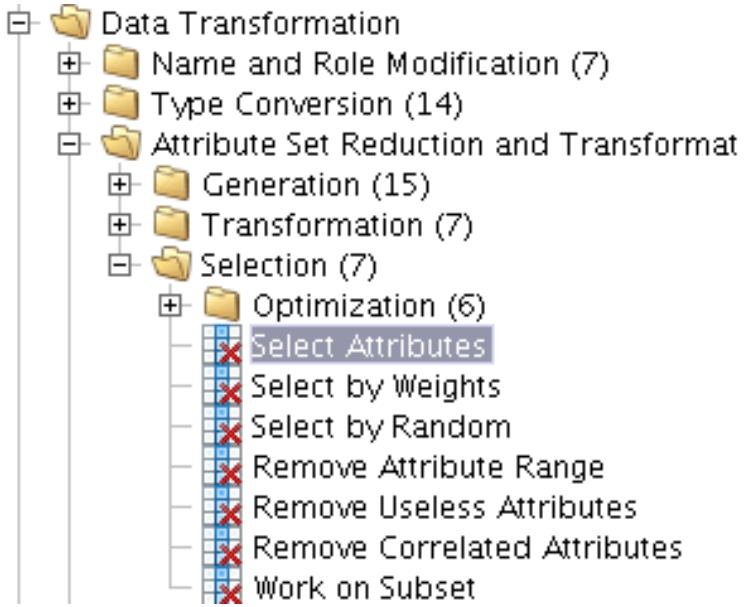


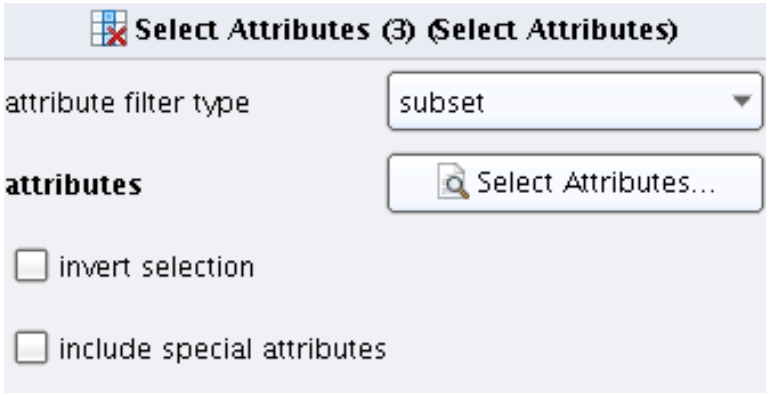
Transformación de Atributos

RapidMiner v5.0
Seleccionar un atributo en particular [1]

- Si deseamos seleccionar un atributo (o un conjunto de atributos) en particular para procesamiento específico, es posible hacerlo.
  - No olvidar el atributo ID (o incluso el label) para hacer un procesamiento y una posterior re-incluir el atributo en la base de datos.
  - El operador para re-incluir los atributos procesados de manera independiente es el "join"

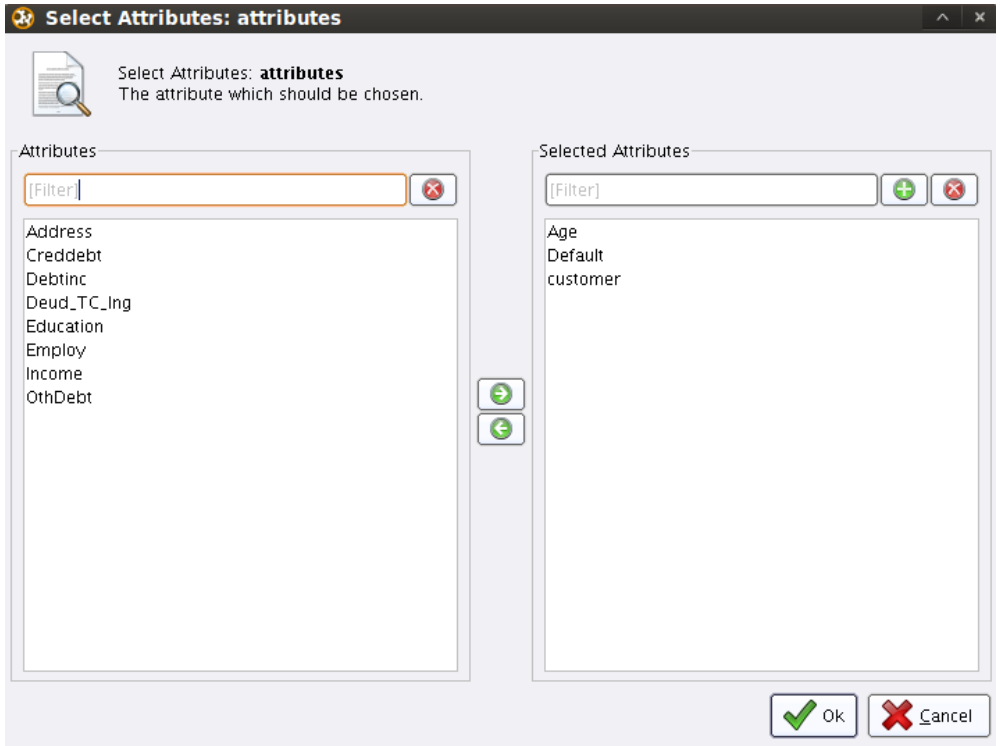
RapidMiner v5.0
Seleccionar un atributo en particular [2]



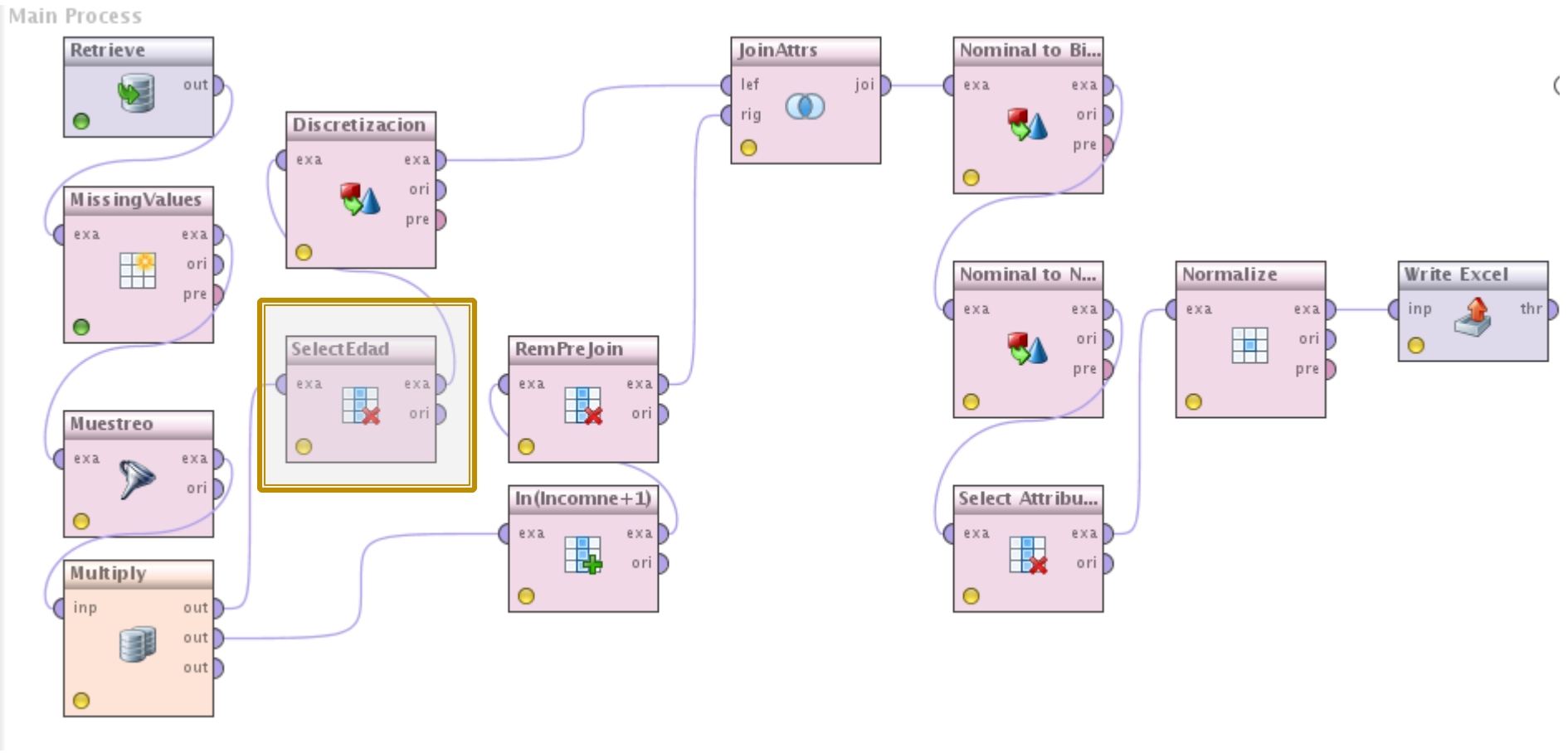


RapidMiner v5.0
Seleccionar un atributo en particular [3]

- Si queremos procesar la Edad de manera independiente:

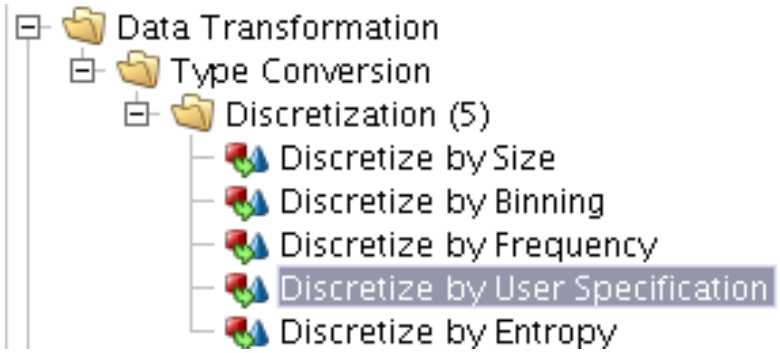


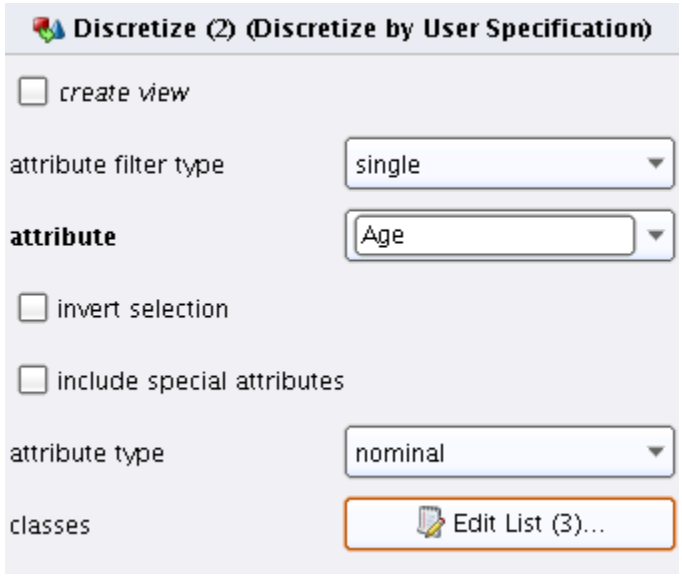
Proceso Transformación



RapidMiner v5.0
Discretización de atributos [1]

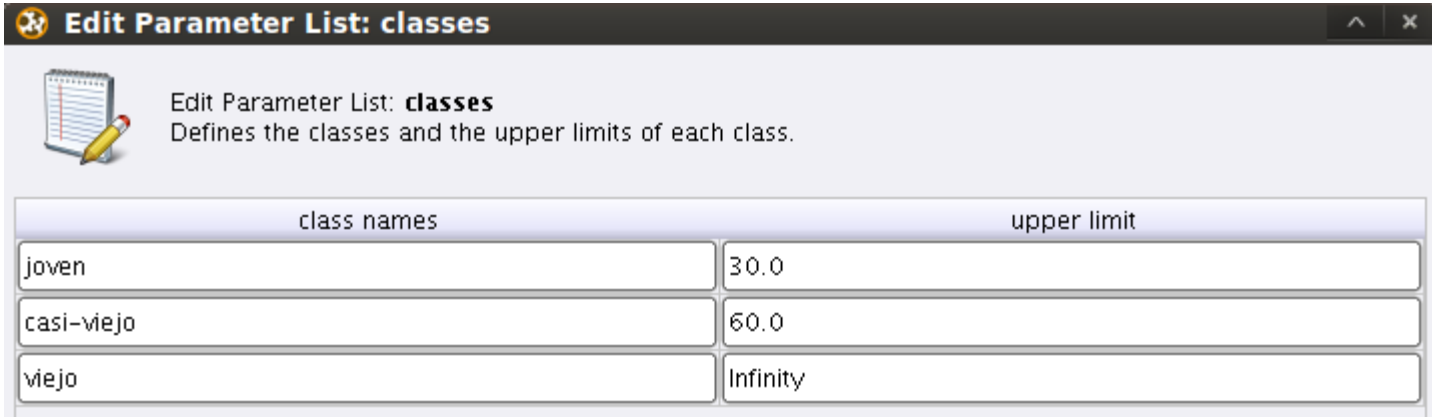
- En ciertos casos, es necesario aplicar discretización sobre atributos numéricos (e.g. Edad)



