



Universidad Tecnológica de Morelia

# ANALISIS DE DATASET ! UTILIZANDO LA HERRAMIENTA KNIME

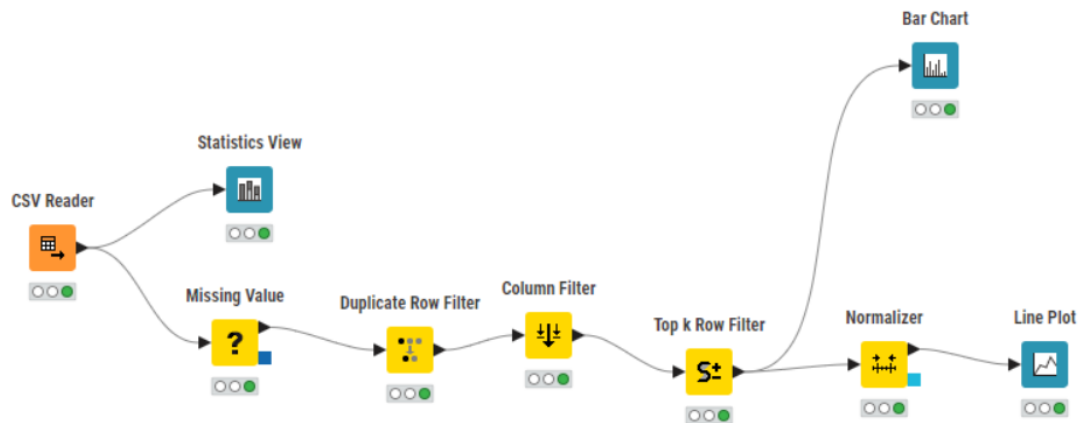
EXTRACCION DE CONOCIMIENTO  
DE BASE DE DATOS

Presentado por:  
**NICOLÁS LEMUS CISNEROS**

# INTRODUCCIÓN

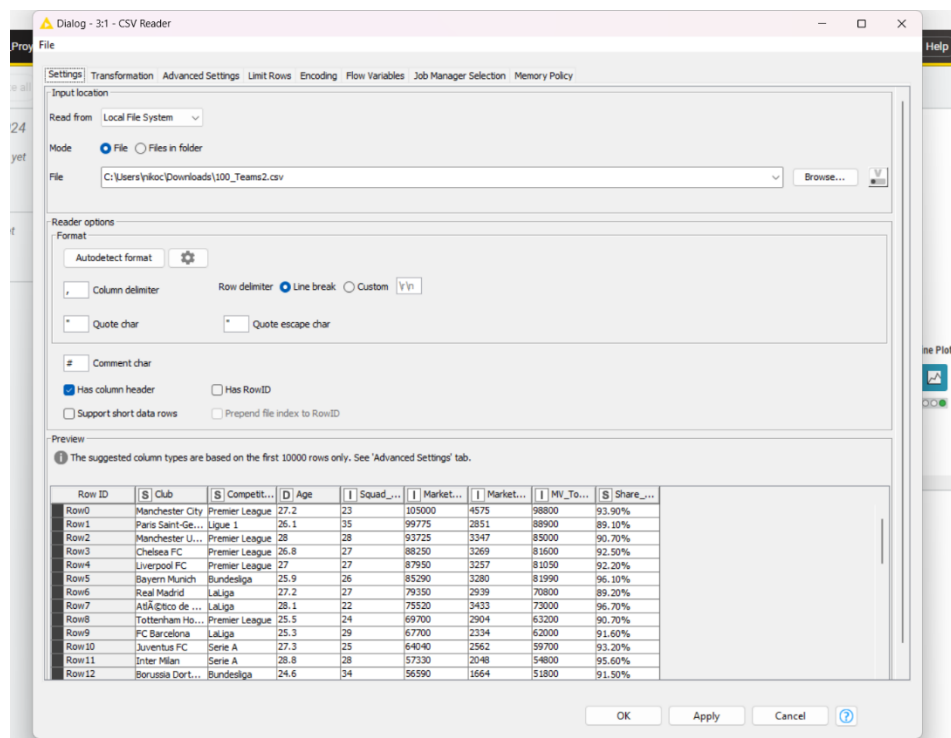
El fútbol es uno de los deportes más populares y lucrativos del mundo, atrayendo a millones de fanáticos y generando ingresos significativos a través de derechos de transmisión, patrocinios, ventas de entradas y mercancías. En este contexto, el valor de los equipos de fútbol se ha convertido en un indicador crucial de su éxito tanto dentro como fuera del campo. Este trabajo se enfoca en el análisis de un dataset que contiene información sobre los 100 equipos más valiosos del fútbol.

El objetivo principal de este análisis es explorar y entender los factores que contribuyen al valor de un equipo de fútbol. Utilizando el software KNIME, una plataforma de análisis de datos basada en flujos de trabajo, realizaremos un análisis exhaustivo de los datos para identificar patrones, tendencias y correlaciones que puedan proporcionar una visión más profunda del valor de los equipos.

