ANÁLISIS DE DATOS

PROYECTO VGSALES & WHEAT SEEDS

https://www.kaggle.com/gregorut/videogamesales
https://www.kaggle.com/dongeorge/seed-from-uci

GRUPO 9:

JUAN MIGUEL BLANCO FERREIRA

ALEJANDRO GUERRERO DÍAZ

CÉSAR GARCÍA PASCUAL

ALFONSO BRAVO LLANOS

{juablafer,aleguedia,cesgarpas,alfbralla}@alum.us.es

MÁSTER EN INGENIERÍA DEL SOFTWARE:
CLOUD, DATOS Y GESTIÓN TI

FUNDAMENTOS DE INGENIERÍA DE DATOS (FID)

CURSO 2020-2021

ÍNDICE

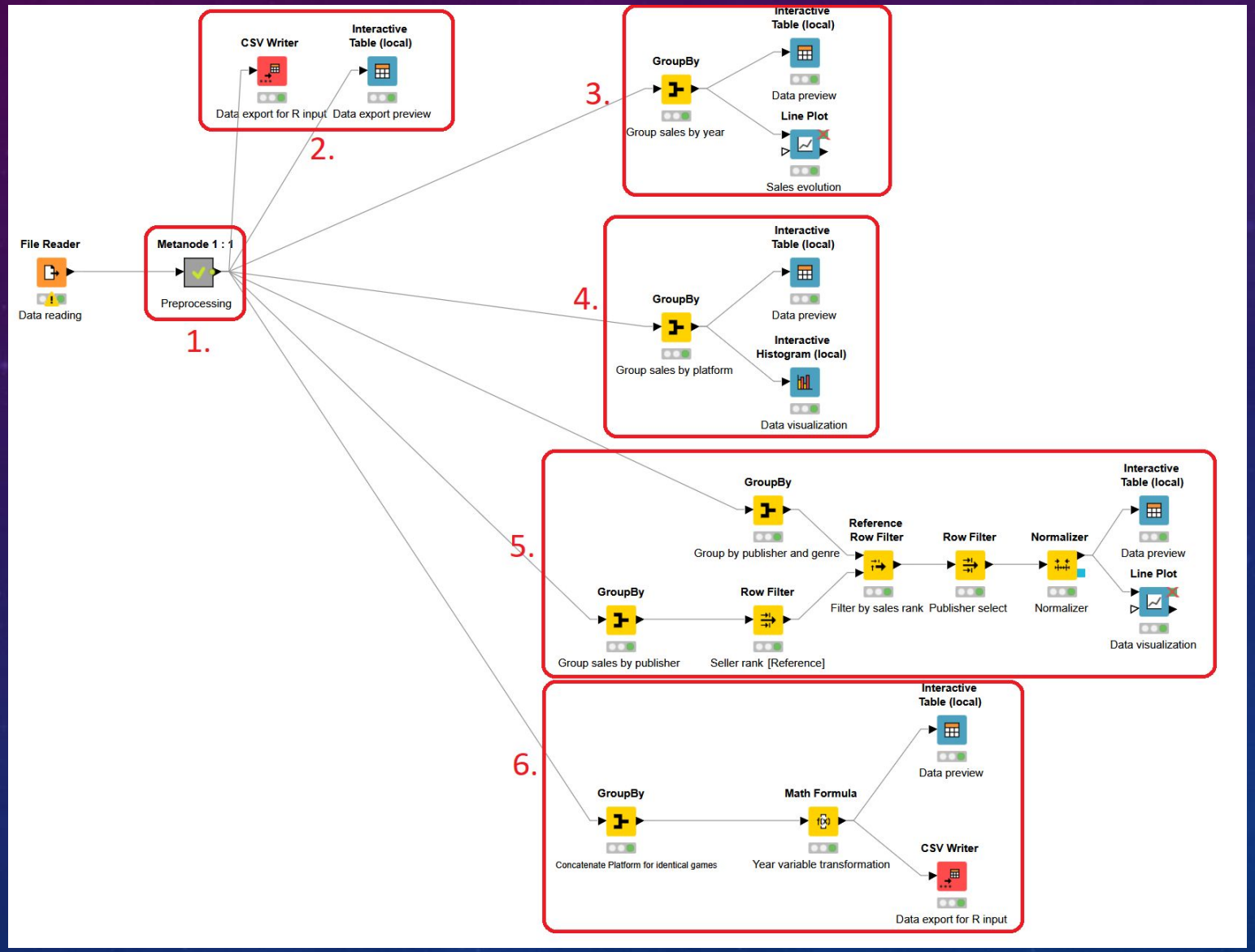
1. DATASET - VGSALES

	A	B	C	D	E	F	G	H	I	J	K
1	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sa
2	1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
3	2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
4	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
5	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	
6	5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1	31.37
7	6	Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26
8	7	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.38	9.23	6.5	2.9	30.01
9	8	Wii Play	Wii	2006	Misc	Nintendo	14.03	9.2	2.93	2.85	29.02
10	9	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.59	7.06	4.7	2.26	28.62
11	10	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31
12	11	Nintendogs	DS	2005	Simulation	Nintendo	9.07	11	1.93	2.75	24.76
13	12	Mario Kart DS	DS	2005	Racing	Nintendo	9.81	7.57	4.13	1.92	23.42
14	13	Pokemon Gold/Pokemon Silver	GB	1999	Role-Playing	Nintendo	9	6.18	7.2	0.71	23.1
15	14	Wii Fit	Wii	2007	Sports	Nintendo	8.94	8.03	3.6	2.15	22.72
16	15	Wii Fit Plus	Wii	2009	Sports	Nintendo	9.09	8.59	2.53	1.79	
17	16	Kinect Adventures!	X360	2010	Misc	Microsoft Ga	14.97	4.94	0.24	1.67	21.82
18	17	Grand Theft Auto V	PS3	2013	Action	Take-Two Int	7.01	9.27	0.97	4.14	21.4
19	18	Grand Theft Auto: San Andreas	PS2	2004	Action	Take-Two Int	9.43	0.4	0.41	10.57	20.81
20	19	Super Mario World	SNES	1990	Platform	Nintendo	12.78	3.75	3.54	0.55	20.61
21	20	Brain Age: Train Your Brain in Minutes a Day	DS	2005	Misc	Nintendo	4.75	9.26	4.16	2.05	20.22
22	21	Pokemon Diamond/Pokemon Pearl	DS	2006	Role-Playing	Nintendo	6.42	4.52	6.04	1.37	18.36
23	22	Super Mario Land	GB	1989	Platform	Nintendo	10.83	2.71	4.18	0.42	18.14

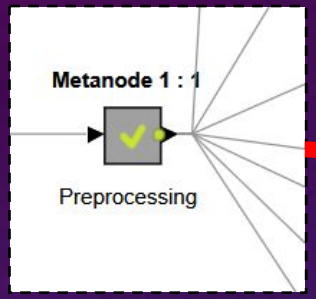
1. DATASET - WHEAT SEEDS

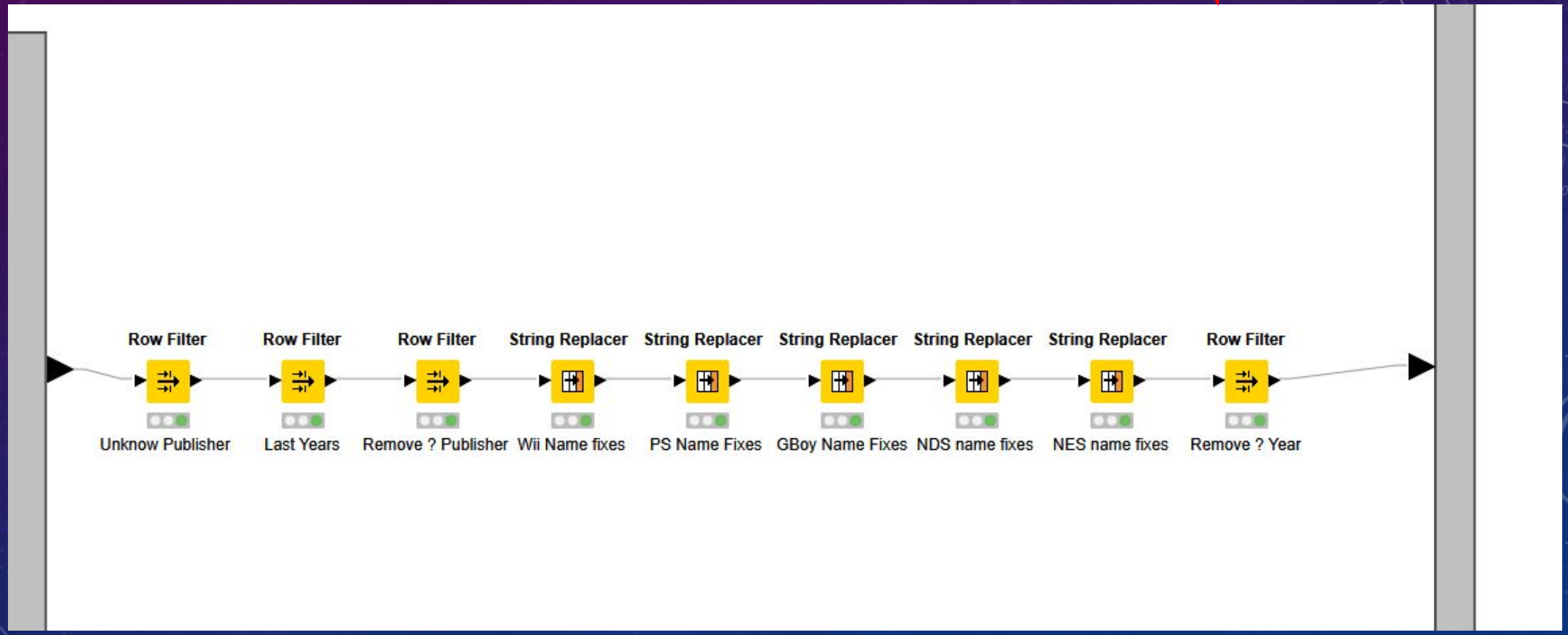
1	A	P	C	LK	WK	A_Coef	LKG	target
2	15.26	14.84	0.871	5.763	3.312	2.221	5.22	0
3	14.88	14.57	0.8811	5.554	3.333	1.018	4.956	0
4	14.29	14.09	0.905	5.291	3.337	2.699	4.825	0
5	13.84	13.94	0.8955	5.324	3.379	2.259	4.805	0
6	16.14	14.99	0.9034	5.658	3.562	1.355	5.175	0
7	14.38	14.21	0.8951	5.386	3.312	2.462	4.956	0
8	14.69	14.49	0.8799	5.563	3.259	3.586	5.219	0
9	14.11	14.1	0.8911	5.42	3.302	2.7	5	0
10	16.63	15.46	0.8747	6.053	3.465	2.04	5.877	0
11	16.44	15.25	0.888	5.884	3.505	1.969	5.533	0
12	15.26	14.85	0.8696	5.714	3.242	4.543	5.314	0
13	14.03	14.16	0.8796	5.438	3.201	1.717	5.001	0
14	13.89	14.02	0.888	5.439	3.199	3.986	4.738	0
15	13.78	14.06	0.8759	5.479	3.156	3.136	4.872	0
16	13.74	14.05	0.8744	5.482	3.114	2.932	4.825	0
17	14.59	14.28	0.8993	5.351	3.333	4.185	4.781	0
18	13.99	13.83	0.9183	5.119	3.383	5.234	4.781	0