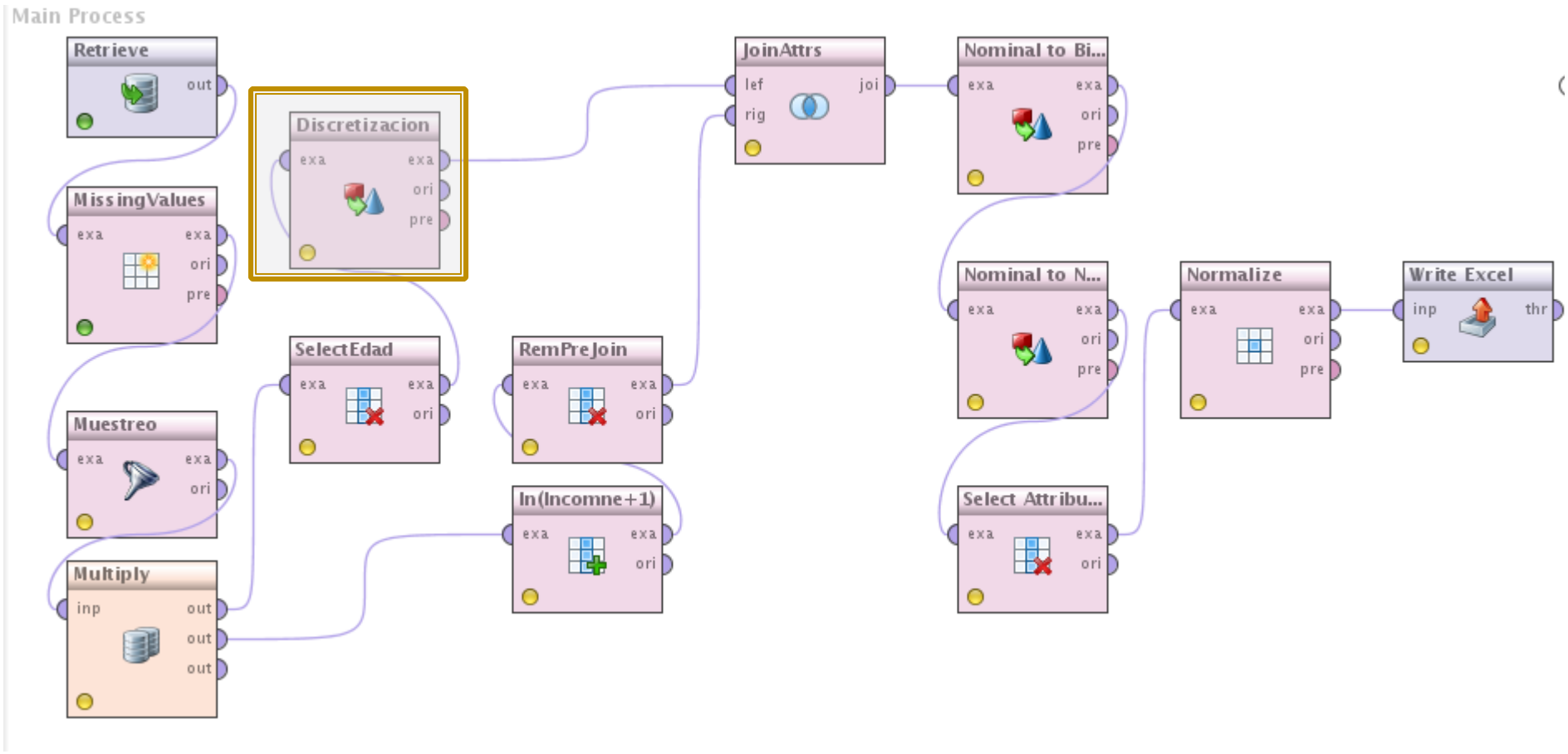


RapidMiner v5.0
Discretización de atributos [2]

- Separemos los rangos de valores en jóvenes, viejos y muy viejos.



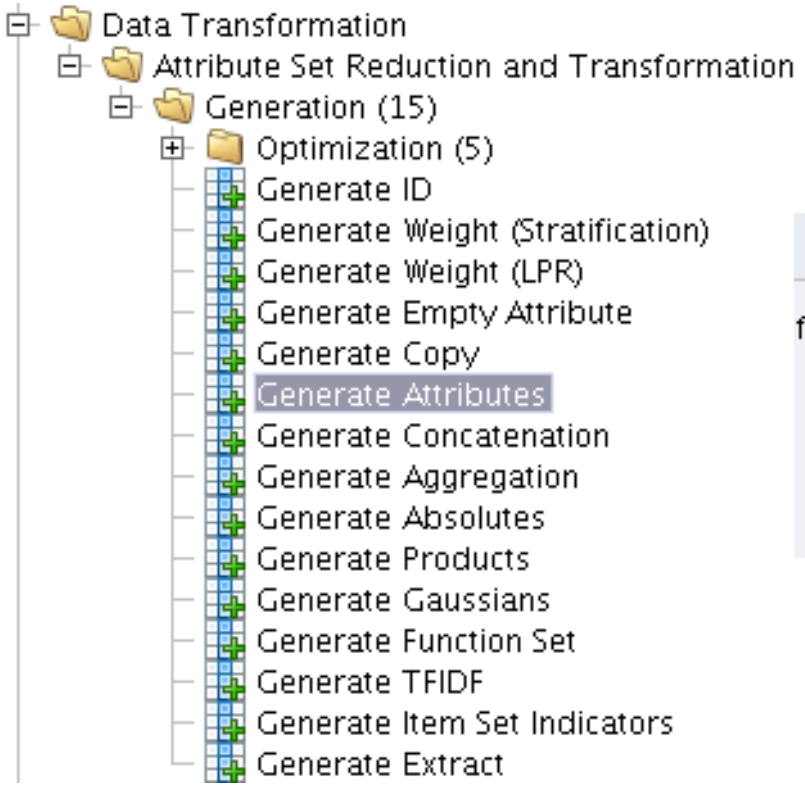
Proceso Transformación

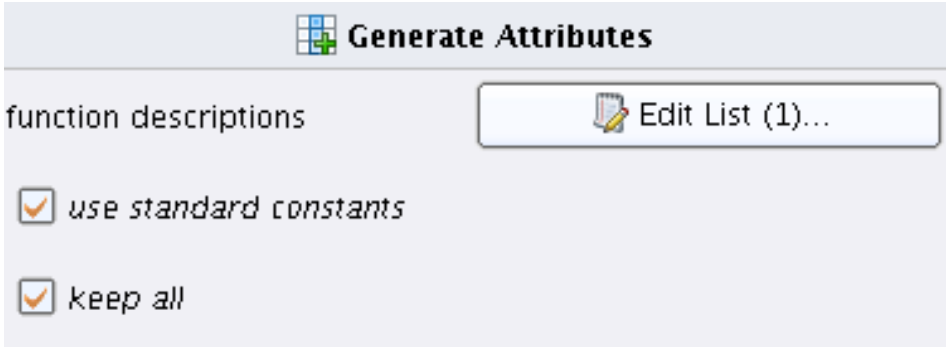


RapidMiner v5.0
Generación de atributos [1]

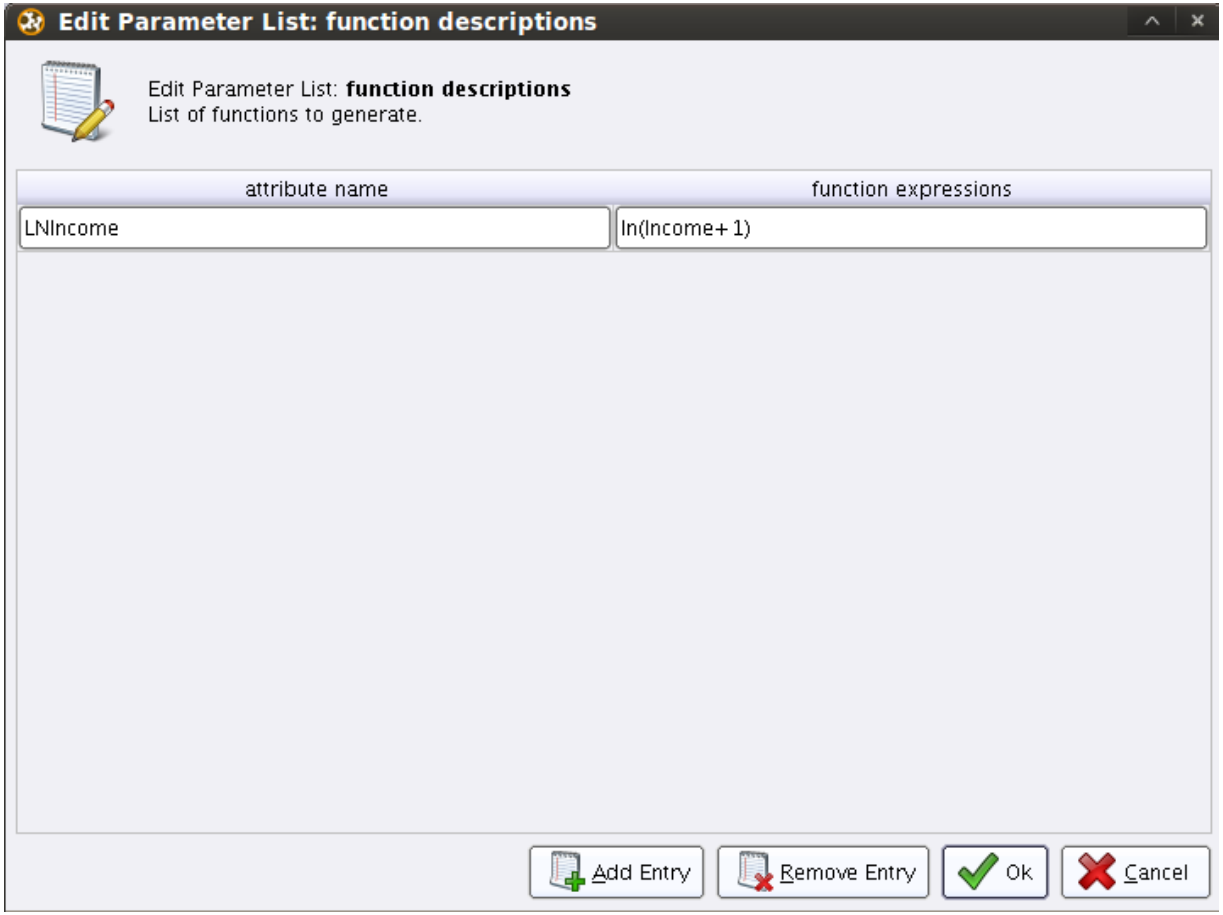
- Es posible generar nuevos atributos a partir de los anteriores.
- Un ejemplo, es el caso de aquellas variables con una distribución que se puede suavizar con transformaciones.
- Ejemplo: La variable "Ingreso" está muy centrada en valores bajos, con un decrecimiento exponencial.
- Idea: Transformarla en ln(Ingreso+1)

RapidMiner v5.0
Generación de atributos [2]