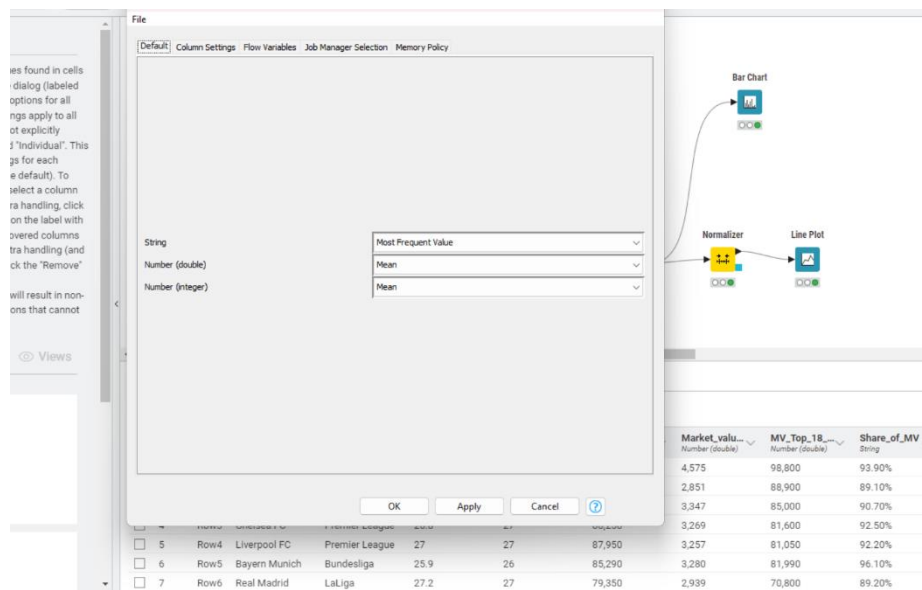


Esta es la vista general de todos los nodos que utilice para la preparación de los datos para su modelaje.

La imagen de a continuación muestra el dataset insertado en knime.



Una vez insertado, agregamos el nodo statistics view, este nos permitirá ver los datos de una mejor manera para saber que tipo de dato son los campos además si hay campos vacíos o no.



Esta imagen muestra la configuración que decidí colocar en el nodo missing value, esto para evitar que nuestro data set tenga campos vacíos y los autocomplete sin necesidad de borrarlos u omitirlos.

Column Filter

This node allows columns to be filtered from the input table while only the remaining columns are passed to the output table. Within the dialog, columns can be moved between the include and Exclude list.

External resources

- KNIME E-Learning Course: Column Filter
- Java API documentation about regex patterns
- Java API documentation about regex matching

Ports Options Views

Input ports

- Type: Table to be filtered
- Table from which columns are to be excluded.

Output ports

- Type: Filtered table
- Table excluding selected columns.

Dialog - 3/5 - Column Filter

Manual Wildcard Regex Type

Search

Excludes

- Share\_of\_MV

Includes

- Competition
- Age
- Squad\_size
- Market\_value
- Market\_value\_of\_players
- MV\_Top\_18\_players
- Any unknown column

Cancel OK

Top k Row Filter Normalizer Line Plot

Market_value	Market_value...	MV_Top_18...	Share_of_MV
Number (double)	Number (double)	Number (double)	String
105,000	4,575	98,800	93.90%
99,775	2,851	88,900	89.10%
93,725	3,347	85,000	90.70%
88,250	3,269	81,600	92.50%
87,950	3,257	81,050	92.20%
85,290	3,280	81,990	96.10%
79,350	2,939	70,800	89.20%

Los nodos column filter y row filter los utilizamos para filtrar los datos repetidos, además de las columnas que no nos interesen en nuestro análisis por ejemplo el ID o número de equipo, por ejemplo.

Top k Row Filter

The node behaves the same as a combination of the Sorter node followed by a Row Filter that only keeps the first k rows of the table except for the order of the rows which depends on the Output order settings. Note, however, that the implementation of this node is more efficient than the node combination above. In the dialog, select the columns according to which the data should be selected. For each column you can also specify whether a larger or smaller value is considered as superior.

Ports Options Views

Input ports

- Type: Input Table
- Table to select rows from.

Output ports

- Type: Top k Table
- A table containing the top k rows.

Dialog - 3/7 - Top k Row Filter

File Settings Advanced Settings Flow Variables Job Manager Selection Memory Policy