2. PREPROCESAMIENTO

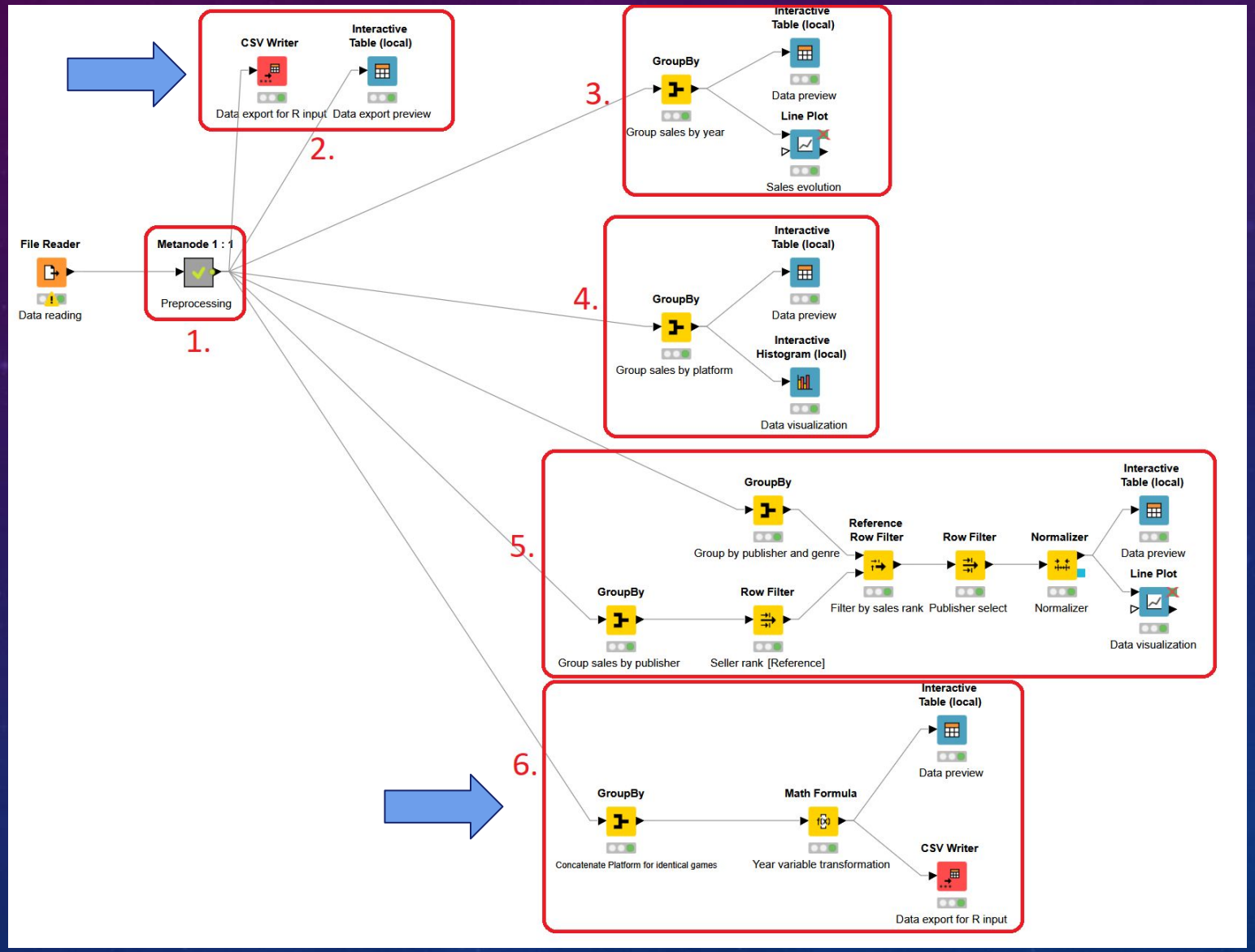


2. PREPROCESAMIENTO

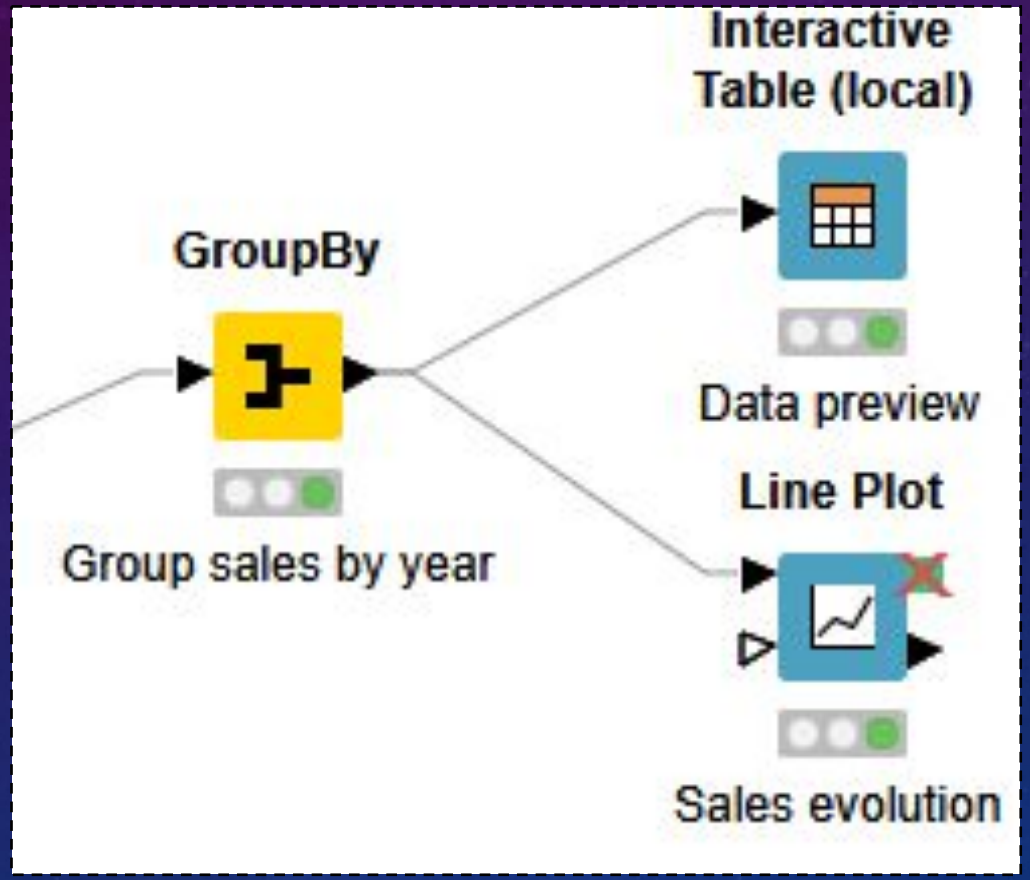




2. PREPROCESAMIENTO - EXPORTADO

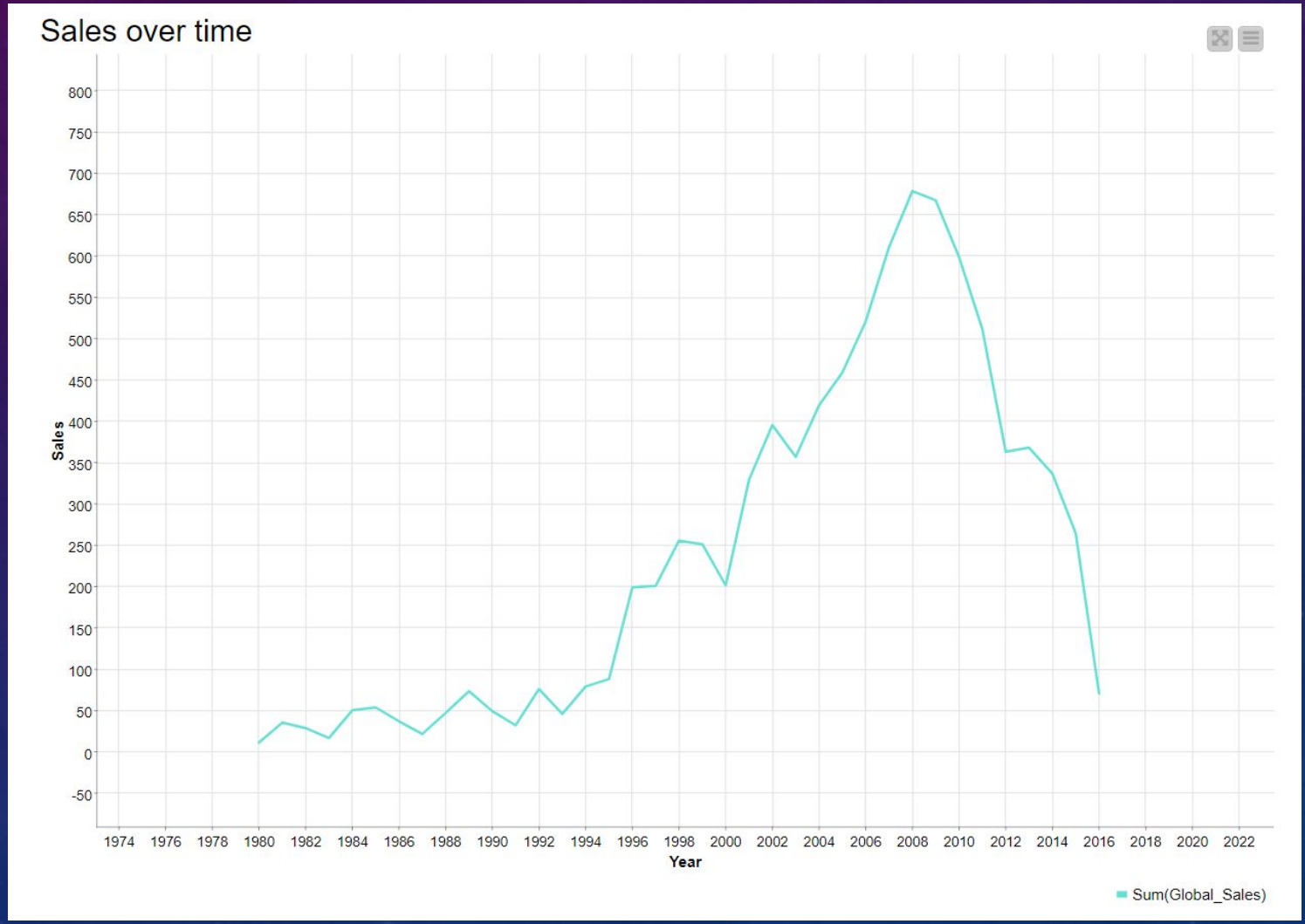


3. VISUALIZACIÓN - KNIME

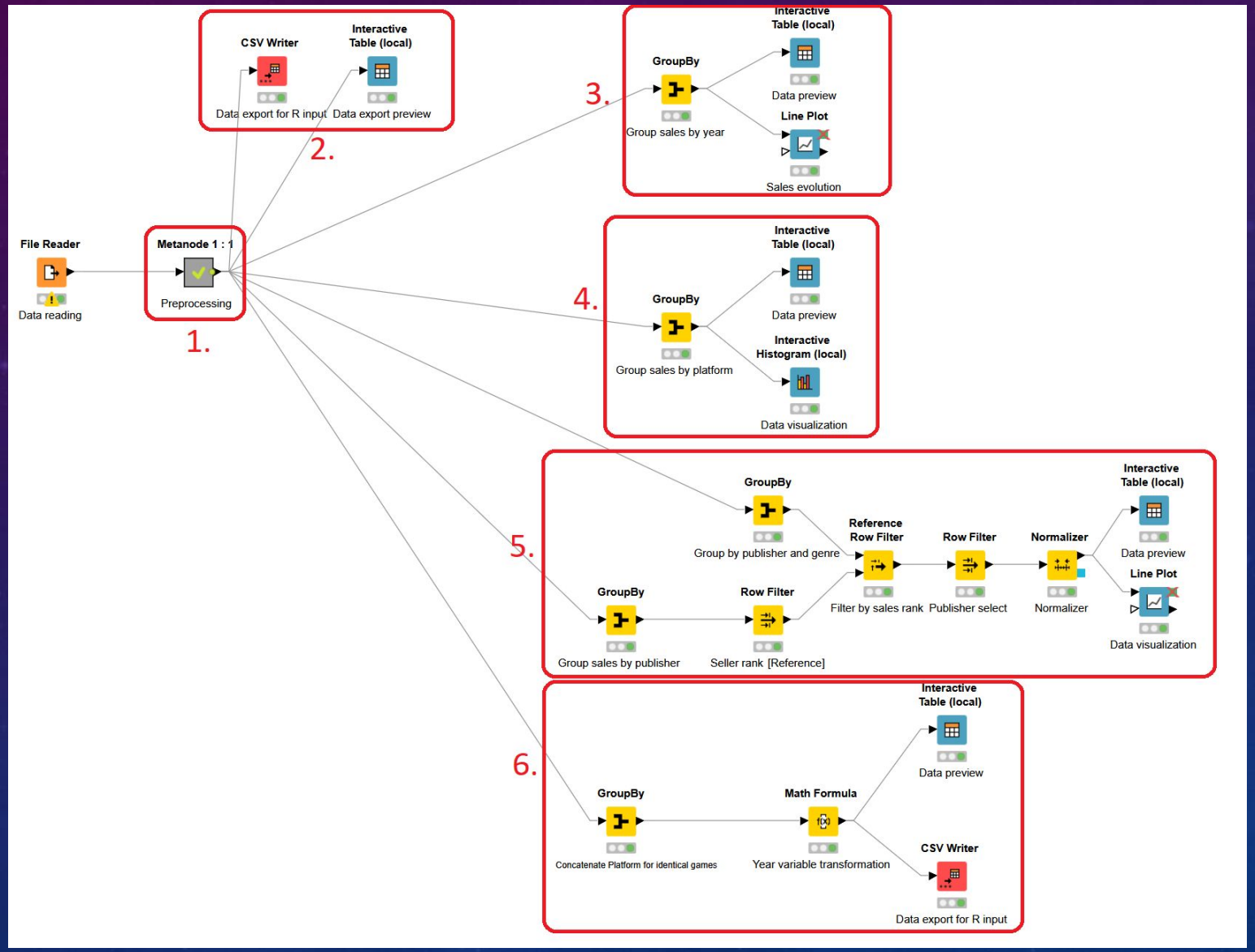


3. VISUALIZACIÓN - KNIME

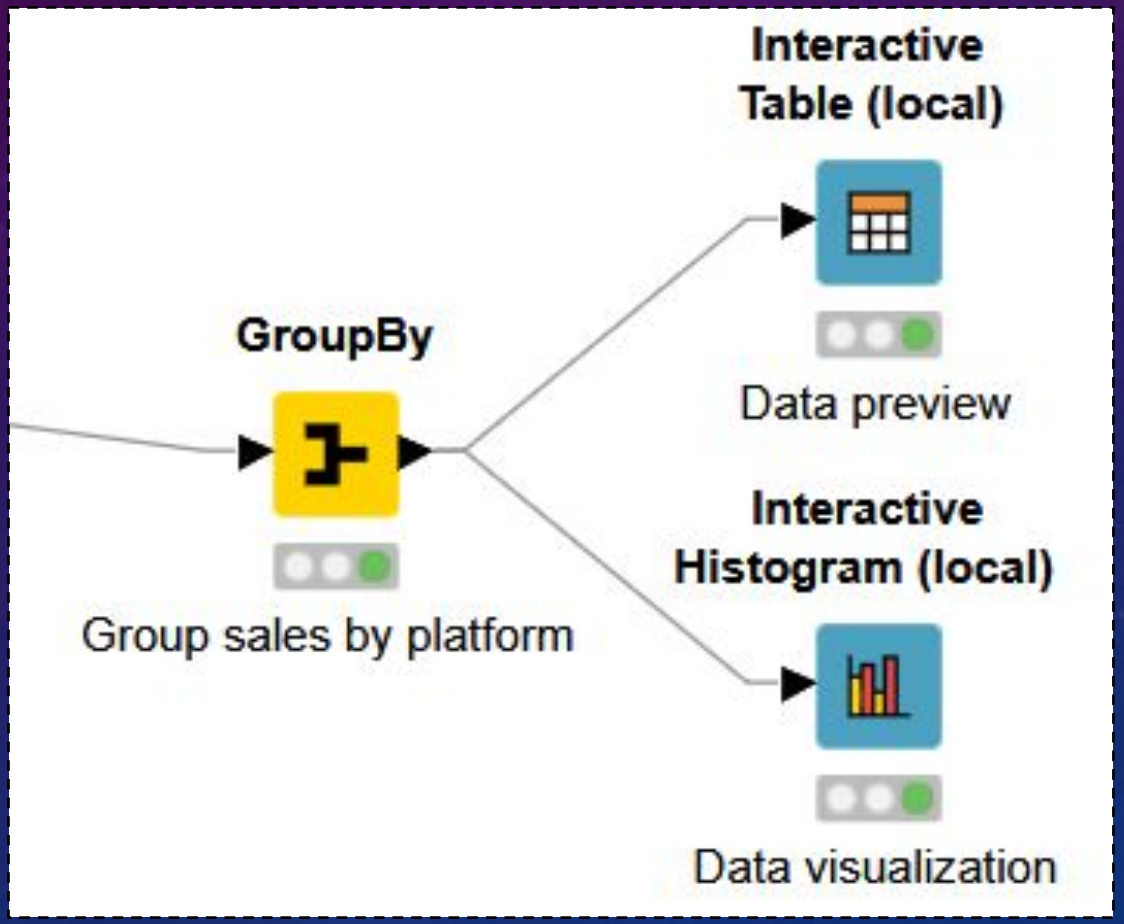
VENTAS DE JUEGO POR AÑO



3. VISUALIZACIÓN - KNIME

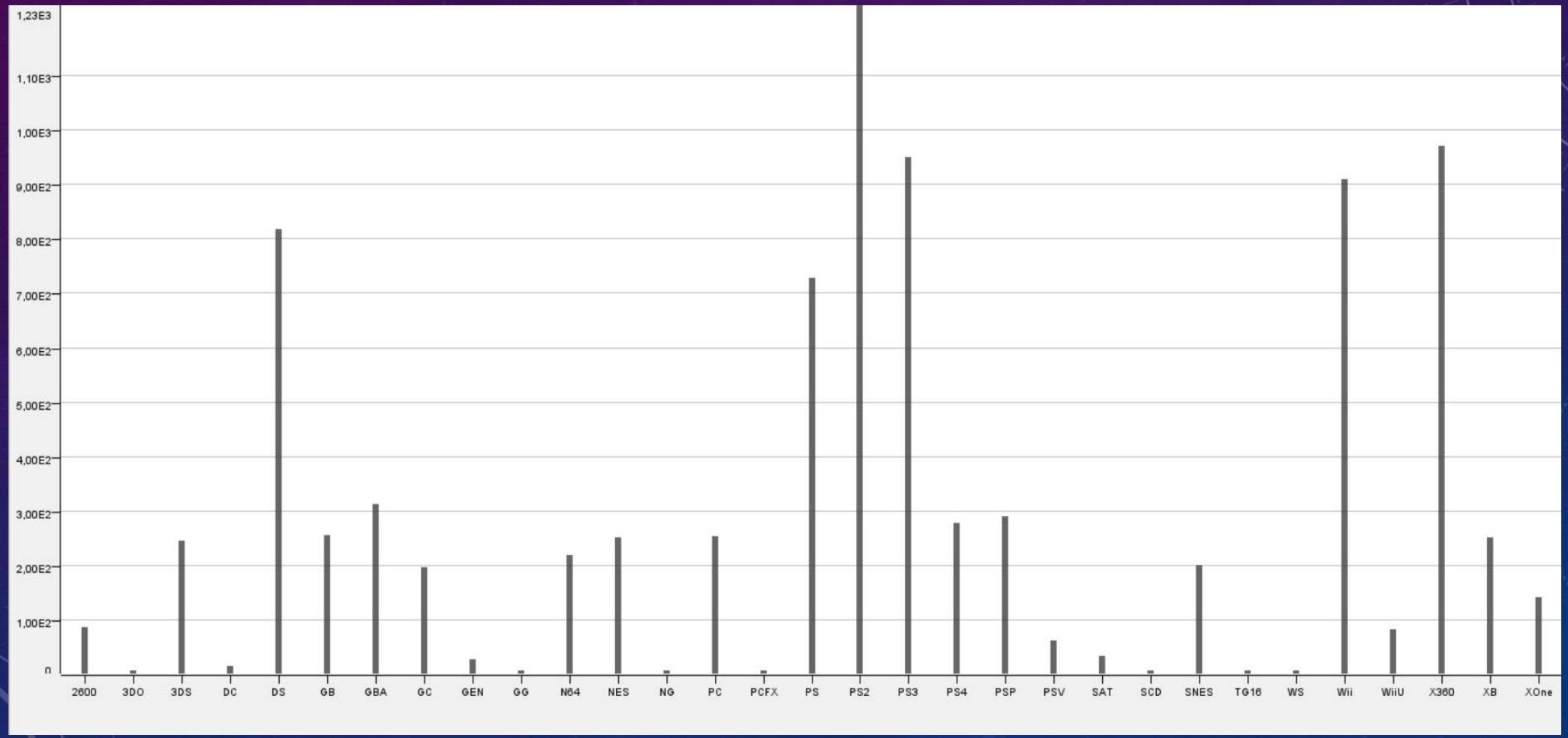


3. VISUALIZACIÓN - KNIME

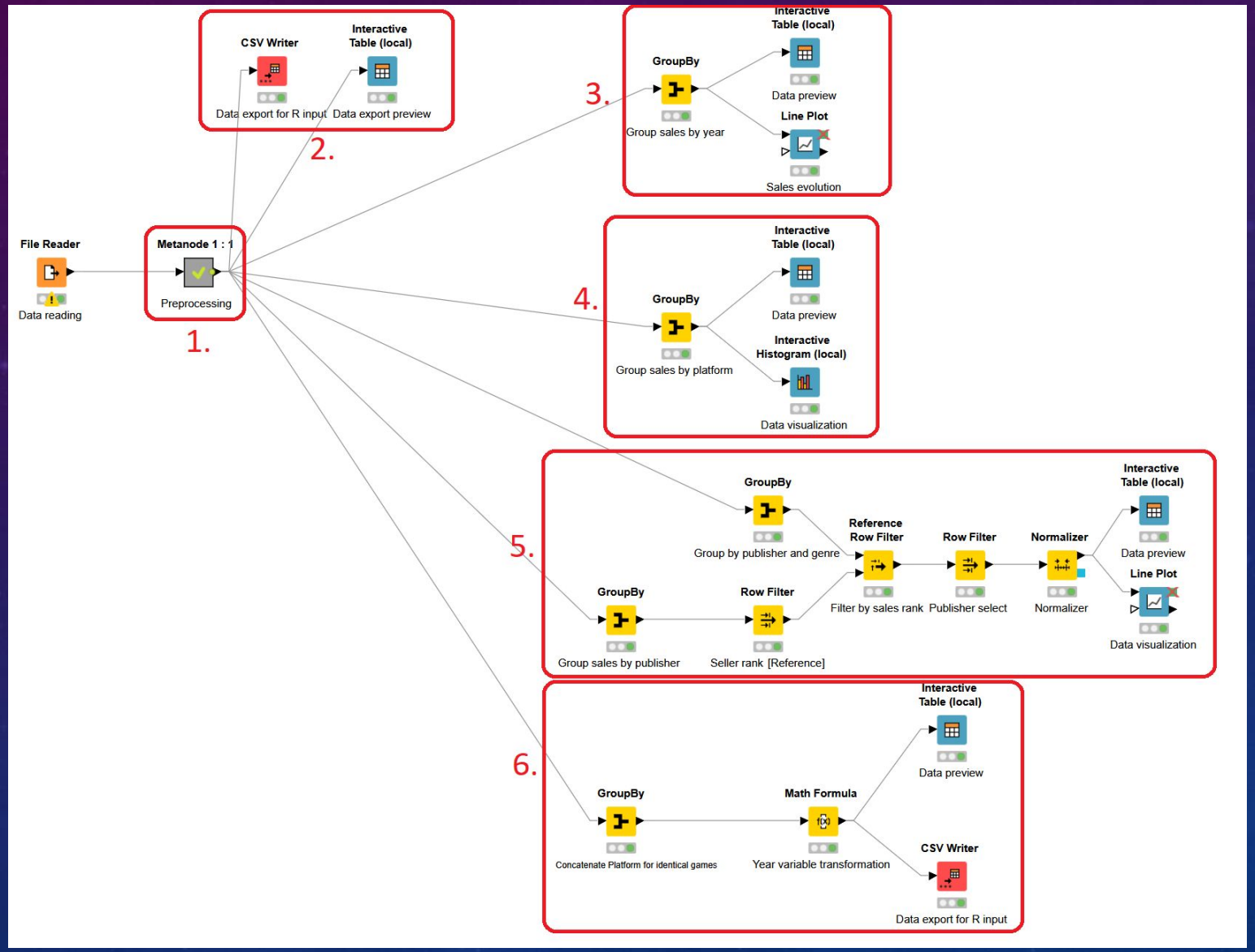


3. VISUALIZACIÓN - KNIME

VENTAS DE JUEGOS SEGÚN PLATAFORMA



3. VISUALIZACIÓN - KNIME



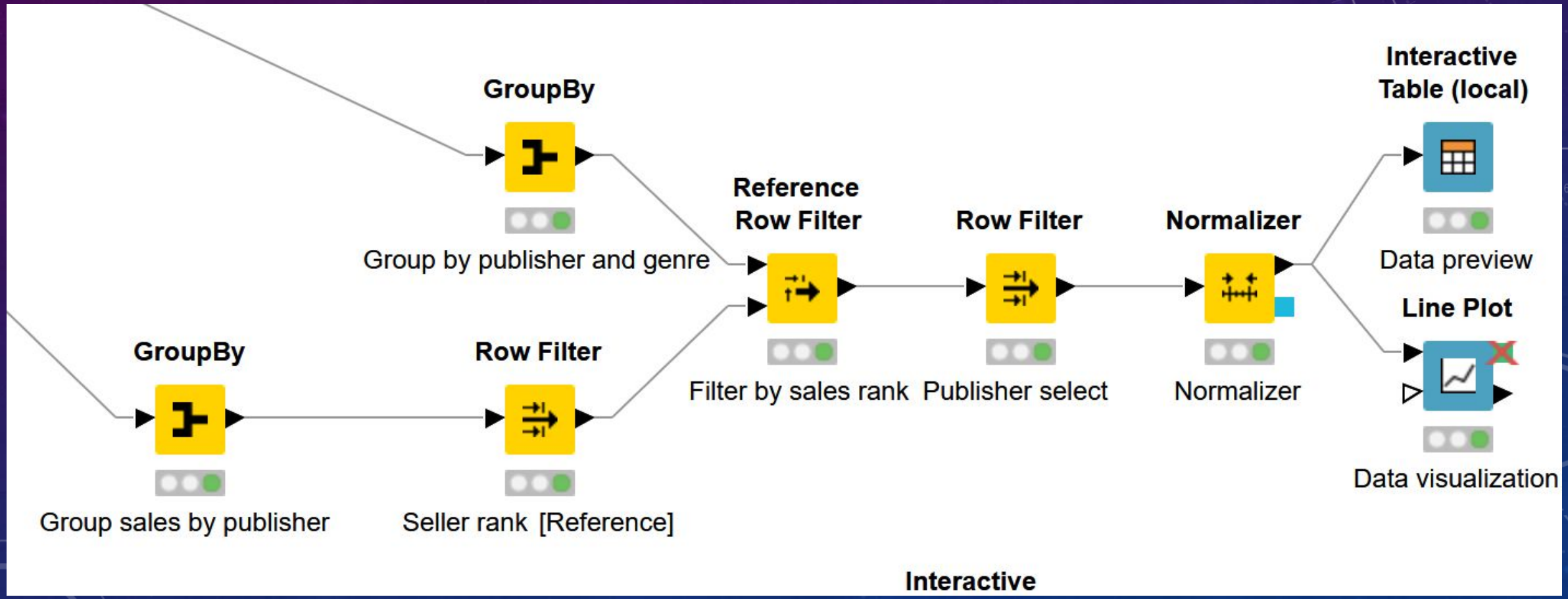
3. VISUALIZACIÓN - KNIME

ANALIZANDO VENTAS POR CATEGORÍA

EJEMPLO: NINTENDO

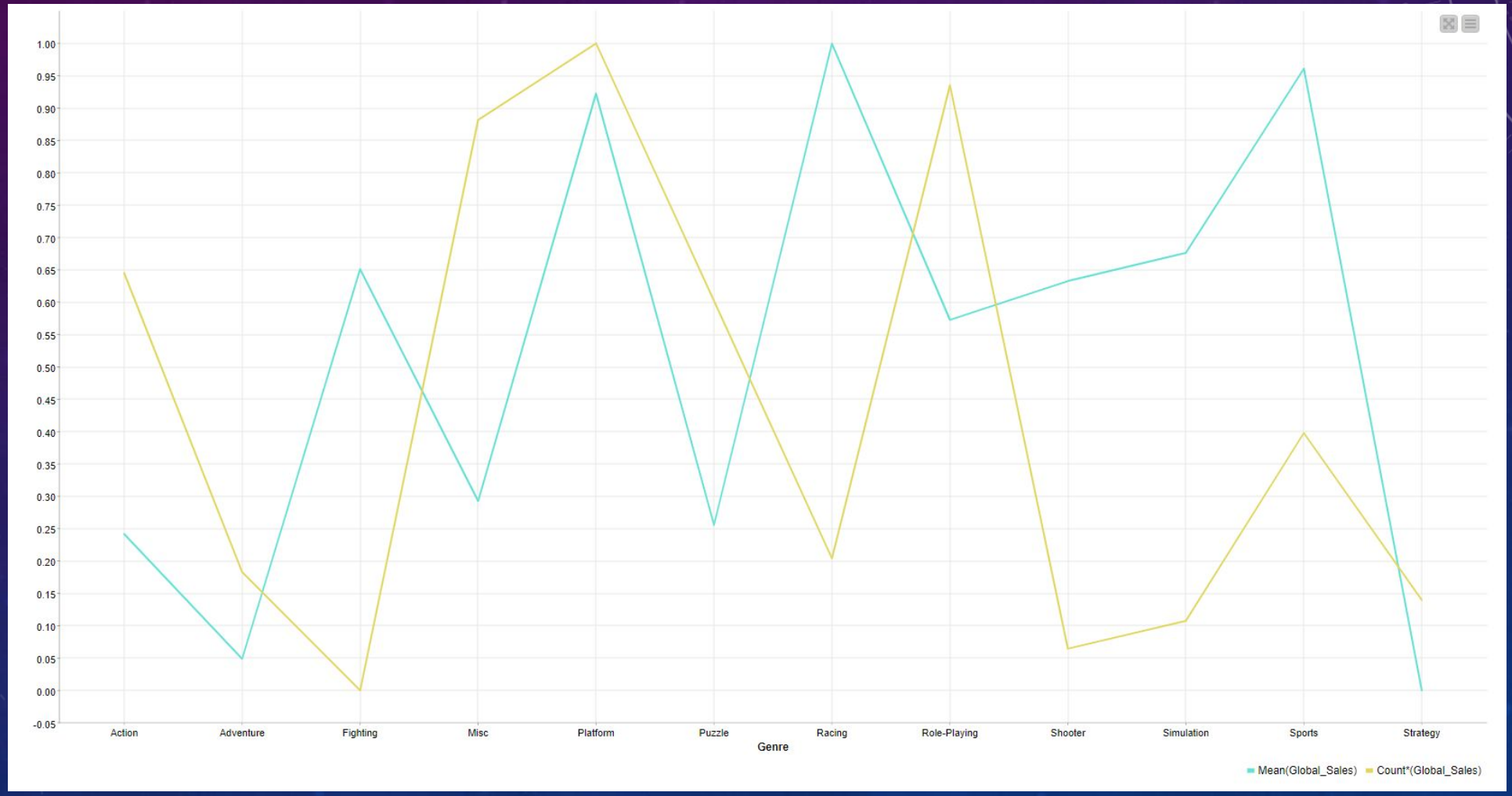
3. VISUALIZACIÓN - KNIME

ANÁLISIS DE VENTAS POR CATEGORÍA

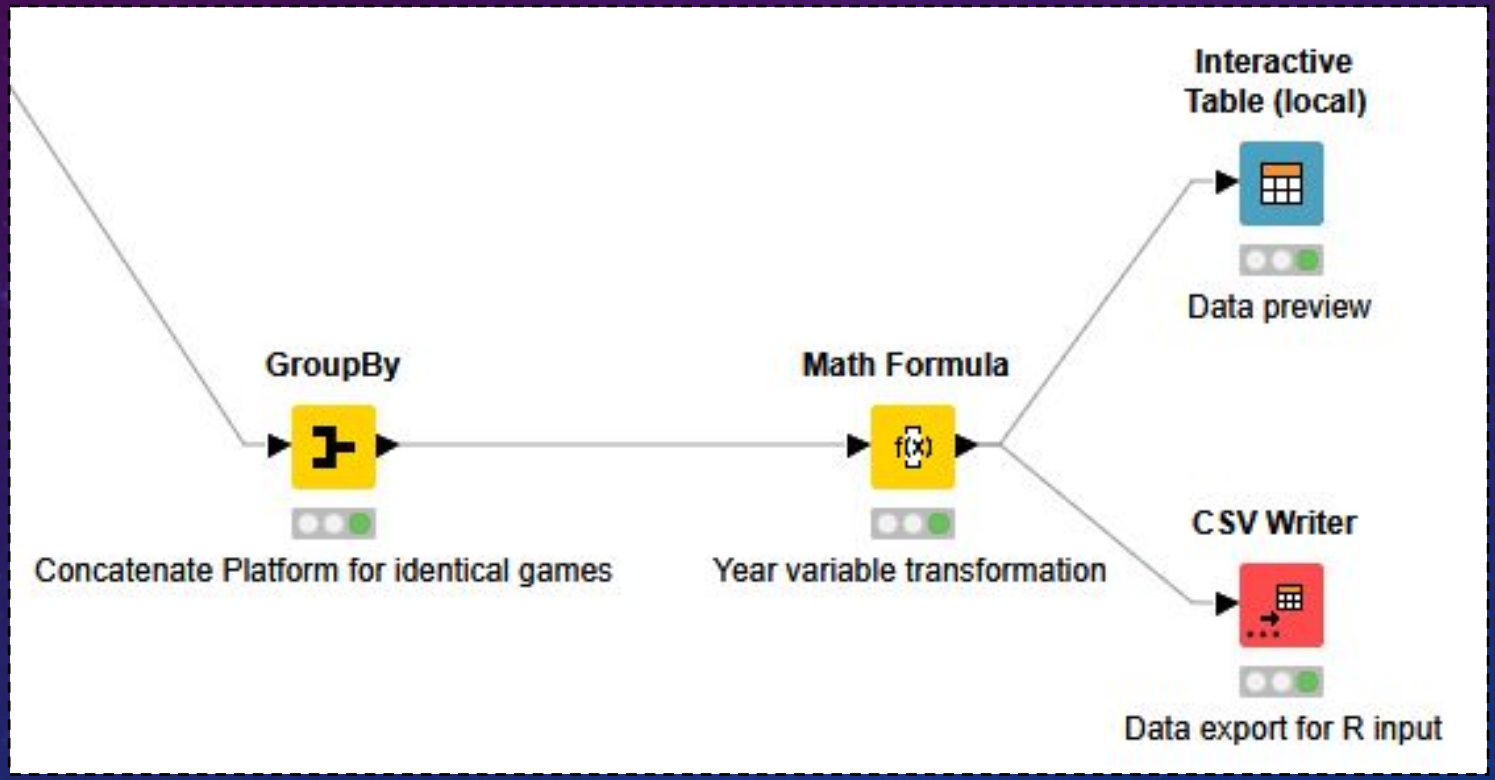


3. VISUALIZACIÓN - KNIME

ANÁLISIS DE VENTAS POR CATEGORÍA



3. VISUALIZACIÓN - KNIME



2. DATASET PREPROCESADO

	A	B	C	D	E	F	G	H	I	J	K
1	Name	Concatenate(Platform)	Min*(Year)	First(Genre)	First(Publisher)	Sum(NA_Sale	Sum(EU_Sale	Sum(JP_Sales	Sum(Other_S	Sum(Global_S	YearCount
2	'98 Koshien	PSX	1998	Sports	Magical Company	0.15	0.1	0.12	0.03	0.41	23