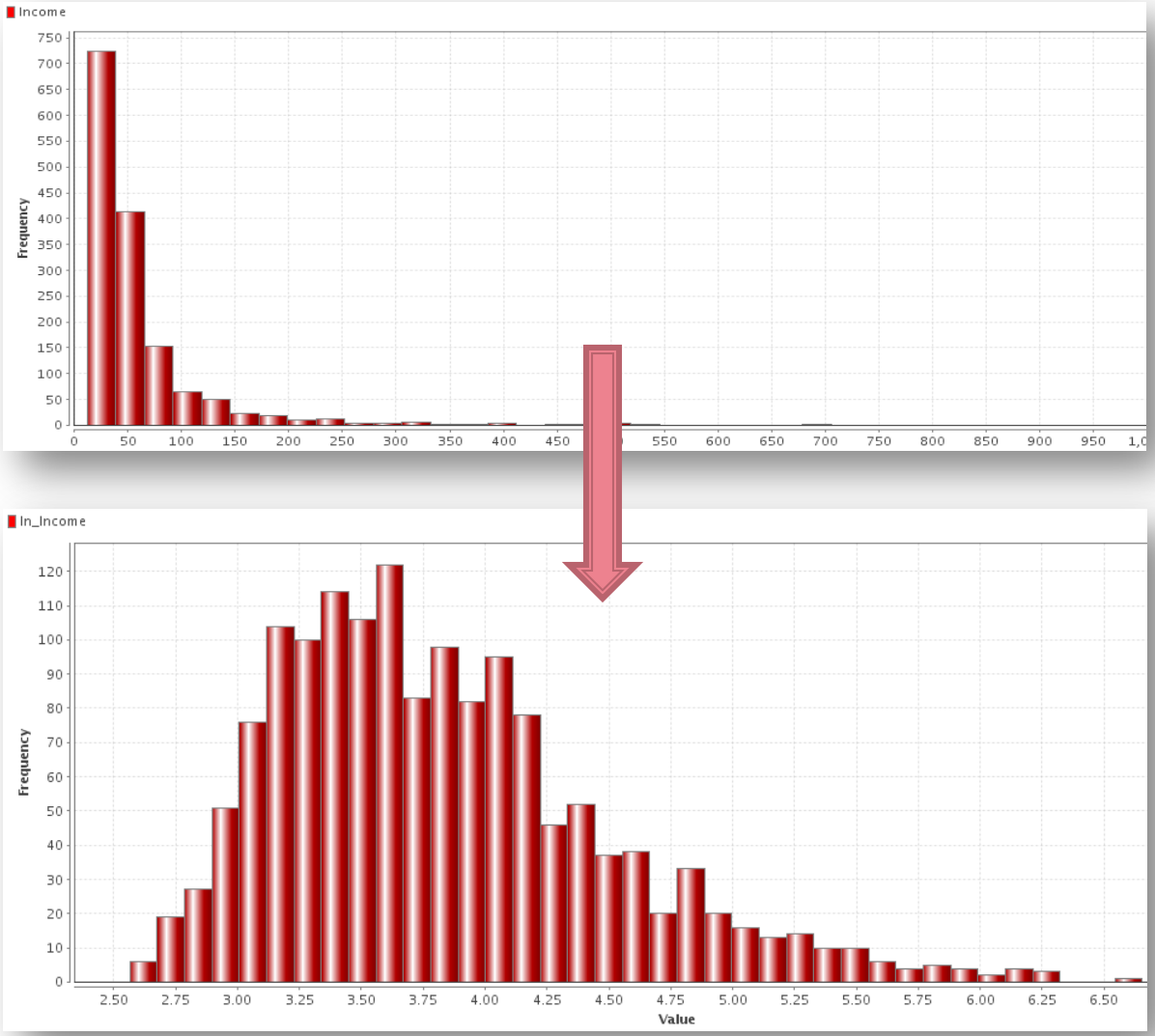




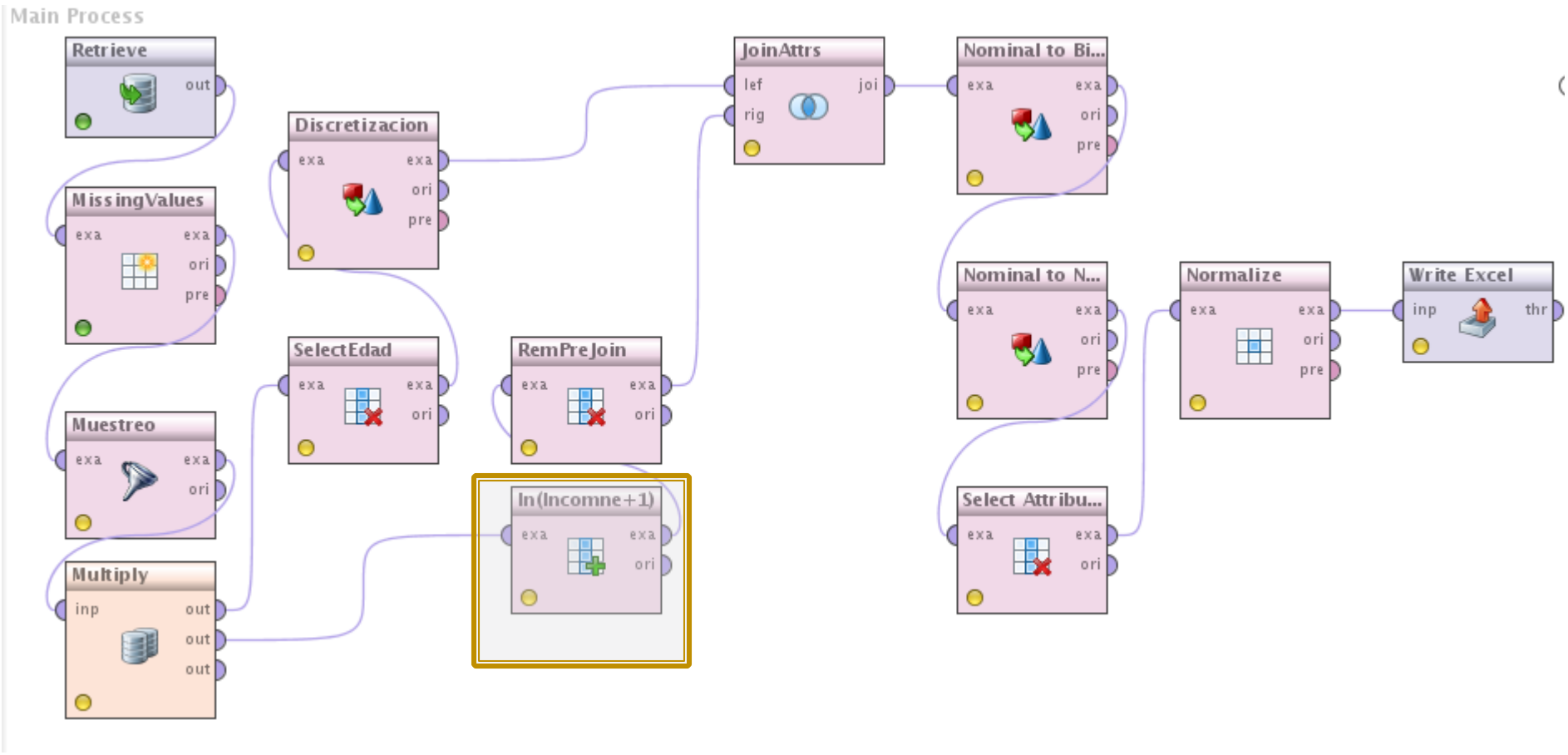
RapidMiner v5.0
Generación de atributos [3]



RapidMiner v5.0
Generación de atributos [4]



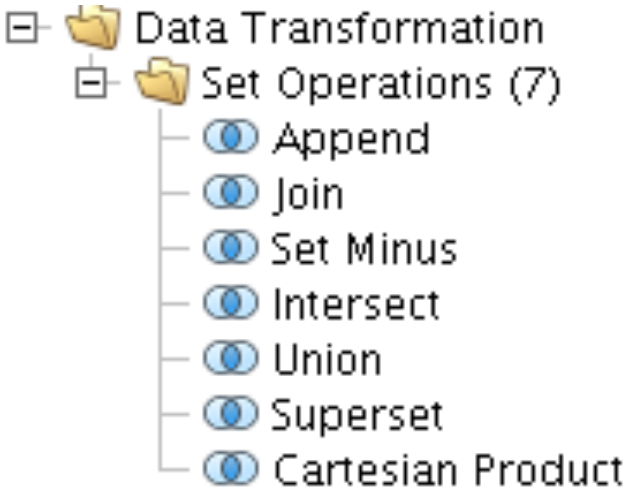
Proceso Transformación



RapidMiner v5.0
Unión de Conjuntos [1]

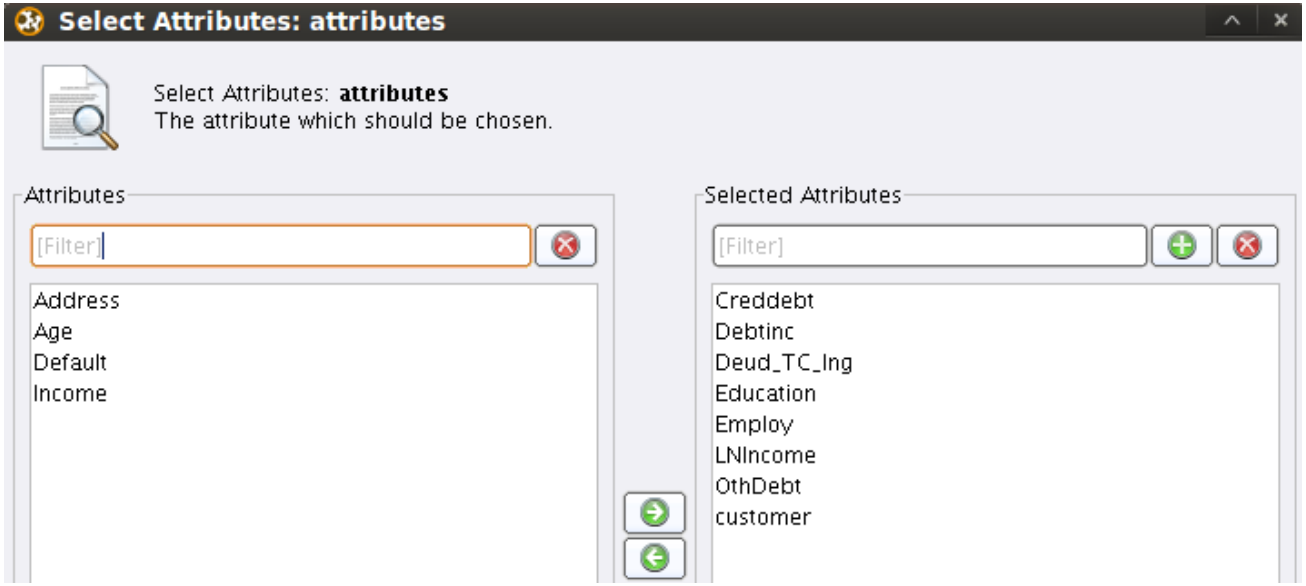
- Es posible utilizar la teoría de conjuntos para unir distintos atributos relacionados a la misma base de datos.
  - Un ejemplo de esto: "Join"
  - Existen distintos sabores: Inner y Outer (full, left, right)
  - Para un Inner Join, es relevante tener un ID claramente definido para ambos conjuntos, que además hagan referencia a los mismo objetos.