Number of rows 10

Sorting criteria

Sort by

- Market\_value
- Alphanumeric string comparison

Ascending Descending

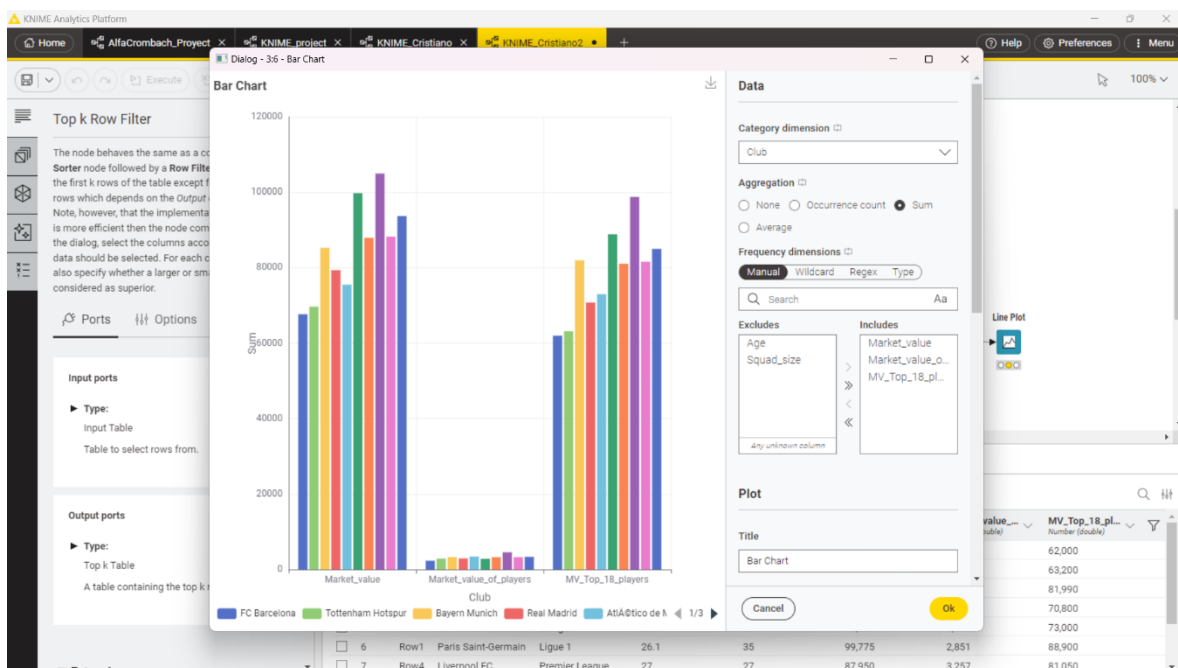
+ Add sorting criterion

OK Apply Cancel ?

Top k Row Filter Normalizer Line Plot

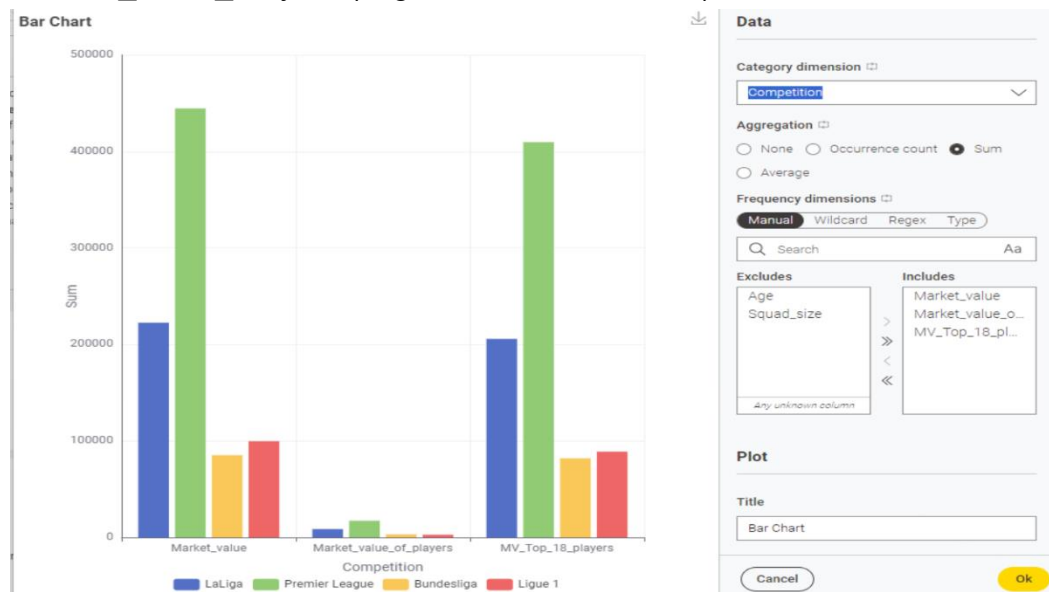
RowID	Club	Competition	Age	Squad_size	Market_value	Market_value...	MV_Top_18...	Share_of_MV
String	String	String	Number (double)	Number (double)	Number (double)	Number (double)	Number (double)	String
Row90	Udinese Calcio	Serie A	27.7	25	10,145	406	9,520	
Row91	Feyenoord Rotterdam	Eredivisie	23.5	28	10,140	362	9,890	
Row93	RCD Espanyol Barc...	LaLiga	27.6	25	10,030	401	9,030	
Row92	1.FSV Mainz 05	Bundesliga	25.5	25	10,050	402	9,630	
Row95	Clube Atlético M...	Série A	26.9	31	9,845	318	8,495	
Row94	Genoa CFC	Serie A	27.5	34	9,920	292	8,600	
Row98	Trabzonspor	Süper Lig	26.7	29	9,630	332	8,900	

Si siguiendo con los filtros el nodo Top k Row Filter, lo configure para que solo muestre los 10 mejores resultados de los 100 equipos, ya que al hacer la visualización que sean 100 datos dificulta la amabilidad del análisis de las graficas