3	.hack//G.U. Vol.1//Rebirth	PS2	2006	Role-Playing	Namco Bandai Gam	0	0	0.17	0	0.17	15
4	.hack//G.U. Vol.2//Reminisce	PS2	2006	Role-Playing	Namco Bandai Gam	0.11	0.09	0	0.03	0.23	15
5	.hack//G.U. Vol.2//Reminisce (jp sales)	PS2	2006	Role-Playing	Namco Bandai Gam	0	0	0.16	0	0.16	15
6	.hack//G.U. Vol.3//Redemption	PS2	2007	Role-Playing	Namco Bandai Gam	0	0	0.17	0	0.17	14
7	.hack//Infection Part 1	PS2	2002	Role-Playing	Atari	0.49	0.38	0.26	0.13	1.27	19
8	.hack//Link	PSP	2010	Role-Playing	Namco Bandai Gam	0	0	0.14	0	0.14	11
9	.hack//Mutation Part 2	PS2	2002	Role-Playing	Atari	0.23	0.18	0.2	0.06	0.68	19
10	.hack//Outbreak Part 3	PS2	2002	Role-Playing	Atari	0.14	0.11	0.17	0.04	0.46	19
11	.hack//Quarantine Part 4: The Final Chapter	PS2	2003	Role-Playing	Atari	0.09	0.07	0	0.02	0.18	18
12	.hack: Sekai no Mukou ni + Versus	PS3	2012	Action	Namco Bandai Gam	0	0	0.03	0	0.03	9
13	007 Racing	PSX	2000	Racing	Electronic Arts	0.3	0.2	0	0.03	0.53	21
14	007: Quantum of Solace	X360, PS3, Wii1, PS2, NDS, F	2008	Action	Activision	1.84	1.35	0.04	0.68	3.92	13
15	007: The World is not Enough	N64, PSX	2000	Action	Electronic Arts	1.64	0.73	0.02	0.09	2.47	21
16	007: Tomorrow Never Dies	PSX	1999	Shooter	Electronic Arts	1.72	1.33	0	0.16	3.21	22