RapidMiner v5.0
Unión de Conjuntos [2]

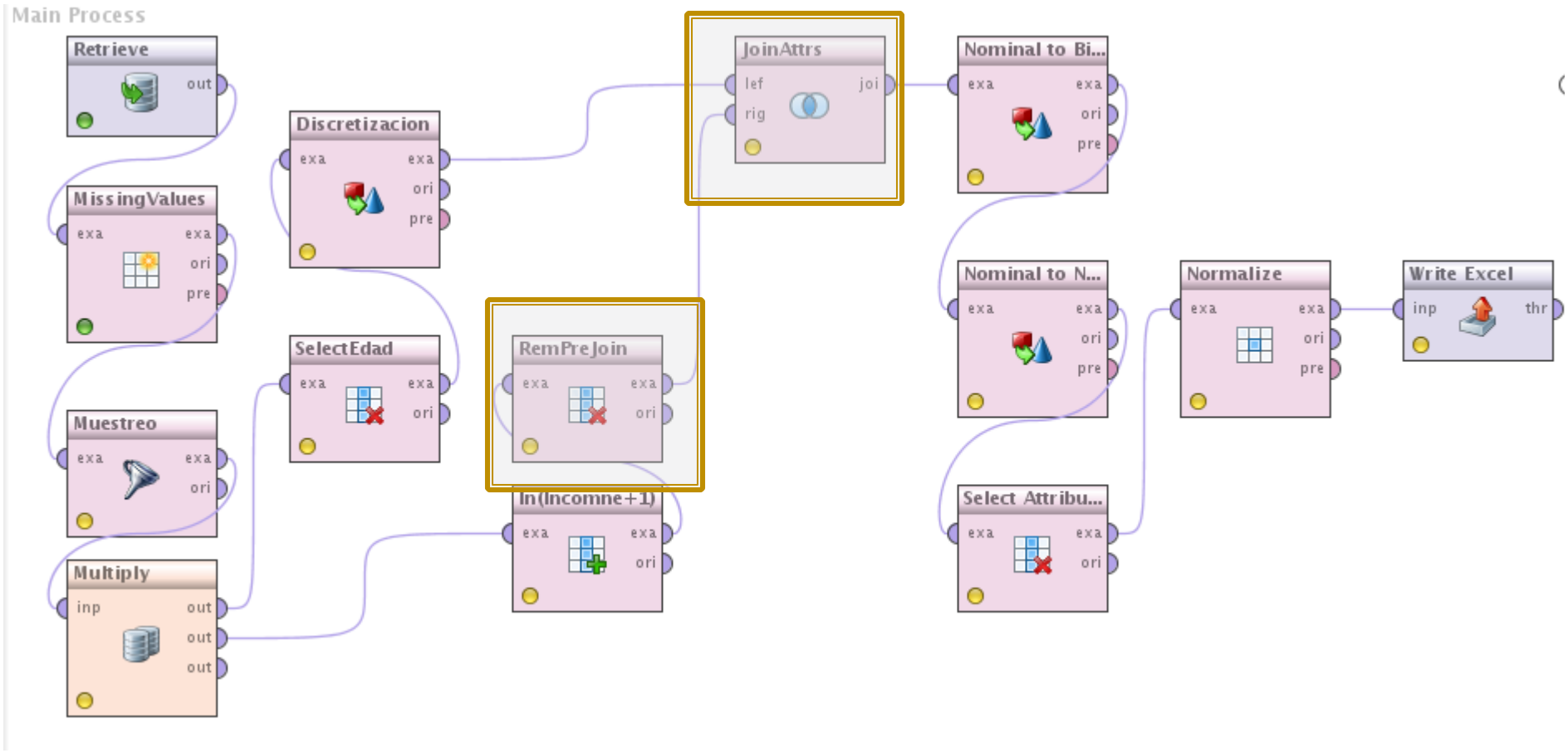


RapidMiner v5.0
Unión de Conjuntos [3]

- Eliminamos algunos atributos de uno de los conjuntos, para que no tengamos problemas en la unión de ambos conjuntos.



Proceso Transformación



RapidMiner v5.0
Transformación de datos categóricos

- En general, los datos categóricos no pueden ser utilizados para los modelos
- Ejemplo:
  - Ciudad, profesión, etapa de crecimiento, etc.
- Existen 2 estrategias para su transformación.
  - Si no hay un orden implícito (e.g. ciudad) es necesaria la creación de variables "dummies".
  - Si hay un orden implícito (etapa de crecimiento), se puede transformar en una escala de valores.

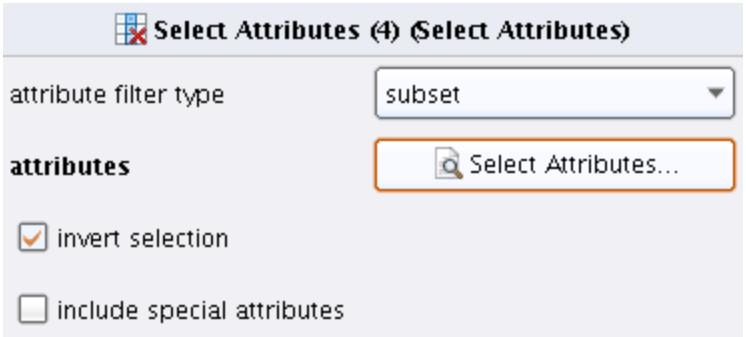
RapidMiner v5.0
Transformación de datos categóricos

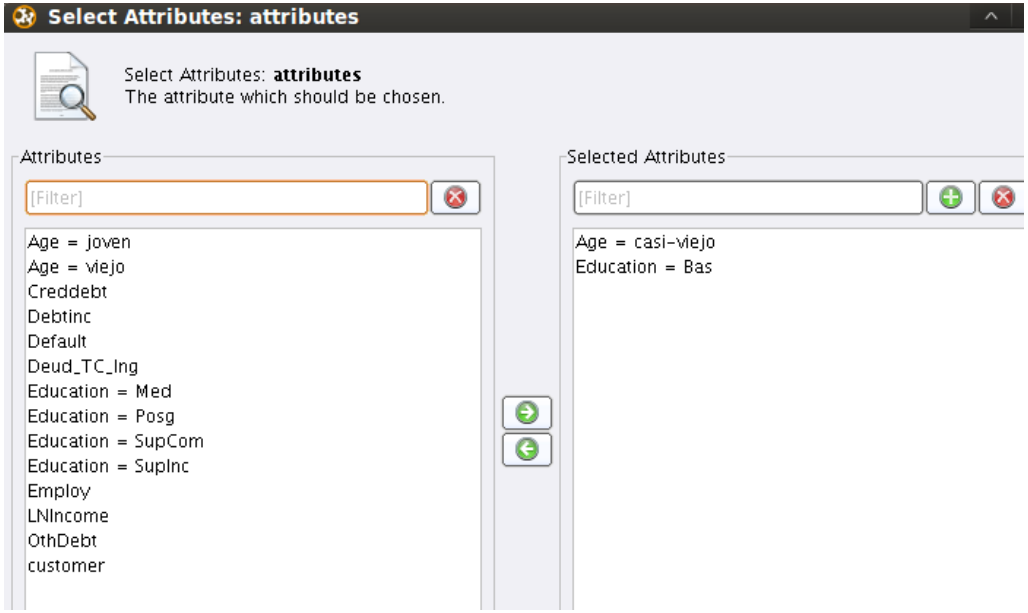
- Por inspección en la base de datos tenemos la variable categórica "Education" con 5 valores distintos.
- Es necesario transformarla en 5 nuevas variables, pero posteriormente debemos eliminar una dado que tendríamos información redundante (la 5ta categoría se puede formar con las 4 anteriores)

RapidMiner v5.0
Transformación de datos categóricos

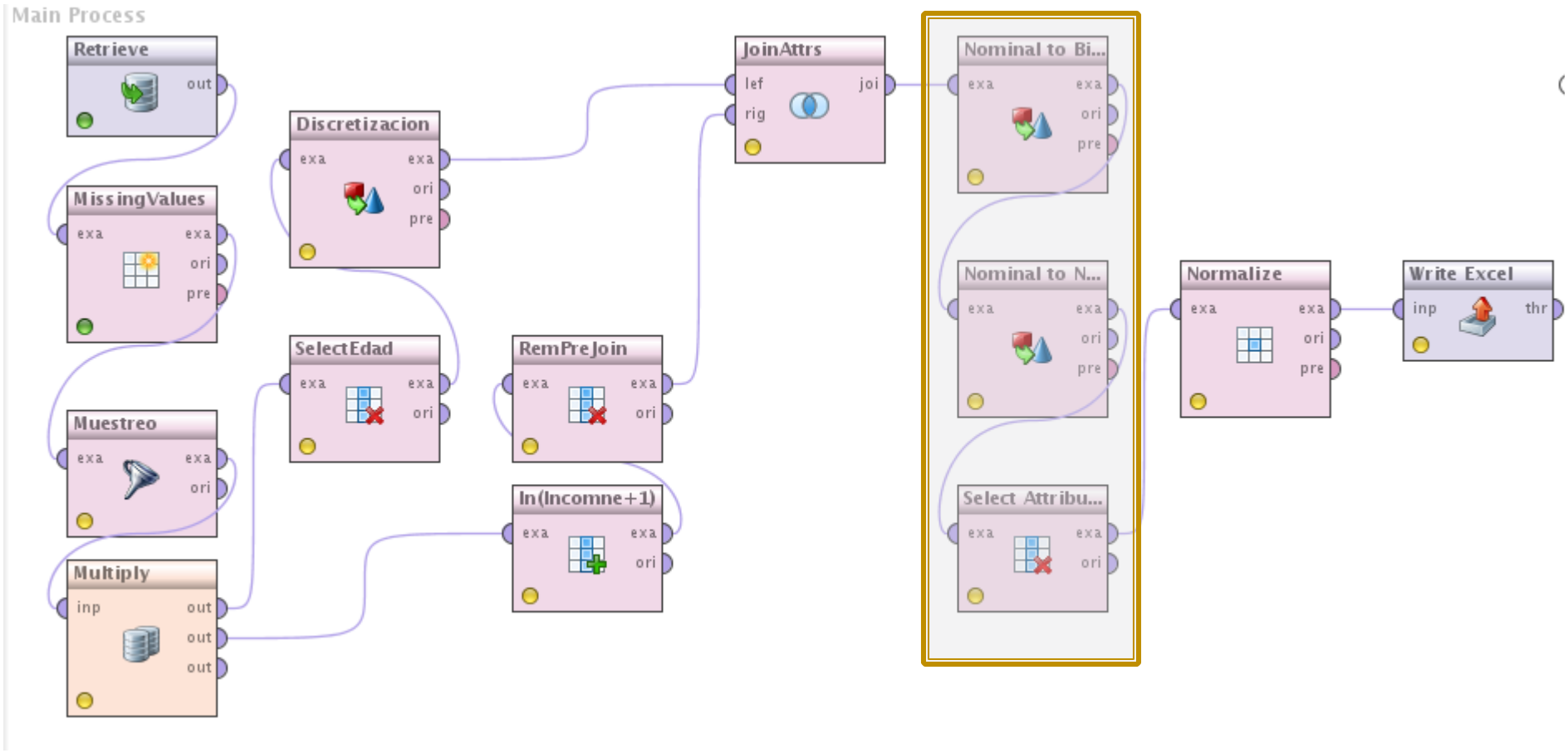
- En general, si tenemos una variable con N categorías distintas, es necesario crear N-1 variables discretas con valores {0,1}
- En RapidMiner se deben utilizar los operadores:
  - Nominal to Binominal: Dado una variable con N categorías, permite crear N variables con valores True o False.
  - Nominal to Numerical: Transforma valores nominales en valores numéricos (True => 1, False =>0)
  - Select Attributes: Nos permite eliminar aquellas variables que pueden ser representadas por las anteriores.

RapidMiner v5.0
Transformación de datos categóricos





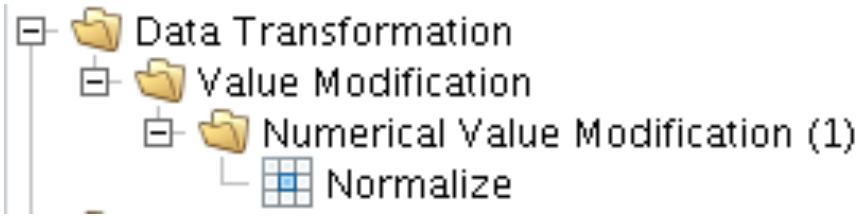
Proceso Transformación

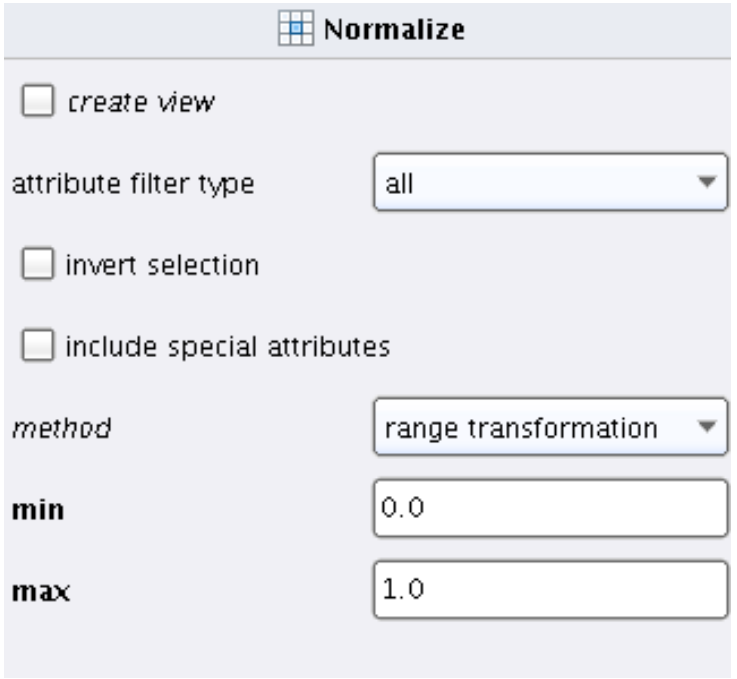


RapidMiner v5.0
Normalización [1]

