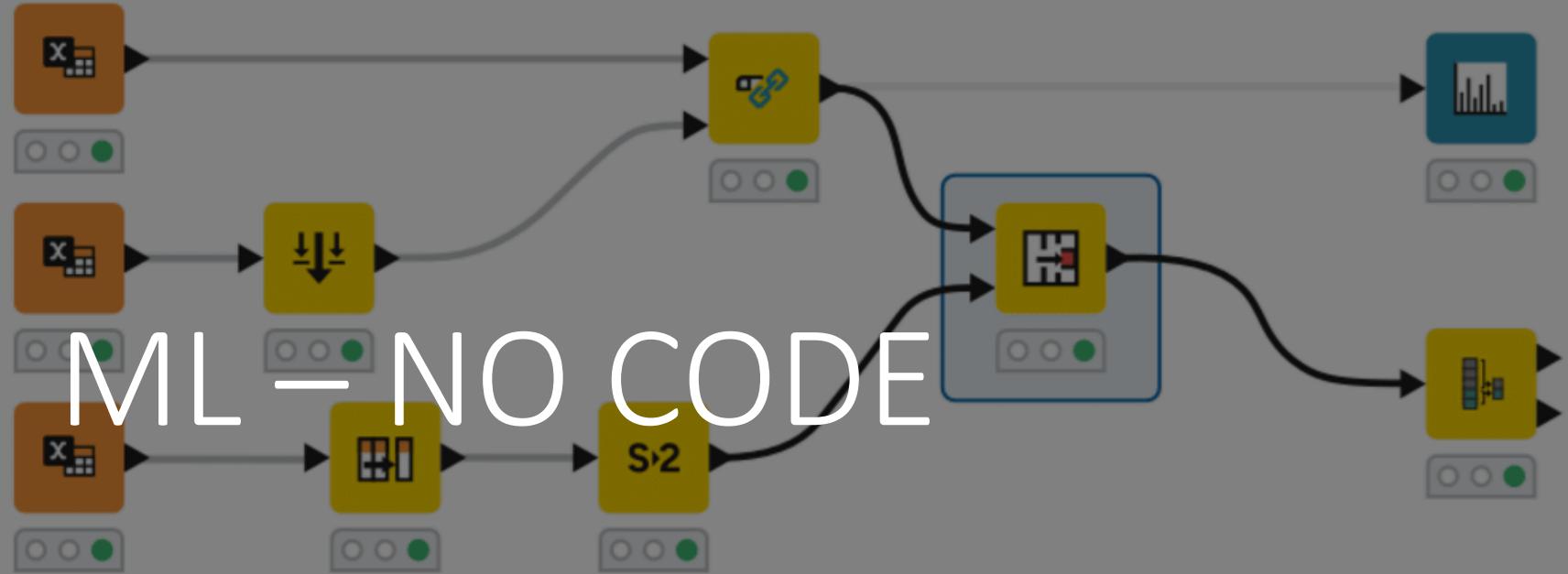


# ML - NO CODE



# SQLite Connector

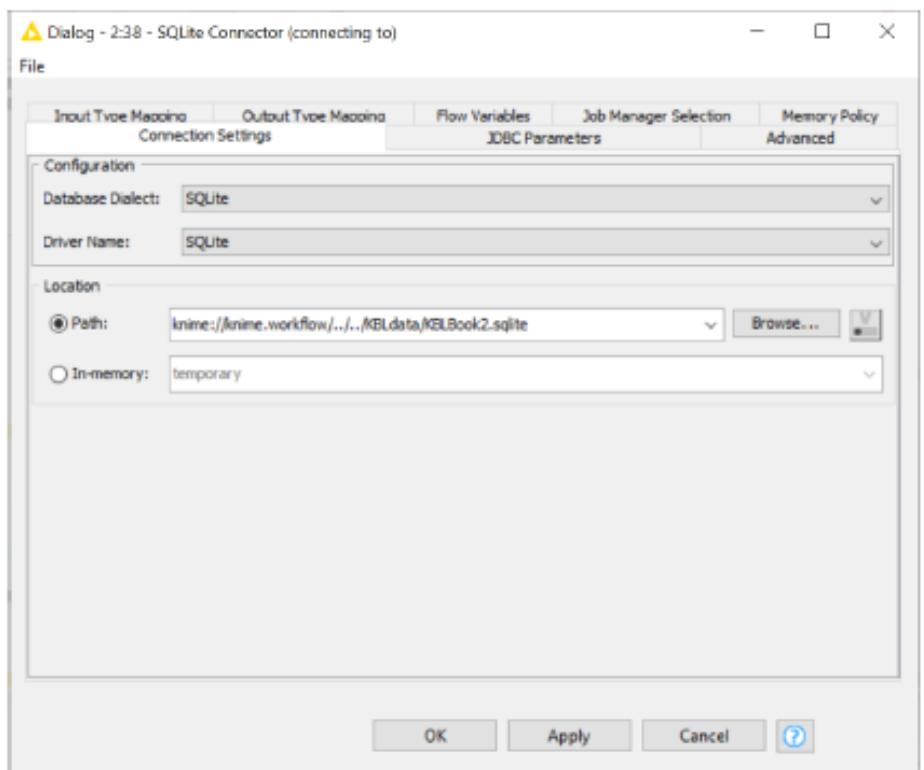
Para la base de datos SQLite, existe un nodo conector dedicado: el nodo Conector SQLite. Su ventana de configuración sólo requiere la ruta del archivo sqlite y ninguna contraseña. El controlador JDBC para la base de datos SQLite ya está precargado.

En la ventana de configuración de todos los nodos del conector, hay cinco pestañas: “Connection Settings”, “JDBC Parameters”, “Advanced”, “Input Type Mapping”, “Output Type Mapping”.

“Connection Settings” contiene todos los ajustes necesarios para conectarse a la base de datos: Controlador JDBC, nombre de host, puerto, nombre de la base de datos y credenciales completas cuando sea necesario.

Las pestañas “JDBC Parameters” y “Advanced” permiten establecer comandos específicos para conectarse a la base de datos; mientras que las pestañas “Input Type Mapping” y “Output Type Mapping” permiten mapear correctamente todos los tipos de datos de KNIME a la base de datos y viceversa.

3.20. Nodo “SQLite Connector” : pestaña “Connection Settingsen la ventana de configuración



# MySQL Connector

El nodo Conector MySQL se conecta a una base de datos MySQL y requiere:

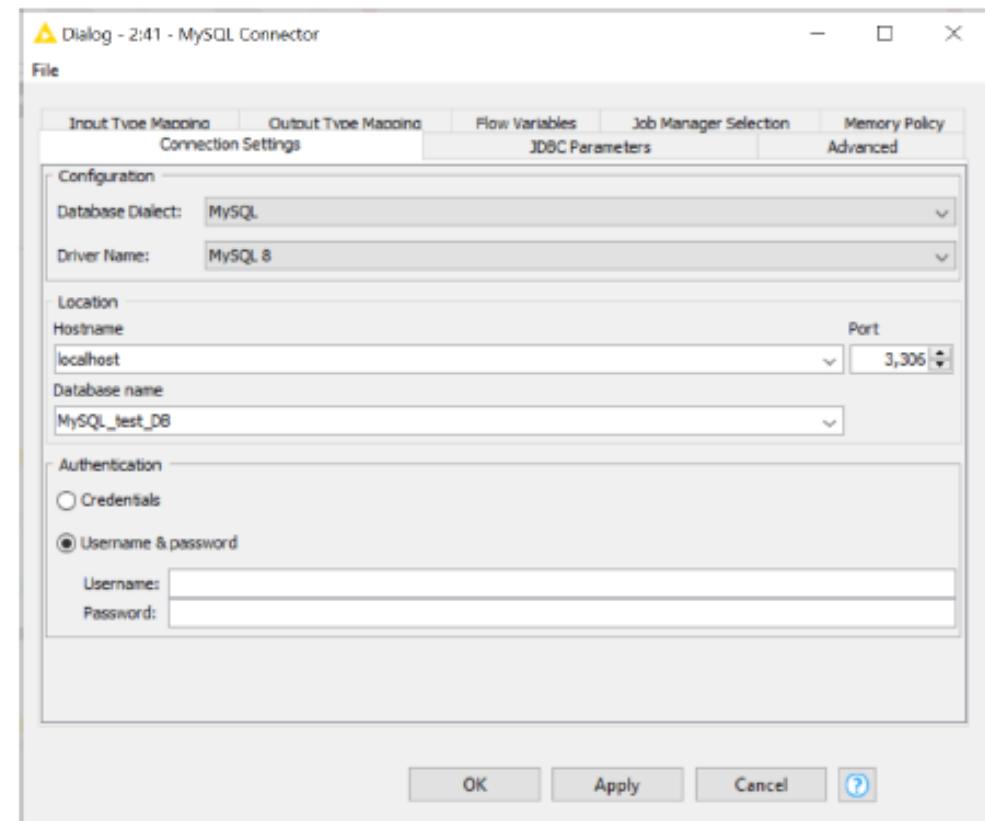
- El controlador de la base de datos (precargado)
- El nombre del host y de la base de datos
- El nombre de usuario y la contraseña para la autenticación

Las credenciales pueden ser suministradas como nombre de usuario y contraseña habilitando la opción “Username & password”.

Otra opción es definirlas como credenciales a nivel de flujo de trabajo.

Las otras pestañas: “JDBC Parameters”, “Advanced”, and “Mappings” incluye las mismas funcionalidades que el nodo SQLite Connector

3.21. Nodo “MySQL Connector” : pestaña “Connection Settings”

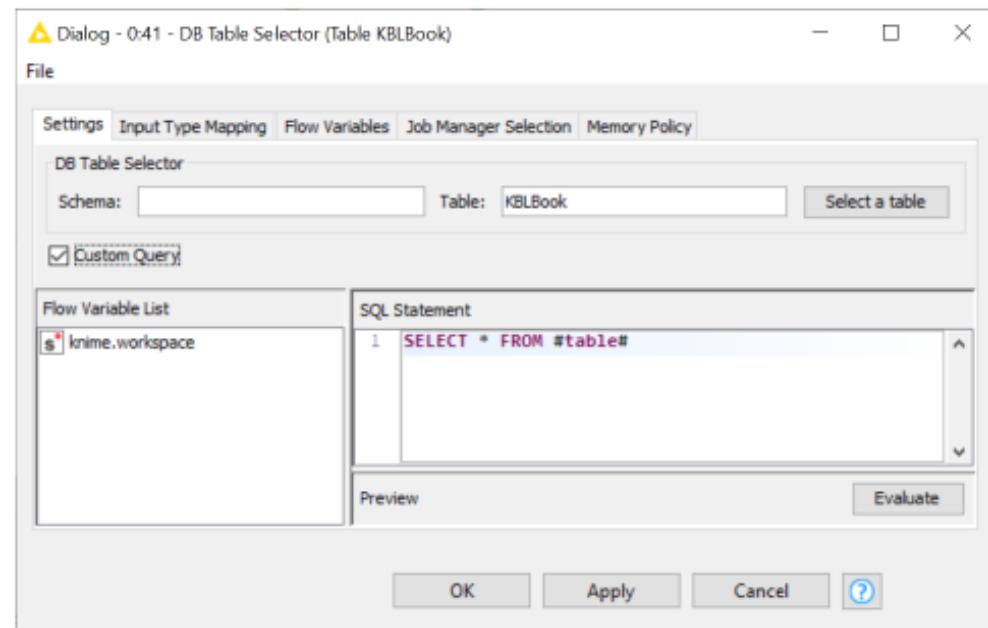


## DB Table Selector

El nodo "DB Table Selector", situado en "DB"/"Query", selecciona una tabla de la base de datos proporcionada con la conexión a la base de datos en su puerto de entrada. Se necesitan los ajustes de configuración siguientes:

- El nombre y opcionalmente el esquema de la tabla de la base de datos. El botón "Select a table" permite navegar por el contenido de la base de datos para seleccionar la tabla deseada.
- La consulta por defecto es "SELECT \* FROM #table#" donde #table# es la tabla seleccionada, que incluye todas las filas de datos y todas las columnas de datos de la tabla.
- También es posible extraer una parte o una transformación de la #table# original, creando una consulta personalizada. El editor de consultas aparece cuando se selecciona "Custom Query".
- En el editor SQL (si está disponible) puede entonces escribir su consulta personalizada para extraer los datos de la #table#

3.29. Configuration of the "Database Writer" node when following a Database Connector node



# Contenedores numéricos (Numeric Binner)

El nodo "Numeric Binner" – localizado en "Node Repository" : "Manipulation" → "Column" → "Binning" category - define una serie de intervalos (es decir, bins) y asigna cada valor de columna a su bin.

3.31. Ventana del nodo Numeric Binner\*

La ventana de configuración requiere lo siguiente:

- La columna numérica que se va a procesar
- La lista de los intervalos de las ubicaciones
- Un indicador que señale si los valores de los recipientes deben aparecer en una nueva columna o reemplazar la columna original

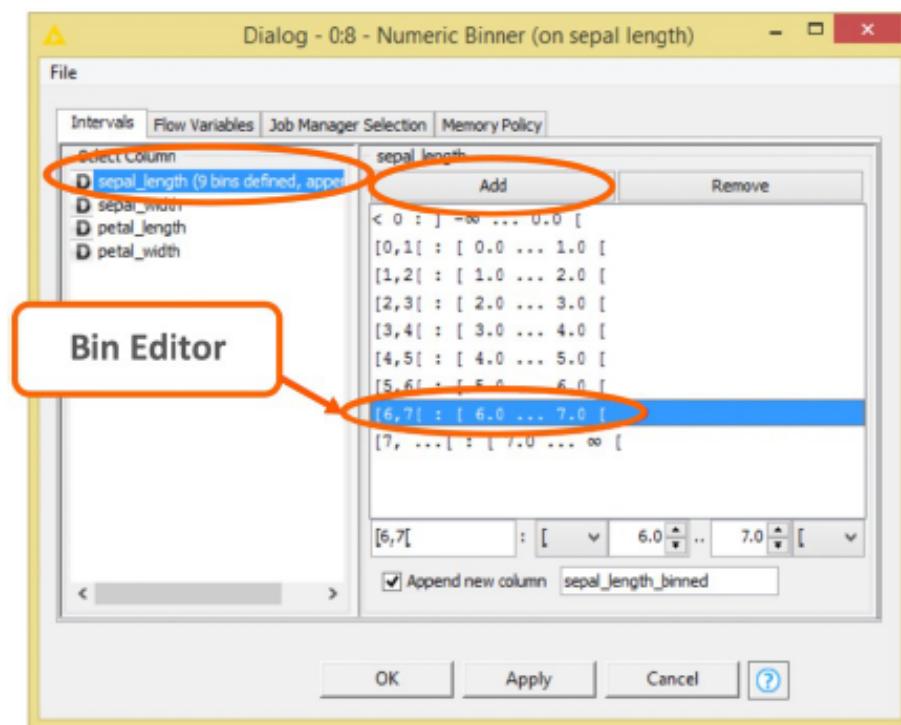
Para definir un nuevo intervalo de contenedores:

- Haga Click en "Add"
- Personalice el rango de contenedores en el Bin Editor

Para editar un nuevo intervalo de contenedores:

- Seleccione el intervalo de recipientes en la lista de intervalos de recipientes
- Personalice el intervalo de recipientes en el Editor de recipientes

Puede construir una nueva representación de recipientes seleccionando otra columna y repitiendo el procedimiento de agrupación de recipientes.



## Agrupando (GroupBy: "Groups" tab)

El nodo "GroupBy" encuentra grupos de filas de datos utilizando la combinación de valores en una o más columnas (**Group Columns**); posteriormente agrega los valores en otras columnas (**Aggregation Columns**) a través de esos grupos. Los valores de las columnas pueden agregarse en forma de suma, media, sólo un recuento de ocurrencias, o utilizando otros métodos de agrupación (**Aggregation Method**).

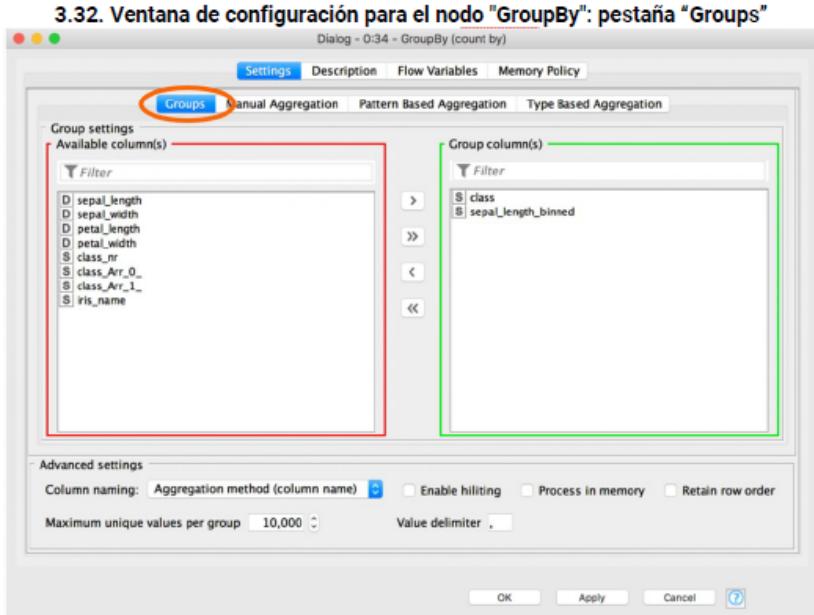
La ventana de configuración del nodo "GroupBy" consta de varias pestañas. Aquí marcamos la pestaña denominada "**Groups**".

La pestaña "**Groups**". define las opciones de agrupación. Es decir, selecciona la(s) columna(s) del grupo mediante un cuadro "Excluir"/"Incluir":

- Las columnas aún disponibles para la agrupación aparecen en el cuadro "Available column(s)". Las columnas seleccionadas aparecen en el marco "Group column(s)".
- Para pasar del marco "Available column(s)" al marco "Group column(s)" y viceversa, utilice los botones "add y "remove". Para mover todas las columnas a un marco u otro, utilice los botones ""add all" y "remove all".

La parte inferior de la ventana de configuración

- establece el nombre de la nueva columna
- mantiene el orden de las filas o las resitúa en orden alfabético
- rechaza las columnas con demasiados valores distintos (por defecto 10000),
- la opción "Enable hiliting" hace referencia a una función disponible en el antiguo nodo "Data Views".



## GroupBy: Aggregation tabs

El resto de pestañas de la ventana de configuración definen los ajustes de agregación, es decir

- *La(s) columna(s) de agregación*
- *El método de agregación (uno para cada columna de agregación)*

Las diferentes pestañas seleccionan las columnas sobre las que realizar la agregación utilizando diferentes criterios:

- Manualmente, una por una, a través de un cuadro "Exclude"/"Include": todas las columnas seleccionadas se utilizarán para la agregación
- Basado en un patrón regex o comodín: todas las columnas cuyo nombre coincide con el patrón se utilizarán para la agregación
- Basado en el tipo de columna: todas las columnas del tipo seleccionado se utilizarán para la agregación

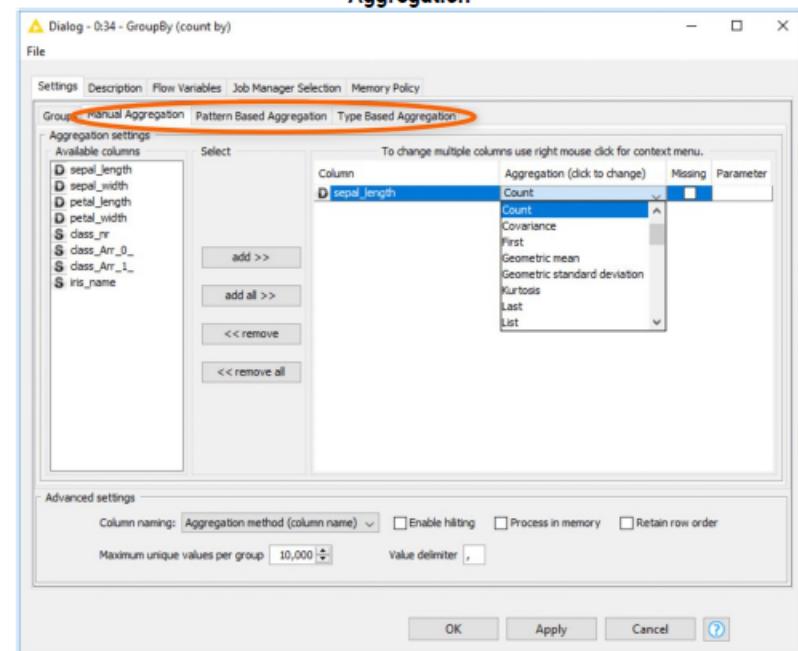
Hay varios métodos de agregación disponibles en todas las fichas de agregación. Todos los métodos de agregación disponibles se describen en detalle en el "Description" tab.

Los métodos de agregación difieren para las columnas numéricas (incluyendo aquí las medidas estadísticas, como la media, la varianza, la asimetría, la mediana, etc...) y para las columnas de cadenas (incluyendo el recuento único, por ejemplo).

Observe que los métodos de agregación "Count" y "Percent" sólo cuentan el número de filas de datos de un grupo y su valor porcentual con respecto al conjunto de datos. Esto significa que cualquiera que sea la columna de agregación asociada a estos dos métodos de agregación, los resultados no cambiarán, ya que el recuento de filas de datos de un grupo y su porcentaje no depende de la columna de agregación, sino sólo del grupo de datos.

Los métodos de agregación "First" and "Last" extraen respectivamente la primera y la última fila de datos del grupo actual.

3.33. Ventana de configuración del nodo "GroupBy": pestaña "Manual Aggregation"



## Gráfico de dispersión: vista interactiva

Esta es la vista del nodo "Scatter Plot", donde se pueden ver los puntos del gráfico de dispersión.

Hay tres botones en la esquina superior derecha. Son los botones de interactividad.

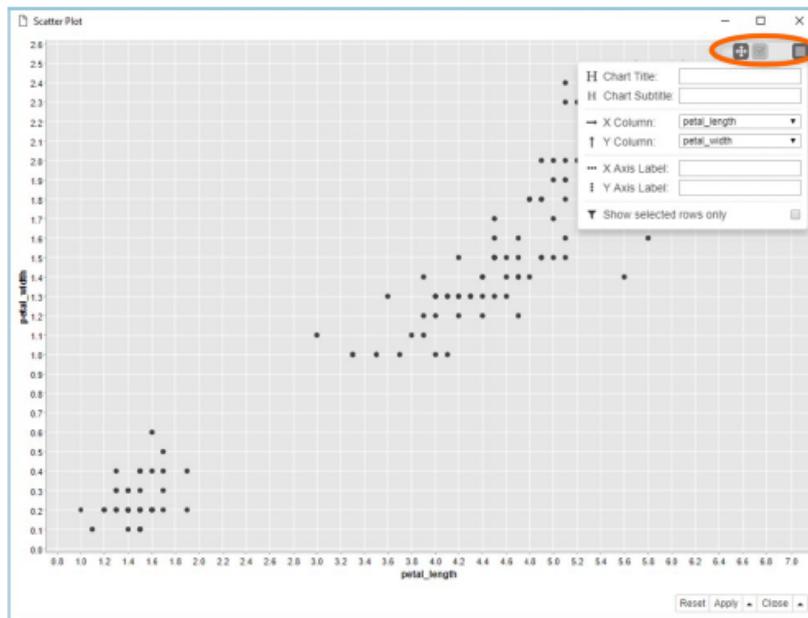
Empezando por el extremo derecho, tenemos el botón que permite cambiar la configuración del gráfico, como las etiquetas de los ejes, las columnas para el eje x y el eje y, y el título.