17	1 vs. 100	NDS	2008	Misc	DSI Games	0.08	0	0	0.01	0.09	13
18	1/2 Summer +	PSP	2013	Adventure	Kaga Create	0	0	0.01	0	0.01	8
19	10 Minute Solution	Wii1	2010	Sports	Activision	0.06	0.01	0	0.01	0.08	11
20	100 All-Time Favorites	NDS	2009	Puzzle	Ubisoft	0.35	0.12	0	0.04	0.51	12
21	100 Classic Books	NDS	2008	Misc	Nintendo	0.13	0.52	0	0.02	0.67	13
22	100 Classic Games	NDS	2011	Misc	Rondomedia	0	0.03	0	0	0.04	10
23	1000 Cooking Recipes from ELLE Ã table	NDS	2010	Misc	Nintendo	0	0.02	0	0	0.03	11
24	1001 Touch Games	NDS	2011	Action	Avanquest	0.12	0.17	0	0.04	0.33	10
25	101-in-1 Explosive Megamix	NDS	2008	Puzzle	Nordcurrent	0.05	0.13	0	0.02	0.2	13
26	101-in-1 Party Megamix Wii	Wii1	2009	Misc	Nordcurrent	0.19	0.01	0	0.02	0.23	12
27	101-in-1 Sports Megamix	NDS	2010	Sports	Nordcurrent	0.08	0	0	0.01	0.08	11
28	101-in-1 Sports Party Megamix	Wii1	2010	Sports	Nordcurrent	0.02	0	0	0	0.03	11
29	1080Â°: TenEighty Snowboarding	N64	1998	Sports	Nintendo	1.25	0.61	0.13	0.05	2.03	23
30	11eyes: CrossOver	X360, PSP	2009	Adventure	5pb	0	0	0.04	0	0.04	12
31	12-Sai. Honto no Kimochi	3DS	2014	Adventure	Happinet	0	0	0.07	0	0.07	7
32	12-Sai. Koisuru Diary	3DS	2016	Adventure	Happinet	0	0	0.04	0	0.04	5