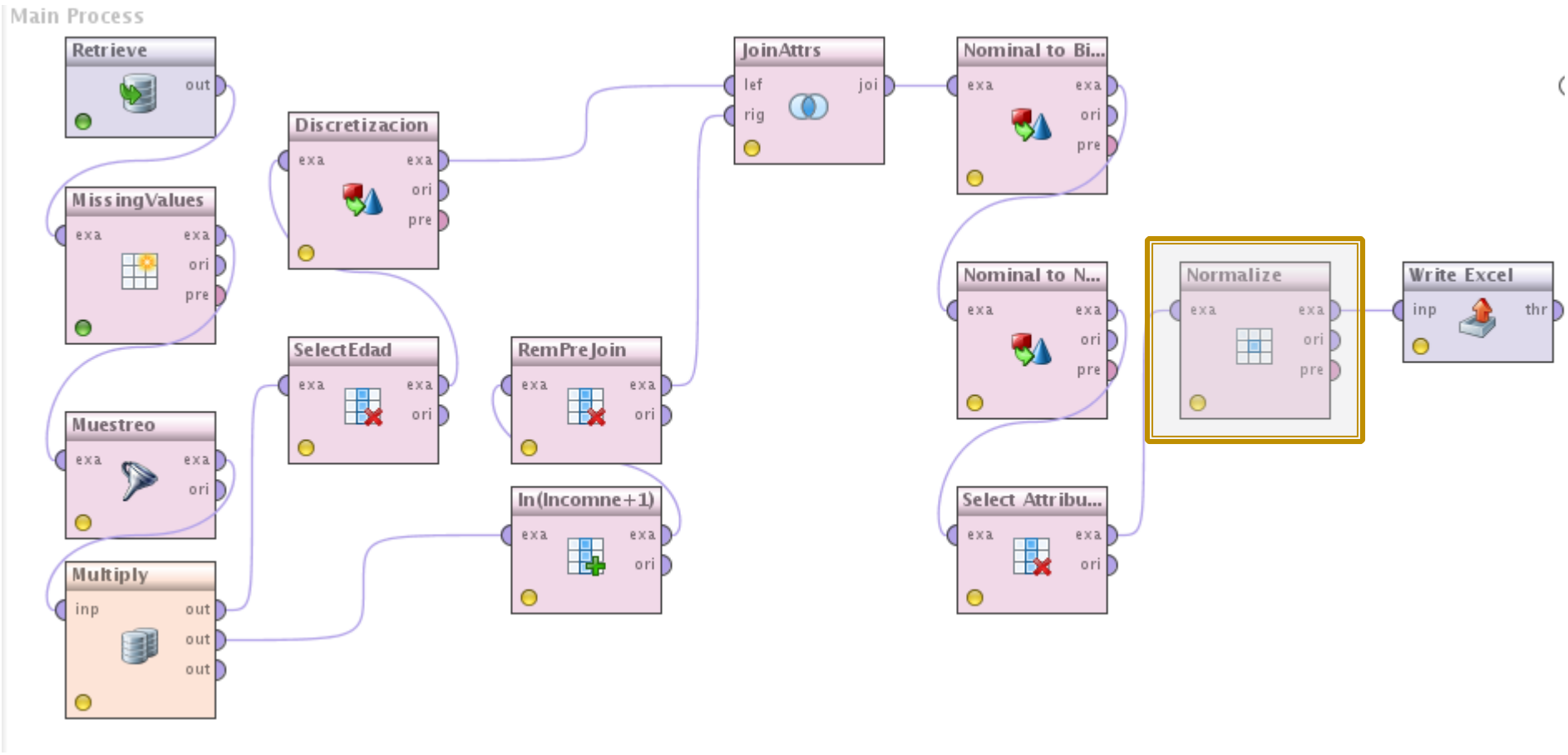
- Es posible normalizar los atributos de acuerdo a ciertas operaciones matemáticas para minimizar el ruido que puedan tener ciertos atributos, además de estandarizarlos y ajustarlos a distribuciones similares.
  - Normalización [0,1]: escala entre 0 y 1 todos los valores de un determinado atributo.
  - Estandarización (Z-transformation): la media de un atributo es 0 y su desviación estándar 1

RapidMiner v5.0
Normalización [2]

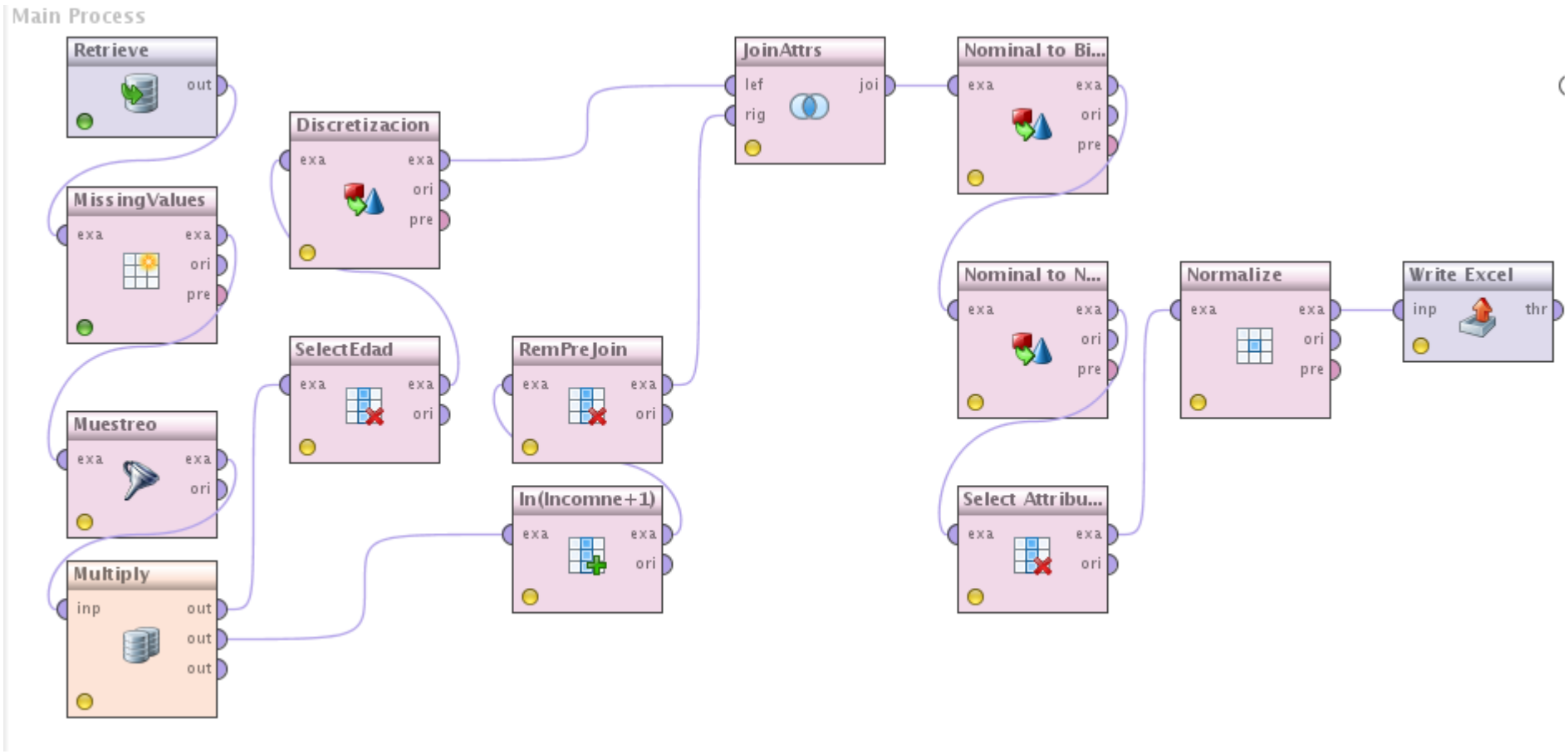




Proceso Transformación



Proceso Transformación

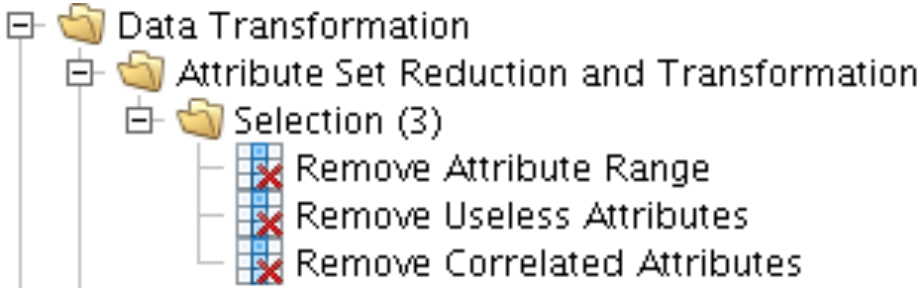


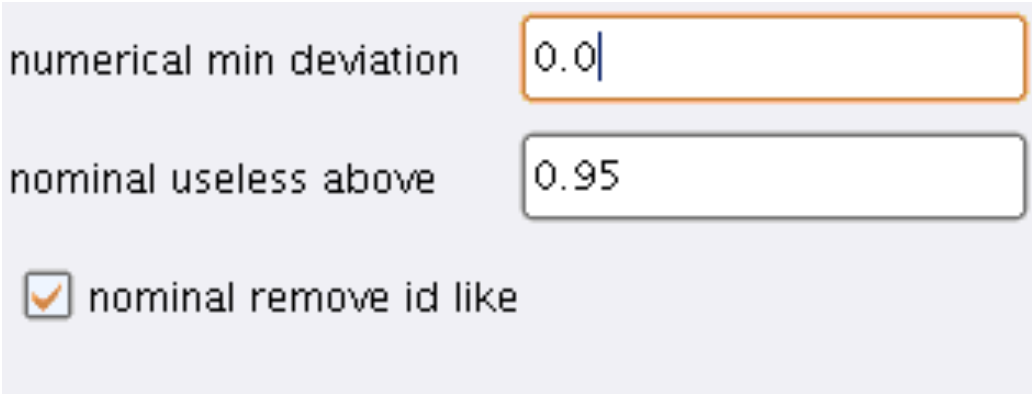
Selección de Atributos

Eliminación de Variables Concentradas

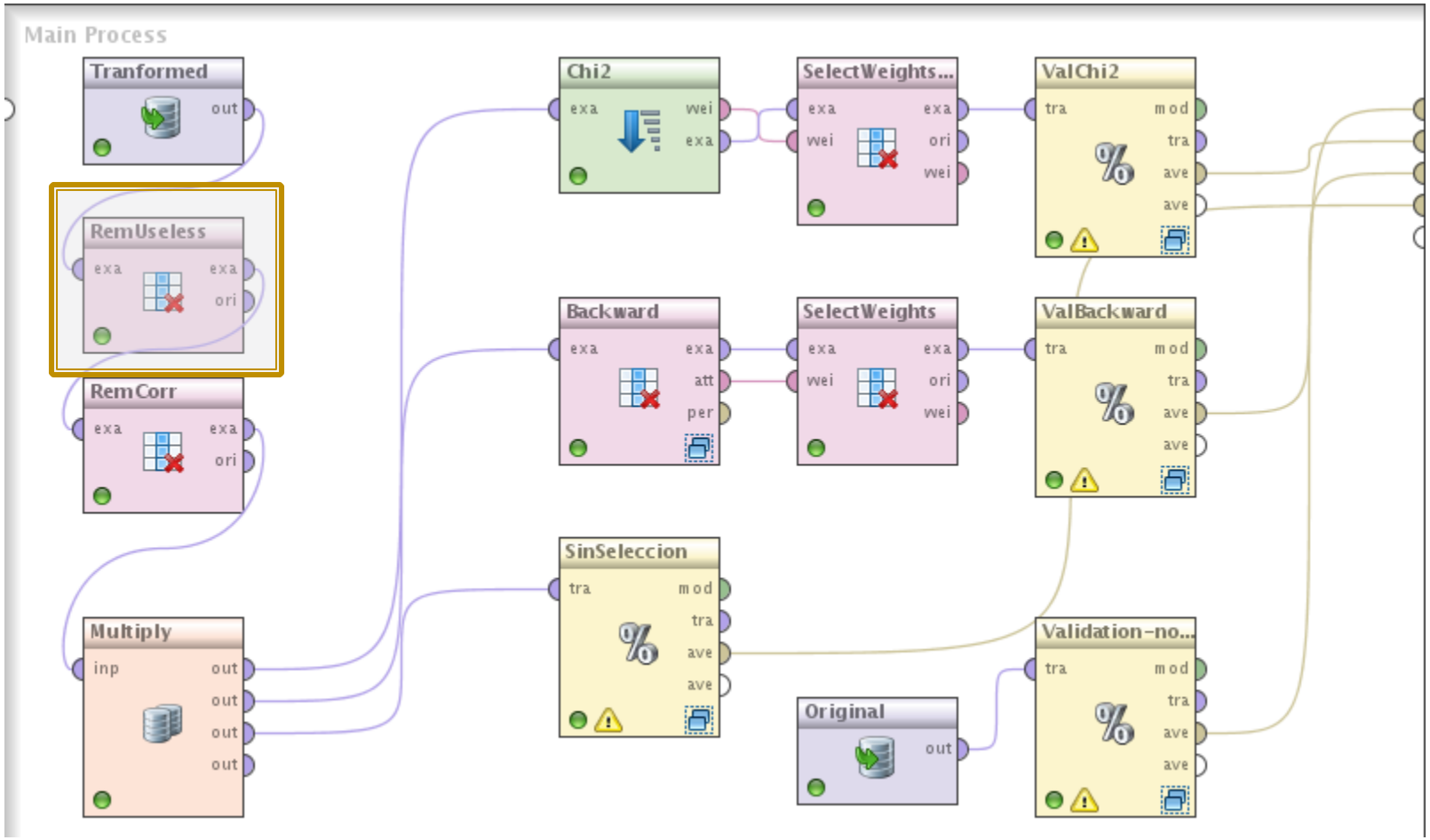
- Si una variable posee muchos valores concentrados en un solo valor (e.g. cero), es recomendable eliminarla.
- Variable sin varianza, no posee información relevante, por lo que puede ser seguramente descartada.
- Criterios:
  - Nominales: Si se concentra en más de un cierto porcentaje (95% de repetidos).
  - Numéricas: Si desviación estándar es menor a un cierto valor