El segundo botón desde la derecha pone el clic del ratón en modo de selección. Cuando está activado, al hacer clic en un punto o dibujar un rectángulo en el gráfico se seleccionan los puntos correspondientes.

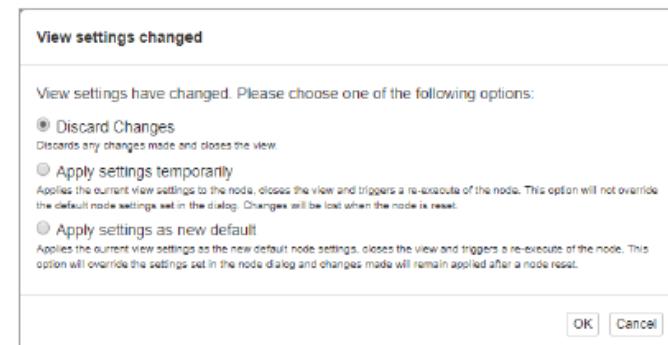
Después de seleccionar los puntos o cambiar los ajustes, si hacemos clic en el botón "Close" de la esquina inferior derecha, aparece una ventana que nos pregunta si queremos mantener los nuevos ajustes, es decir, los puntos seleccionados, de forma temporal o permanente. Con esta última opción prácticamente sobrescribimos los ajustes del nodo.

El último botón de la derecha permite hacer un paneo, es decir, hacer zoom y desplazarse por el gráfico.

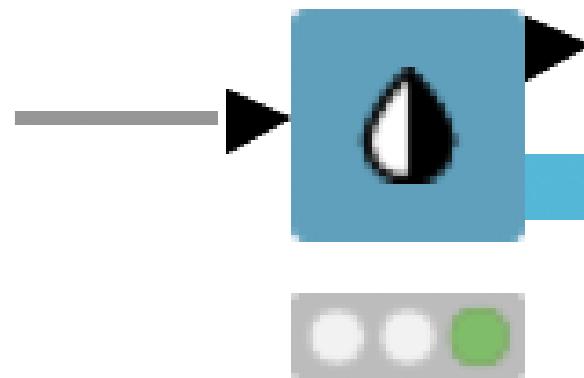
3.40. Vista del nodo "Scatter Plot"



3.41. Confirmación de los cambios tras pulsar el botón "Cerrar"

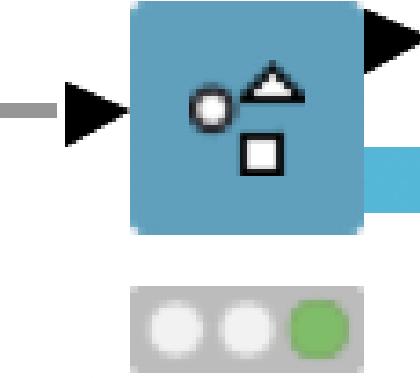


## **Color Manager**



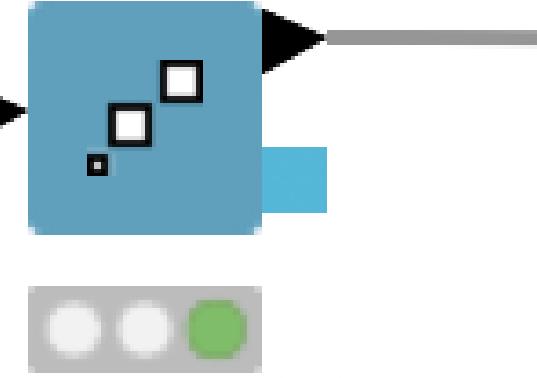
**set colors**

## **Shape Manager**



**set shapes**

## **Size Manager**



**set sizes**

# Gestor de color “Color Manager”

El nodo “Gestor de colores” asigna un color a cada fila de una tabla de datos en función de su valor en una columna determinada.

Si se selecciona una columna nominal en el diálogo de configuración, se asignan colores a cada uno de los valores nominales.

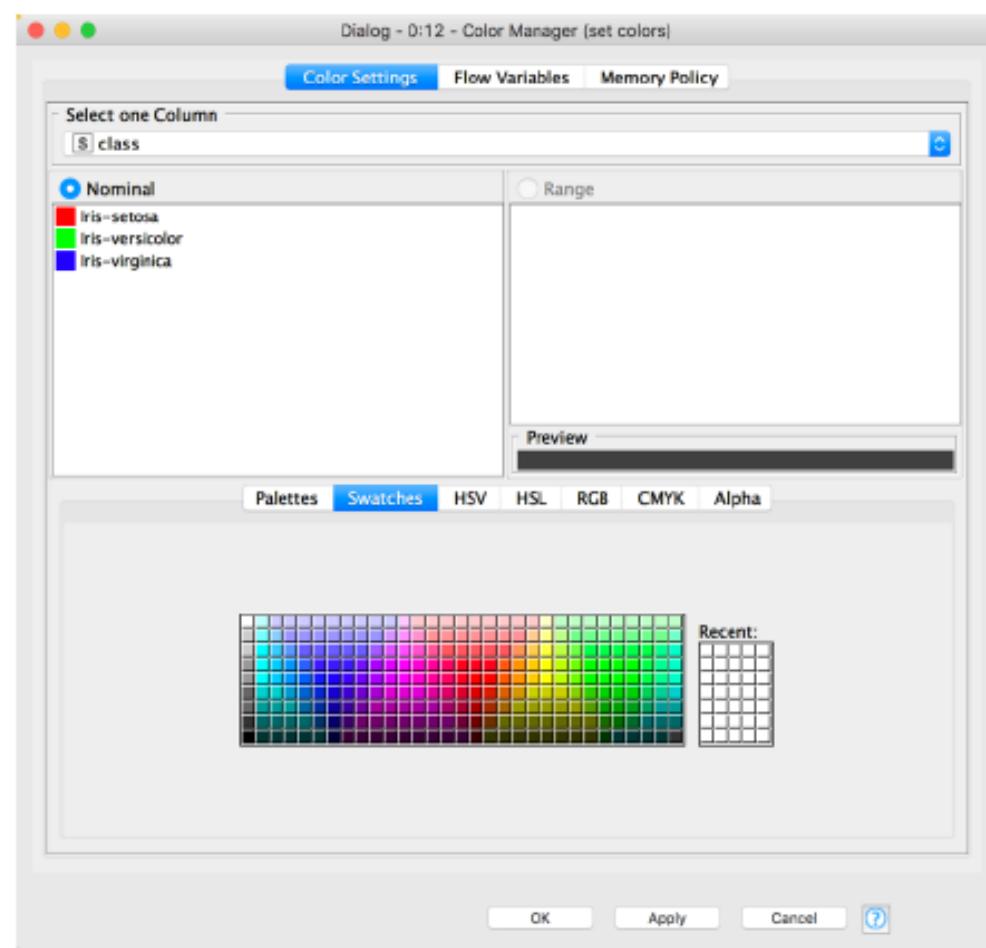
Si se selecciona una columna numérica, un mapa de calor de colores abarca el rango numérico de la columna.

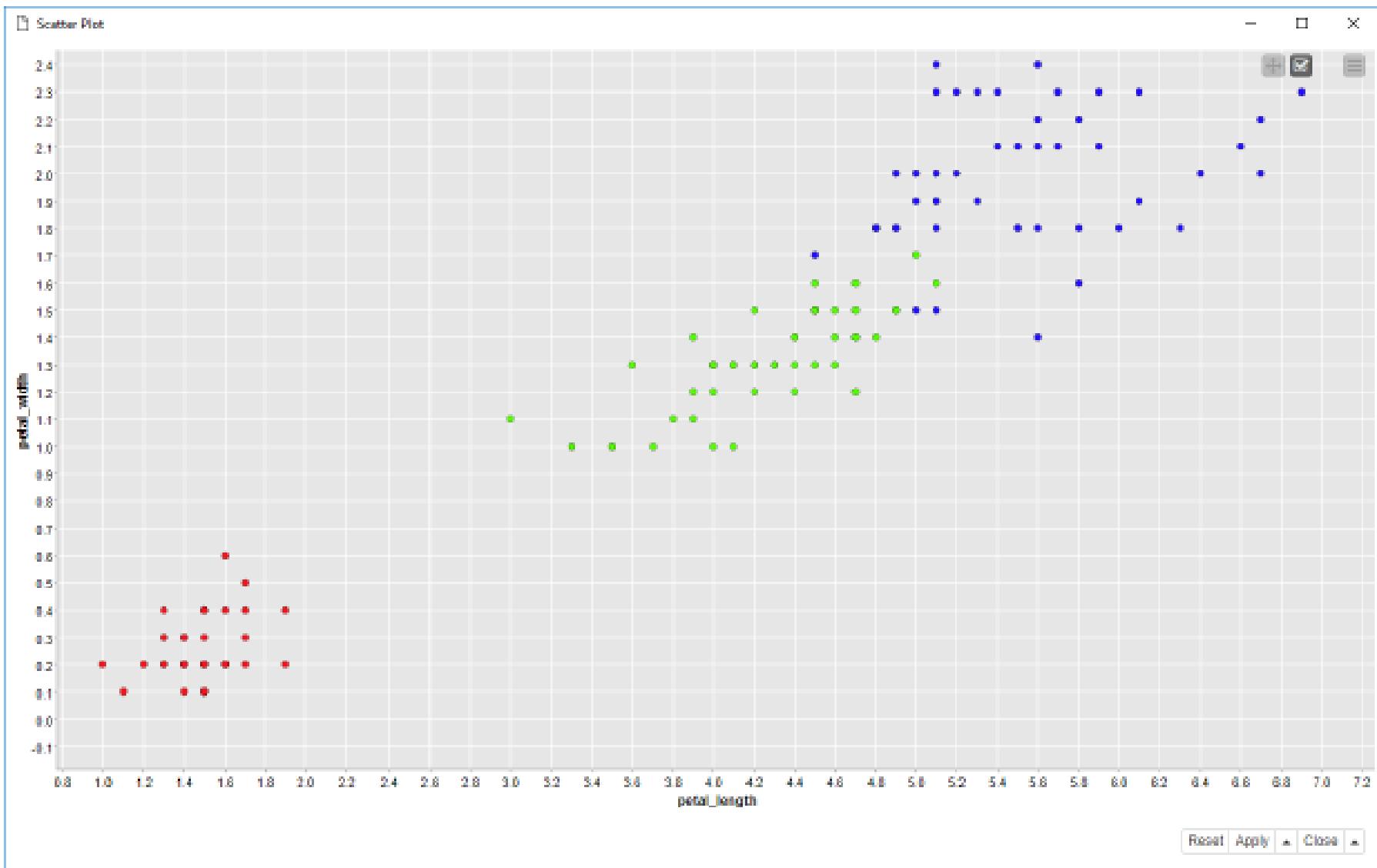
La ventana de configuración requiere:

- La columna de la que extraer valores (columnas nominales) o rangos (columnas numéricas)
- El mapa de colores para cada lista de valores o rango de valores

Se asigna por defecto un mapa de colores a la lista / rango de valores. Esto puede cambiarse seleccionando el valor / rango y luego asignando un color diferente del mapa de colores que aparece en la parte inferior de la ventana de configuración.

3.44. Ventana de configuración del nodo “Color Manager”





## Line Plot

El nodo "Line Plot" muestra un gráfico de líneas, utilizando una columna como eje X y uno o más valores de columna como eje Y.

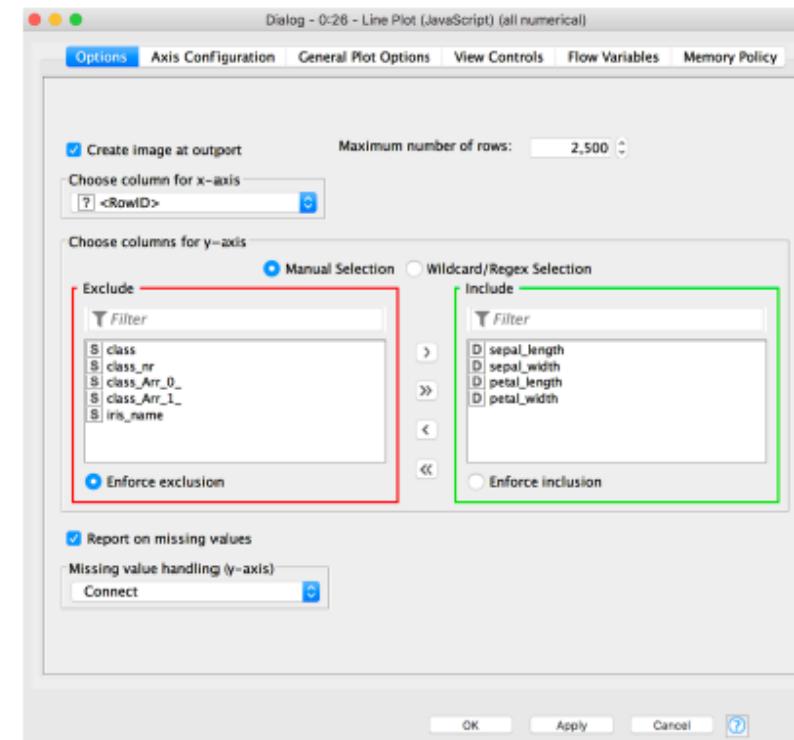
Al igual que en los anteriores nodos de visualización basados en Javascript, la ventana de configuración del nodo "Line Plot" tiene cuatro pestañas: "**Options**" para los datos; "**Axis Configuration**" y "**General Plot Options**" para los detalles del trazado; y "**View Controls**" para las características de interactividad.

La principal diferencia está en la pestaña "Options", y "Include"/"Exclude" en donde se permite seleccionar las columnas para el gráfico.

También hay disponibles varias estrategias de gestión de valores perdidos: ignorar el valor perdido y conectar los dos más cercanos; dejar un hueco vacío; o eliminar toda la columna si contiene valores perdidos.

A diferencia de otros nodos de visualización basados en JavaScript, el nodo "Line Plot" tiene un segundo puerto de entrada opcional para el esquema de colores. En este mapa de entrada los nombres de las columnas se asocian a los colores. En el gráfico final, cada columna se dibujará utilizando el color asociado en el mapa.

3.46. Ventana de configuración, nodo "Line Plot" node: "Options"



# Coordenadas paralelas (Parallel Coordinates)

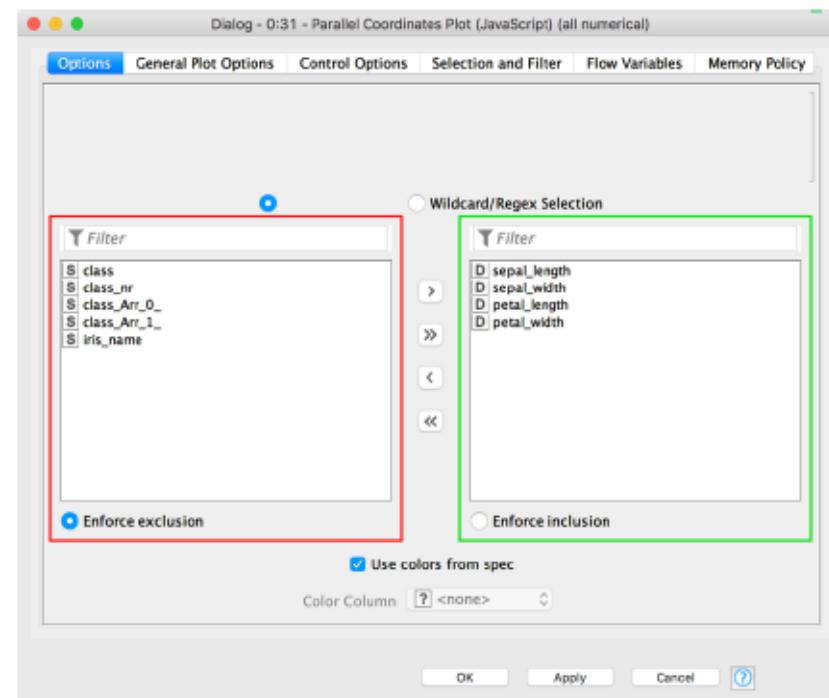
El nodo "Coordenadas paralelas" muestra la tabla de datos de entrada en un gráfico de coordenadas paralelas. Un gráfico de coordenadas paralelas despliega los nombres de las columnas a lo largo del eje X y muestra cada valor de la columna en un eje Y separado. Como resultado, un punto de datos se mapea como una línea que conecta los valores a través de los atributos.

La ventana de configuración de este nodo tiene tres pestañas.

- "**Options**" contiene un marco de "Exclude/Include" para insertar/remover mas columnas(por ejemplo. Y-axis) en/desde el gráfico de coordenadas paralelas.
- "**General Plot Options**" La pestaña define los ajustes generales para el trazado y la imagen de salida
- "**Control Options**" establece el nivel de interactividad de la vista final

Los colores de las líneas pueden proceder de una columna específica que contenga el color como propiedad gráfica (que es el resultado del nodo "Extraer color") o simplemente de la propiedad gráfica asociada a cada fila (flag "use color from spec").

3.48. Vista del nodo "Parallel Coordinates"



# Bar Chart

El nodo "Gráfico de barras" crea un gráfico de barras genérico. Para ello, necesita

- Una columna de categoría, que en el caso de un histograma es la columna de la jerarquía.
- Una columna de agregación y un método de agregación.

En el caso de un histograma, el método de agregación es el "Recuento de ocurrencias", que sólo cuenta las filas de datos que caen en cada casilla y, por lo tanto, no requiere una columna de agregación específica.

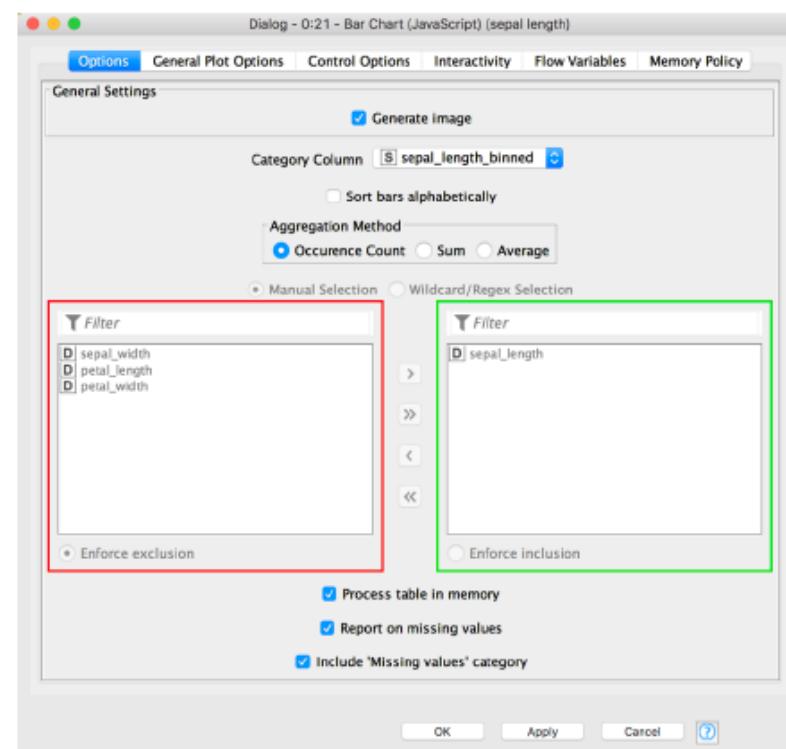
Todos estos ajustes se definen en la pestaña "Opciones" de la ventana de configuración. Dos pestañas adicionales "Opciones generales de ploteo" y "Opciones de control" definen respectivamente los detalles gráficos del ploteo y los controles de vista habilitados.

La pestaña "Opciones generales de ploteo" incluye las preferencias para el título, las etiquetas de los ejes, la orientación del ploteo, la leyenda y el tamaño de la imagen de salida.

La pestaña "Opciones de control" incluye el zoom, el cambio de orientación del gráfico, la edición de títulos y etiquetas, el apilamiento/agrupamiento de barras y el apilamiento de etiquetas.

El nodo "Gráfico de barras" tiene un puerto de entrada opcional para un mapa de colores.

3.50. Ventana de configuración del nodo "Bar Chart": Pestaña "Opciones" configurada para dibujar un histograma



# MODELADO EN KNIME

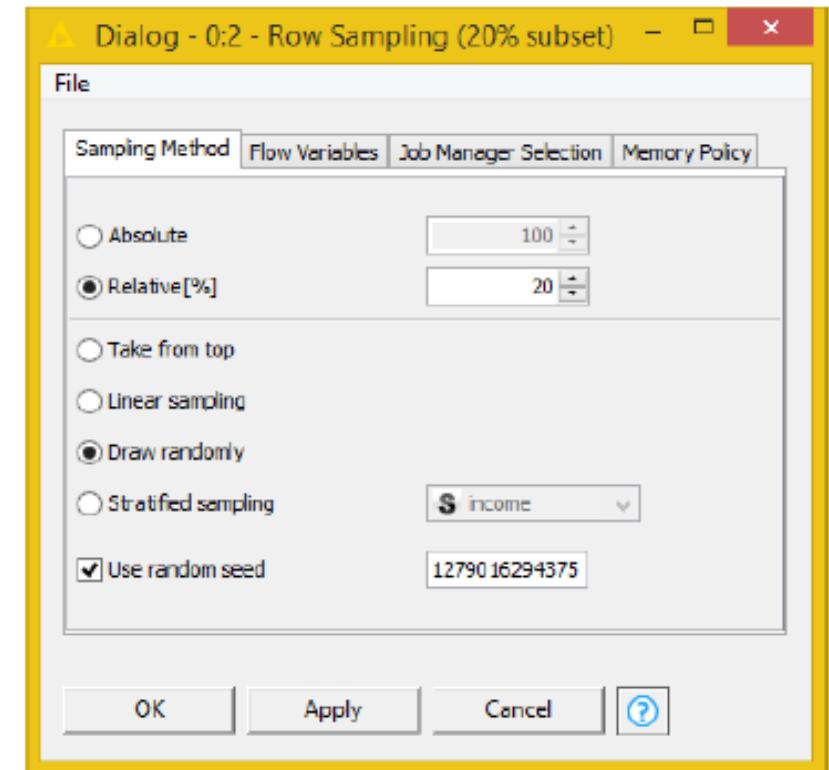
# Muestreo por filas (Row Sampling)

El nodo “Row Sampling” extrae una muestra (= un subconjunto de filas) de los datos de entrada. La ventana de configuración le permite especificar:

- El tamaño de la muestra como número absoluto de filas o como porcentaje del conjunto de datos original
- El modo de extracción
  - o “Take from the top” significa las filas superiores del conjunto de datos original
  - o “Linear Sampling” toma la primera y la última fila y muestrea entre estas filas en pasos regulares
  - o “Draw randomly” extrae filas al azar
  - o “Stratified sampling” extrae filas de forma aleatoria por lo que la distribución de los valores en la columna seleccionada se mantiene aproximadamente en la tabla de salida

Para “Draw randomly” y “Stratified sampling” se puede definir una semilla aleatoria para que la extracción aleatoria sea reproducible (nunca se sabe cuándo se necesita recrear exactamente el mismo conjunto de entrenamiento aleatorio).