4. ANÁLISIS SUPERVISADO

1. RPART
2. NAIVE BAYES
3. CTREE2

4. ANÁLISIS SUPERVISADO

F	G	H	I	J
Sum(NA_Sale	Sum(EU_Sale	Sum(JP_Sales	Sum(Other_S	Sum(Global_S
0.15	0.1	0.12	0.03	0.41
0	0	0.17	0	0.17
0.11	0.09	0	0.03	0.23
0	0	0.16	0	0.16
0	0	0.17	0	0.17
0.49	0.38	0.26	0.13	1.27
0	0	0.14	0	0.14
0.23	0.18	0.2	0.06	0.68
0.14	0.11	0.17	0.04	0.46
0.09	0.07	0	0.02	0.18
0	0	0.03	0	0.03
0.3	0.2	0	0.03	0.53
1.84	1.35	0.04	0.68	3.92
1.64	0.73	0.02	0.09	2.47
1.72	1.33	0	0.16	3.21
0.08	0	0	0.01	0.09
0	0	0.01	0	0.01
0.06	0.01	0	0.01	0.08

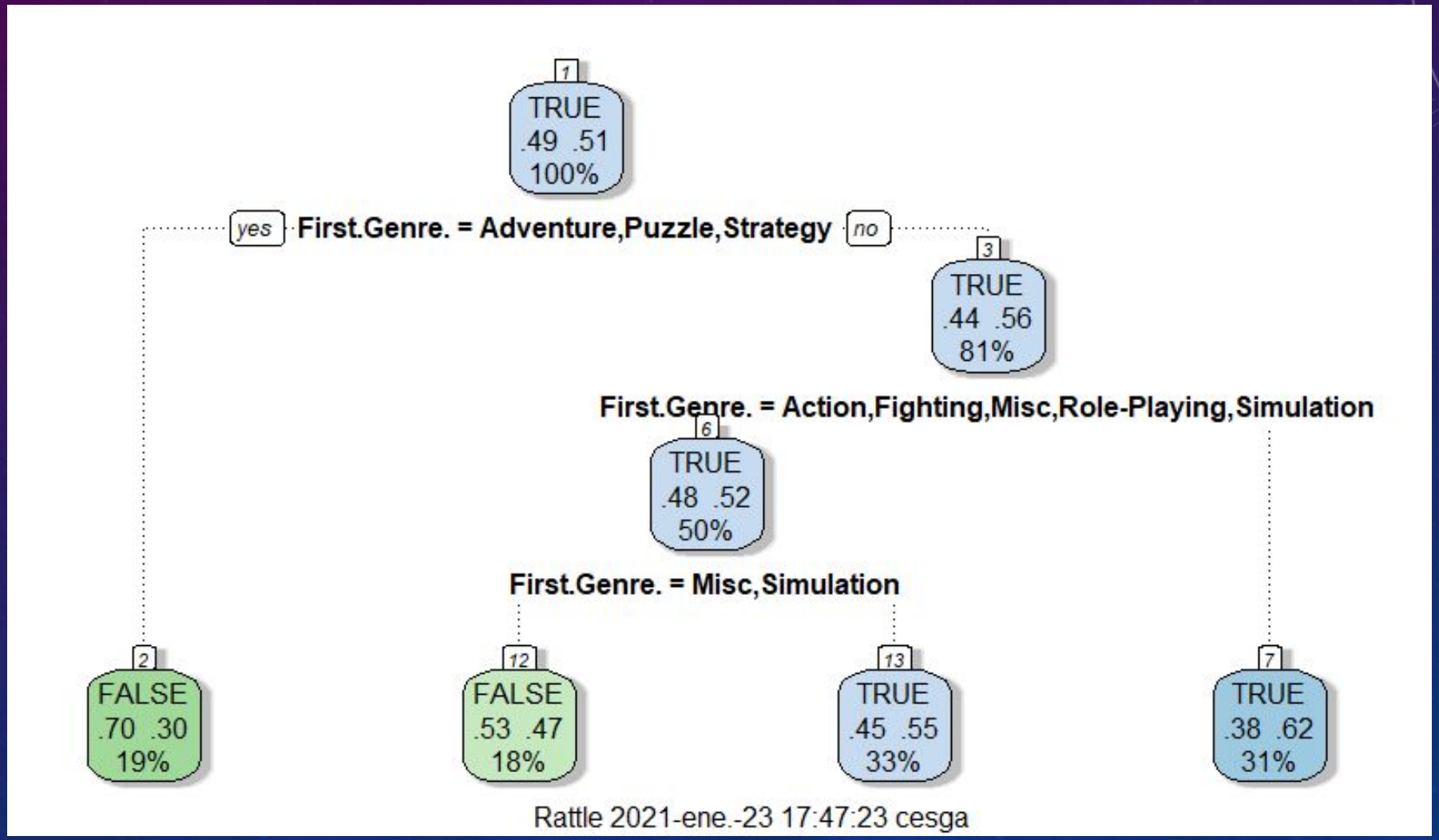
4.1. ANÁLISIS SUPERVISADO - RPart

Género

	C	D	E	F	
	Min*(Year)	First(Genre)	First(Publisher)	Sum(NA_Sale	Sum(
	1998	Sports	Magical Company	0.15	0.1
	2006	Role-Playing	Namco Bandai Gam	0	
	2006	Role-Playing	Namco Bandai Gam	0.11	0.09
	2006	Role-Playing	Namco Bandai Gam	0	
	2007	Role-Playing	Namco Bandai Gam	0	
	2002	Role-Playing	Atari	0.49	0.38
	2010	Role-Playing	Namco Bandai Gam	0	
	2002	Role-Playing	Atari	0.23	0.18
	2002	Role-Playing	Atari	0.14	0.11
	2003	Role-Playing	Atari	0.09	0.07
	2012	Action	Namco Bandai Gam	0	
	2000	Racing	Electronic Arts	0.3	0.2
S, F	2008	Action	Activision	1.84	1.35
	2000	Action	Electronic Arts	1.64	0.73
	1999	Shooter	Electronic Arts	1.72	1.33
	2008	Misc	DSI Games	0.08	
	2013	Adventure	Kaga Create	0	
	2010	Sports	Activision	0.06	0.01

4.1. ANÁLISIS SUPERVISADO - RPart

Género - Acc: 57,06%



4.1. ANÁLISIS SUPERVISADO - RPart

Plataforma

007 Racing	PSX
007: Quantum of Solace	X360, PS3, Wii1, PS2, NDS, F
007: The World is not Enough	N64, PSX
007: Tomorrow Never Dies	PSX
1 vs. 100	NDS
1/2 Summer +	PSP
10 Minute Solution	Wii1
100 All-Time Favorites	NDS
100 Classic Books	NDS
100 Classic Games	NDS
1000 Cooking Recipes from ELLE Ã table	NDS
1001 Touch Games	NDS
101-in-1 Explosive Megamix	NDS
101-in-1 Party Megamix Wii	Wii1
101-in-1 Sports Megamix	NDS
101-in-1 Sports Party Megamix	Wii1
1080Â°: TenEighty Snowboarding	N64
11eyes: CrossOver	X360, PSP
12-Sai. Honto no Kimochi	3DS
12-Sai. Koisuru Diary	3DS

4.1. ANÁLISIS SUPERVISADO - RPart

Plataforma

	Name	Concatenate.Platform.	Min..Year.	First.Genre.	First.Publisher.	Sum.NA_Sales.	Sum.EU_Sales.	Sum.JP_Sales.	Sum.Other_Sales.
1	Wii Sports	Wii1	2006	Sports	Nintendo	41.49	29.02	3.77	8.46
2	Grand Theft Auto V	PS3, X360, PS4, XOne, PC	2013	Action	Take-Two Interactive	23.46	23.04	1.39	8.03
3	Super Mario Bros.	Famicom, GBoy	1985	Platform	Nintendo	32.48	4.88	6.96	0.99
4	Tetris	GBoy, Famicom	1988	Puzzle	Nintendo	26.17	2.95	6.03	0.69
5	Mario Kart Wii	Wii1	2008	Racing	Nintendo	15.85	12.88	3.79	3.31
6	Wii Sports Resort	Wii1	2009	Sports	Nintendo	15.75	11.01	3.28	2.96

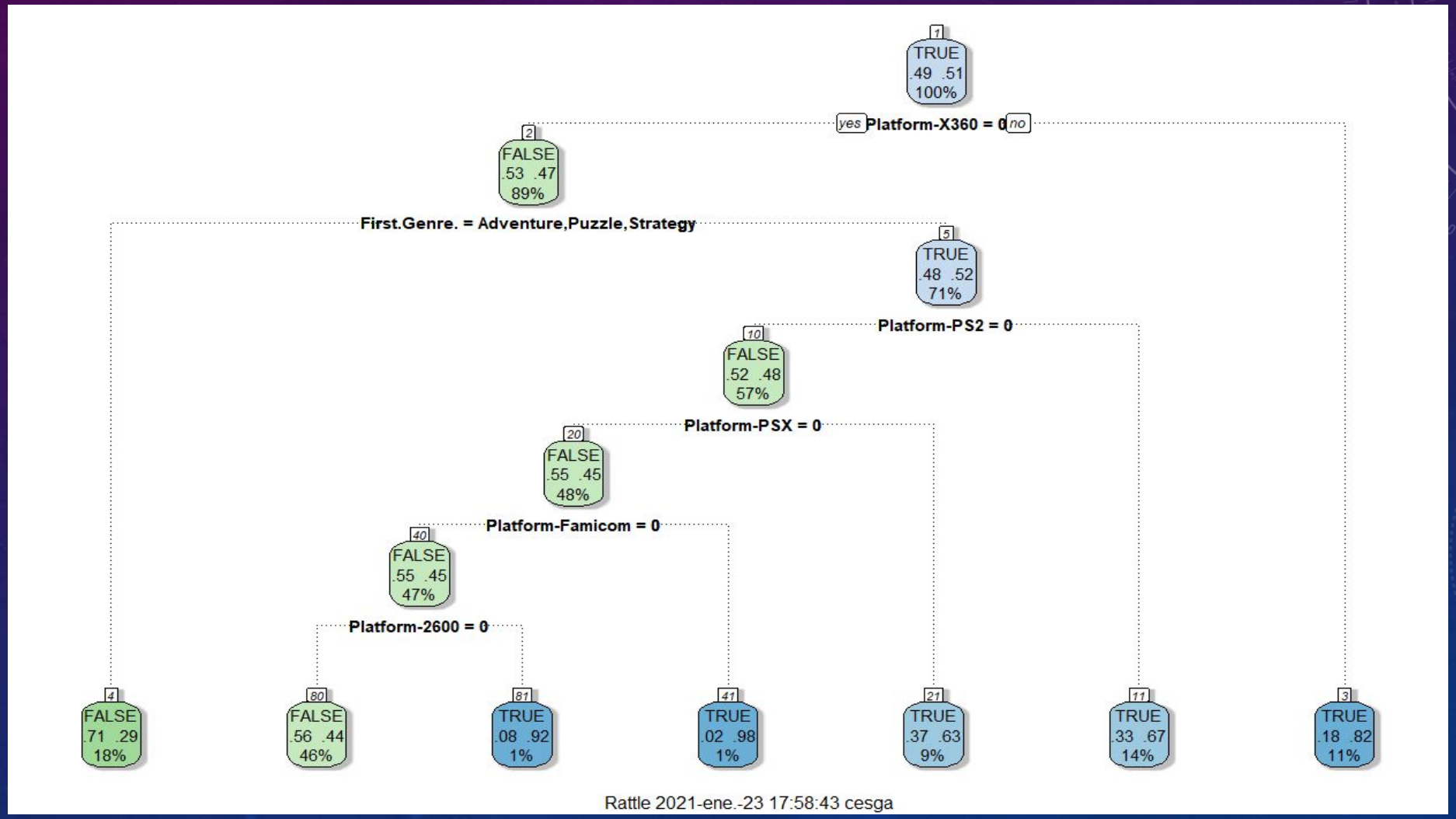
	Sum.Global_Sales.	YearCount	Supersale	Platform-Wii1	Platform-Famicom	Platform-GBoy	Platform-NDS	Platform-X360	Platform-PS3	Platform-PS2	Platform-SNES	Platform-GBA
1	82.74	15	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	55.92	8	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
3	45.31	36	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	35.84	33	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	35.82	13	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	33.00	12	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

	Platform-3DS	Platform-PS4	Platform-N64	Platform-PSX	Platform-XB	Platform-PC	Platform-2600	Platform-PSP	Platform-XOne	Platform-GC	Platform-WiiU	Platform-GEN
1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