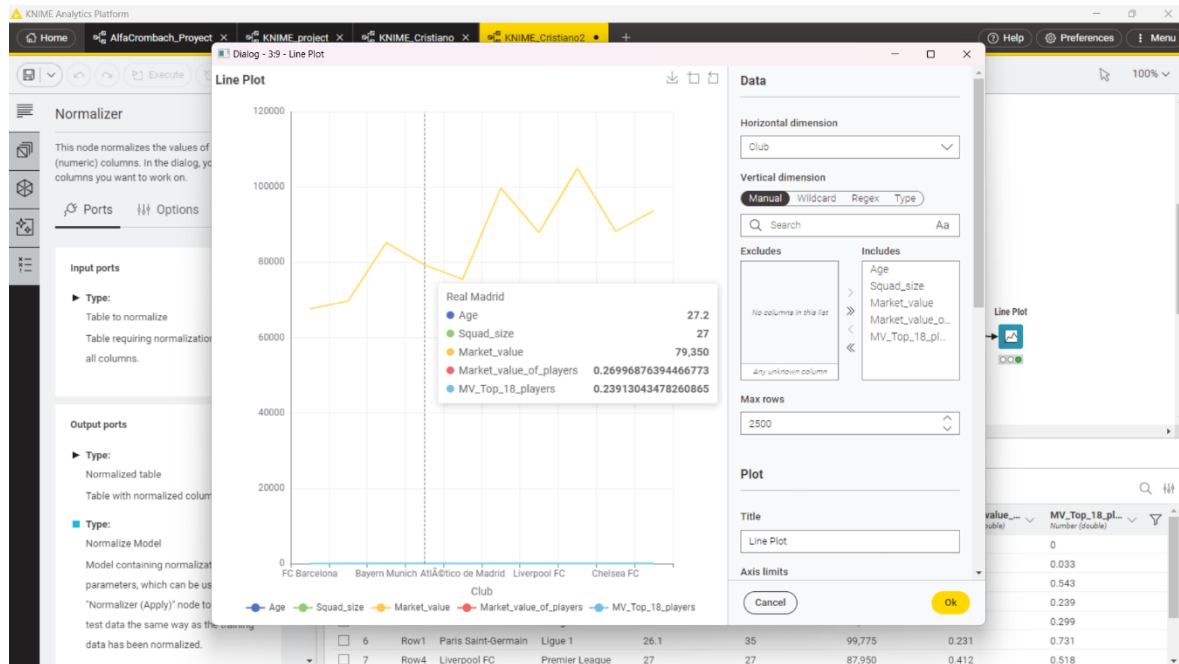


En esta primera página podemos observar el visualizador usado Barchart los clubes como parámetro a medir y debajo como comparaciones están los 3 valores a analizar:

- Most\_Value (Valor del equipo)
- Value\_Market\_Player (Valor del mercado de los jugadores)
- Most\_Value\_Players (Jugadores más valiosos)



En la anterior grafica cambie el parámetro a comparar club/competición en el cual veremos que ligas tienen un mayor valor de esos 100 clubes de la lista al igual que los jugadores más valuados de ese top 100.



Para concluir con este análisis y preprocesamiento de datos, busque una opción para poder mostrar todos los datos de los mejores 10 equipos, para obtener un mejor conocimiento de los mismo, que aportan, su numero de plantilla además de la edad promedio de sus jugadores.

Esto fue posibles gracias a la adición de un nodo para evitar perder los datos con poco valor entre los datos de millones de unidades.

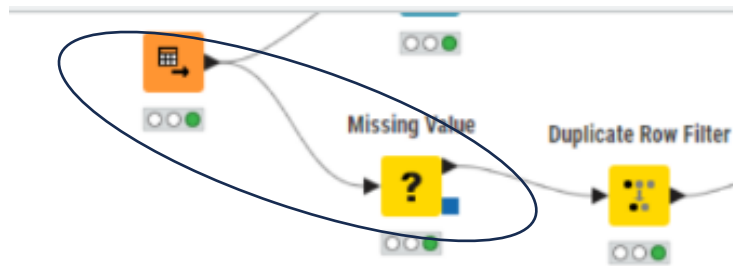
La función se llama Normalizer y filtre únicamente los datos string para que en la grafica me mostrara algo como se muestra en la imagen anterior:

- Nombre del club
- Edad
- Tamaño de plantilla
- Valor de mercado
- Representación del valor del mercado de jugadores
- Cantidad que aporta sobre 1 que representa los top 18 jugadores más valiosos.