4.1. Ventana de configuración del nodo “Row Sampling”



# Partitioning

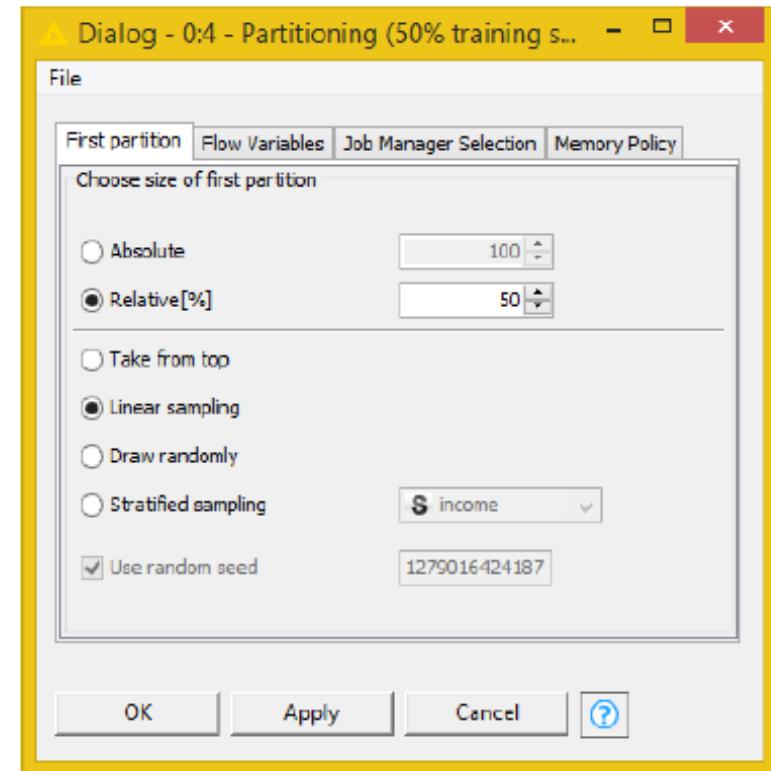
El nodo "Partitioning" realiza la misma tarea que el nodo "Row Sampling": extrae una muestra (= un subconjunto de filas) de los datos de entrada. También construye un segundo conjunto de datos con las filas restantes y lo pone a disposición en el puerto de salida inferior.

La ventana de configuración permite especificar

- El tamaño de la muestra como número absoluto de filas o como porcentaje del conjunto de datos original
- El modo de extracción
  - o "Take from the top" se refiere a las primeras filas del conjunto de datos original
  - o "Linear sampling" toma la primera y la última fila y muestrea entre las filas a pasos regulares
  - o "Draw randomly" extrae filas al azar
  - o El "Stratified sampling" extrae filas en las que la distribución de los valores de la columna seleccionada se mantiene aproximadamente en la tabla de salida

Para "Draw randomly" y "Stratified sampling" puede definirse una semilla aleatoria para que la extracción aleatoria sea reproducible (nunca se sabe cuándo hay que volver a crear el mismo conjunto de aprendizaje).

4.2. Ventana de configuración del nodo "Partitioning"



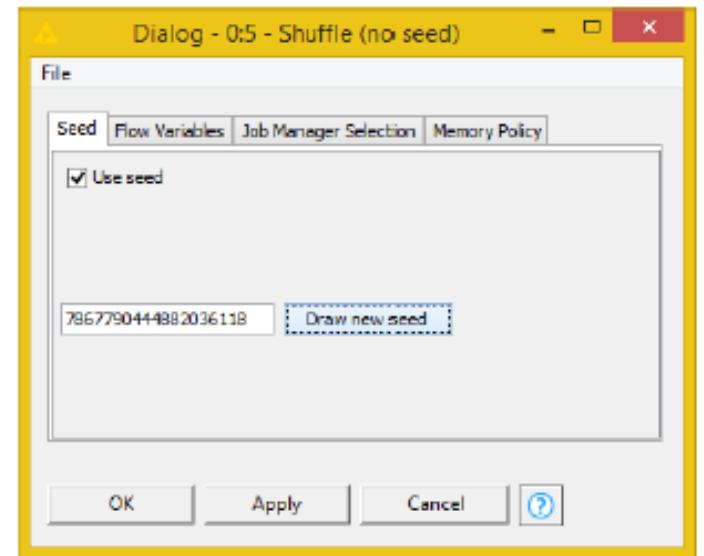
# Shuffle

El nodo "Shuffle" baraja las filas de la tabla de entrada poniéndolas en un orden aleatorio.

En general, el nodo "Shuffle" no necesita ser configurado. Si queremos ser capaces de repetir exactamente el mismo barajado aleatorio de las filas, necesitamos usar una semilla, como sigue:

- Marque la casilla "Use seed".
- Haga clic en el botón "Draw new seed" para crear una semilla para el barajado aleatorio y volver a crearla en cada ejecución

4.3. Configuration window for the "Shuffle" node



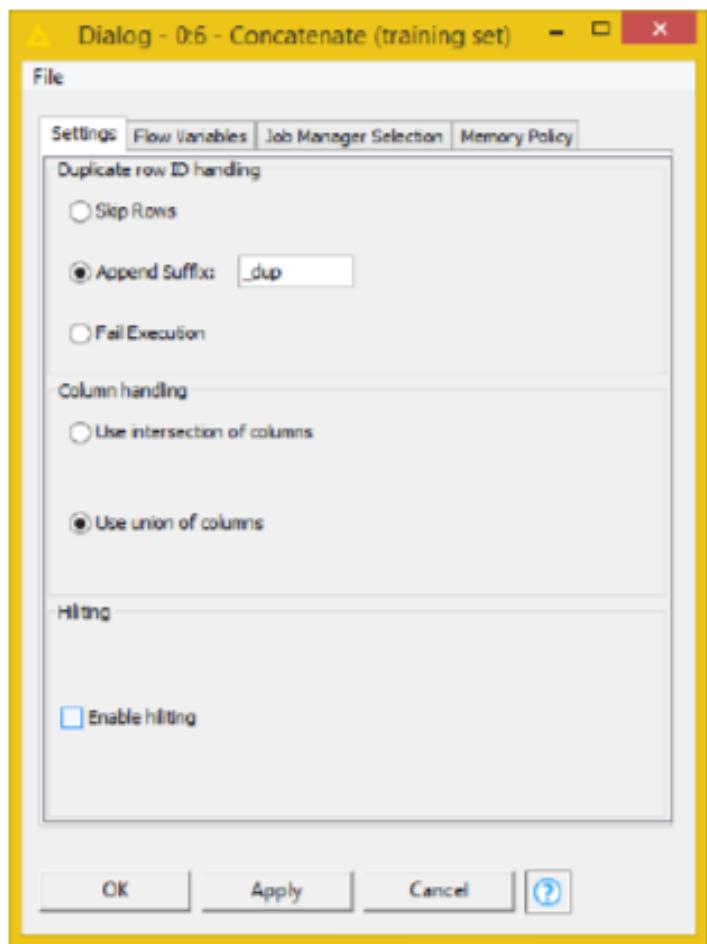
# Concaternar (Concatenate)

El nodo "Concatenate" tiene dos puertos de entrada, cada uno para un conjunto de datos. El nodo "Concatenate" añade el conjunto de datos del puerto de entrada inferior al conjunto de datos del puerto de entrada superior.

La ventana de configuración se ocupa de lo siguiente

- Qué hacer con las filas con el mismo ID
  - o omitir las filas del segundo conjunto de datos
  - o renombrar el RowID con un sufijo añadido
  - o abortar la ejecución con un error (Esta opción puede utilizarse para comprobar si los RowIDs son únicos)
- Qué columnas conservar
  - o todas las columnas del segundo y primer conjunto de datos (unión de columnas)
  - o sólo la intersección de las columnas de los dos conjuntos de datos (es decir, las columnas contenidas en ambas tablas)
- La opción "Enable hiliting" hace referencia a la propiedad hiliting disponible en los antiguos nodos "Data Views".

4.4. Ventana de configuración del nodo "Concatenate"



# Valores faltantes (Missing Value)

El nodo "Missing Value" sustituye los valores ausentes en un conjunto de datos en todas partes o sólo en las columnas seleccionadas por un valor de su elección.

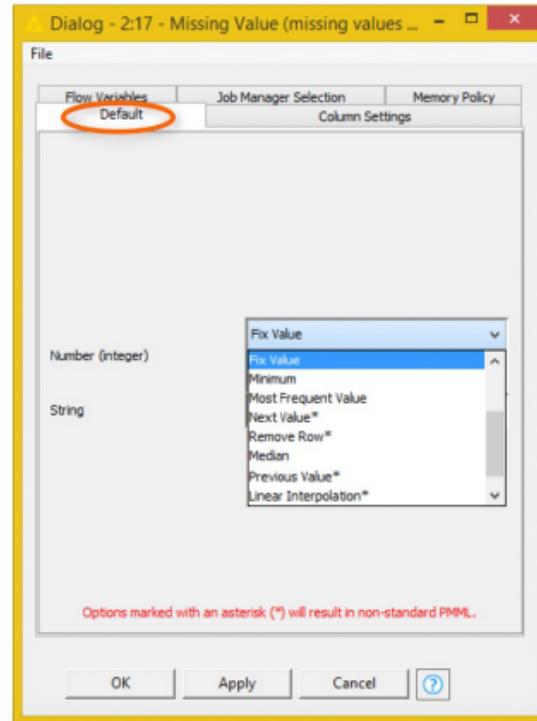
En la pestaña "**Default**", los valores de sustitución se definen por separado para las columnas de tipo numérico y de cadena y se aplican a todas las columnas de datos del mismo tipo.

En la pestaña "**Column Settings**", se define un valor de sustitución específicamente para cada columna de datos seleccionada y se aplica sólo a esa columna. Para definir el valor de sustitución de una columna:

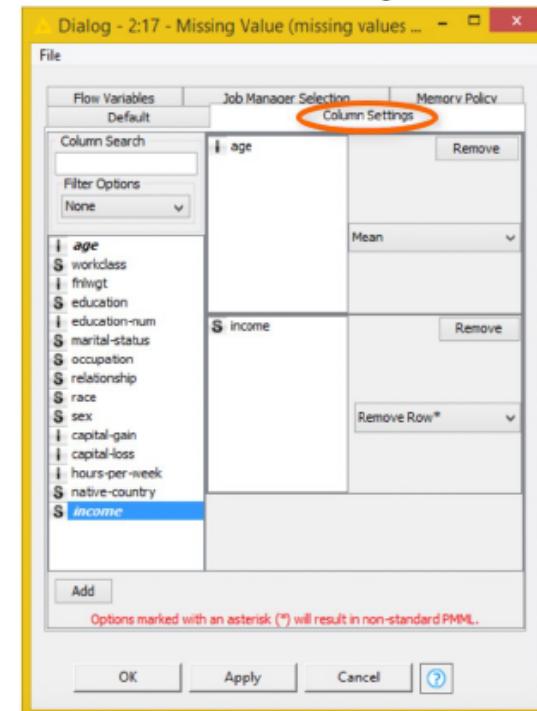
- aga doble clic en la columna de la lista de la izquierda
- ó
- Seleccione la columna en la lista de la izquierda
- Haga clic en el botón "Add" debajo de la lista

A continuación, seleccione la estrategia de tratamiento de los valores perdidos que deseé.

4.7. Configuración para la ventana "Missing Value"  
: Tab "Default"



4.8. Configuración para la ventana "Missing Value"  
: Tab "Column Settings"



# Normalizer (normalizador)

El nodo "Normalizer" normaliza los datos; es decir, transforma los datos para que caigan en un intervalo determinado o para que sigan una distribución estadística determinada.

El nodo "Normalizer" se encuentra en el panel "Node Repository" en la categoría "Manipulation" → "Column" → "Transform"

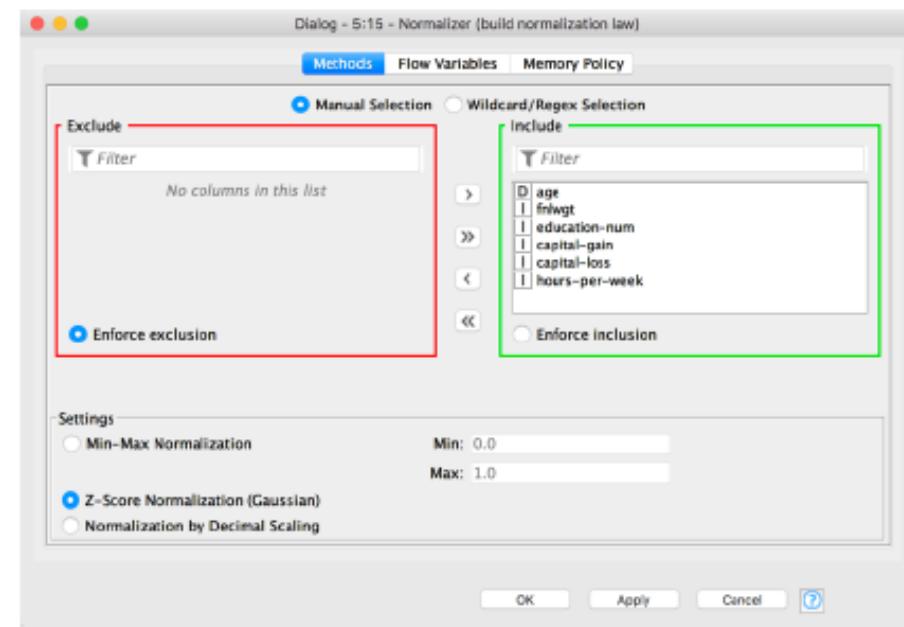
La ventana de configuración requiere:

- la lista de columnas de datos numéricos a normalizar
- el método de normalización

La selección de la columna se realiza mediante un cuadro "Exclude"/"Include", por selección manual o por selección Wildcard/RegEx. Para la selección manual:

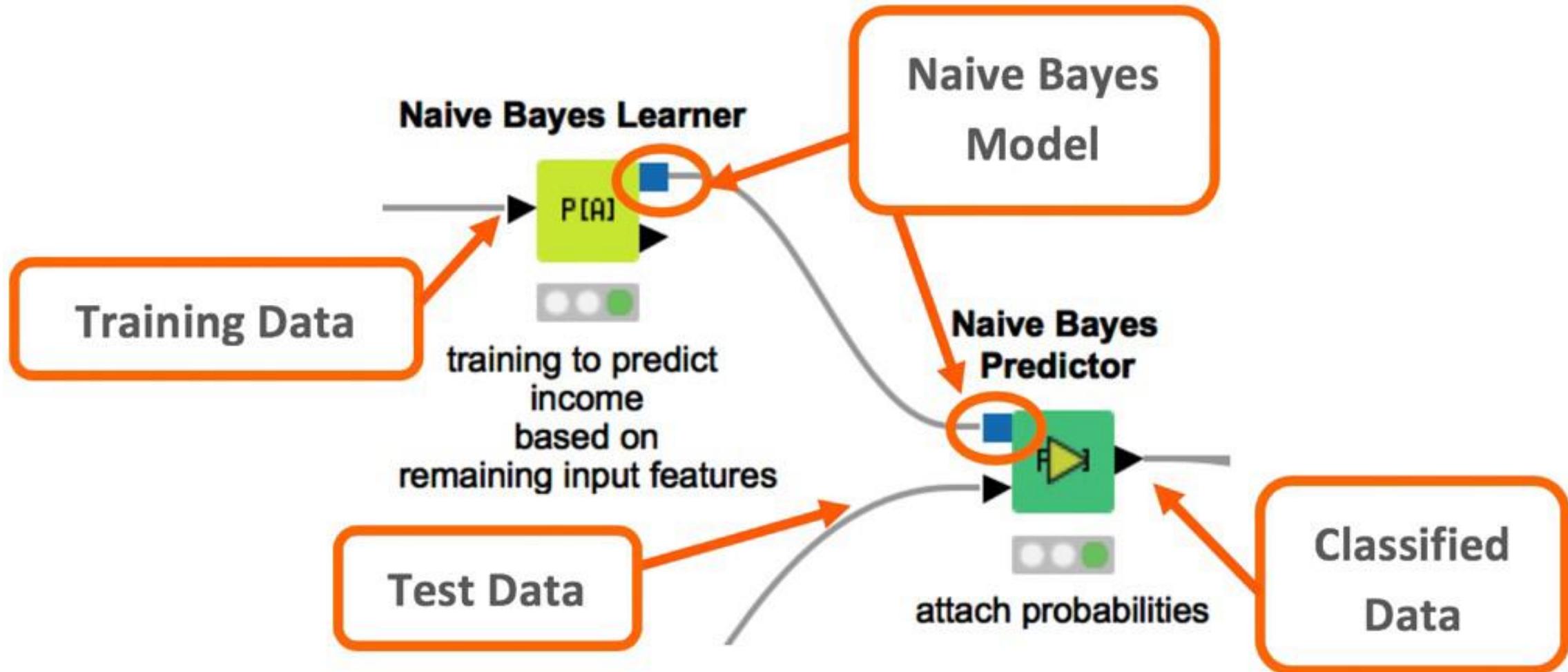
- Las columnas que se van a normalizar aparecen en el cuadro "Normalize". Todas las demás columnas aparecen en el marco "No normalizar".
- Para pasar del marco "Normalize" al marco "Do not normalize" y viceversa, utilice los botones "add" y "remove". Para mover todas las columnas a un marco u otro, utilice los botones "add all" y "remove all".

4.9. Configuración del nodo Normalizer<sup>6</sup>



El nodo "Normalizer" tiene 2 puertos de salida:

- En el puerto superior se encuentran los datos normalizados
- En el puerto inferior se proporcionan los parámetros de transformación para repetir la misma normalización en otros datos (puerto cuadrado azul claro/azul oscuro)



## Scorer (Javascript)

El nodo "Scorer" compara los valores de dos columnas (columna de destino y columna de predicción) en la tabla de datos; basándose en esta comparación muestra la matriz de confusión y algunas medidas de precisión.

Este nodo produce tres tablas de datos de salida: la matriz de confusión, las estadísticas de filas correctamente identificadas para cada clase, y las medidas de precisión globales establecidas en la ventana de configuración.

Este nodo tiene una opción de Vista, donde se muestra la matriz de confusión y algunas medidas de precisión.

La ventana de configuración tiene tres pestañas: "Scorer Options", "Statistics Options", "Control Options".

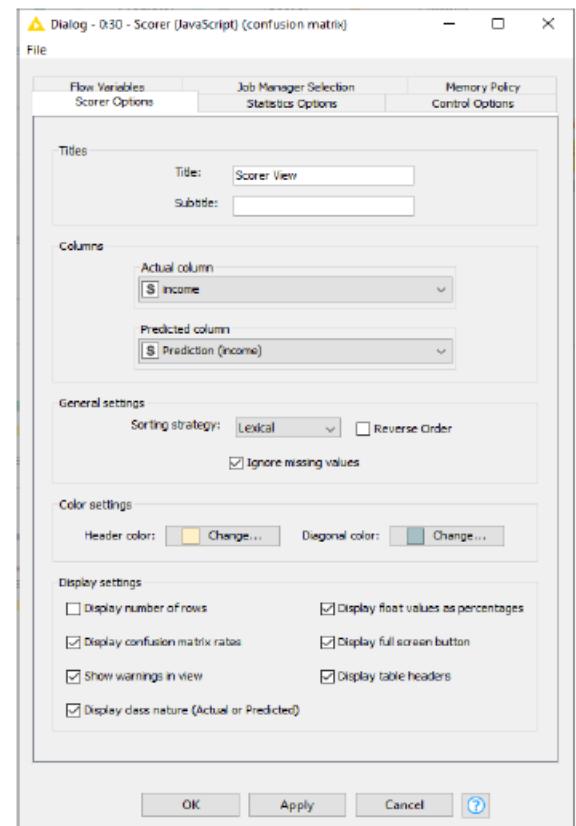
La ventana de configuración tiene tres pestañas: "Opciones del calificador", "Opciones de estadísticas", "Opciones de control".

La pestaña "Scorer Options" requiere la selección de las dos columnas a comparar ("Actual Column" y "Predicted Column") y la ordenación que se utilizará en la estrategia de evaluación. La bandera "Ignorar valores perdidos", si no está marcada, hace que el nodo falle si se encuentran valores perdidos en una de las dos columnas a comparar. Todas las demás opciones están relacionadas con la visualización de la vista del nodo.