RapidMiner v5.0
Eliminación de variables concentradas





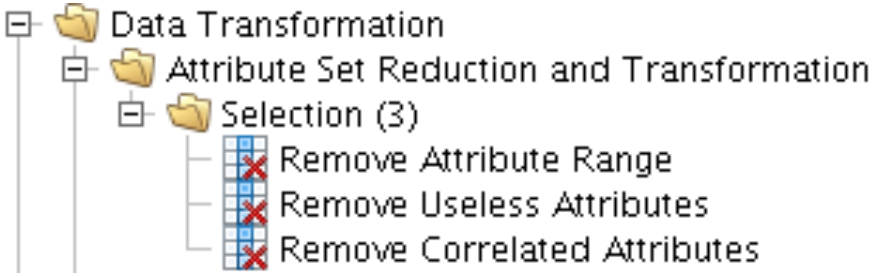
Selección de atributos

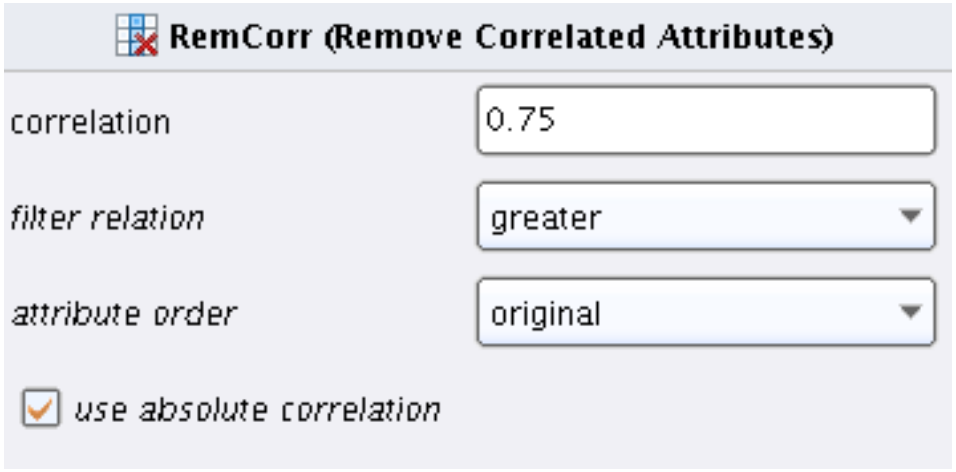


Eliminación de Variables Correlacionadas

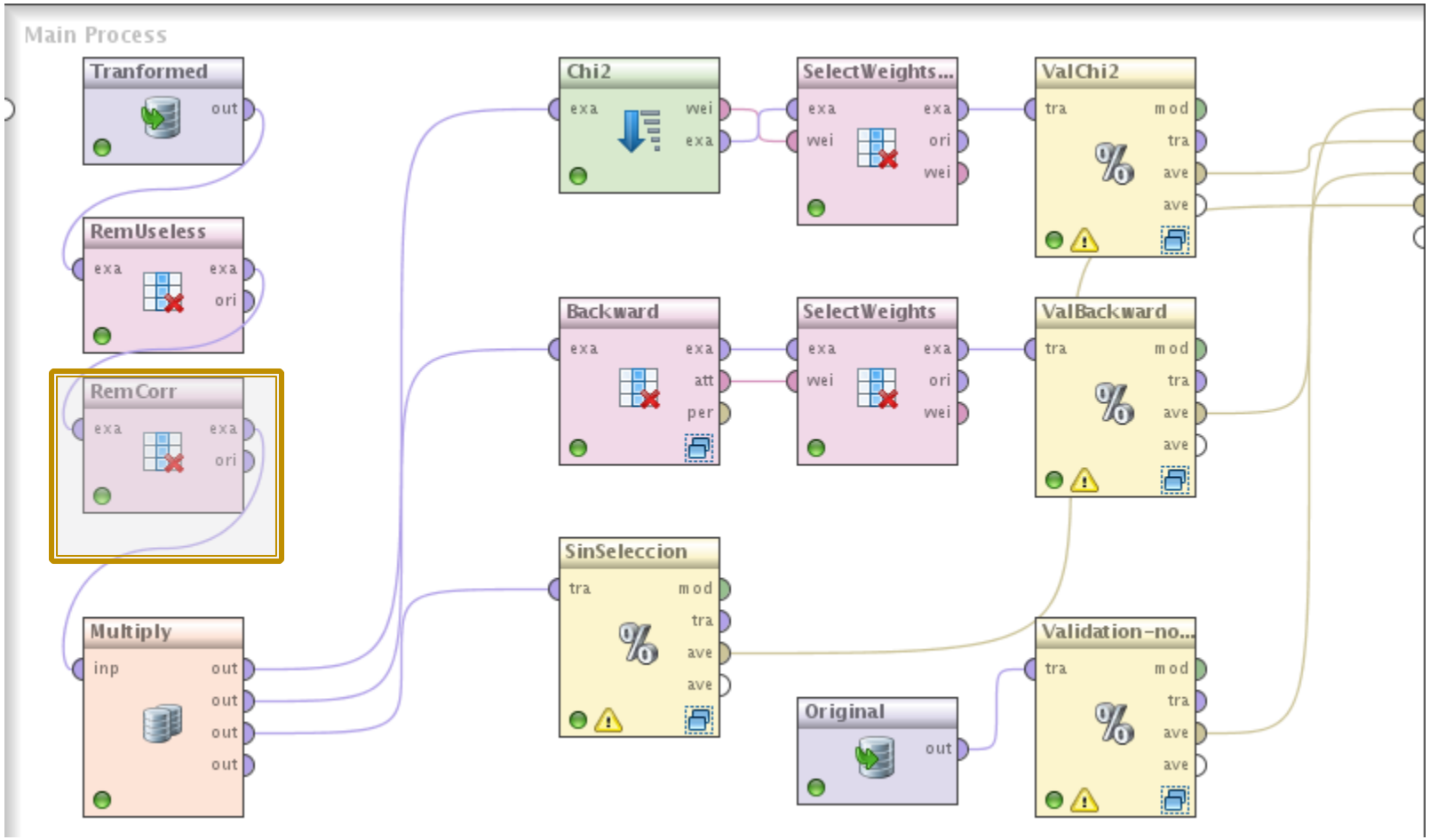
- Si un par de variables está altamente correlacionada, es posible que con una de ellas sea suficiente para la construcción de un modelo.
  - Redundancia de información.
  - Una regla aceptable para definir una "correlación alta" en problemas reales, es Corr > 0.75

RapidMiner v5.0
Eliminación de variables correlacionadas





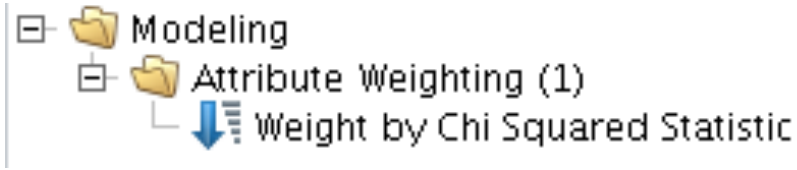
Selección de atributos

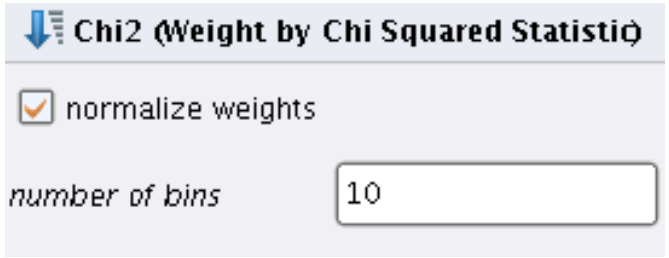


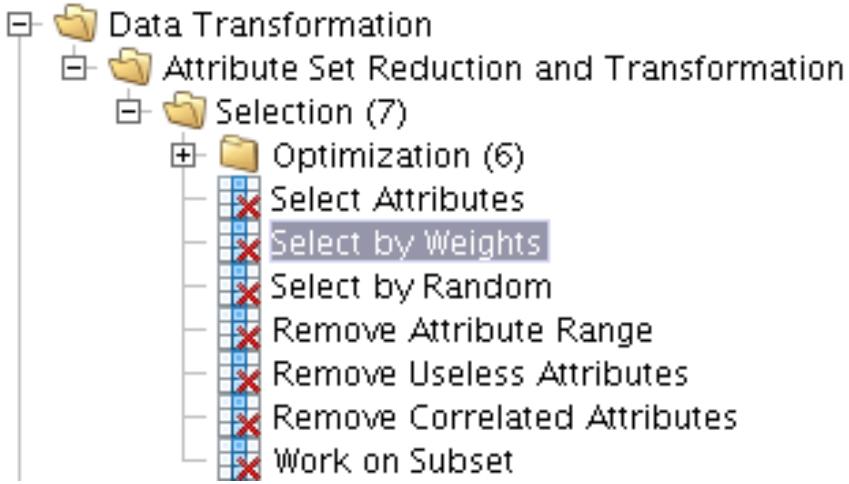
Selección de Variables por pesos

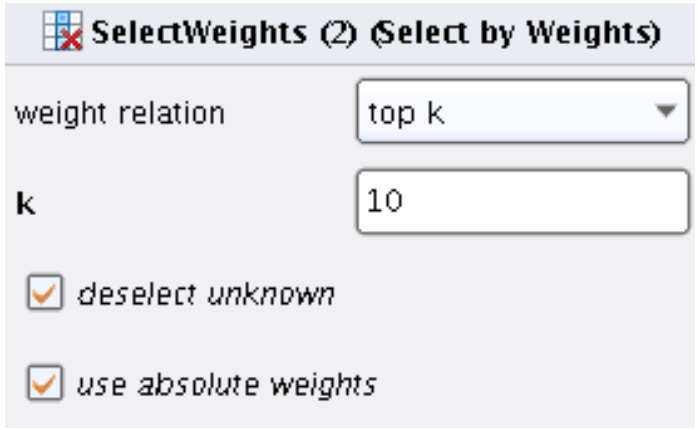
- Es posible asignar por cada atributo cuanto es su peso asociado al problema de clasificación particular.
- Según esto, varias técnicas cómo Ganancia de Información, test Chi^2, Coeficiente de correlación, entre otras, se pueden utilizar para definir cuanto es el peso de una variable dada, con respecto a la variable objetivo.