## Aplicación de modelo matemático

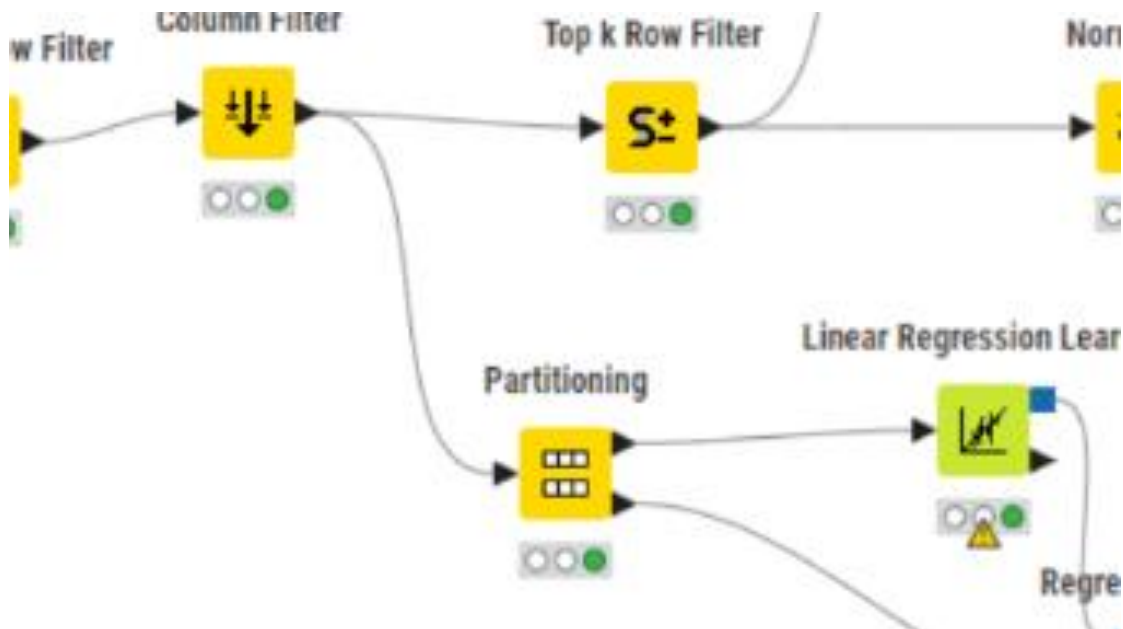
### Paso 1: Preparación de los Datos

1. **Importar los datos:** Utilizamos nuevamente CSV Reader para importar tus datos a KNIME.
2. **Limpieza de datos:** Hay que asegurarnos de que los datos no contengan valores nulos o inconsistencias que puedan afectar el modelo. Utilizamos nodos como Missing Value para manejar valores faltantes.



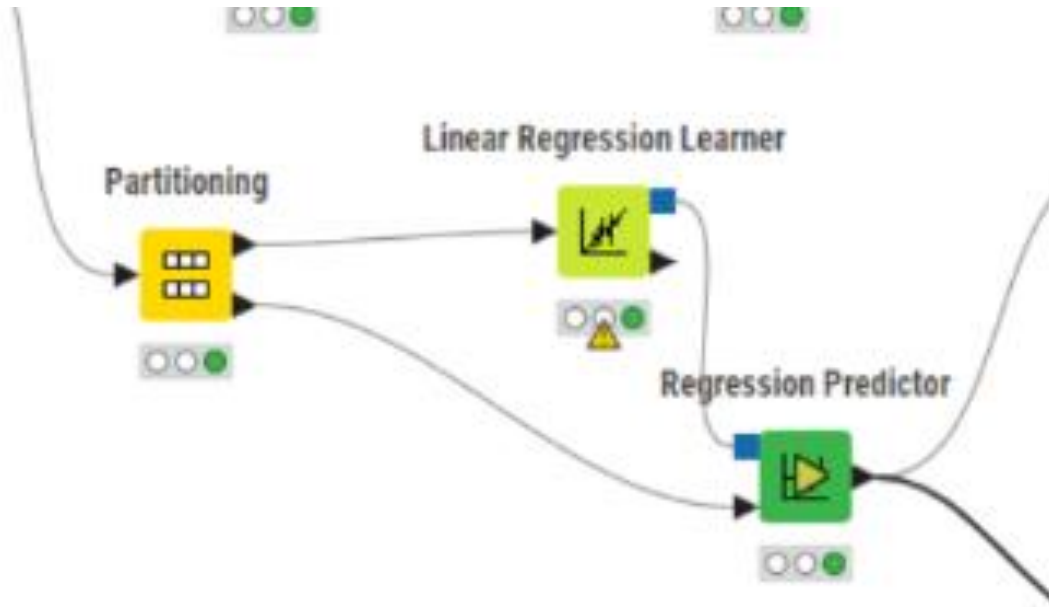
### Paso 2: División de los Datos

1. **Dividir los datos en entrenamiento y prueba:** Se utiliza el nodo Partitioning para dividir tus datos en un conjunto de entrenamiento y un conjunto de prueba.



### Paso 3: Creación del Modelo

1. **Seleccionar las características y la variable objetivo:** Utiliza el nodo Column Filter para seleccionar las columnas que serán utilizadas como características (variables independientes) y la columna que será variable objetivo (variable dependiente).
2. **Aplicar el modelo de regresión lineal:** Utilizé el nodo Linear Regression Learner. Este nodo entrenará un modelo de regresión lineal utilizando tus datos de entrenamiento.

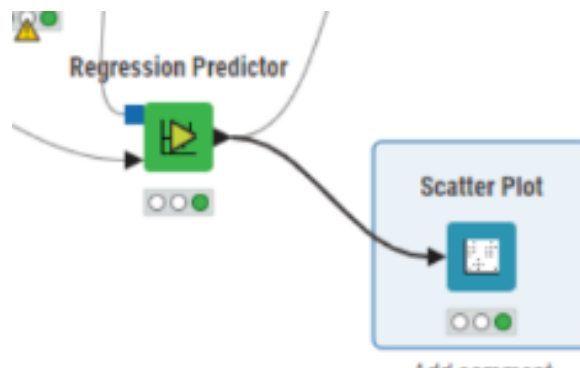


### Paso 4: Evaluación del Modelo

1. **Aplicar el modelo a los datos de prueba:** Utilizamos el nodo Linear Regression Predictor para aplicar el modelo entrenado del conjunto de datos de prueba.