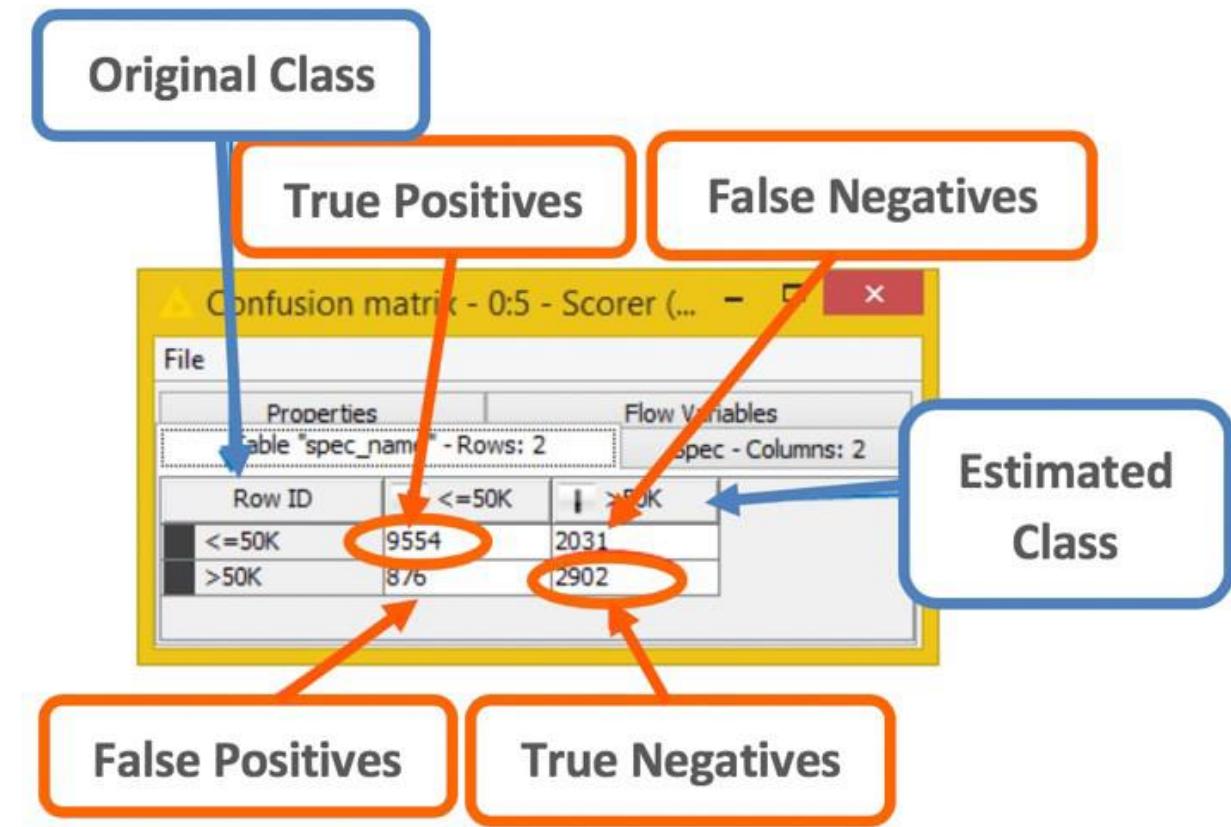
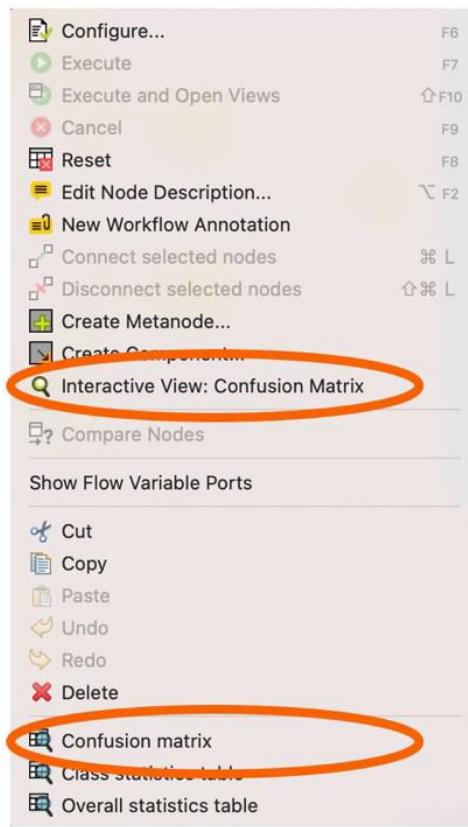
La pestaña "Opciones de estadísticas" incluye todas las medidas de precisión y los números falsos/verdaderos positivos/negativos a calcular.

Como todos los nodos basados en JavaScript, este nodo produce una vista con cierto grado de interactividad. Las opciones de interactividad se definen en la pestaña "Opciones de control".

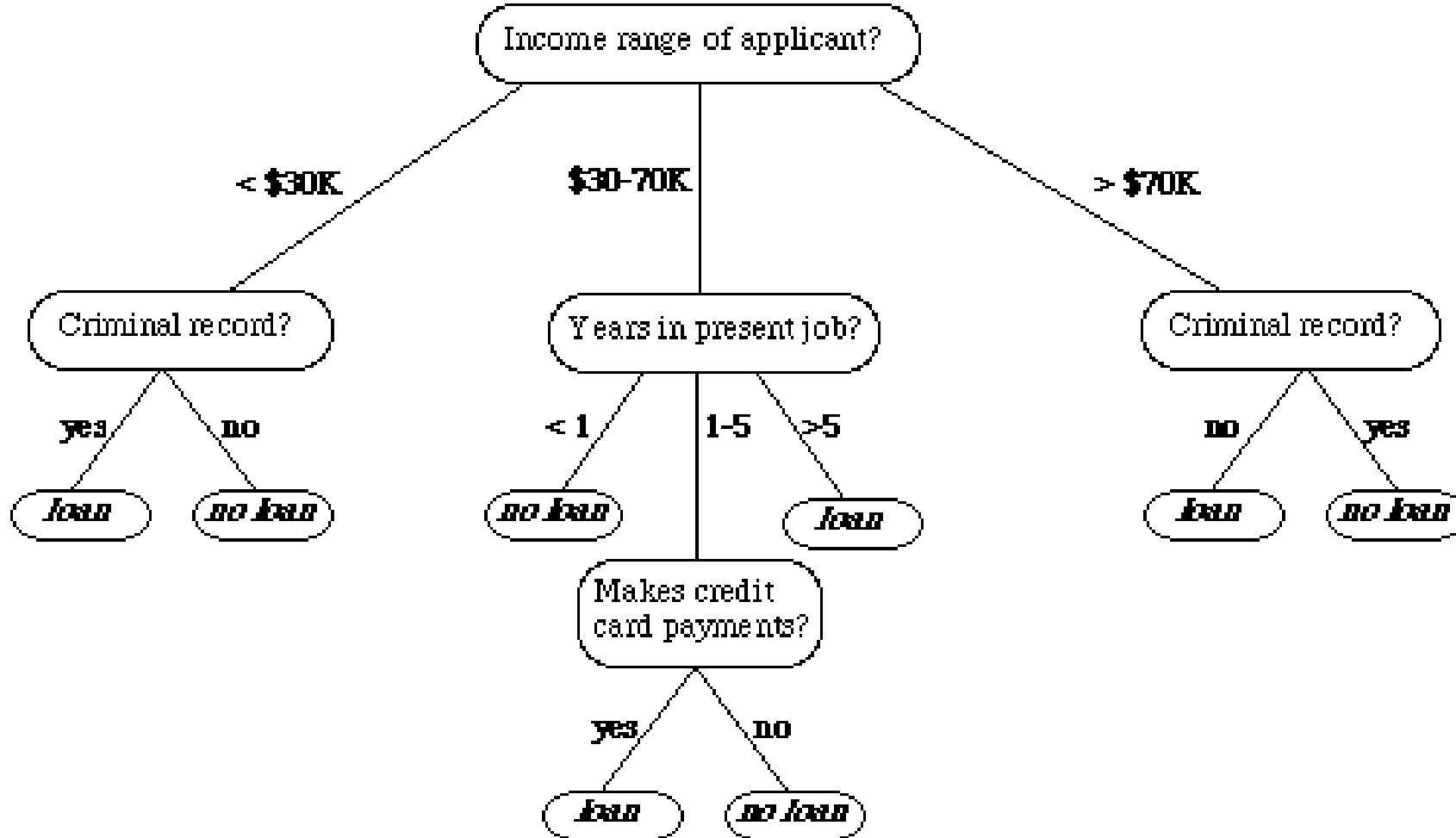
4.15. Ventana de configuración del nodo "Scorer (Javascript)"



# MATRIZ DE CONFUSIÓN



Matriz de confusión		Estimado por el modelo			
		Negativo (N)	Positivo (P)		
Real	Negativo	a: (TN)	b: (FP)		
	Positivo	c: (FN)	d: (TP)	Precisión ("precision") Porcentaje predicciones positivas correctas:	d/(b+d)
		Sensibilidad, exhaustividad ("Recall") Porcentaje casos positivos detectados	Especificidad (Specificity) Porcentaje casos negativos detectados	Exactitud ("accuracy") Porcentaje de predicciones correctas <i>(No sirve en datasets poco equilibrados)</i>	
		d/(d+c)	a/(a+b)	(a+d)/(a+b+c+d)	



# ARBOLES DE DECISIÓN

## Pestaña del nodo Decision Tree Learner: "Options"

El nodo "Decision Tree Learner" construye un árbol de decisión a partir de los datos de entrenamiento de entrada. En la ventana de configuración hay que especificar:

### General

La *class column*. El atributo de destino debe ser nominal (String).

La *quality measure* para el cálculo de la división: "Índice de Gini" o "Ratio de ganancia".

El *pruning method*: "No Pruning" o una poda basada en el principio de "Minimum Description Length (MDL)" [8] [9]. La opción "Reduced Error Pruning", si está marcada, aplica una poda simple de posprocesamiento.

El *stopping criterion*: el número mínimo de registros en cada nodo del árbol de decisión. Si un nodo tiene menos registros que este número mínimo, el algoritmo detiene la división de esta rama. Cuanto mayor sea el número, menos profundo será el árbol.

El número de registros a almacenar para la vista: el número máximo de filas a almacenar para la funcionalidad de la "hilite". Un número elevado ralentiza la ejecución del algoritmo.

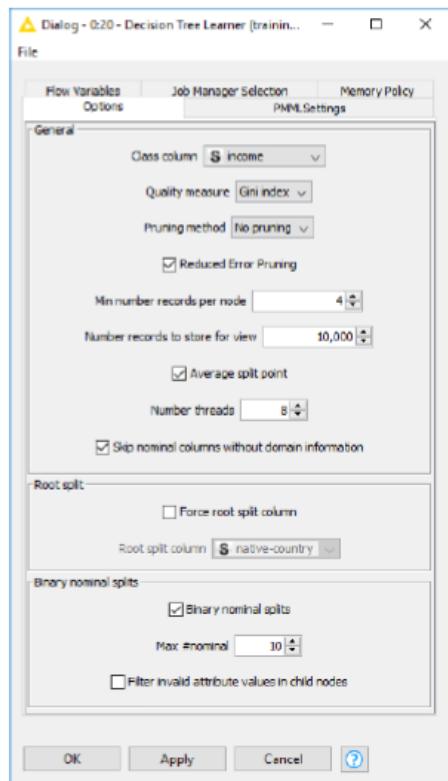
La casilla "Average Split Point". Para los atributos numéricos, el usuario debe elegir una de las dos estrategias de división:

- El punto de división se calcula como el valor medio entre los atributos de las dos particiones (casilla "Average Split Point" activada)
- El punto de división se fija en el mayor valor de la partición inferior (indicador "Average Split Point" desactivado)

El número de hilos en los que se ejecutará el nodo (número de hilos por defecto = 2 \* número de procesadores disponibles para KNIME).

**Root Split**: Si sabe que un atributo debe ser importante para la clasificación, puede forzarlo en el nodo raíz del árbol, activando "Force root split column" y seleccionando la "Root split column".

4.21. Pestaña del nodo Decision Tree Learner:  
"Options"



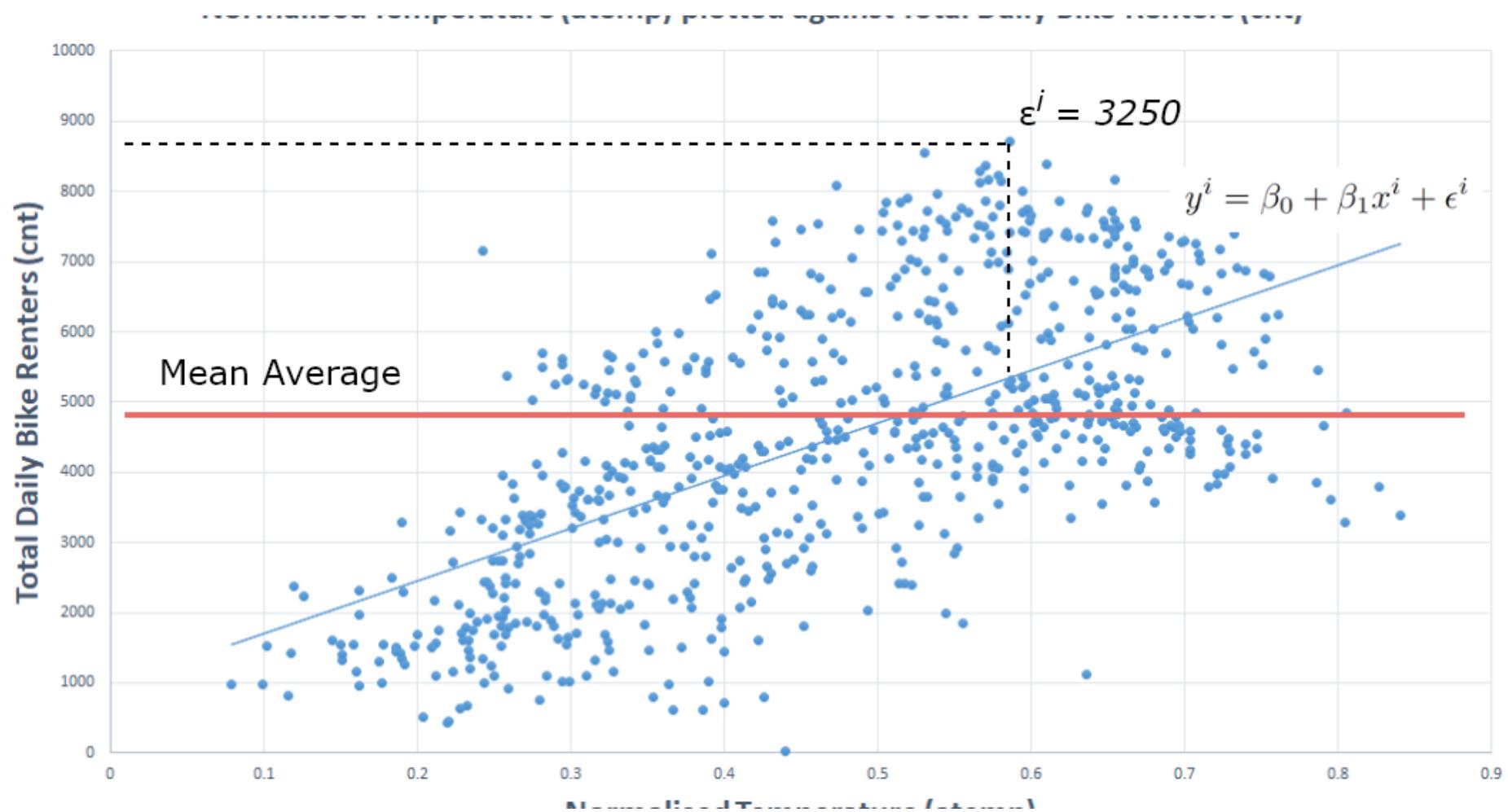
## Normalized class distributions

## Final prediction

Classified Data - 0:16 - Decision Tree Predictor (attach class)

File Table "test\_set.csv" - Rows: 15363 Spec - Columns: 18 Properties Flow Variables

Row ID	Urs-p...	S native-...	S income	D P (income=<=50K)	D P (income=>50K)	S Prediction (income)
Row0		United-States	<=50K	0.333	0.667	>50K
Row1		United-States	<=50K	0.962	0.038	<=50K
Row2		United-States	<=50K	0.132	0.868	>50K
Row3		United-States	>50K	1	0	<=50K
Row4		United-States	>50K	0	1	>50K
Row5		India	>50K	0.853	0.147	<=50K
Row6		United-States	<=50K	0.956	0.044	<=50K
Row7		Mexico	<=50K	0.958	0.042	<=50K
Row8		United-States	<=50K	0.996	0.004	<=50K
Row9		United-States	>50K	0.889	0.111	<=50K
Row10		United-States	<=50K	0.996	0.004	<=50K
Row11		United-States	<=50K	1	0	<=50K
Row12		United-States	>50K	n	1	>50K



# REGRESIÓN LINEAL

## Nodo “Linear Regression Learner”

El nodo “Linear Regression Learner” realiza una regresión lineal multivariante sobre una columna objetivo, es decir, la respuesta.

El nodo “Linear Regression (Learner)” se encuentra en: “Node Repository” “Analytics” → “Mining” → “Regression”.

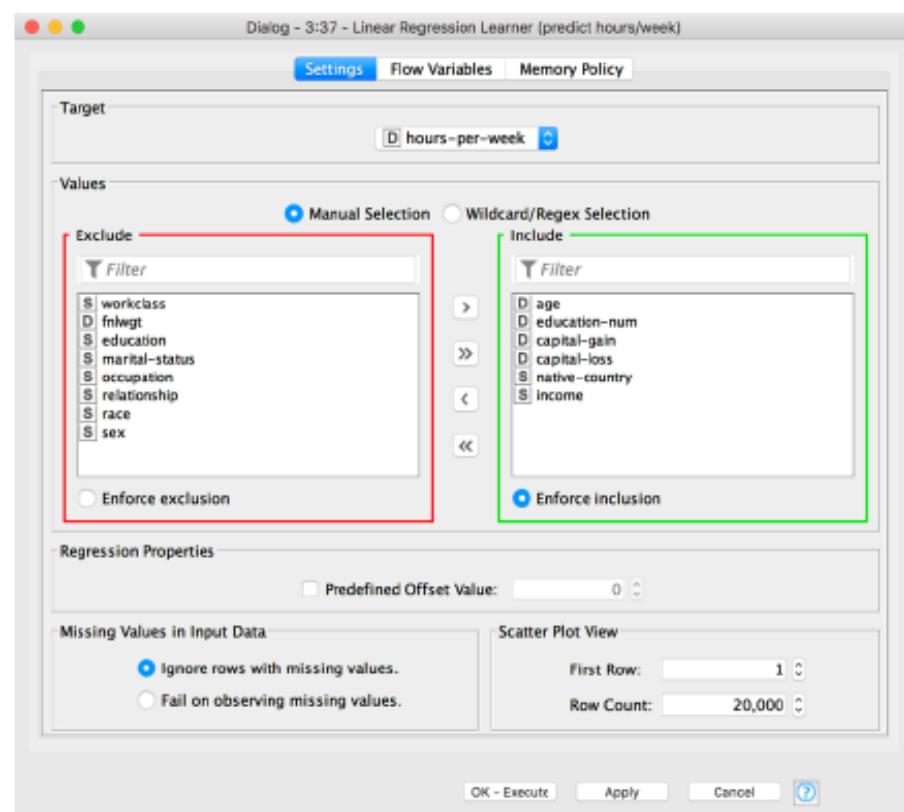
En la ventana de configuración hay que especificar:

- La columna para la que se calcula la regresión
- Las columnas que se utilizarán como variables independientes en la regresión lineal
- El número de la fila inicial y el número de filas que se visualizarán en la vista del gráfico de dispersión del nodo
- La estrategia de gestión de los valores perdidos
- Un valor de desplazamiento por defecto a utilizar (si lo hay)

La selección de las columnas de datos de entrada se realiza mediante el marco de selección de columnas: por selección manual con “Include/Exclude” ; por selección de tipo, por selección de expresión Wildcard/Regex.

El nodo produce el modelo de regresión, así como los coeficientes y estadísticas del modelo.

4.43. Configuration window for the “Linear Regression Learner” node



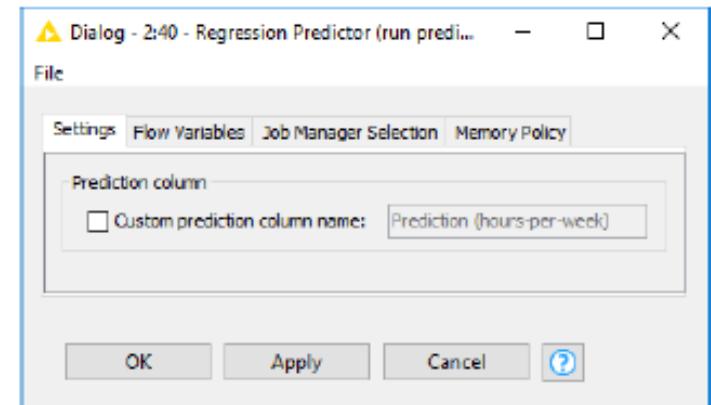
## Nodo “Regression Predictor”

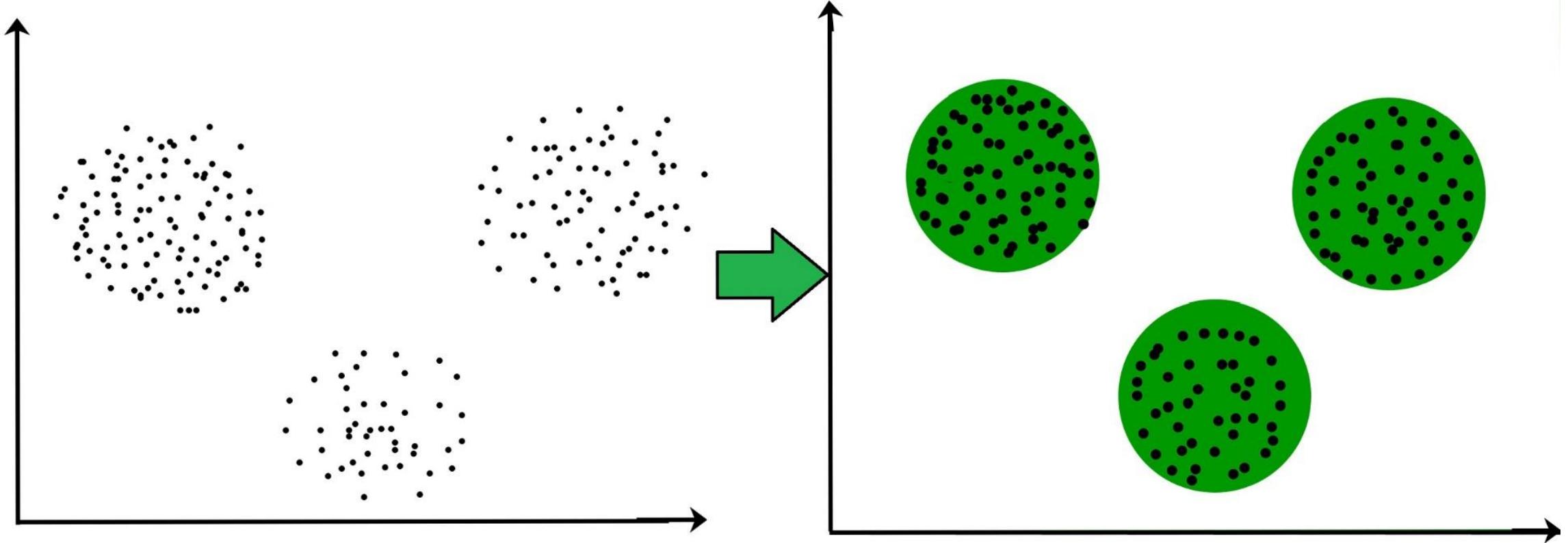
El nodo “Predictor de regresión” obtiene un modelo de regresión de uno de sus puertos de entrada (cuadrado azul) y datos del otro puerto de entrada (triángulo negro). Utiliza el modelo y los datos para hacer una predicción basada en los datos.