RapidMiner v5.0
Selección por Peso

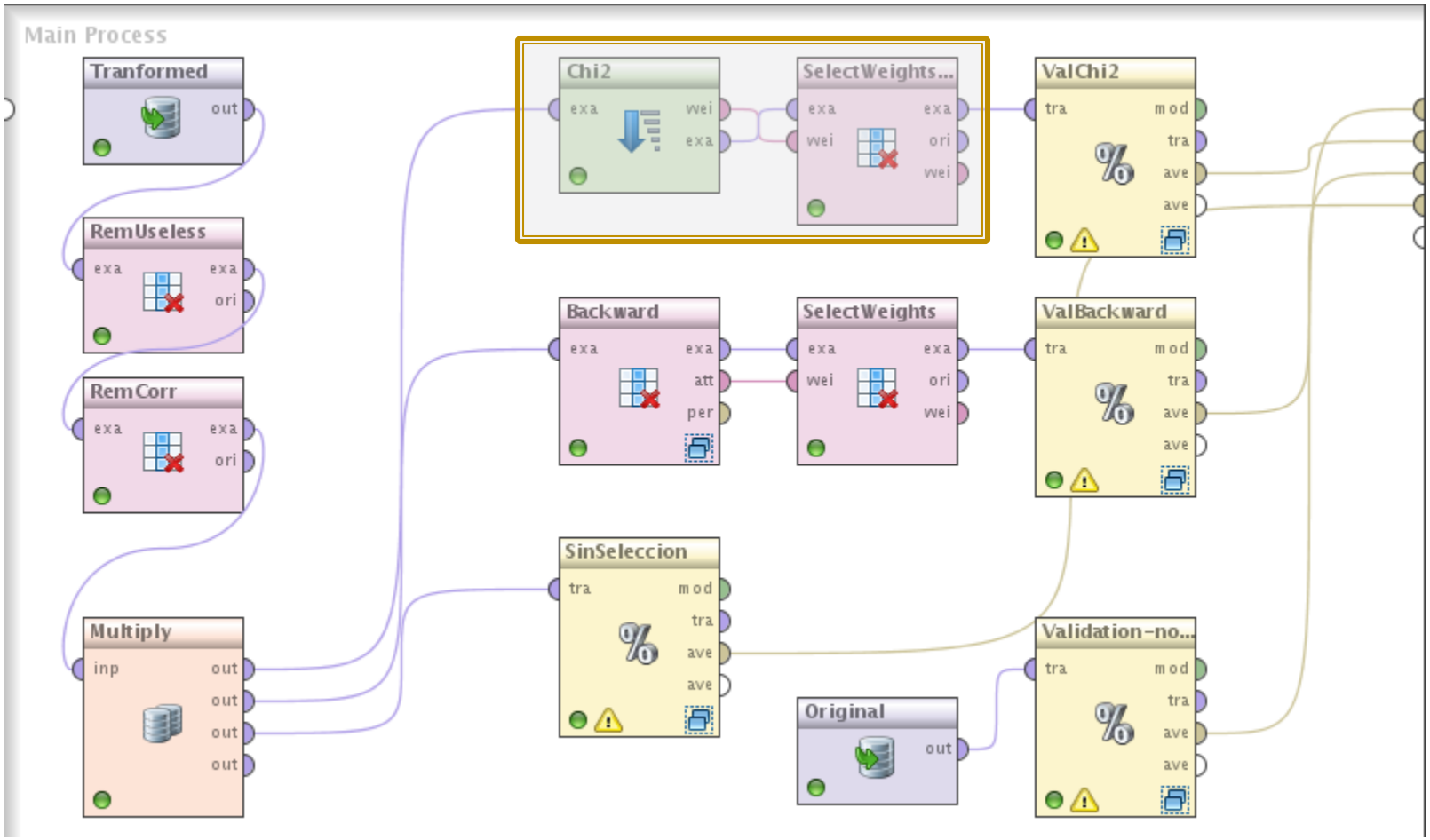








Selección de atributos



Selección de Variables
Backward o Forward Selection

- Existe una estrategia de selección de atributos basada en la evaluación incremental (o de-cremental) de la performance de cierto atributo con respecto a un modelo predictivo.
- "Backward selection", comienza con la evaluación de un modelo con todos sus atributos, luego va eliminando iterativamente los atributos hasta determinar el conjunto óptimo de atributos.

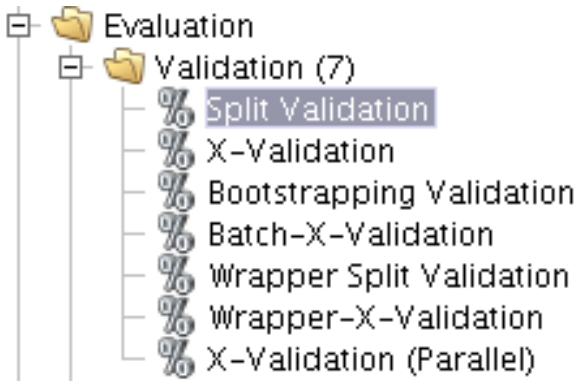
Selección de Variables
Backward o Forward Selection

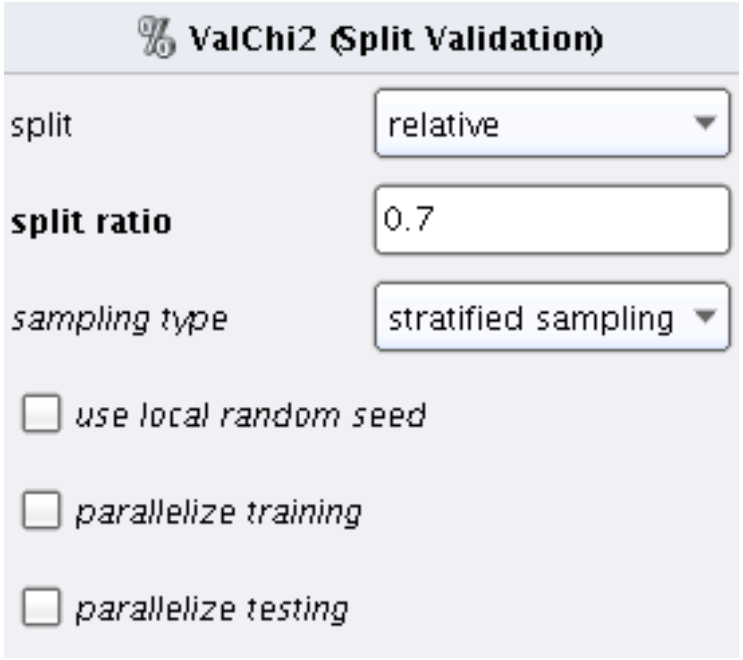
- "Forward selection", comienza con la evaluación de un modelo con un solo atributo. Luego va agregando incrementalmente el resto de los atributos, hasta que determina cuál es el conjunto óptimo.

RapidMiner v5.0
Backward Selection [1]

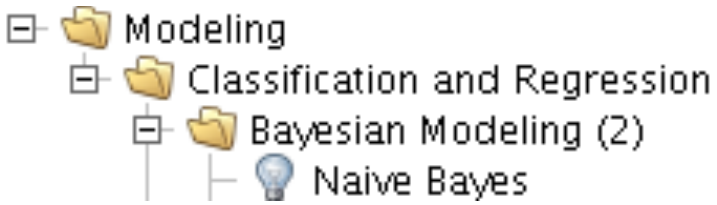
- Es necesario aplicar un modelo predictivo (!)
- Utilizaremos *Naive Bayes*, evaluando un *Hold-out* con 70% de Entrenamiento y 30% de Test.
- Además utilizaremos la *Precision*, *Recall* y la *F-Measure* como métricas de evaluación.

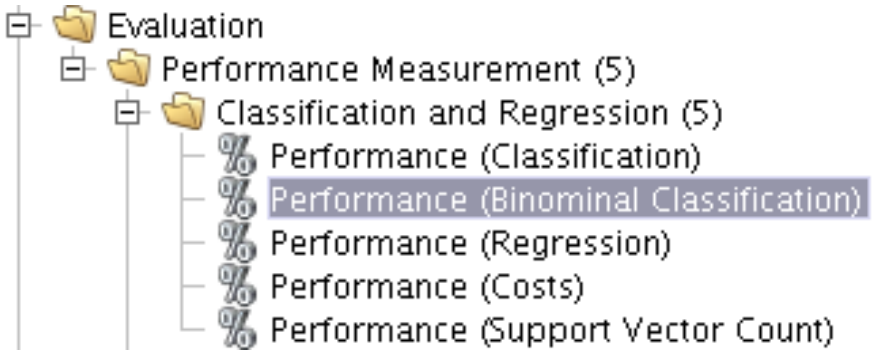
RapidMiner v5.0
Backward Selection [2]

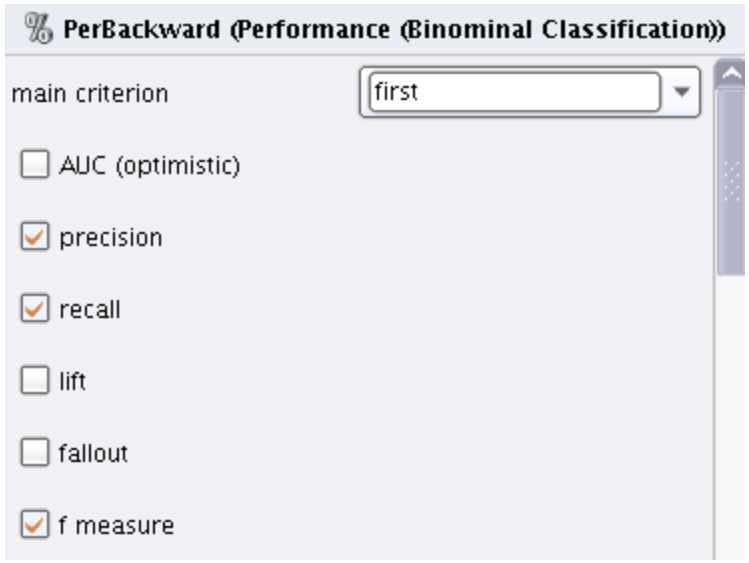




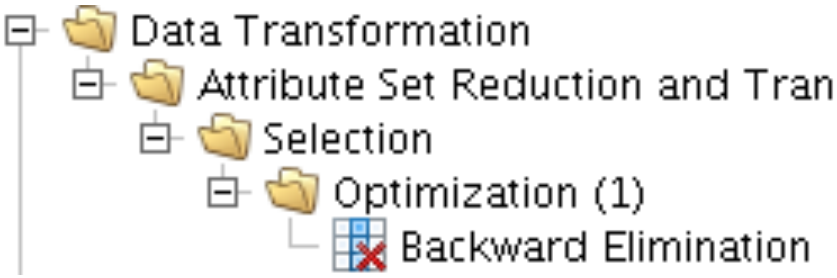
RapidMiner v5.0
Backward Selection [2]