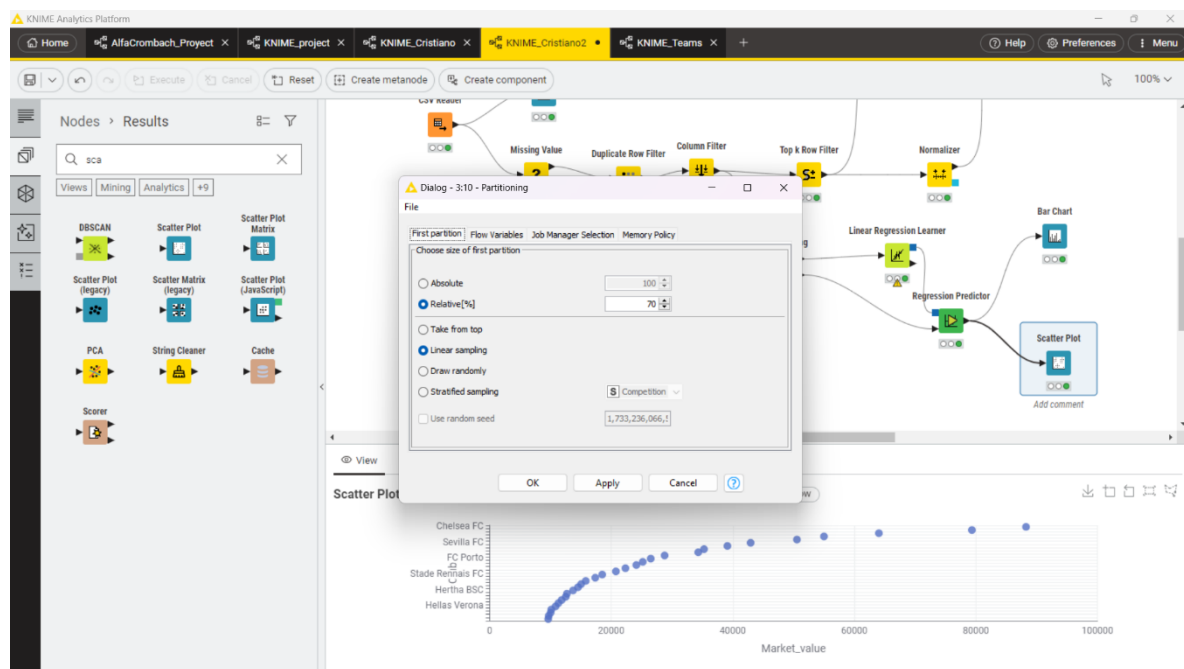
### Paso 5: Visualización y Análisis

1. **Visualizar los resultados:** Utilizamos nodos de visualización como Scatter Plot para visualizar las predicciones del modelo frente a los valores reales.