Como toda la información ya está disponible en el modelo, este nodo sólo necesita los ajustes mínimos del predictor: un nombre alternativo personalizado para la columna de clasificación de salida.

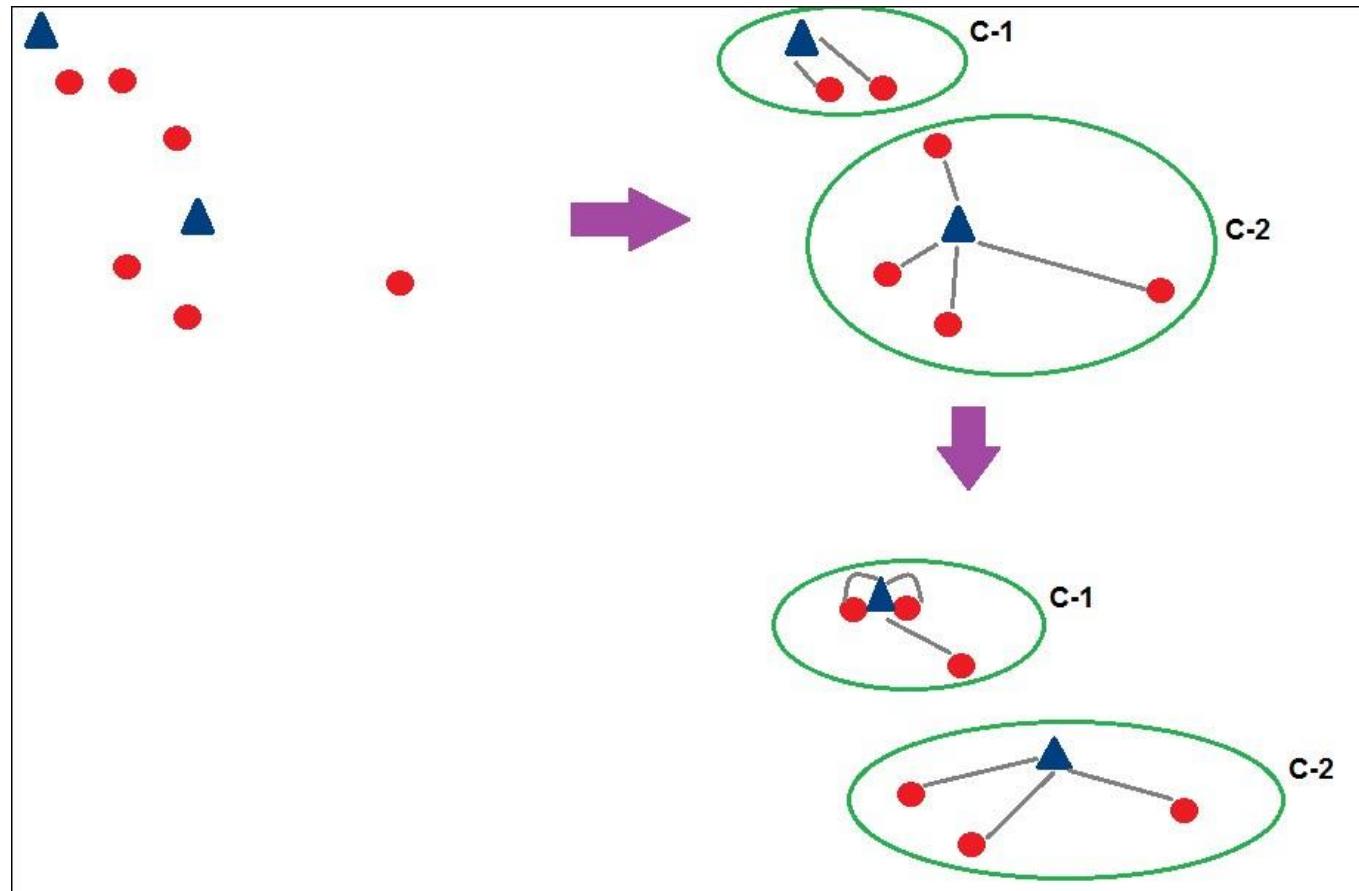
El nodo “Regression Predictor” se encuentra en la sección “Analytics” → “Mining” → “Regression” del panel “Node Repository”

4.44. Configuration window for the “Regression Predictor”

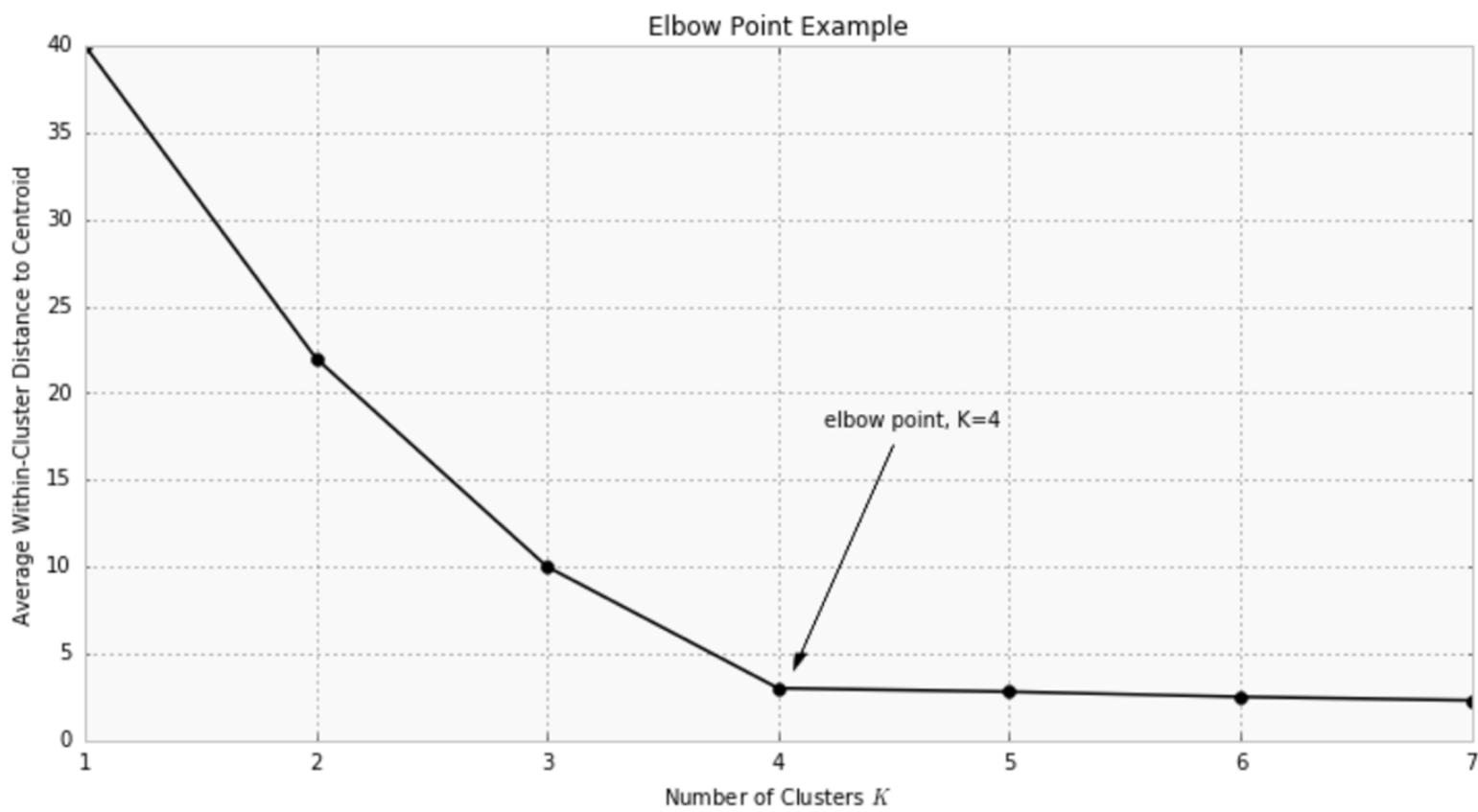




CLUSTERING



K-MEDIAS



# ELBOW METHOD

## k-Means

El nodo "k-Means" agrupa los patrones de entrada en k clusters sobre la base de un criterio de distancia y calcula sus prototipos. Los prototipos se construyen como el valor medio de los patrones de los clusters. Este nodo toma los datos de entrenamiento en el puerto de entrada y presenta el modelo en el puerto de salida cuadrado azul y los datos de entrenamiento con la asignación de clusters en el puerto de salida de datos (triángulo negro).

El nodo "k-Means" se encuentra en el "Node Repository" en "Analytics" → "Mining" → "Clustering".

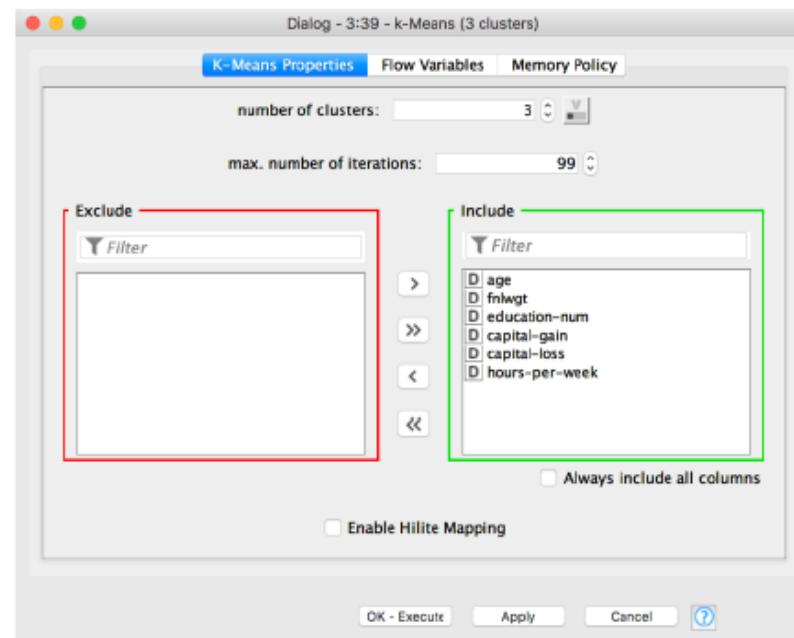
In the configuration window you need to specify:

- El número final de clusters k
- El número máximo de iteraciones para garantizar que la operación de aprendizaje converge en un tiempo razonable
- Las columnas que se utilizarán para calcular la distancia y los prototipos
- casilla "Always include all columns" es una alternativa al marco de selección de columnas.

La selección de columnas se realiza mediante un cuadro "Exclude"/"Include"

- Las columnas que se utilizarán para el cálculo de la distancia se enumeran en el marco "Include". Todas las demás columnas se enumeran en el marco "Exclude"
- Para pasar del marco "Include" al marco "Exclude" y viceversa, utilice los botones "add" y "remove". Para mover todas las

4.45. Ventana de configuración del nodo "k-Means"



## **Nodo “Cluster Assigner”**

El nodo "Cluster Assigner" asigna los datos de la prueba a un conjunto existente de prototipos que han sido calculados por un nodo de clustering como el nodo "k-Means". Cada fila de datos se asigna a su prototipo más cercano.

El nodo toma un modelo de clustering y un conjunto de datos como entradas y produce una copia del conjunto de datos con una columna adicional que contiene las asignaciones de cluster.

El nodo "Cluster Assigner" se encuentra en la categoría "Analytics" → "Mining" → "Clustering" en el panel "Node Repository".

No necesita ningún ajuste de configuración específico para su tarea de asignación de clústeres.

# EJERCICIO 1

Compara las ventas entre estados de EE.UU. en un mapa coroplético

- Ejecute el nodo Lector CSV para cargar los datos de ventas
- Utilice un nodo GroupBy para calcular las ventas por estado
- Visualiza el resultados usando un mapa coroplético



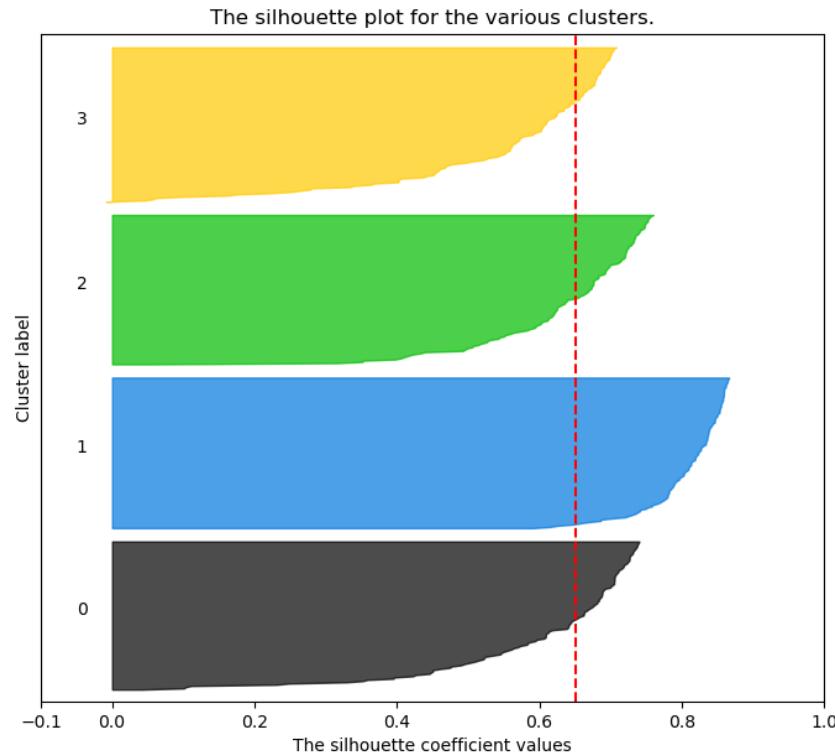
## EJERCICIO 2

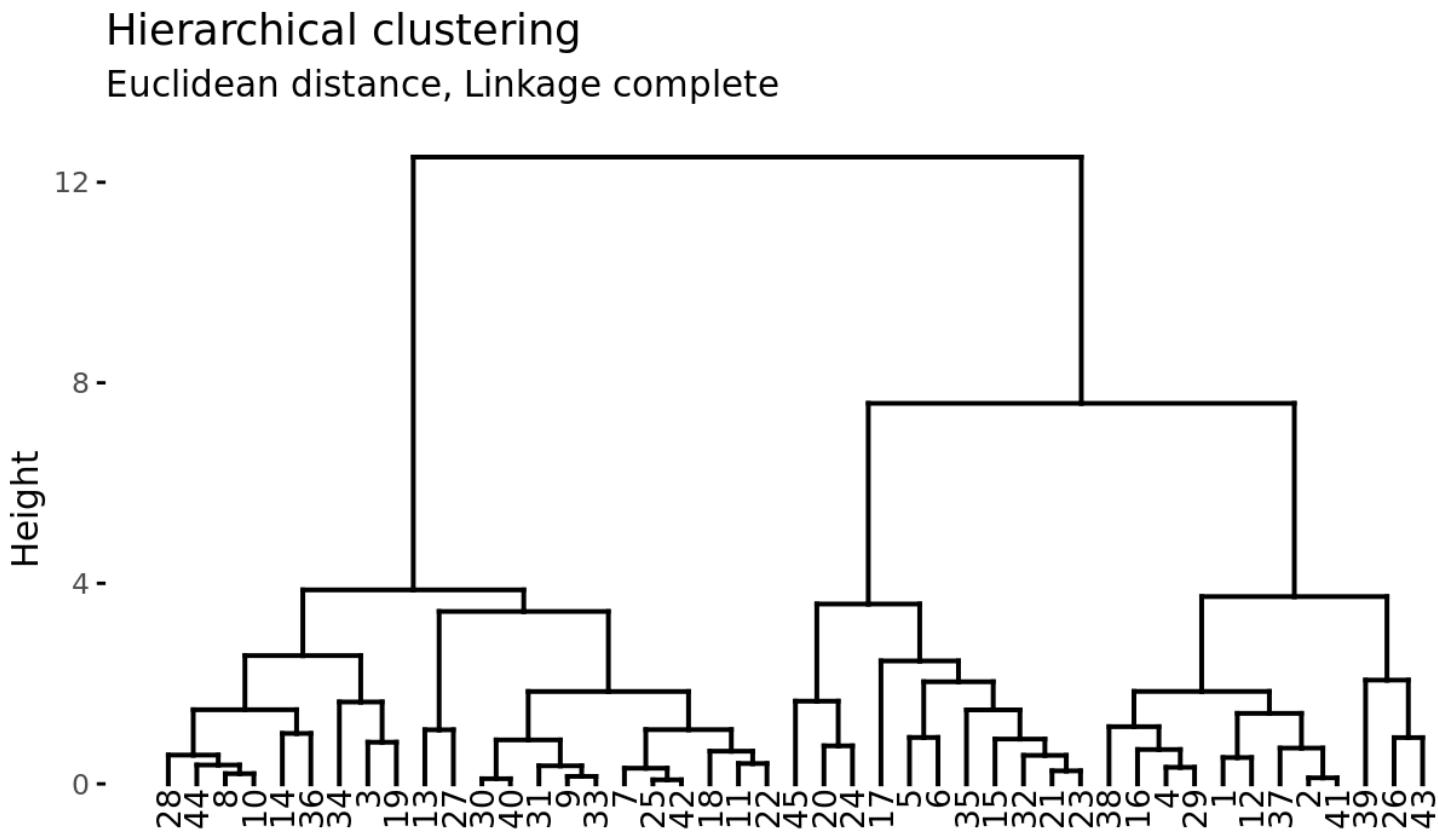
### Árbol de decisión

- Colorea las filas por la columna “class”
- Representa con un Scatter plot: feature 0 vs feature 1
- Representa con un Scatter plot: feature 3 vs feature 1
- Particiona los datos entre entrenamiento y test
- Construye un árbol de decisión para predecir la columna de clase optimizando el índice de Gini
- Construye ahora otro árbol optimizando el ratio de ganancia
- Obtén las matrices de confusión y estadísticos de ambos modelos y comenta cual es mejor

# COEFICIENTE SILHOUETTE

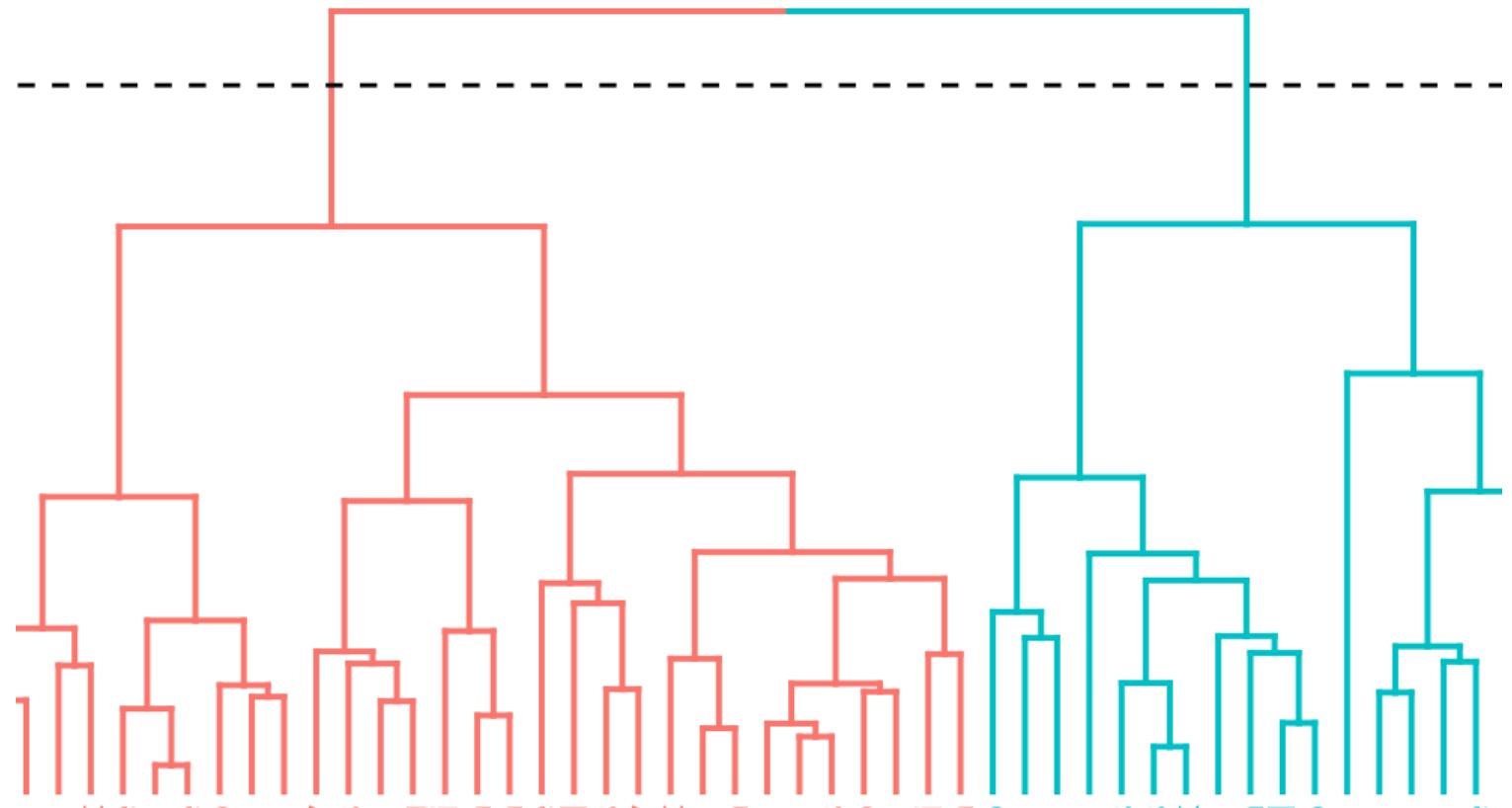
**Silhouette analysis for KMeans clustering on sample data with n\_clusters = 4**