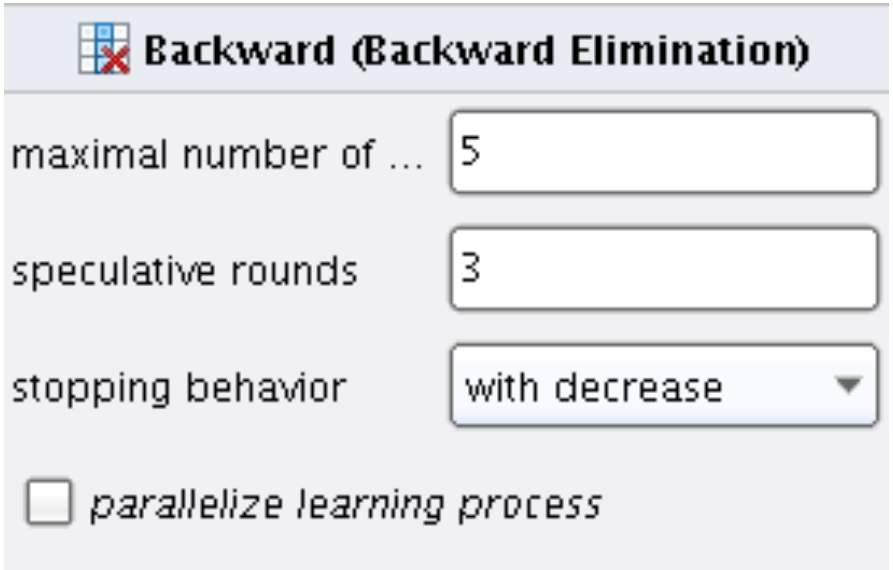




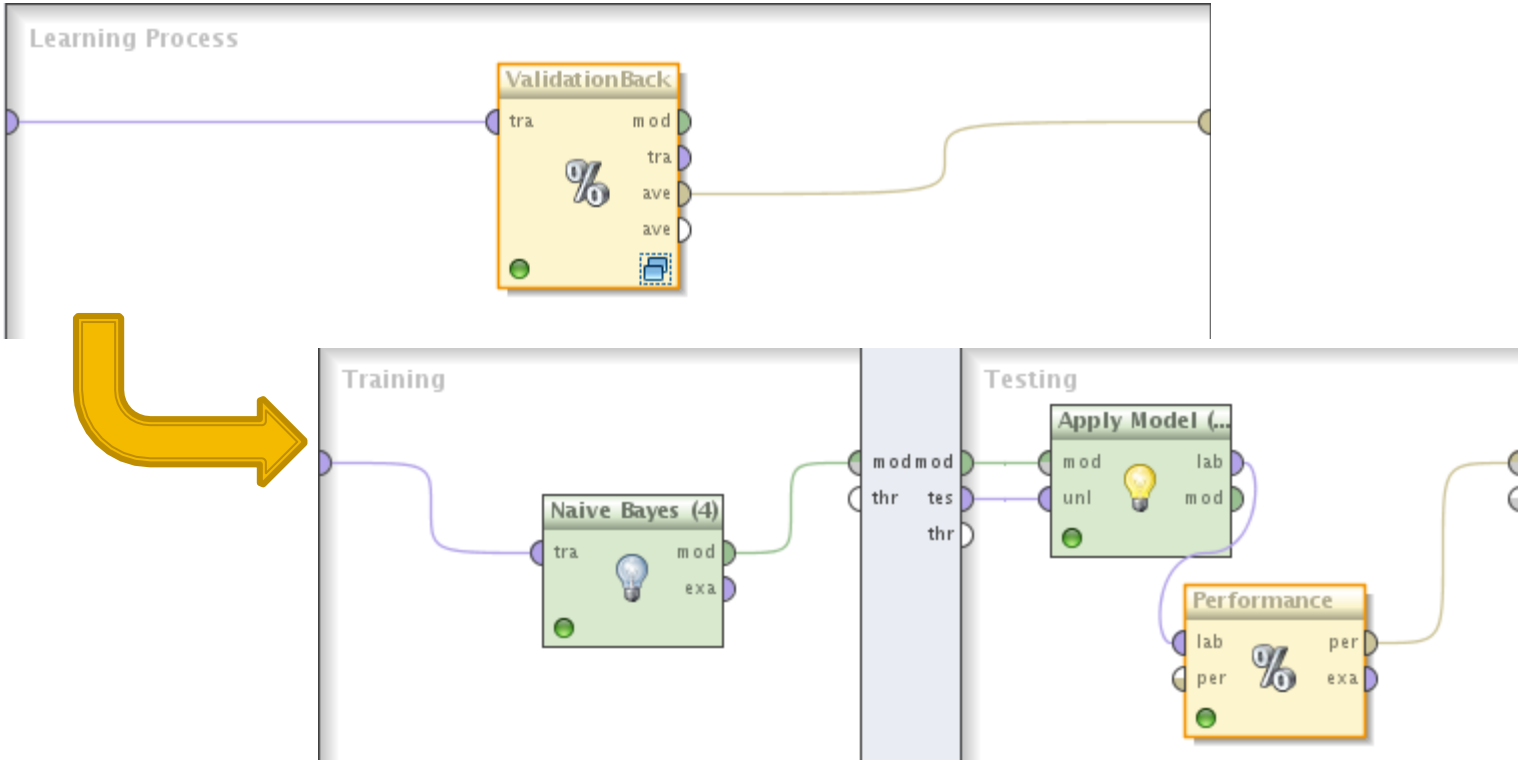


RapidMiner v5.0
Backward Selection [2]

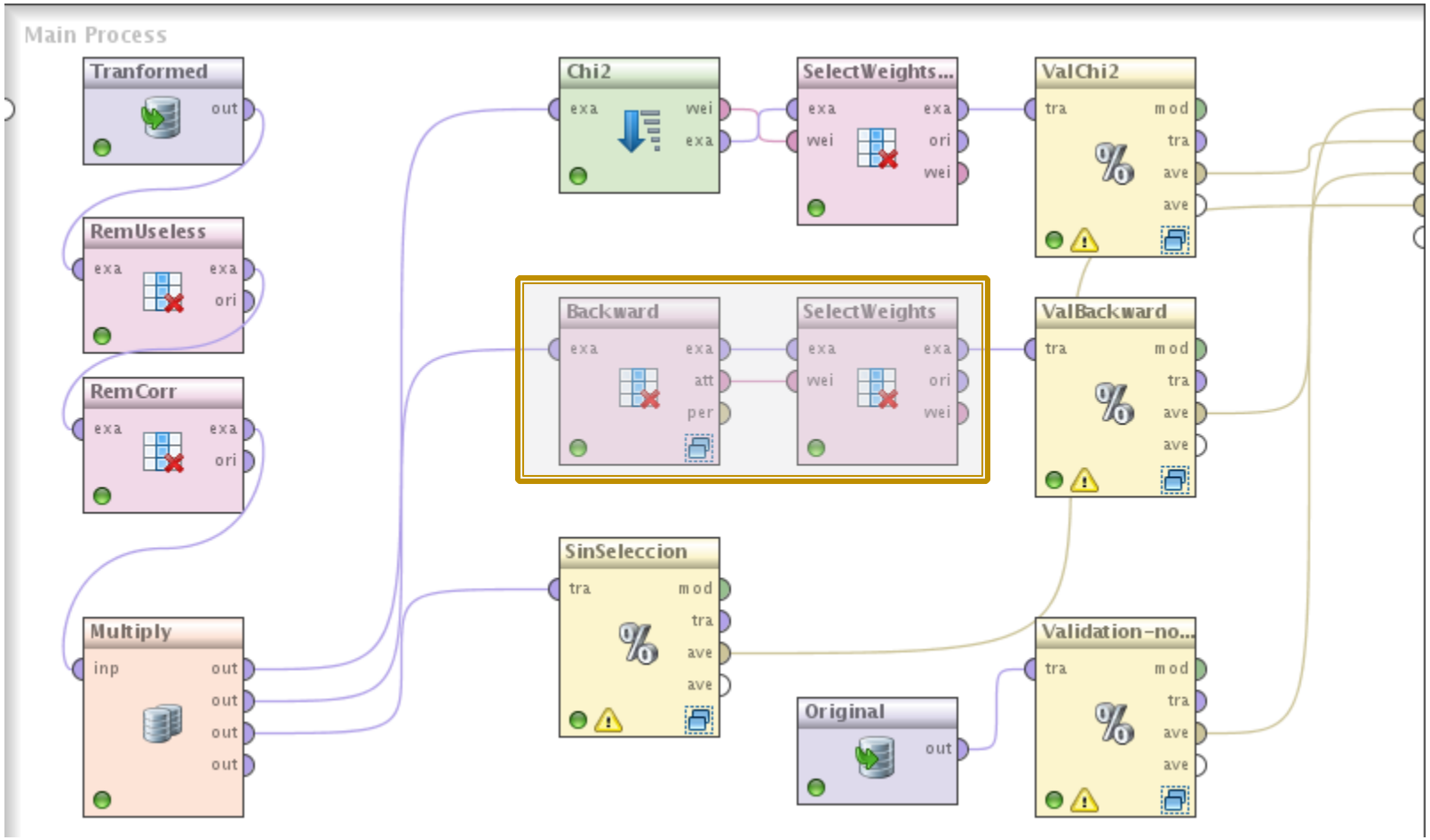




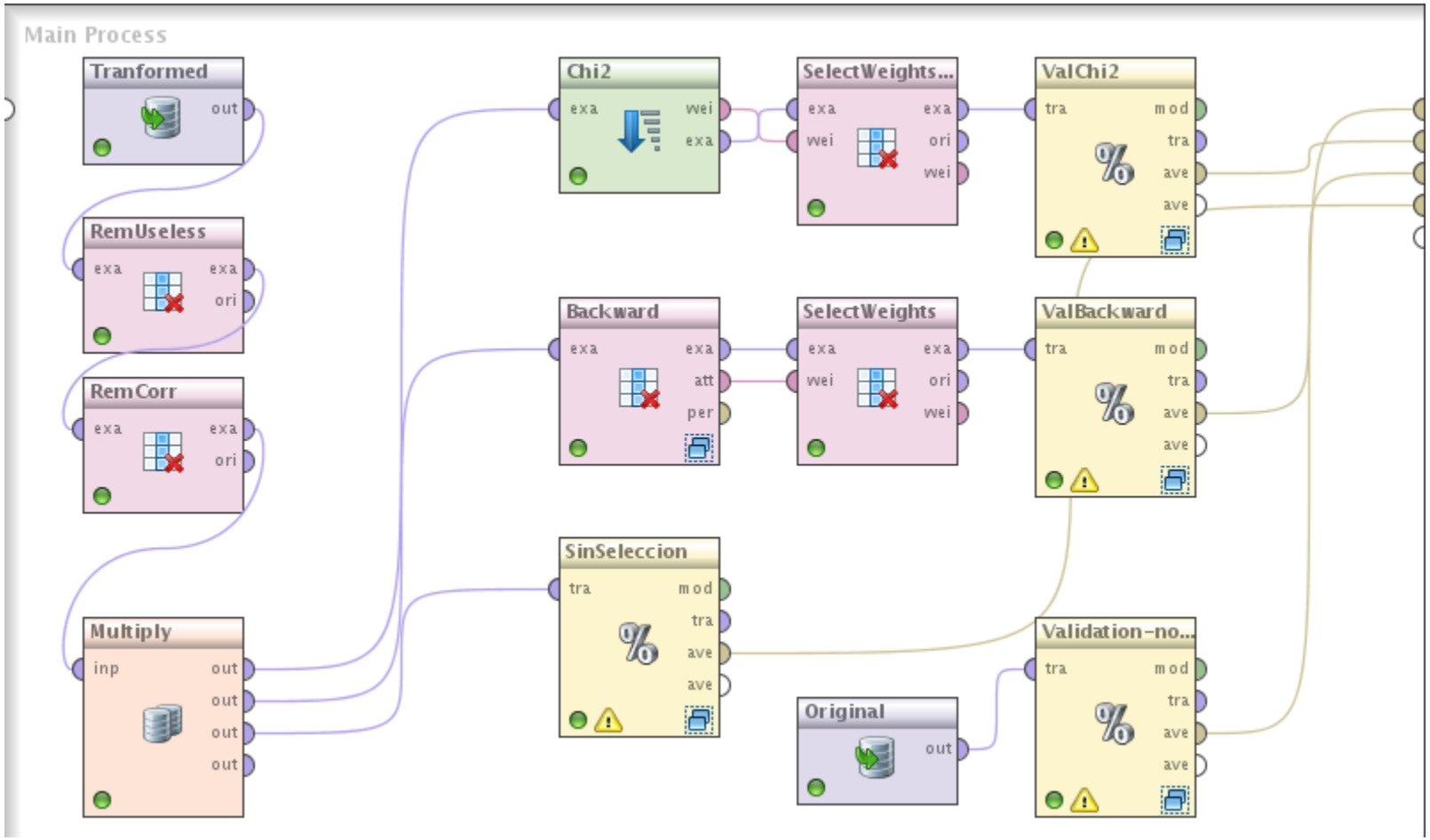
RapidMiner v5.0
Backward Selection [3]



Selección de atributos



RapidMiner v5.0
Evaluación selección de atributos



RapidMiner v5.0
Evaluación selección de atributos

Performance Vector (PerBackward)

```
ConfusionMatrix:
True:   N       S
N:      221     58
S:      65      106
recall: 64.63% (positive class: S)
ConfusionMatrix:
True:   N       S
N:      221     58
S:      65      106
f_measure: 63.28% (positive class: S
ConfusionMatrix:
True:   N       S
N:      221     58
S:      65      106
```

Performance Vector (PerSinPrep)

```
ConfusionMatrix:
True:   N       S
N:      257     87
S:      29      77
recall: 46.95% (positive class: S)
ConfusionMatrix:
True:   N       S
N:      257     87
S:      29      77
f_measure: 57.04% (positive class: S
ConfusionMatrix:
True:   N       S
N:      257     87
S:      29      77
```

Performance Vector (PerSinSelecc)

```
PerformanceVector:
precision: 61.08% (positive class: S
ConfusionMatrix:
True:   N       S
N:      221     62
S:      65      102
recall: 62.20% (positive class: S)
ConfusionMatrix:
True:   N       S
N:      221     62
S:      65      102
f_measure: 61.63% (positive class: S
ConfusionMatrix:
True:   N       S
```

Performance Vector (PerChi2)

```
ConfusionMatrix:
True:   N       S
N:      246     82
S:      40      82
recall: 50.00% (positive class: S)
ConfusionMatrix:
True:   N       S
N:      246     82
S:      40      82
f_measure: 57.34% (positive class: S
ConfusionMatrix:
True:   N       S
N:      246     82
S:      40      82
```

Taller #3

Business Intelligence

Carlos Reveco
creveco@dcc.uchile.cl

Cinthya Vergara
cvergarasilv@ing.uchile.cl