	Platform-DC	Platform-PSV	Platform-SAT	Platform-SCD	Platform-WS	Platform-NG	Platform-TG16	Platform-3DO	Platform-GG	Platform-PCFX
1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

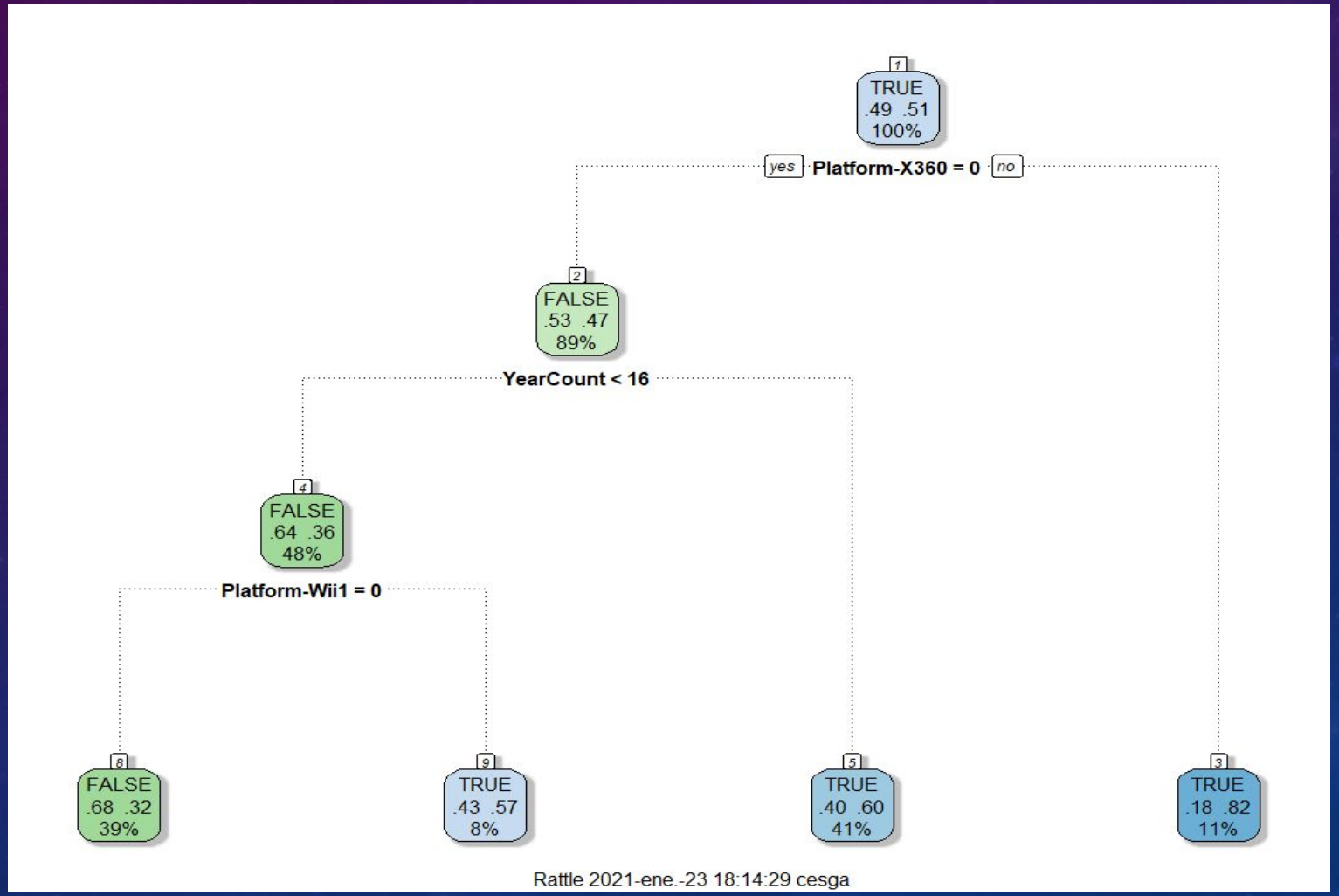
4.1. ANÁLISIS SUPERVISADO - RPart

Plataforma - Acc: 63,72%



4.1. ANÁLISIS SUPERVISADO - RPart

Diferencia de año - Acc: 66,1%



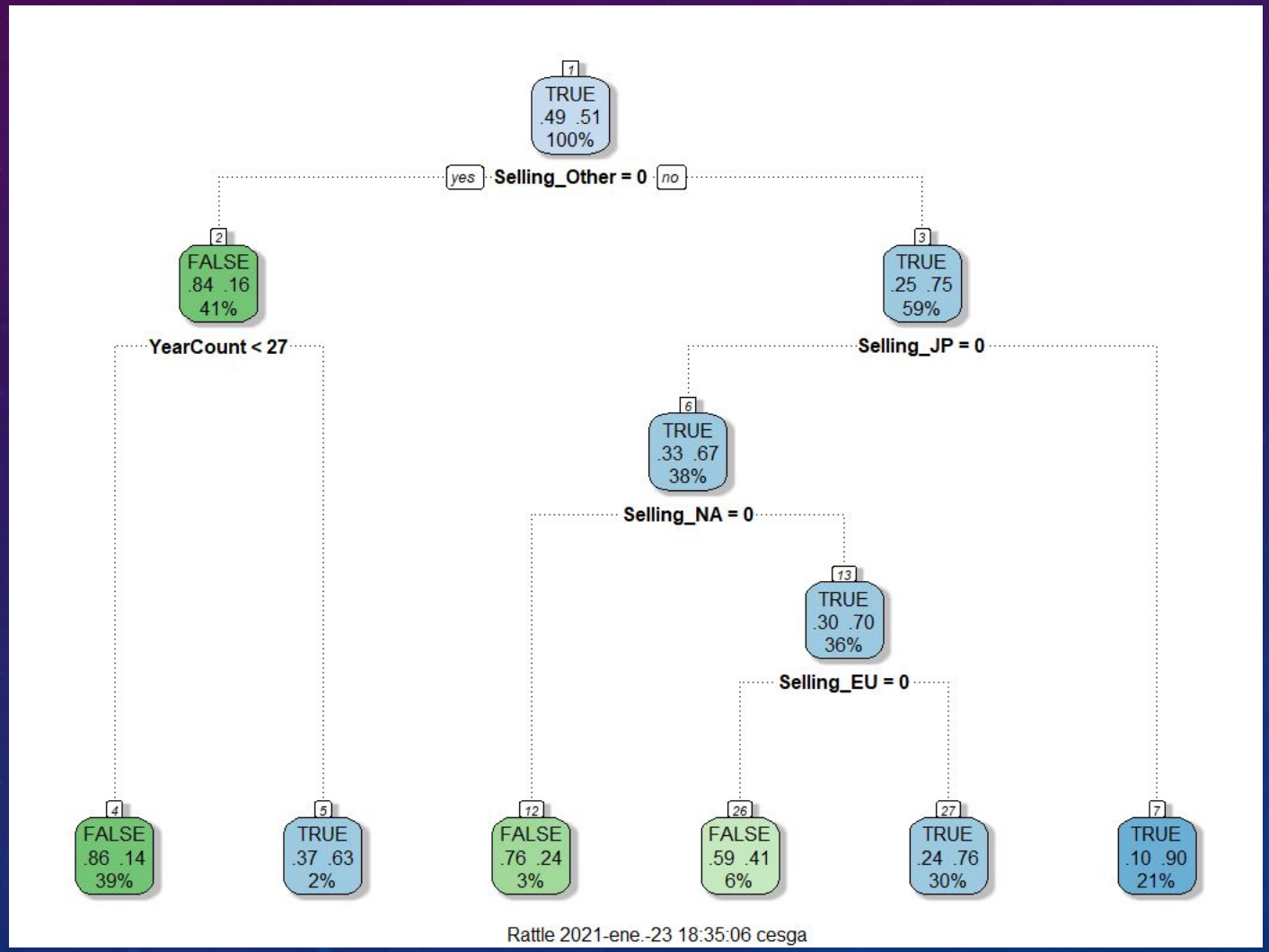
4.1. ANÁLISIS SUPERVISADO - RPart

Mercado

F	G	H	I	J
Sum(NA_Sale	Sum(EU_Sale	Sum(JP_Sales	Sum(Other_S	Sum(Global_S
0.15	0.1	0.12	0.03	0.41
0	0	0.17	0	0.17
0.11	0.09	0	0.03	0.23
0	0	0.16	0	0.16
0	0	0.17	0	0.17
0.49	0.38	0.26	0.13	1.27
0	0	0.14	0	0.14
0.23	0.18	0.2	0.06	0.68
0.14	0.11	0.17	0.04	0.46
0.09	0.07	0	0.02	0.18
0	0	0.03	0	0.03
0.3	0.2	0	0.03	0.53
1.84	1.35	0.04	0.68	3.92
1.64	0.73	0.02	0.09	2.47
1.72	1.33	0	0.16	3.21
0.08	0	0	0.01	0.09
0	0	0.01	0	0.01
0.06	0.01	0	0.01	0.08

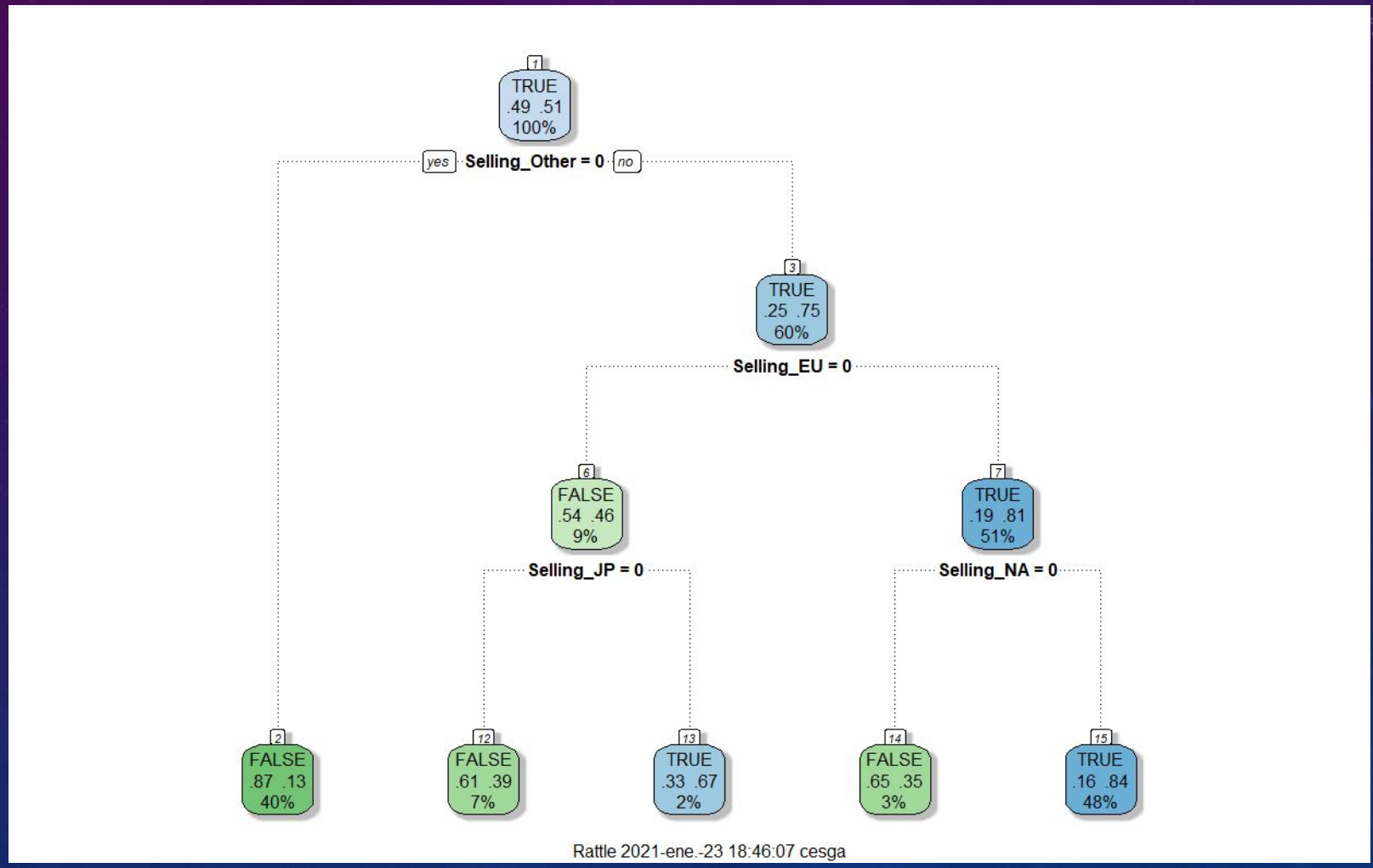
4.1. ANÁLISIS SUPERVISADO - RPart

Mercado - 80,85%



4.1. ANÁLISIS SUPERVISADO - RPart

Filtro año - 82.48%



4.1 ANÁLISIS SUPERVISADO - RPart

Métricas

```
Confusion Matrix and Statistics

       pred
        FALSE TRUE
  FALSE   586  776
  TRUE    435 1023

               Accuracy : 0.5706
                 95% CI : (0.5521, 0.5889)
    No Information Rate : 0.6379
    P-Value [Acc > NIR] : 1

                  Kappa : 0.133

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.5739
            Specificity : 0.5686
         Pos Pred Value : 0.4302
         Neg Pred Value : 0.7016
             Prevalence : 0.3621
         Detection Rate : 0.2078
   Detection Prevalence : 0.4830
      Balanced Accuracy : 0.5713

       'Positive' Class : FALSE
```