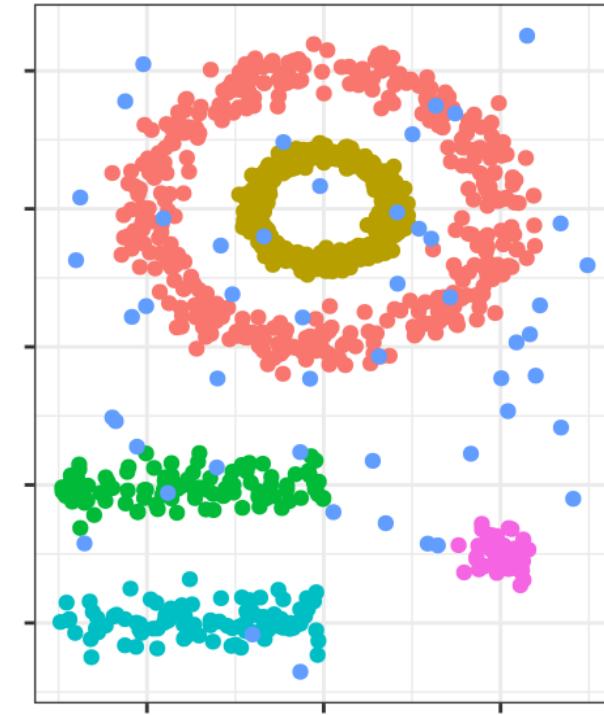
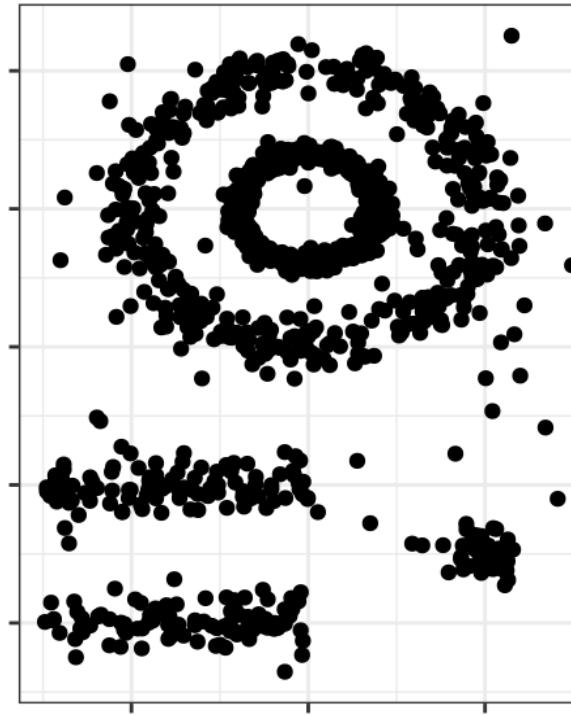
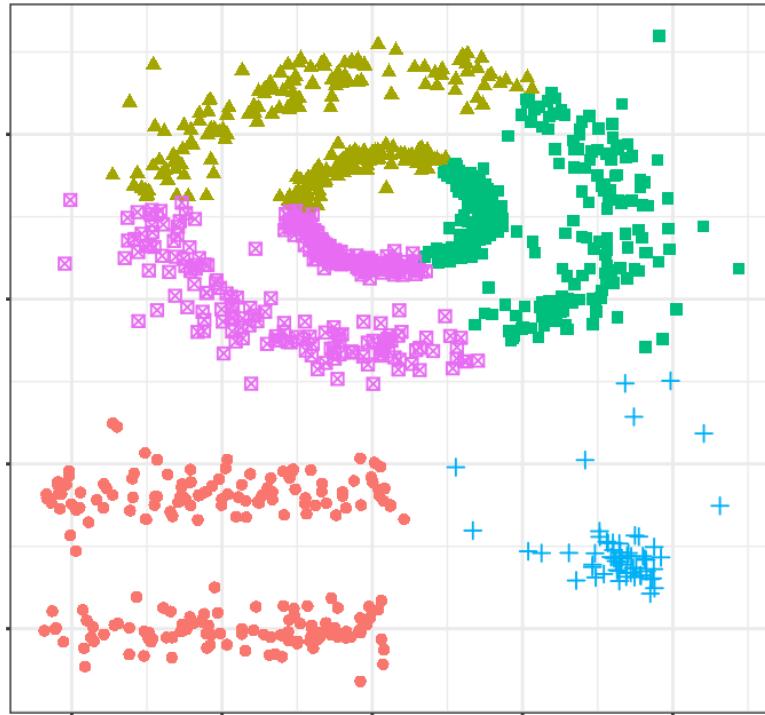


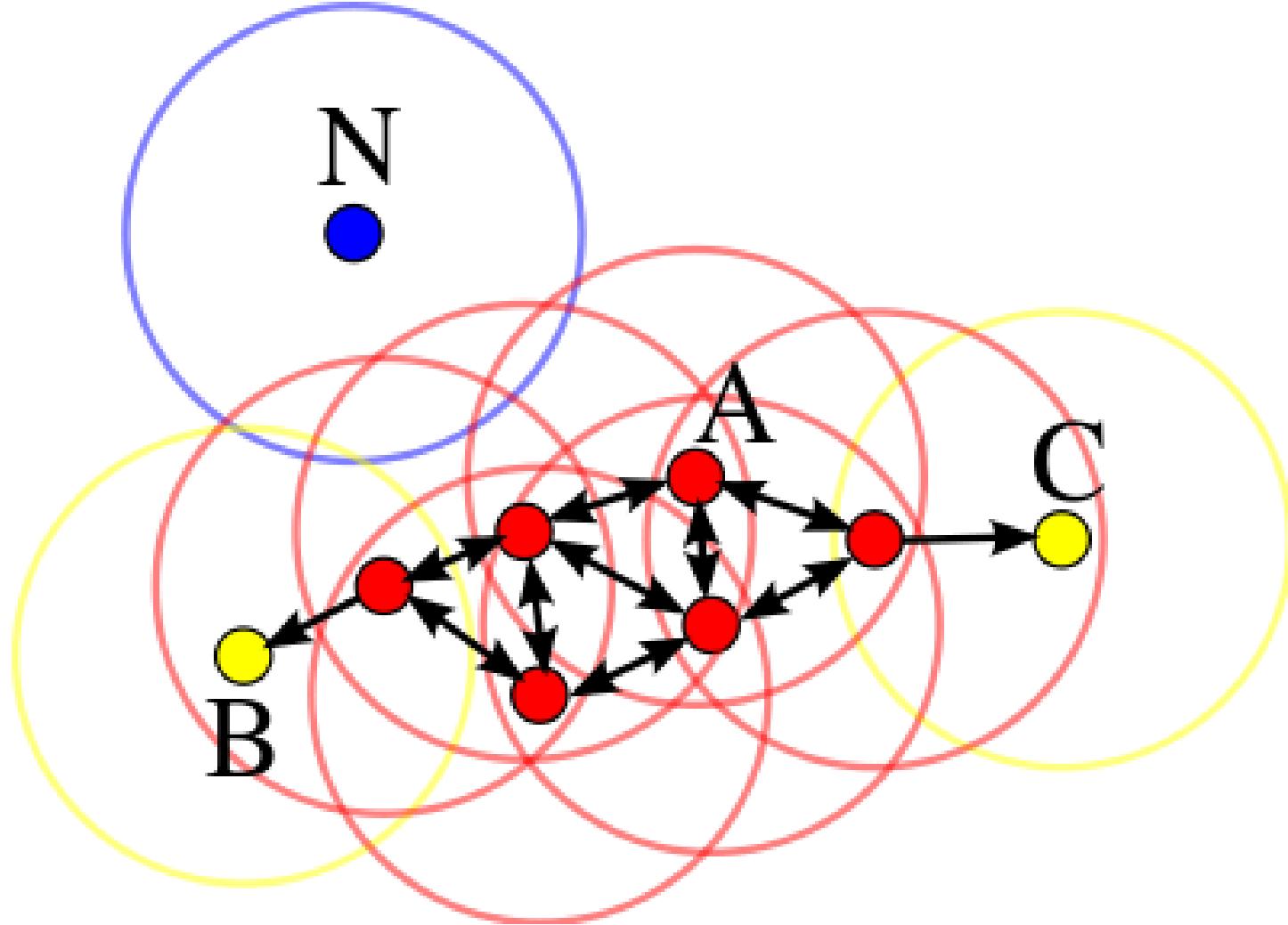
GENERAR  
LOS  
CLUSTERS

archical clustering  
ancia euclídea, Linkage complete, K=2



# DBSCAN







## EJERCICIO 3

- Construye un modelo de K Medias que genere 3 clusters de los datos
- Obten el promedio del coeficiente de Silhouette
- Cambia el numero de clusters a 4 y vuelve a obtener el coeficiente ¿Qué aprecias?
- Construye un modelo de clustering jerárquico para crear 4 clusters
- Asigna colores a los resultados producidos por los dos modelos
- Genera un scatter plot a partir de cada uno de ellos
- Comparalos visualmente



## EJERCICIO 4

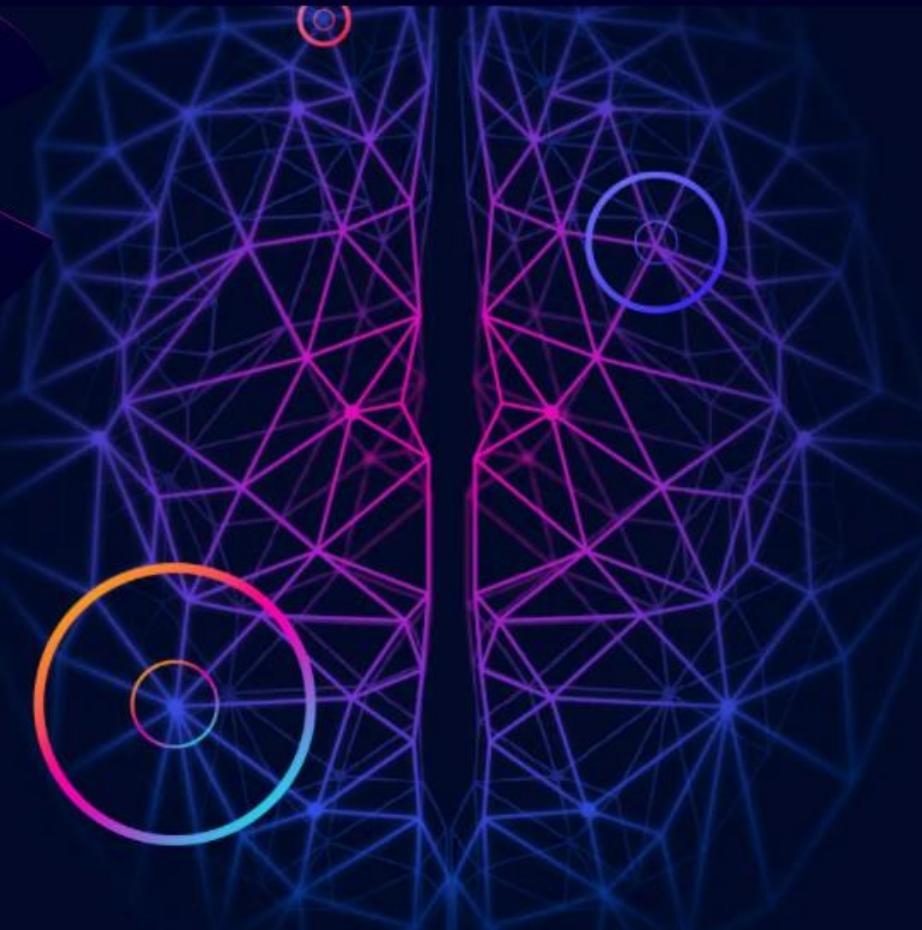
- Particiona los datos en entrenamiento y test
- Predice la columna numérica objetivo con un modelo de regresión lineal y uno de simple regression tree
- Obten el performance de los dos modelos y comparalos



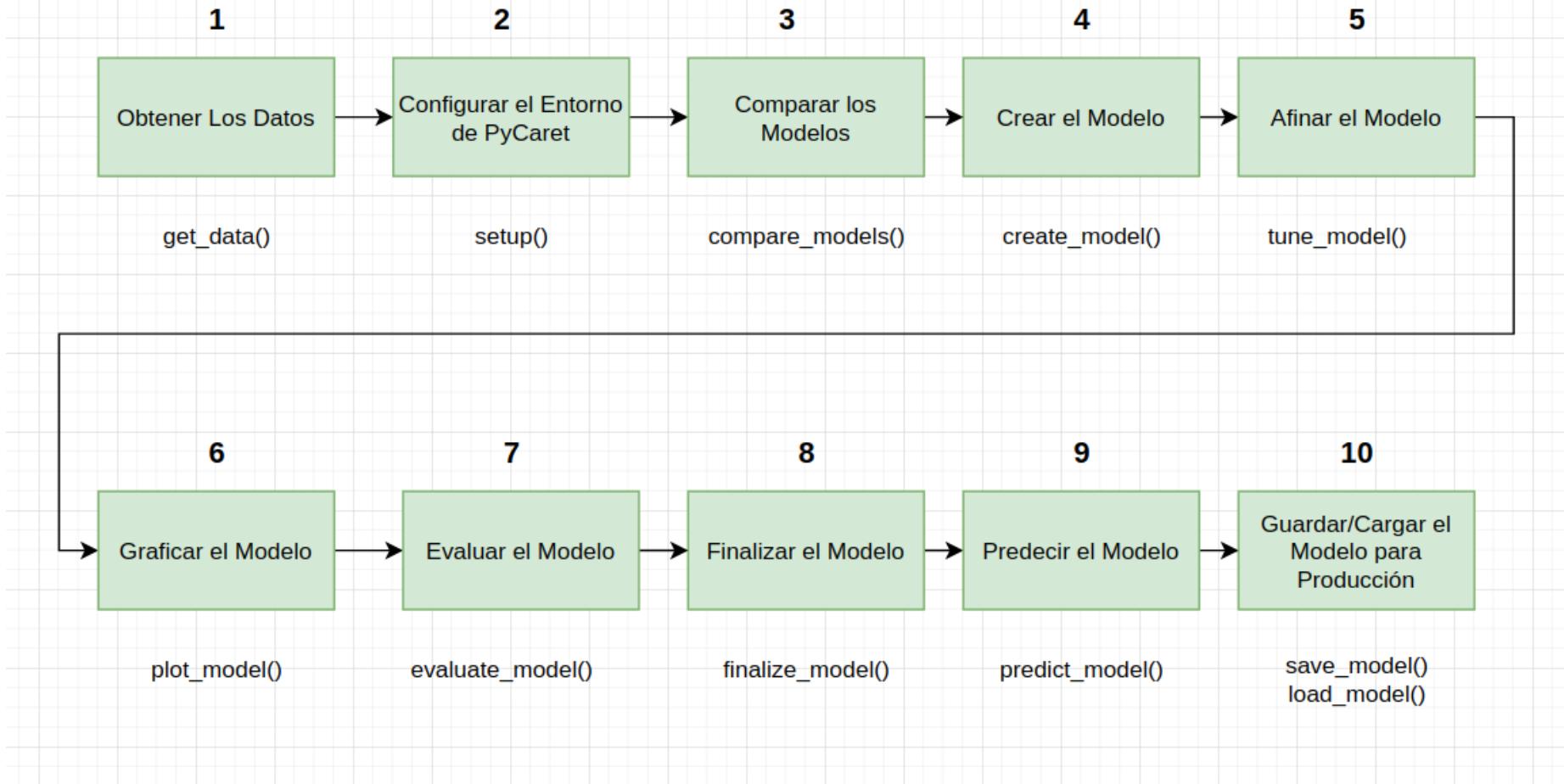
# low-code machine learning

[GET STARTED](#)

PyCaret is an open-source, low-code machine learning library in Python that automates machine learning workflows.

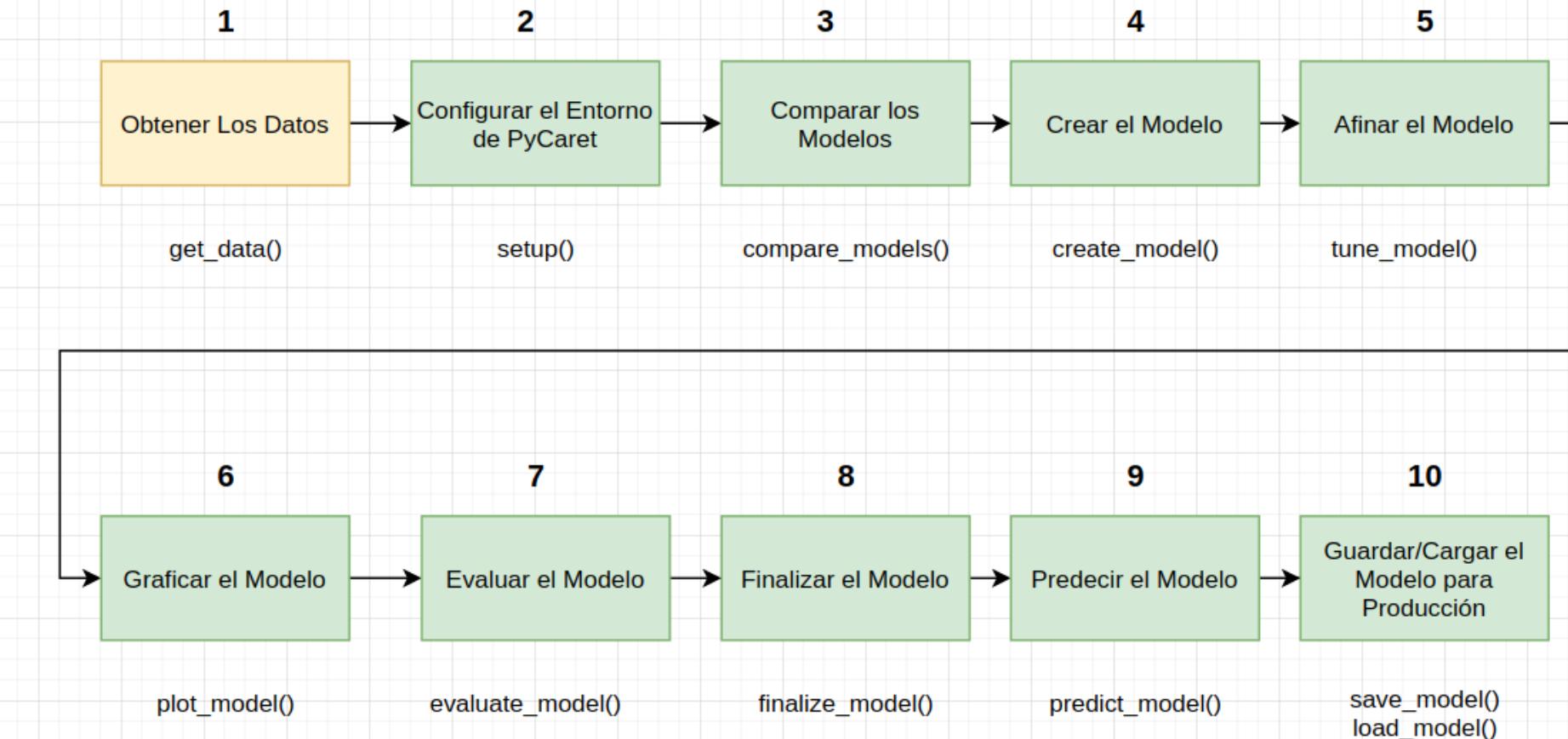


## Trabajando todo el Pipeline con PyCaret

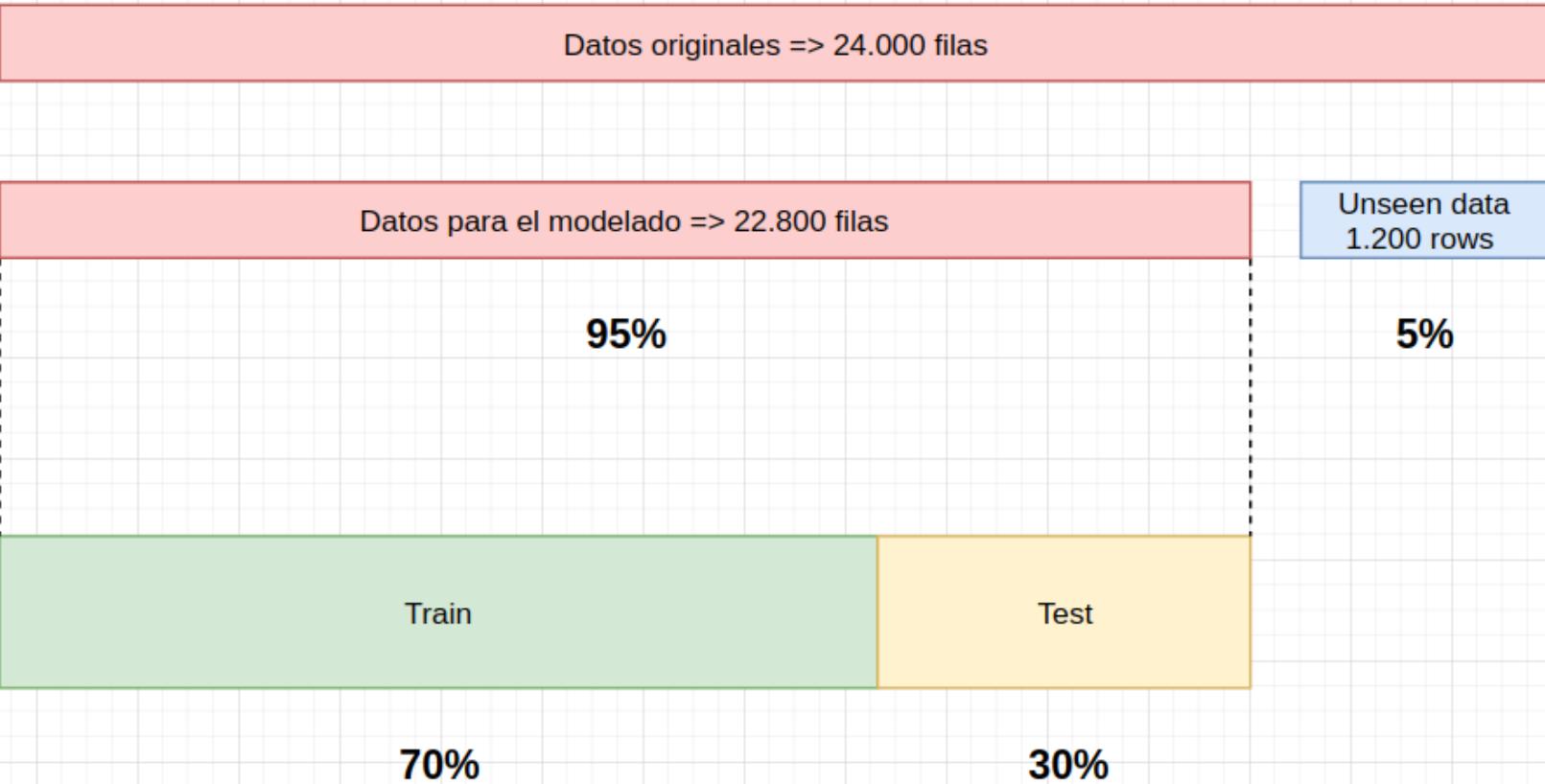


	Name	Reference	Turbo
<b>ID</b>			
<b>lr</b>	Logistic Regression	sklearn.linear_model.LogisticRegression	True
<b>knn</b>	K Neighbors Classifier	sklearn.neighbors.KNeighborsClassifier	True
<b>nb</b>	Naive Bayes	sklearn.naive_bayes.GaussianNB	True
<b>dt</b>	Decision Tree Classifier	sklearn.tree.DecisionTreeClassifier	True
<b>svm</b>	SVM - Linear Kernel	sklearn.linear_model.SGDClassifier	True
<b>rbfsvm</b>	SVM - Radial Kernel	sklearn.svm.SVC	False
<b>gpc</b>	Gaussian Process Classifier	sklearn.gaussian_process.GPC	False
<b>mlp</b>	MLP Classifier	sklearn.neural_network.MLPClassifier	False
<b>ridge</b>	Ridge Classifier	sklearn.linear_model.RidgeClassifier	True
<b>rf</b>	Random Forest Classifier	sklearn.ensemble.RandomForestClassifier	True
<b>qda</b>	Quadratic Discriminant Analysis	sklearn.discriminant_analysis.QDA	True
<b>ada</b>	Ada Boost Classifier	sklearn.ensemble.AdaBoostClassifier	True
<b>gbc</b>	Gradient Boosting Classifier	sklearn.ensemble.GradientBoostingClassifier	True
<b>lda</b>	Linear Discriminant Analysis	sklearn.discriminant_analysis.LDA	True
<b>et</b>	Extra Trees Classifier	sklearn.ensemble.ExtraTreesClassifier	True
<b>xgboost</b>	Extreme Gradient Boosting	xgboost.readthedocs.io	True
<b>lightgbm</b>	Light Gradient Boosting Machine	github.com/microsoft/LightGBM	True
<b>catboost</b>	CatBoost Classifier	catboost.ai	True