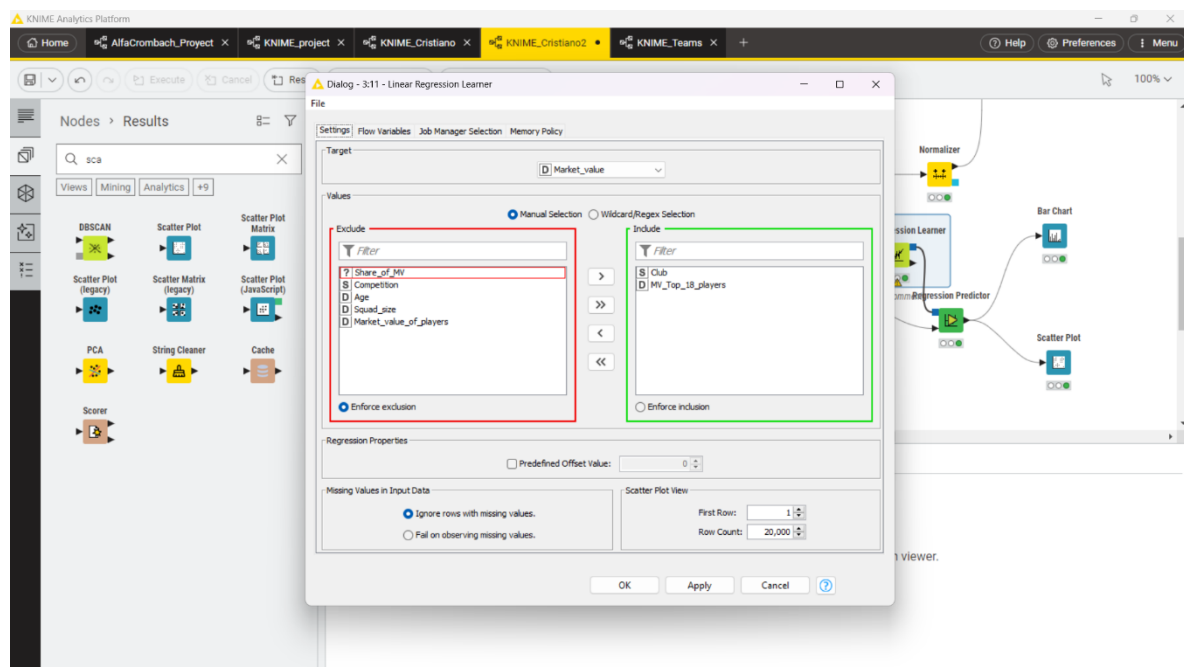




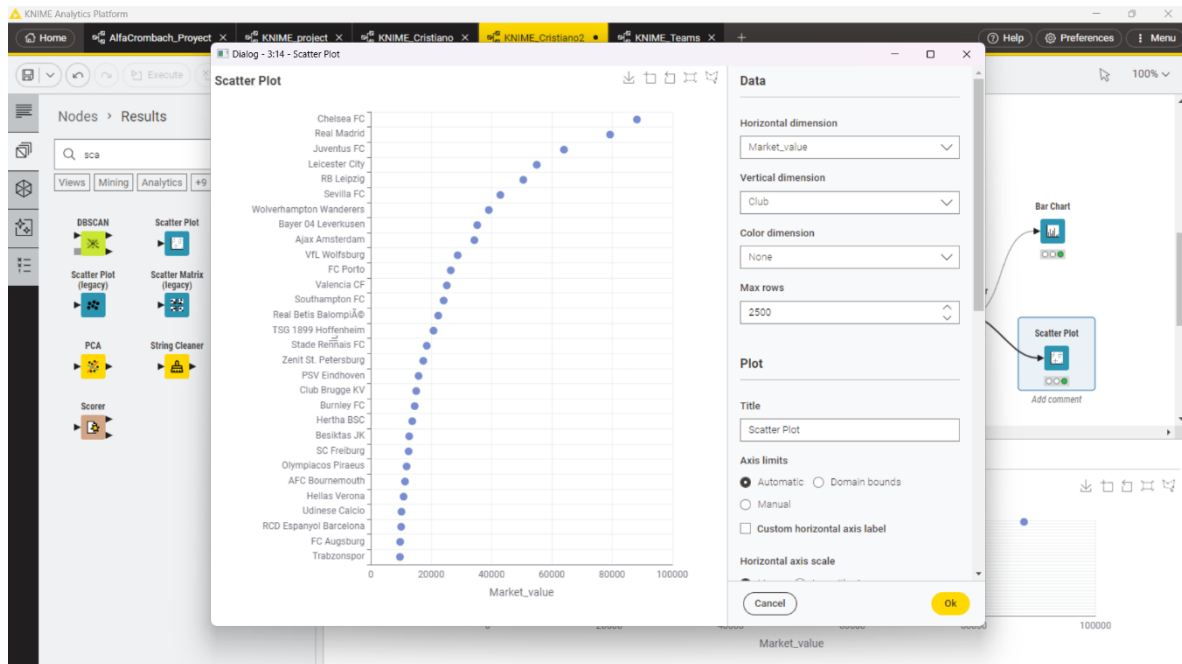
## Configuración de los nodos



Mostramos la configuración del partitioning y la configuraremos en un 100/80 esto para obtener una buena cantidad de datos para el entrenamiento



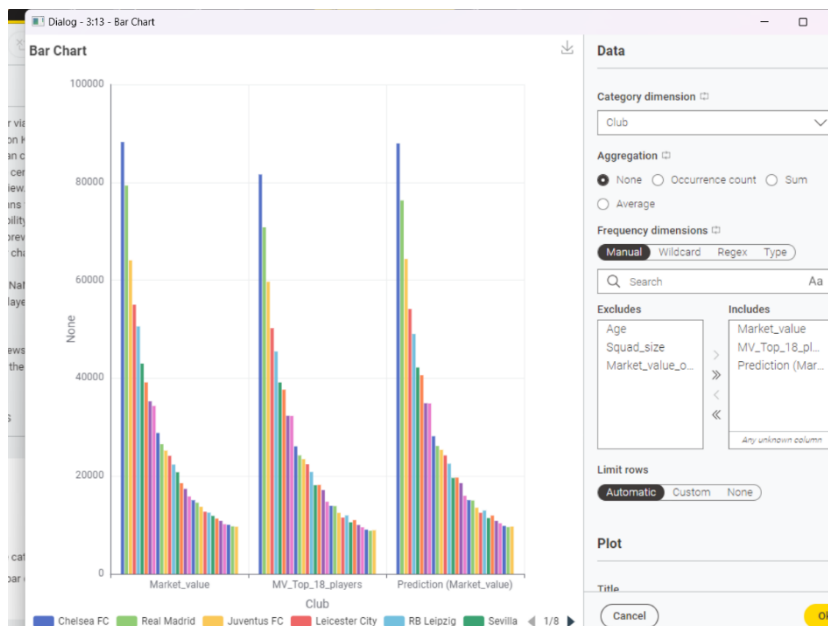
Configuramos el linnear regression learner y excluimos los datos que no queremos valorar