Primer modelo

```
Confusion Matrix and Statistics

       pred
        FALSE TRUE
  FALSE   994  185
  TRUE    221  918

               Accuracy : 0.8248
                 95% CI : (0.8088, 0.8401)
    No Information Rate : 0.5242
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.6494

 Mcnemar's Test P-Value : 0.08238

            Sensitivity : 0.8181
            Specificity : 0.8323
         Pos Pred Value : 0.8431
         Neg Pred Value : 0.8060
             Prevalence : 0.5242
         Detection Rate : 0.4288
   Detection Prevalence : 0.5086
      Balanced Accuracy : 0.8252

       'Positive' Class : FALSE
```

Último modelo

4.1 ANÁLISIS SUPERVISADO - RPart

Métricas

```
Confusion Matrix and Statistics

       pred
        FALSE TRUE
  FALSE   586  776
  TRUE    435 1023

               Accuracy : 0.5706
                 95% CI : (0.5521, 0.5889)
    No Information Rate : 0.6379
    P-Value [Acc > NIR] : 1

                  Kappa : 0.133

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.5739
            Specificity : 0.5686
         Pos Pred Value : 0.4302
         Neg Pred Value : 0.7016
             Prevalence : 0.3621
         Detection Rate : 0.2078
   Detection Prevalence : 0.4830
      Balanced Accuracy : 0.5713

       'Positive' Class : FALSE
```

Primer modelo

```
Confusion Matrix and Statistics

       pred
        FALSE TRUE
  FALSE   994  185
  TRUE    221  918

               Accuracy : 0.8248
                 95% CI : (0.8088, 0.8401)
    No Information Rate : 0.5242
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.6494

 Mcnemar's Test P-Value : 0.08238

            Sensitivity : 0.8181
            Specificity : 0.8323
         Pos Pred Value : 0.8431
         Neg Pred Value : 0.8060
             Prevalence : 0.5242
         Detection Rate : 0.4288
   Detection Prevalence : 0.5086
      Balanced Accuracy : 0.8252

       'Positive' Class : FALSE
```

Último modelo

4.1 ANÁLISIS SUPERVISADO - RPart

Métricas

```
Confusion Matrix and Statistics

       pred
        FALSE TRUE
  FALSE   586  776
  TRUE    435 1023

               Accuracy : 0.5706
                 95% CI : (0.5521, 0.5889)
    No Information Rate : 0.6379
    P-Value [Acc > NIR] : 1

                  Kappa : 0.133

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.5739
            Specificity : 0.5686
         Pos Pred Value : 0.4302
         Neg Pred Value : 0.7016
             Prevalence : 0.3621
         Detection Rate : 0.2078
   Detection Prevalence : 0.4830
      Balanced Accuracy : 0.5713

       'Positive' Class : FALSE
```

Primer modelo

```
Confusion Matrix and Statistics

       pred
        FALSE TRUE
  FALSE   994  185
  TRUE    221  918

               Accuracy : 0.8248
                 95% CI : (0.8088, 0.8401)
    No Information Rate : 0.5242
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.6494

 Mcnemar's Test P-Value : 0.08238

            Sensitivity : 0.8181
            Specificity : 0.8323
         Pos Pred Value : 0.8431
         Neg Pred Value : 0.8060
             Prevalence : 0.5242
         Detection Rate : 0.4288
   Detection Prevalence : 0.5086
      Balanced Accuracy : 0.8252

       'Positive' Class : FALSE
```

Último modelo

4.2 ANÁLISIS SUPERVISADO - Naive Bayes

Métricas

```
Confusion Matrix and Statistics

       pred
        FALSE TRUE
 FALSE   994  185
 TRUE    221  918

               Accuracy : 0.8248
                 95% CI : (0.8088, 0.8401)
    No Information Rate : 0.5242
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.6494

 Mcnemar's Test P-Value : 0.08238

            Sensitivity : 0.8181
            Specificity : 0.8323
         Pos Pred Value : 0.8431
         Neg Pred Value : 0.8060
             Prevalence : 0.5242
         Detection Rate : 0.4288
   Detection Prevalence : 0.5086
      Balanced Accuracy : 0.8252

       'Positive' Class : FALSE
```

Rpart

```
Confusion Matrix and Statistics

       pred
        FALSE TRUE
 FALSE   918  261
 TRUE    182  957

               Accuracy : 0.8089
                 95% CI : (0.7923, 0.8247)
    No Information Rate : 0.5255
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6181

 Mcnemar's Test P-Value : 0.0002106

            Sensitivity : 0.8345
            Specificity : 0.7857
         Pos Pred Value : 0.7786
         Neg Pred Value : 0.8402
             Prevalence : 0.4745
         Detection Rate : 0.3960
   Detection Prevalence : 0.5086
      Balanced Accuracy : 0.8101

       'Positive' Class : FALSE
```

Naive Bayes

4.2 ANÁLISIS SUPERVISADO - Naive Bayes

Métricas

```
Confusion Matrix and Statistics

       pred
        FALSE TRUE
 FALSE    994  185
 TRUE     221  918

               Accuracy : 0.8248
                 95% CI : (0.8088, 0.8401)
    No Information Rate : 0.5242
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.6494

 Mcnemar's Test P-Value : 0.08238

            Sensitivity : 0.8181
            Specificity : 0.8323
         Pos Pred Value : 0.8431
         Neg Pred Value : 0.8060
             Prevalence : 0.5242
         Detection Rate : 0.4288
   Detection Prevalence : 0.5086
      Balanced Accuracy : 0.8252

       'Positive' Class : FALSE
```

Rpart

```
Confusion Matrix and Statistics

       pred
        FALSE TRUE
 FALSE    918  261
 TRUE     182  957

               Accuracy : 0.8089
                 95% CI : (0.7923, 0.8247)
    No Information Rate : 0.5255
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6181

 Mcnemar's Test P-Value : 0.0002106

            Sensitivity : 0.8345
            Specificity : 0.7857
         Pos Pred Value : 0.7786
         Neg Pred Value : 0.8402
             Prevalence : 0.4745
         Detection Rate : 0.3960
   Detection Prevalence : 0.5086
      Balanced Accuracy : 0.8101

       'Positive' Class : FALSE
```

Naive Bayes

4.2 ANÁLISIS SUPERVISADO - Naive Bayes

Métricas

```
Confusion Matrix and Statistics

      pred
       FALSE TRUE
 FALSE   994  185
 TRUE    221  918

               Accuracy : 0.8248
                 95% CI : (0.8088, 0.8401)
    No Information Rate : 0.5242
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.6494

 Mcnemar's Test P-Value : 0.08238

            Sensitivity : 0.8181
            Specificity : 0.8323
         Pos Pred Value : 0.8431
         Neg Pred Value : 0.8060
             Prevalence : 0.5242
         Detection Rate : 0.4288
   Detection Prevalence : 0.5086
      Balanced Accuracy : 0.8252

       'Positive' Class : FALSE
```

Rpart

```
Confusion Matrix and Statistics

      pred
       FALSE TRUE
 FALSE   918  261
 TRUE    182  957

               Accuracy : 0.8089
                 95% CI : (0.7923, 0.8247)
    No Information Rate : 0.5255
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6181

 Mcnemar's Test P-Value : 0.0002106

            Sensitivity : 0.8345
            Specificity : 0.7857
         Pos Pred Value : 0.7786
         Neg Pred Value : 0.8402
             Prevalence : 0.4745
         Detection Rate : 0.3960
   Detection Prevalence : 0.5086
      Balanced Accuracy : 0.8101

       'Positive' Class : FALSE
```

Naive Bayes

4.3 ANÁLISIS SUPERVISADO - CTree2

Métricas

```
Confusion Matrix and Statistics

       pred
        FALSE TRUE
 FALSE    994  185
 TRUE     221  918

               Accuracy : 0.8248
                 95% CI : (0.8088, 0.8401)
    No Information Rate : 0.5242
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.6494

 Mcnemar's Test P-Value : 0.08238

            Sensitivity : 0.8181
            Specificity : 0.8323
         Pos Pred Value : 0.8431
         Neg Pred Value : 0.8060
             Prevalence : 0.5242
         Detection Rate : 0.4288
   Detection Prevalence : 0.5086
      Balanced Accuracy : 0.8252

       'Positive' Class : FALSE
```

Rpart

```
Confusion Matrix and Statistics

       pred
        FALSE TRUE
 FALSE    994  185
 TRUE     221  918

               Accuracy : 0.8248
                 95% CI : (0.8088, 0.8401)
    No Information Rate : 0.5242
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.6494

 Mcnemar's Test P-Value : 0.08238

            Sensitivity : 0.8181
            Specificity : 0.8323
         Pos Pred Value : 0.8431
         Neg Pred Value : 0.8060
             Prevalence : 0.5242
         Detection Rate : 0.4288
   Detection Prevalence : 0.5086
      Balanced Accuracy : 0.8252

       'Positive' Class : FALSE
```

CTree2

4.3 ANÁLISIS SUPERVISADO - CTree2

Métricas

```
      pred
       FALSE TRUE
FALSE   938  213
TRUE    171  996

               Accuracy : 0.8343
                 95% CI : (0.8186, 0.8493)
    No Information Rate : 0.5216
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.6686

 Mcnemar's Test P-Value : 0.03641

            Sensitivity : 0.8458
            Specificity : 0.8238
         Pos Pred Value : 0.8149
         Neg Pred Value : 0.8535
             Prevalence : 0.4784
         Detection Rate : 0.4047
   Detection Prevalence : 0.4965
      Balanced Accuracy : 0.8348

       'Positive' Class : FALSE
```

Rpart

```
      pred
       FALSE TRUE
FALSE   910  241
TRUE    179  988

               Accuracy : 0.8188
                 95% CI : (0.8025, 0.8343)
    No Information Rate : 0.5302
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6375

 Mcnemar's Test P-Value : 0.002916

            Sensitivity : 0.8356
            Specificity : 0.8039
         Pos Pred Value : 0.7906
         Neg Pred Value : 0.8466
             Prevalence : 0.4698
         Detection Rate : 0.3926
   Detection Prevalence : 0.4965
      Balanced Accuracy : 0.8198

       'Positive' Class : FALSE
```