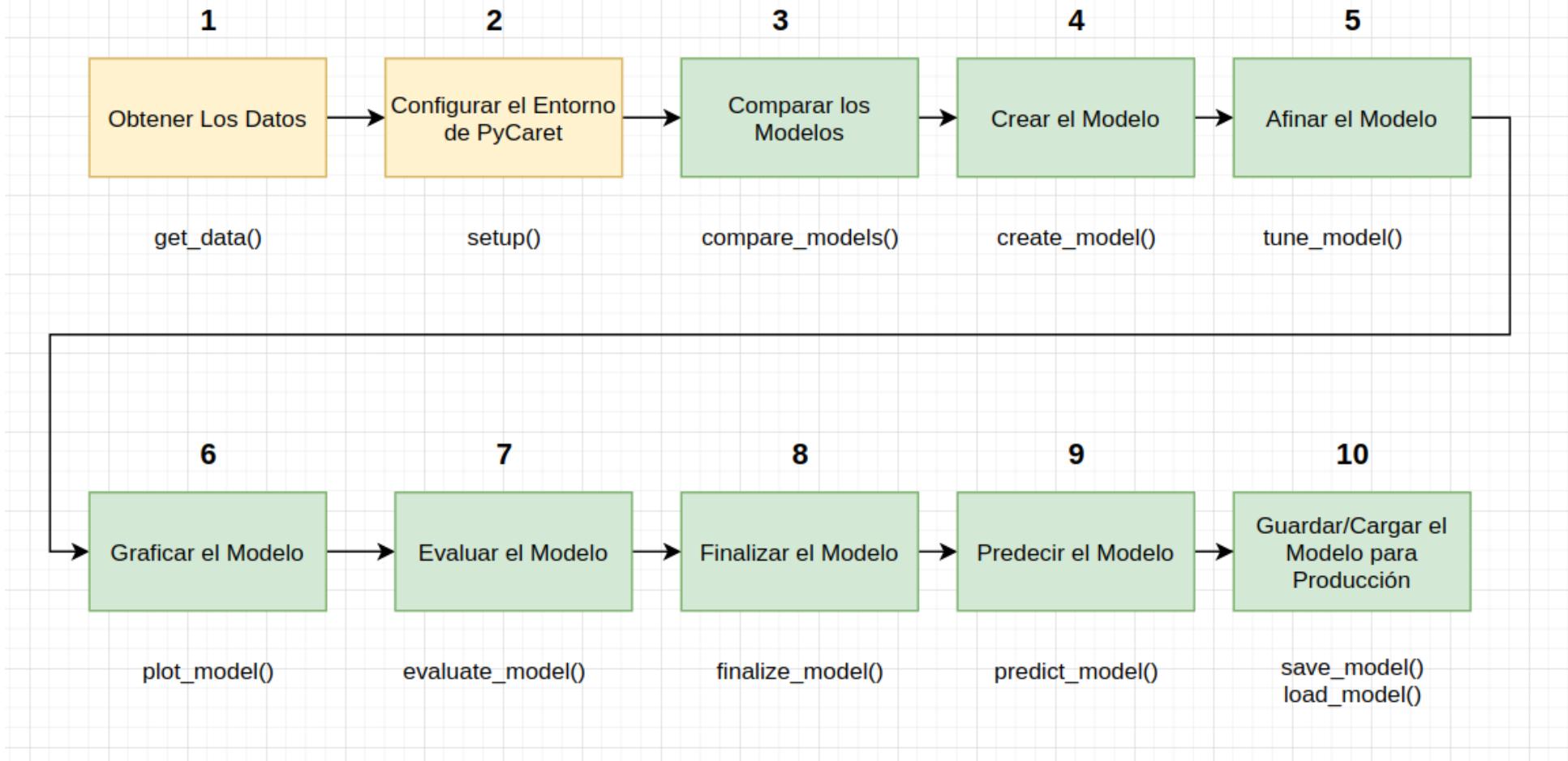
## Trabajando todo el Pipeline con PyCaret



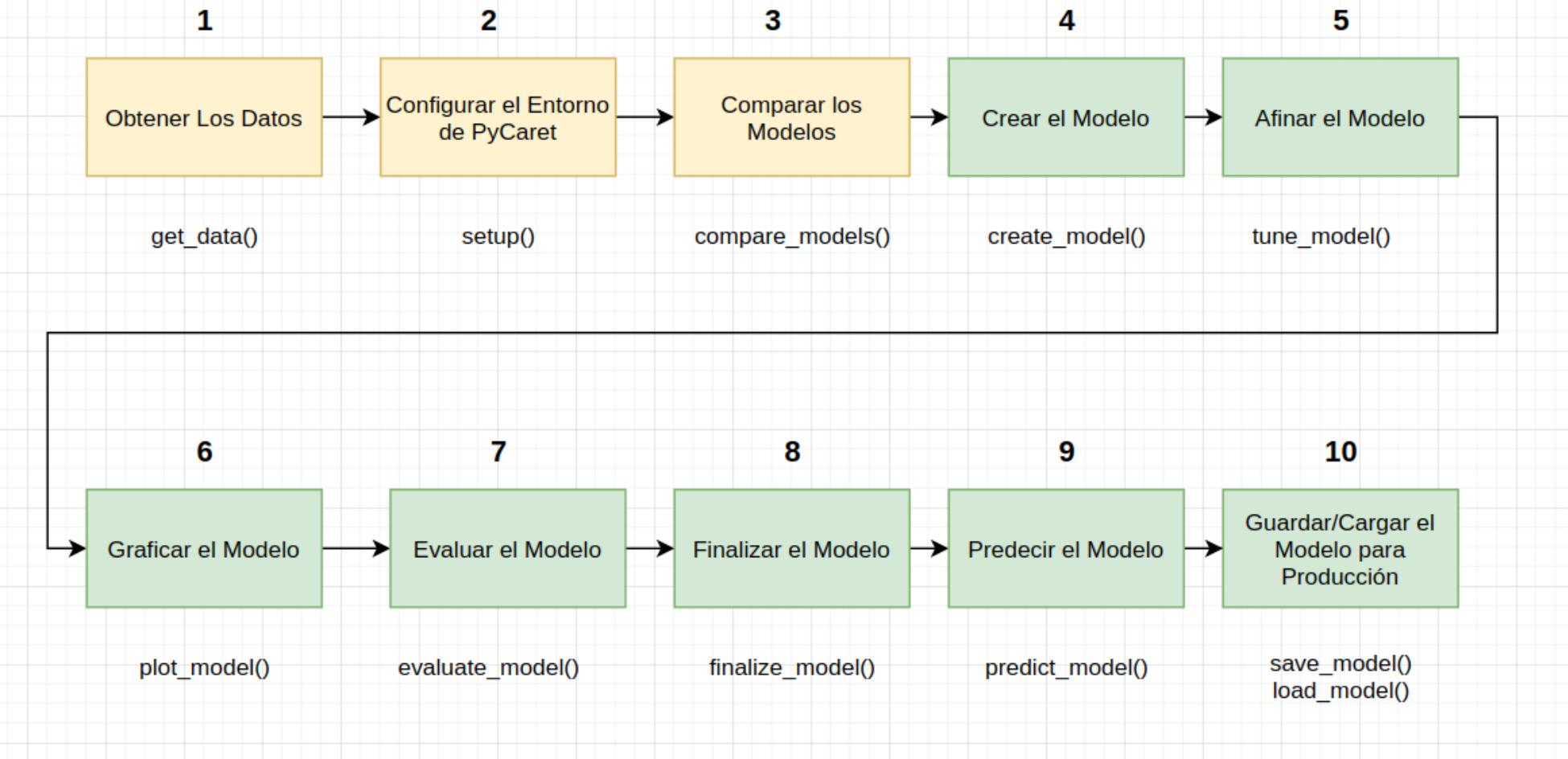
## División de los datos



## Trabajando todo el Pipeline con PyCaret



## Trabajando todo el Pipeline con PyCaret



## 10 Fold Cross Validation

Fold 1



Fold 2



Fold 10

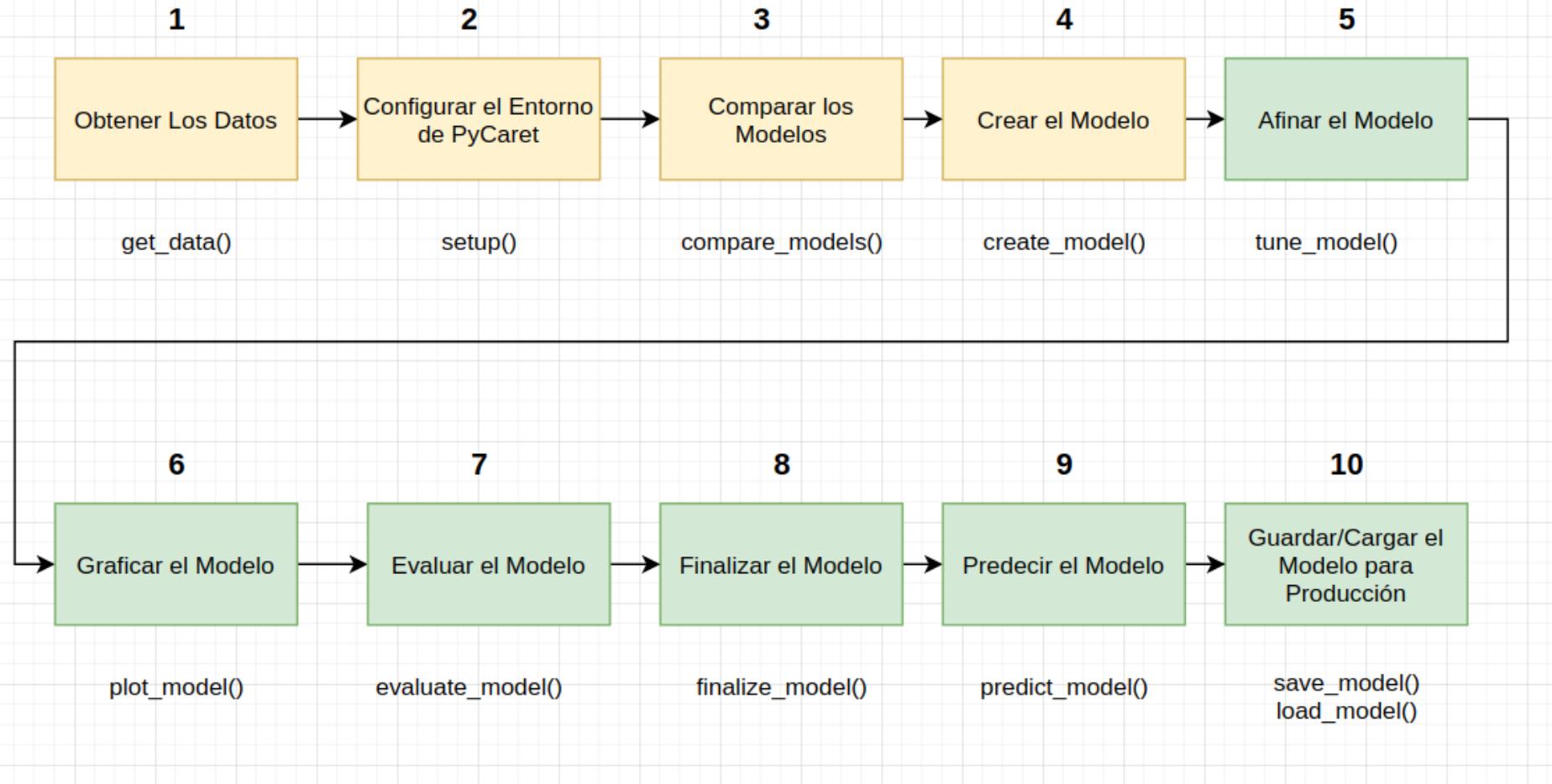


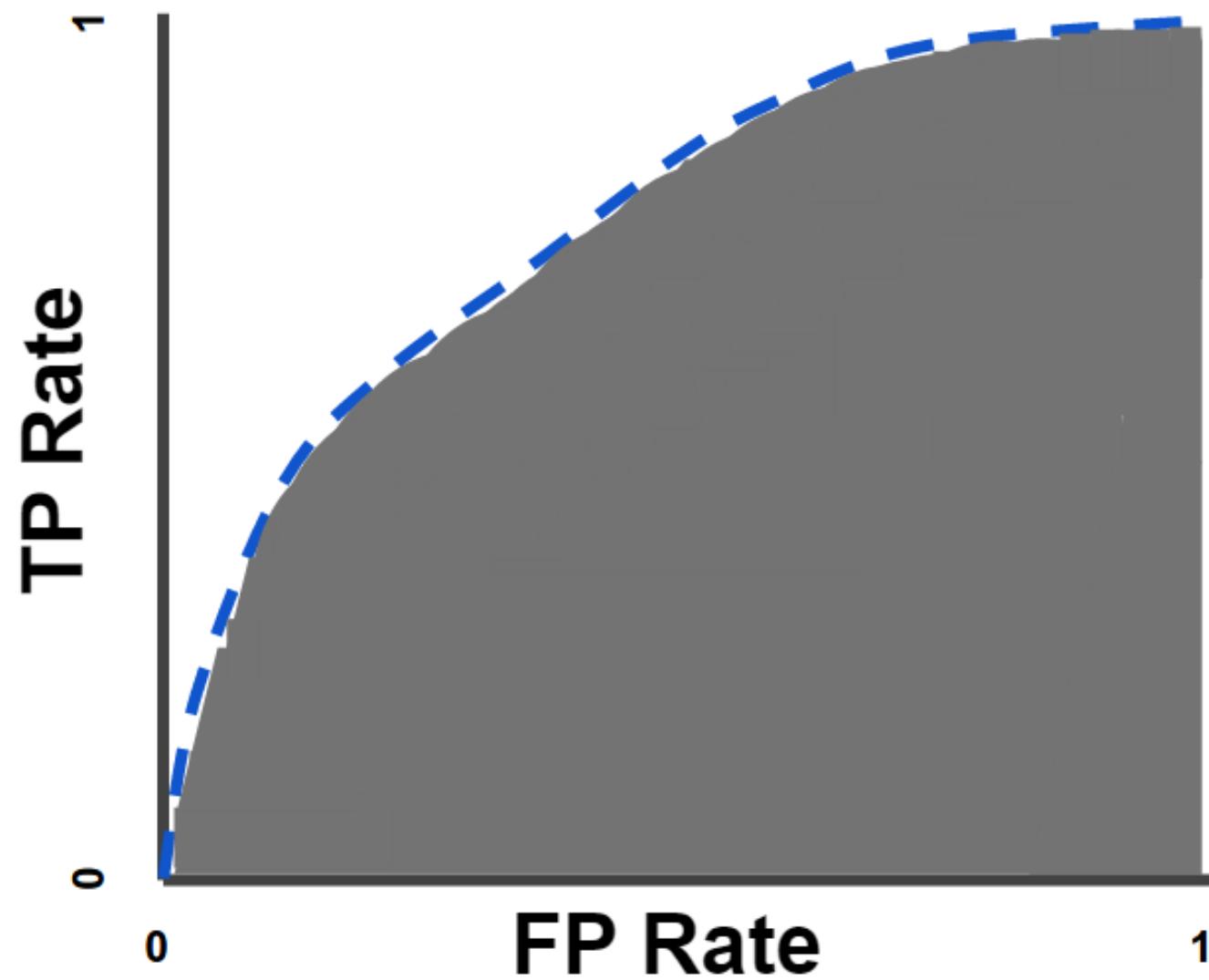
...

● Validation Set

● Training Set

## Trabajando todo el Pipeline con PyCaret





## Trabajando todo el Pipeline con PyCaret

1

Obtener Los Datos

get\_data()

2

Configurar el Entorno de PyCaret

setup()

3

Comparar los Modelos

compare\_models()

4

Crear el Modelo

create\_model()

5

Afinar el Modelo

tune\_model()

6

Graficar el Modelo

plot\_model()

7

Evaluar el Modelo

evaluate\_model()

8

Finalizar el Modelo

finalize\_model()

9

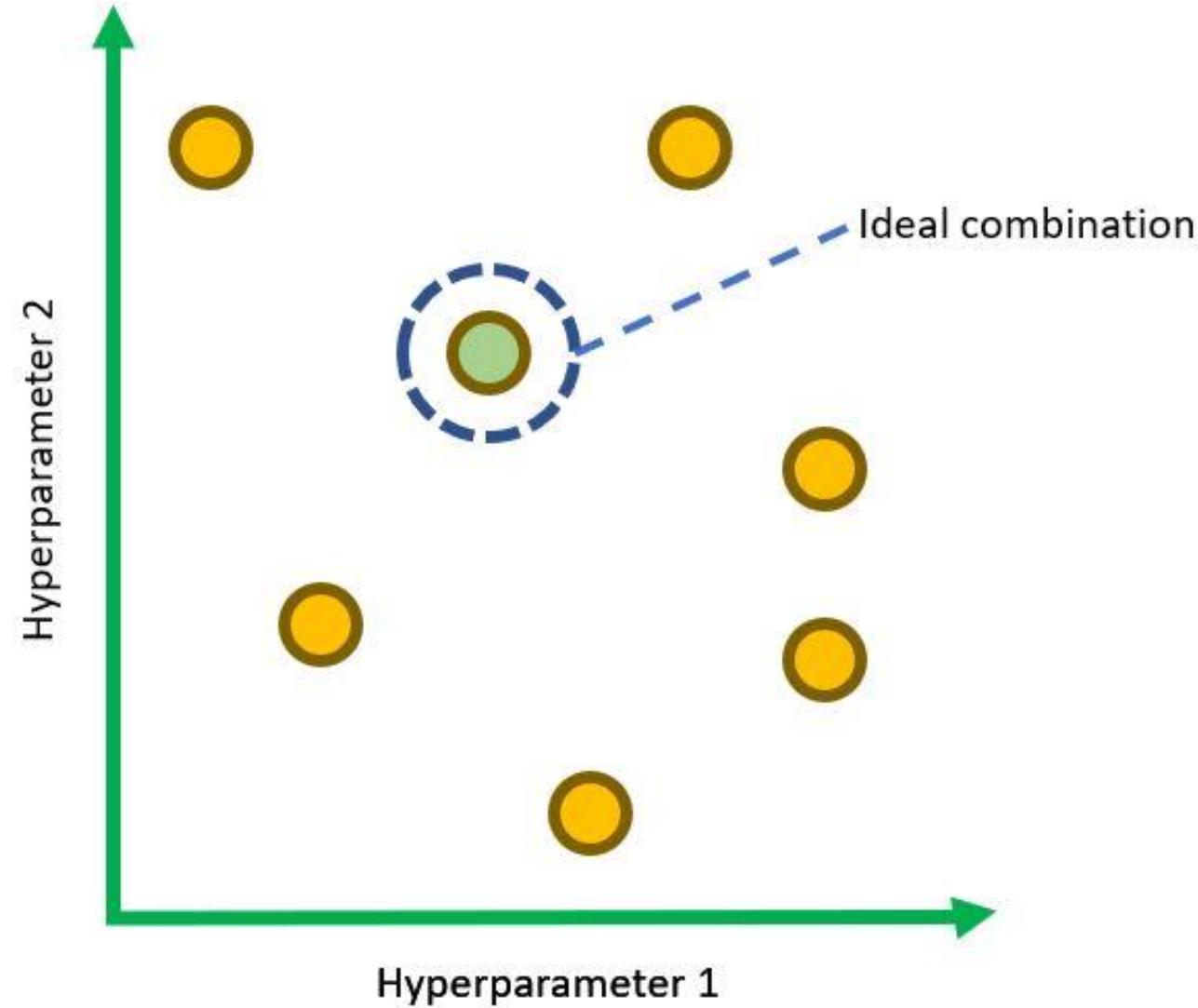
Predecir el Modelo

predict\_model()