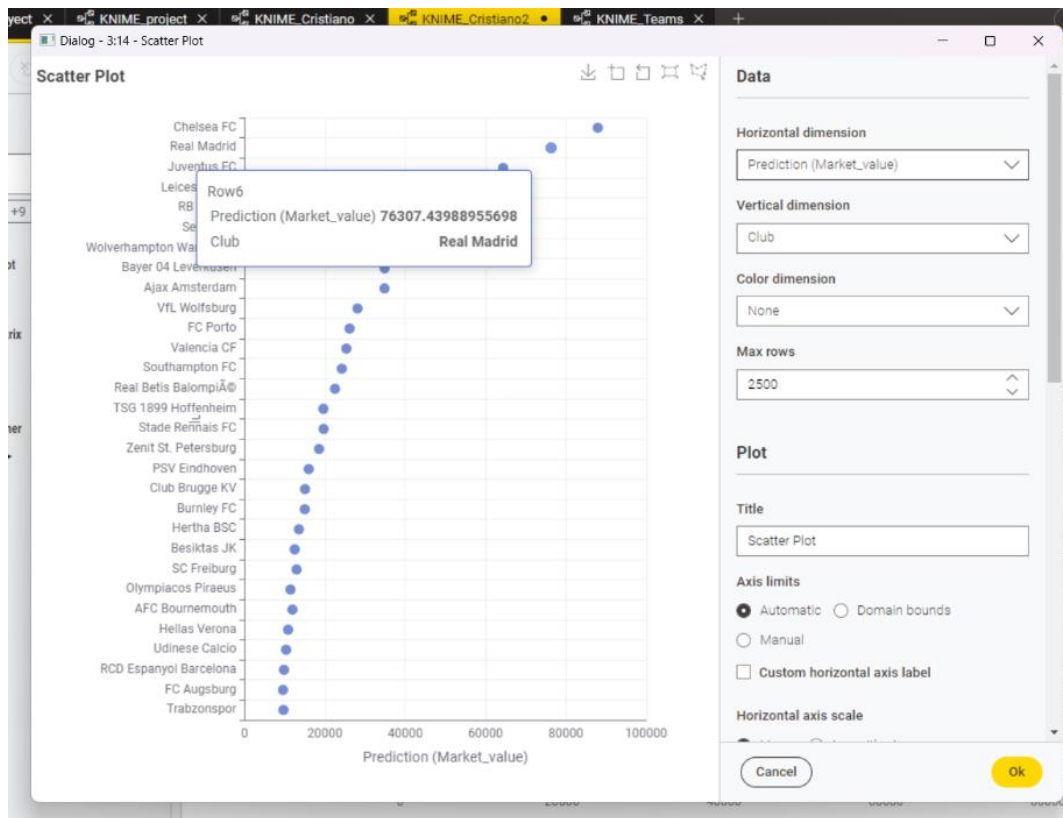
Por último, configuramos la herramienta que utilizaremos para la visualización en este caso utilice Bar Char y la de la imagen Scatter plott ya que con este se mira mejor la comparación entre la predicción y los datos reales.

## Resultados:

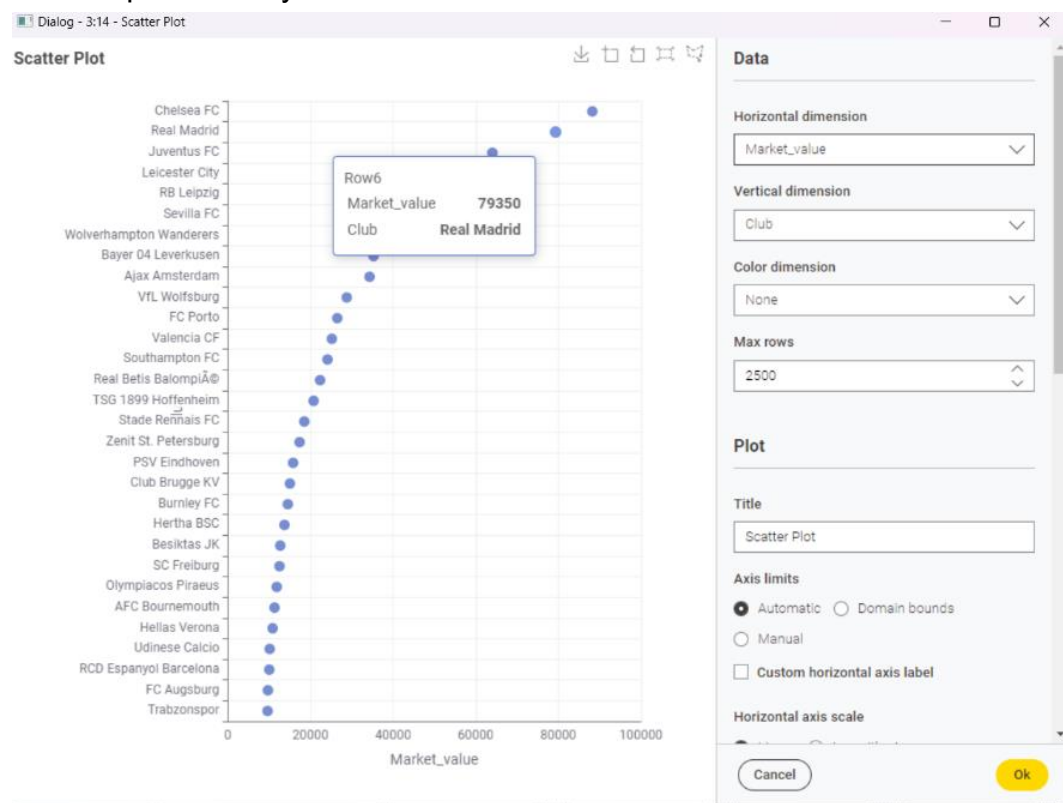
A continuación, mostrare las graficas que obtuve de las 2 visualizaciones



Como podemos observar en la gráfica se muestran 3 parámetros, valor de mercado, valor de mercado de jugadores y la predicción del valor de mercado, logra apreciarse una diferencia entre los 3 valores de mercado.



Ahora utilizando el scatter plot, podemos apreciar a detalle los cambios de la variable de predicción y los datos reales.



## **Conclusión**

Para concluir con este proyecto, hablare sobre mi experiencia y opinión al usar la aplicación y herramientas de Knime, me agrado bastante la interfaz y que es fácil de usar, además despertó un poco más mi interés sobre análisis y extracción de conocimiento de base de datos, espero poder seguir aprendiendo más sobre esta área y poder utilizar más learner y terminar el modelo CRISP en práctica.