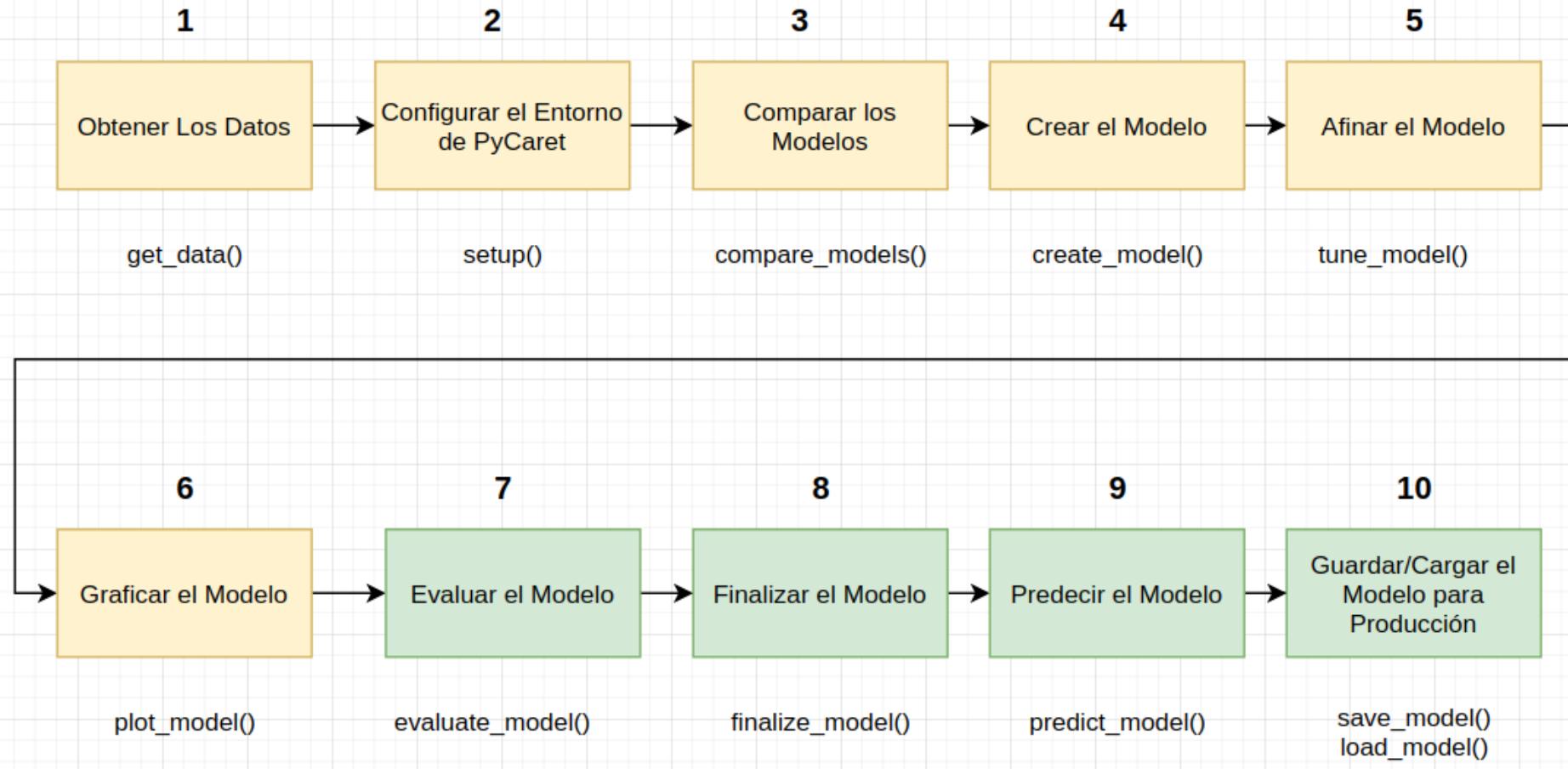
10

Guardar/Cargar el Modelo para Producción

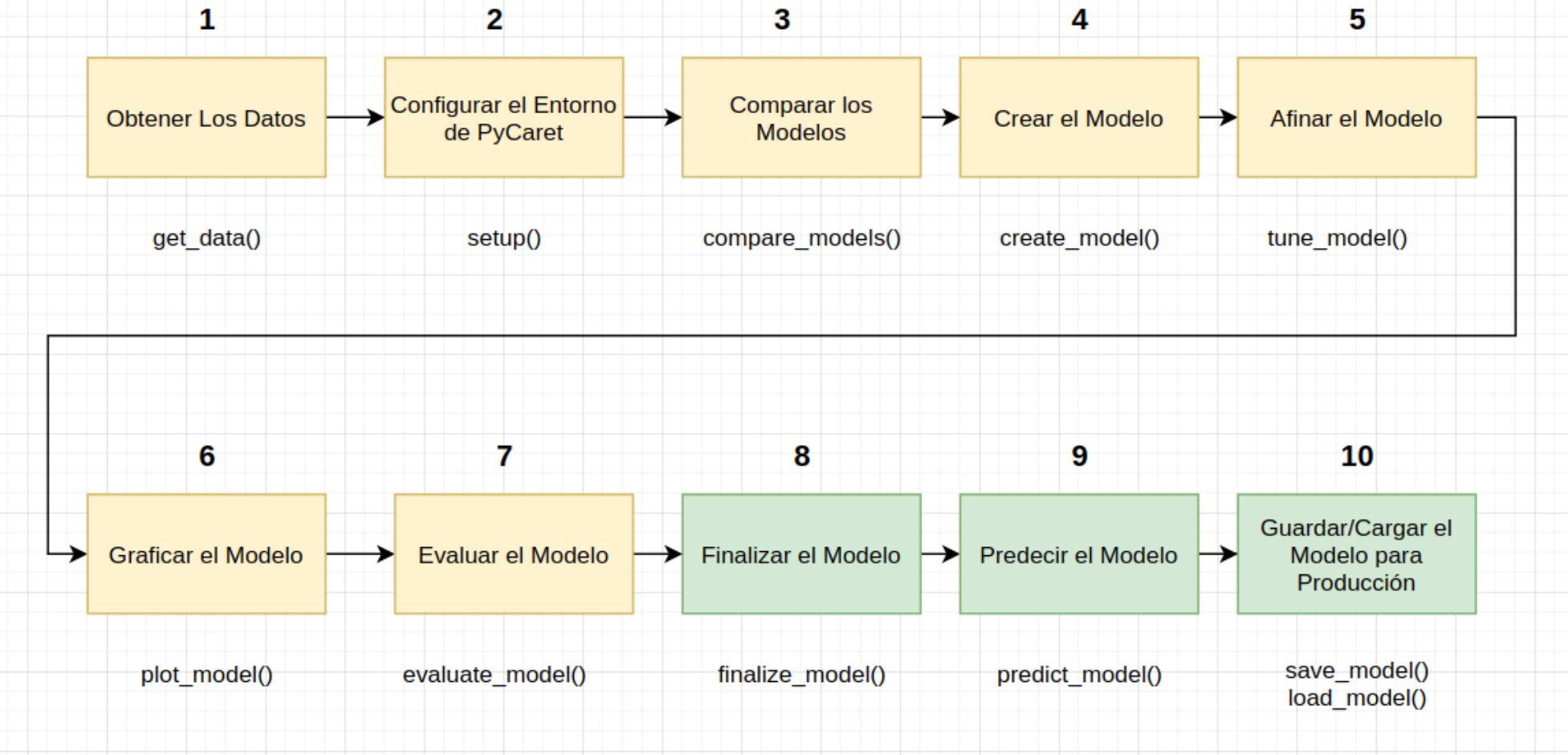
save\_model()  
load\_model()



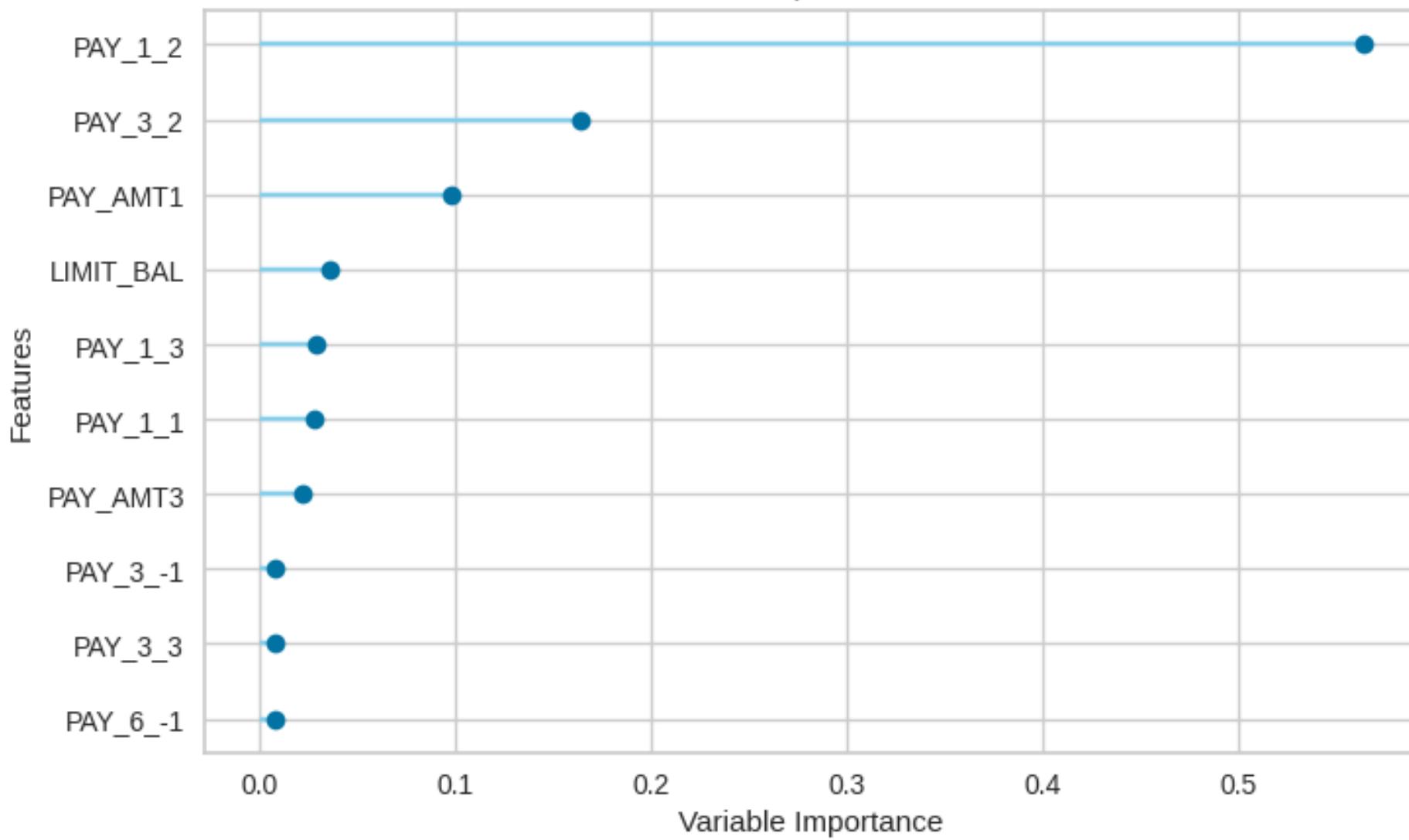
## Trabajando todo el Pipeline con PyCaret



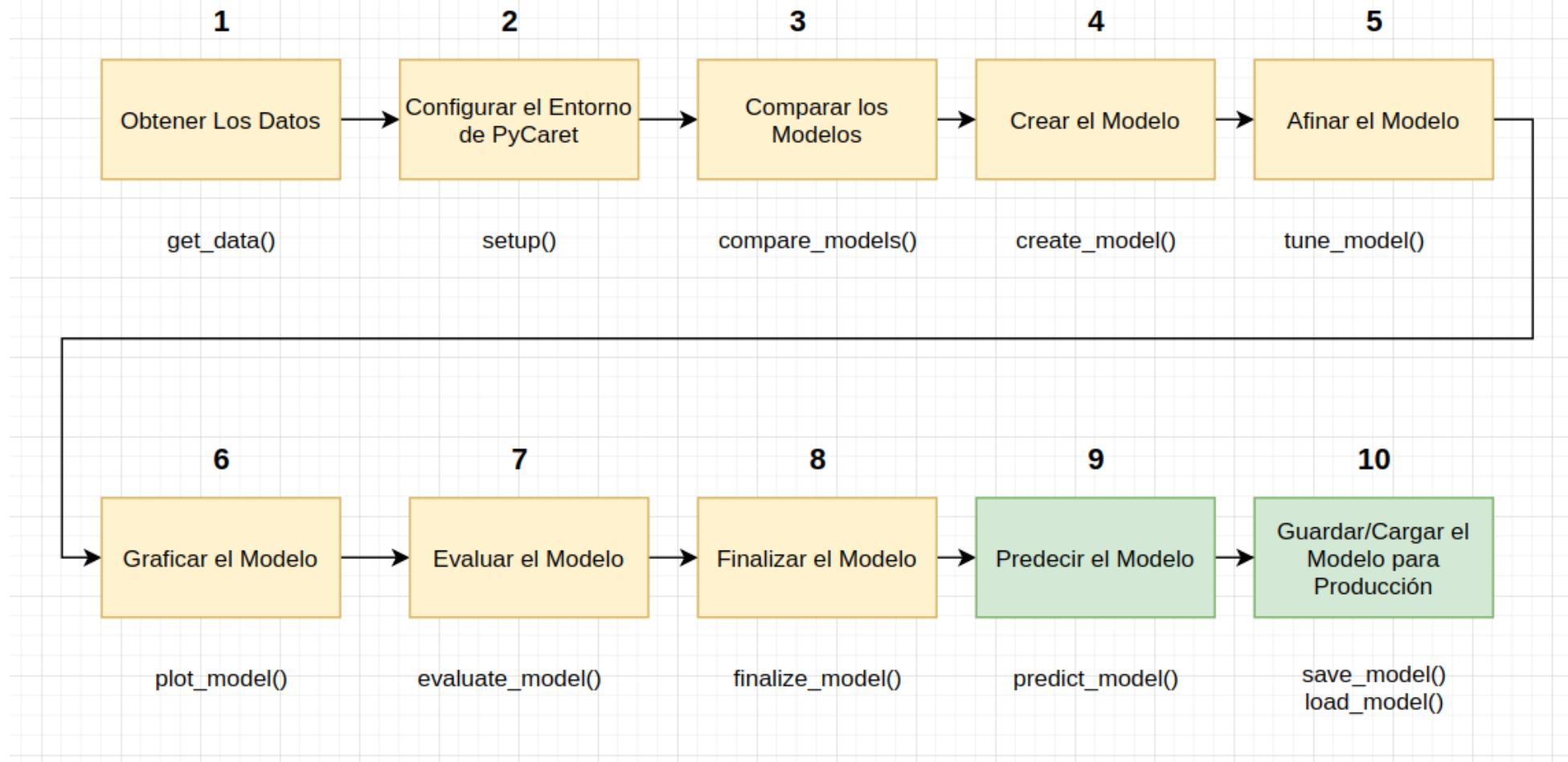
## Trabajando todo el Pipeline con PyCaret



### Feature Importance Plot

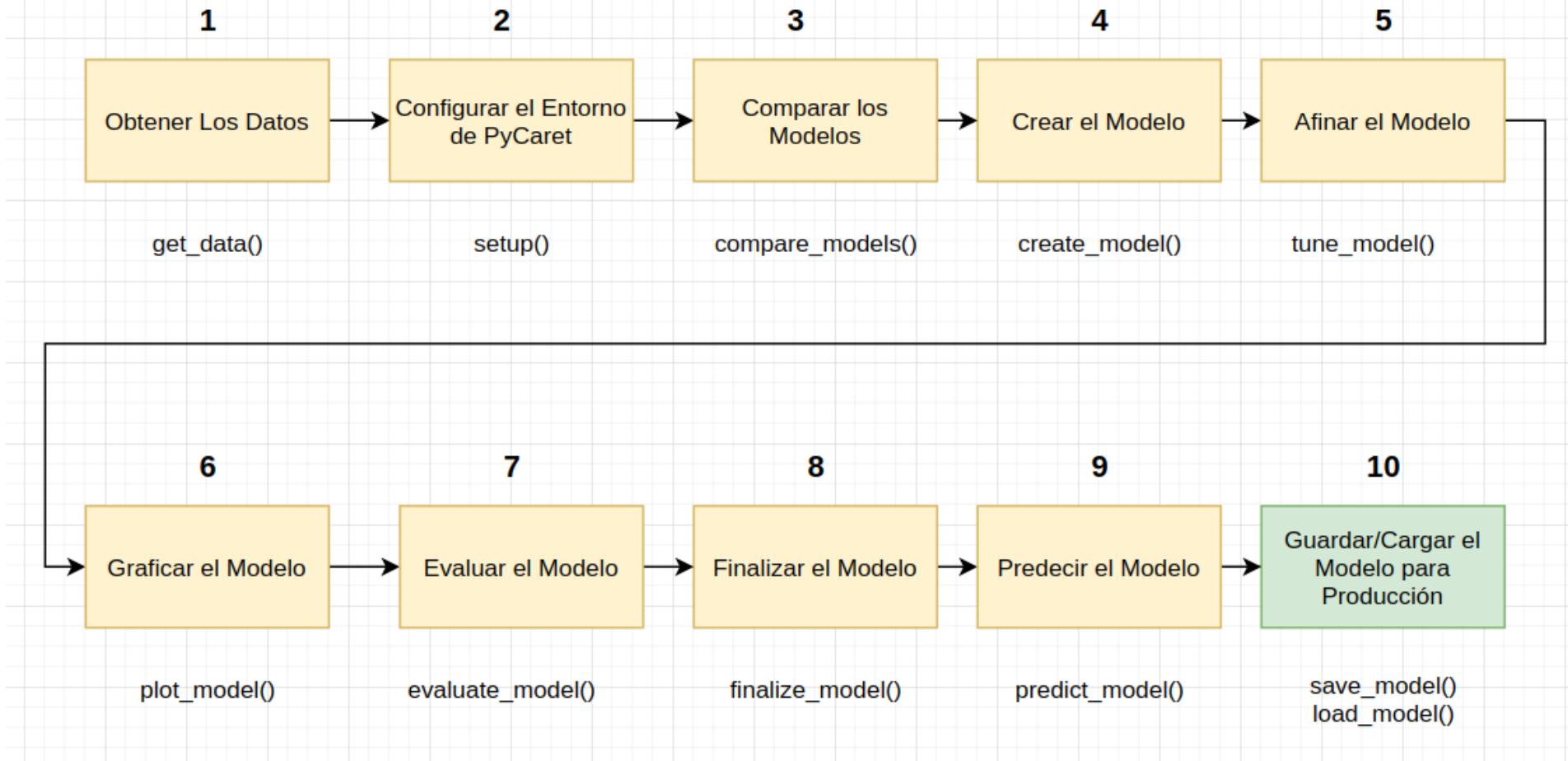


## Trabajando todo el Pipeline con PyCaret



T5	PAY_AMT6	default	Label	Score
7.0	1542.0	0	0	0.8051
5.0	8251.0	0	0	0.9121
5.0	1395.0	0	0	0.8051
0.0	5000.0	1	1	0.7911
7.0	924.0	0	0	0.9121

## Trabajando todo el Pipeline con PyCaret



# EJERCICIO 5

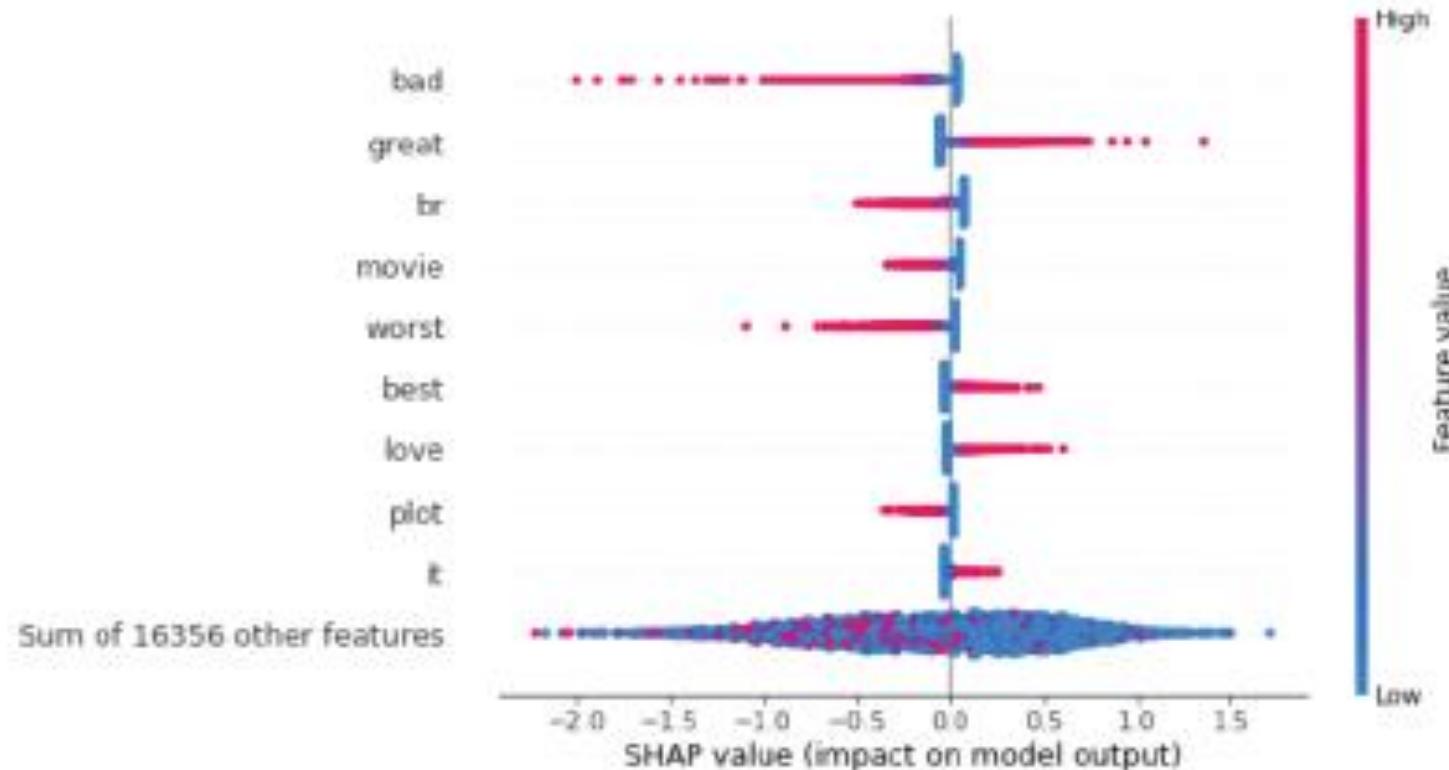
- Utilizando Pycaret construye el mejor modelo posible para trabajar con los datos de

```
from pycaret.datasets import get_data  
data = get_data('juice')
```

- Afina los hiperparametros
- Analiza sus resultados y di si es un modelo correcto



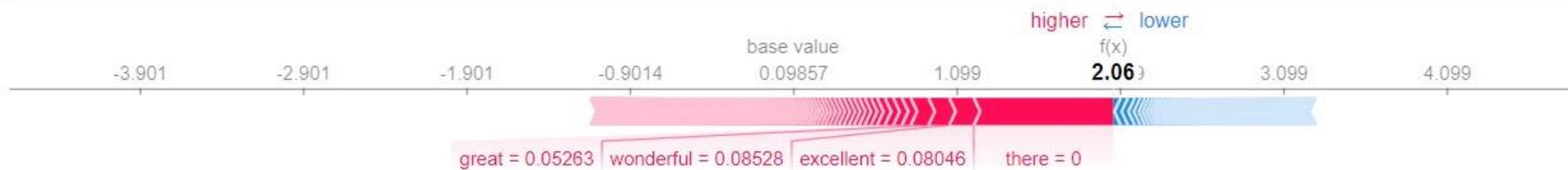
```
explainer = shap.Explainer(model, X_train, feature_names=vectorizer.get_feature_names())
shap_values = explainer(X_test)
shap.plots.beeswarm(shap_values)
```



```
shap.plots.force(shap_values[num_resena])
```



```
shap.plots.force(shap_values[10]) # critica muy positiva la cual incluye palabras como excellent
```



```
shap.plots.force(shap_values[100]) # critica muy negativa la cual incluye palabras como stupid
```