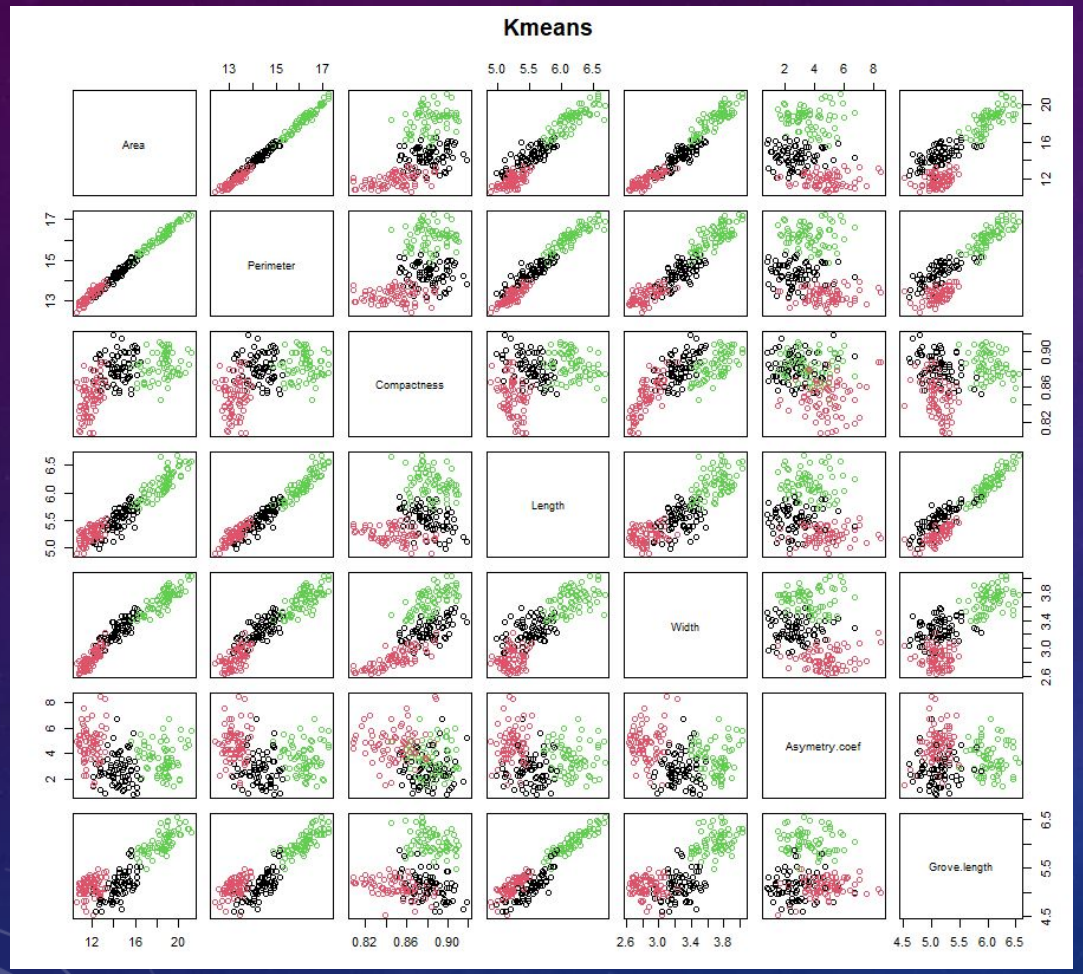
CTree2

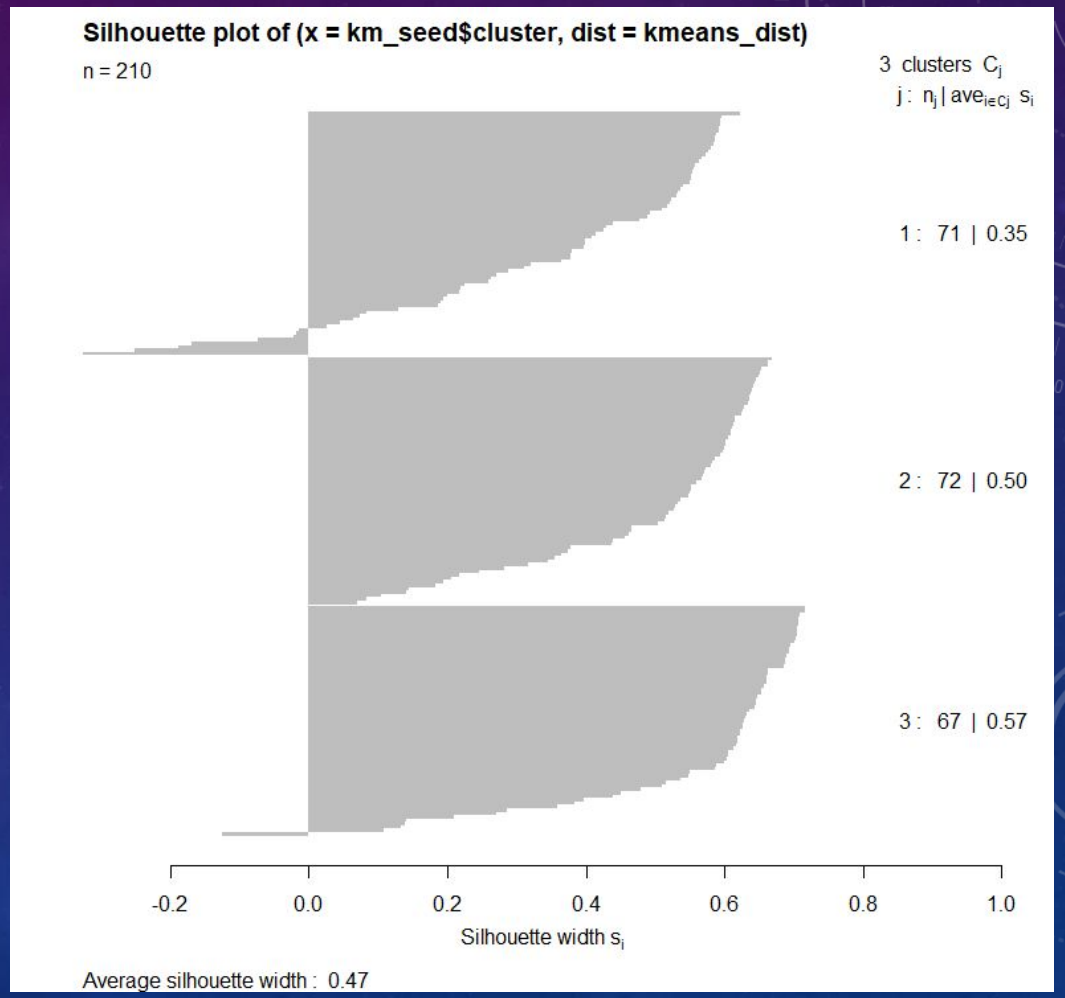
4.4 ANÁLISIS SUPERVISADO - Conclusiones

- Precisión pareja métodos utilizados

- Superventas → Ventas en otros mercados
- Juegos antiguos → Superventas
- Filtrado Año → Mayor precisión

5.1. ANÁLISIS NO SUPERVISADO - KMEANS





4.4 ANÁLISIS SUPERVISADO - Conclusiones

types		
	1	2	3
1	9	68	0
2	61	2	70

K = 2

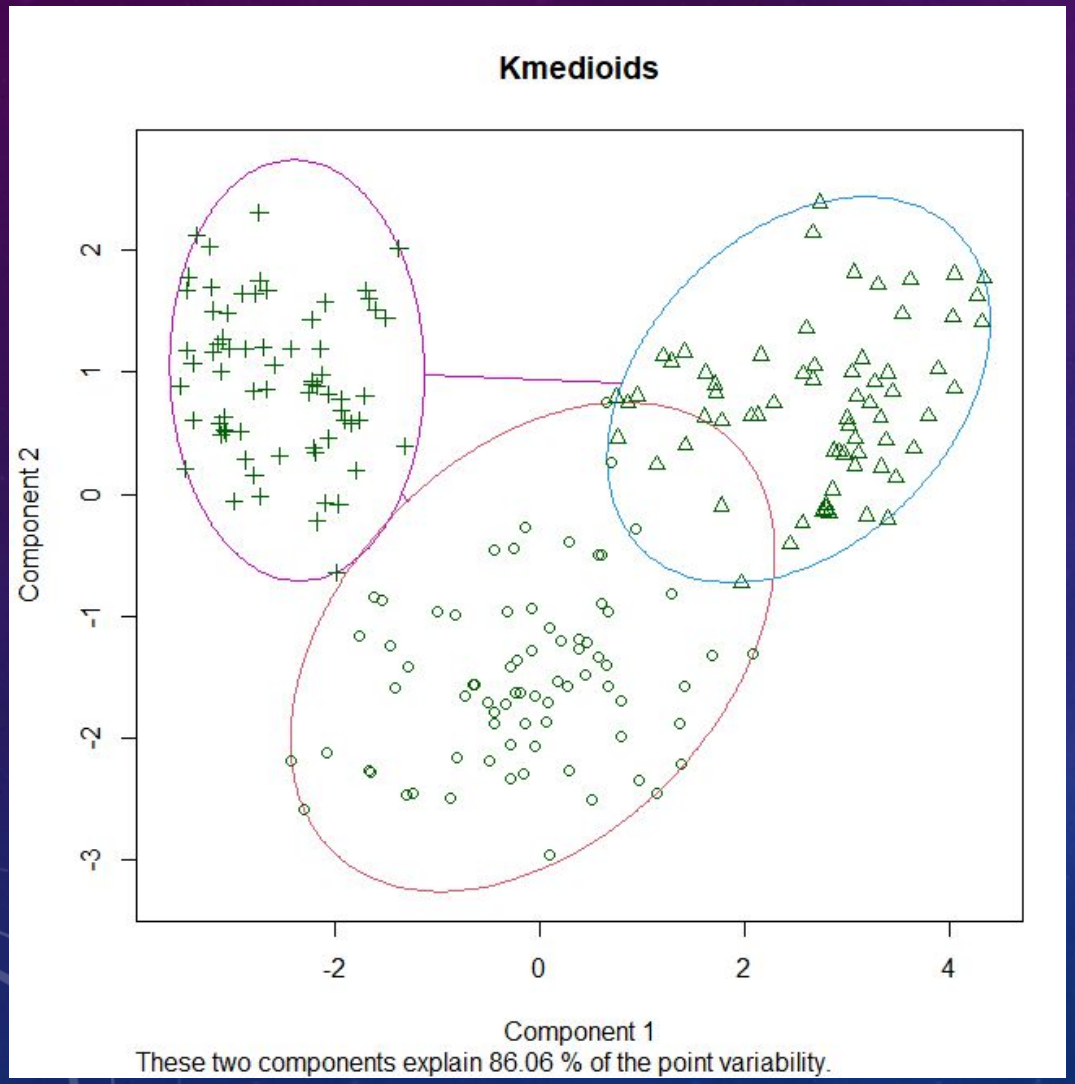
types		
	1	2	3
1	6	0	66
2	62	5	4
3	2	65	0

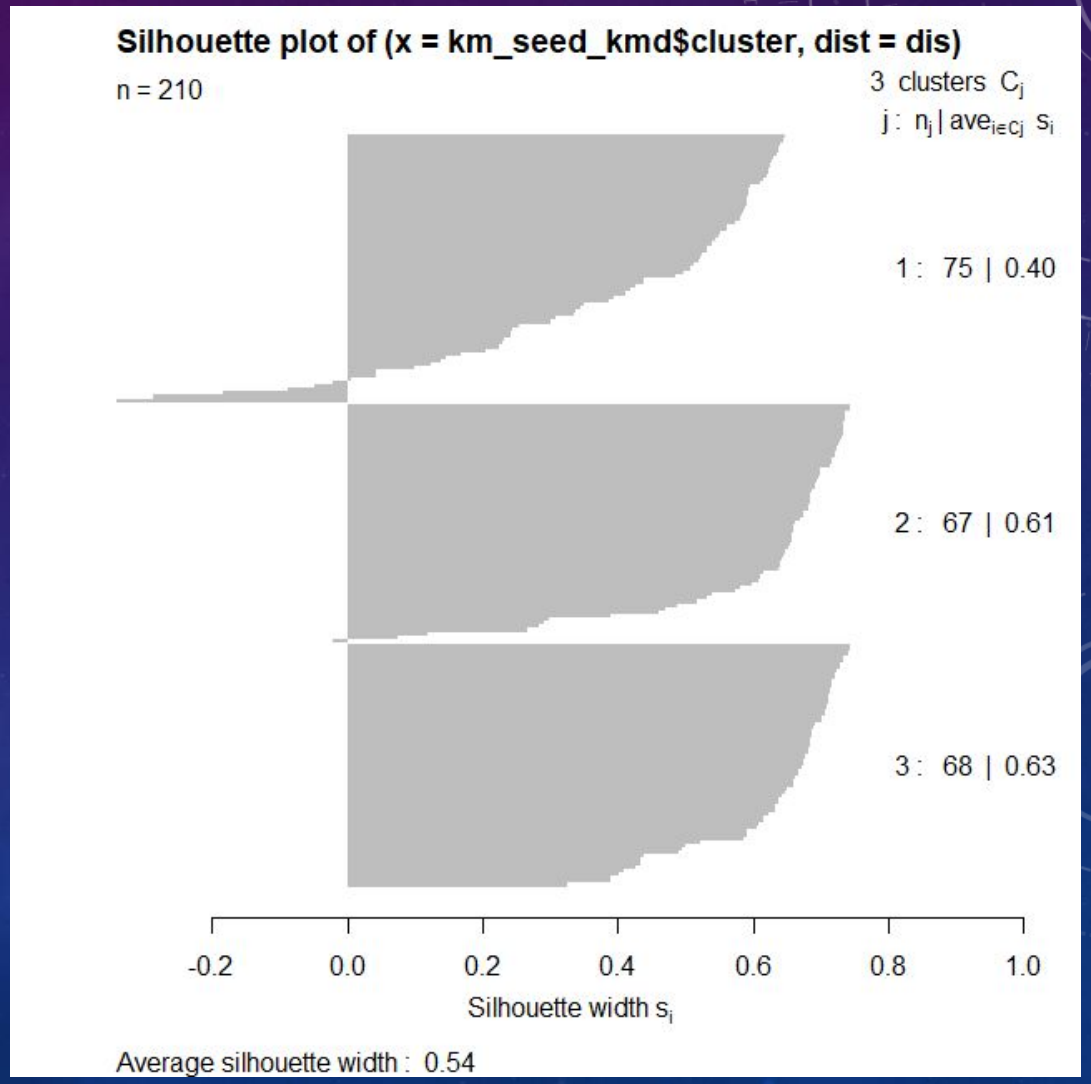
K = 3

types			
	1	2	3
1	2	0	62
2	0	51	0
3	12	18	0
4	56	1	8

K = 4

5.2. ANÁLISIS NO SUPERVISADO - KMEDIOIDS





4.4 ANÁLISIS SUPERVISADO - Conclusiones

types			
	0	1	2
1	53	70	0
2	17	0	70

K = 2

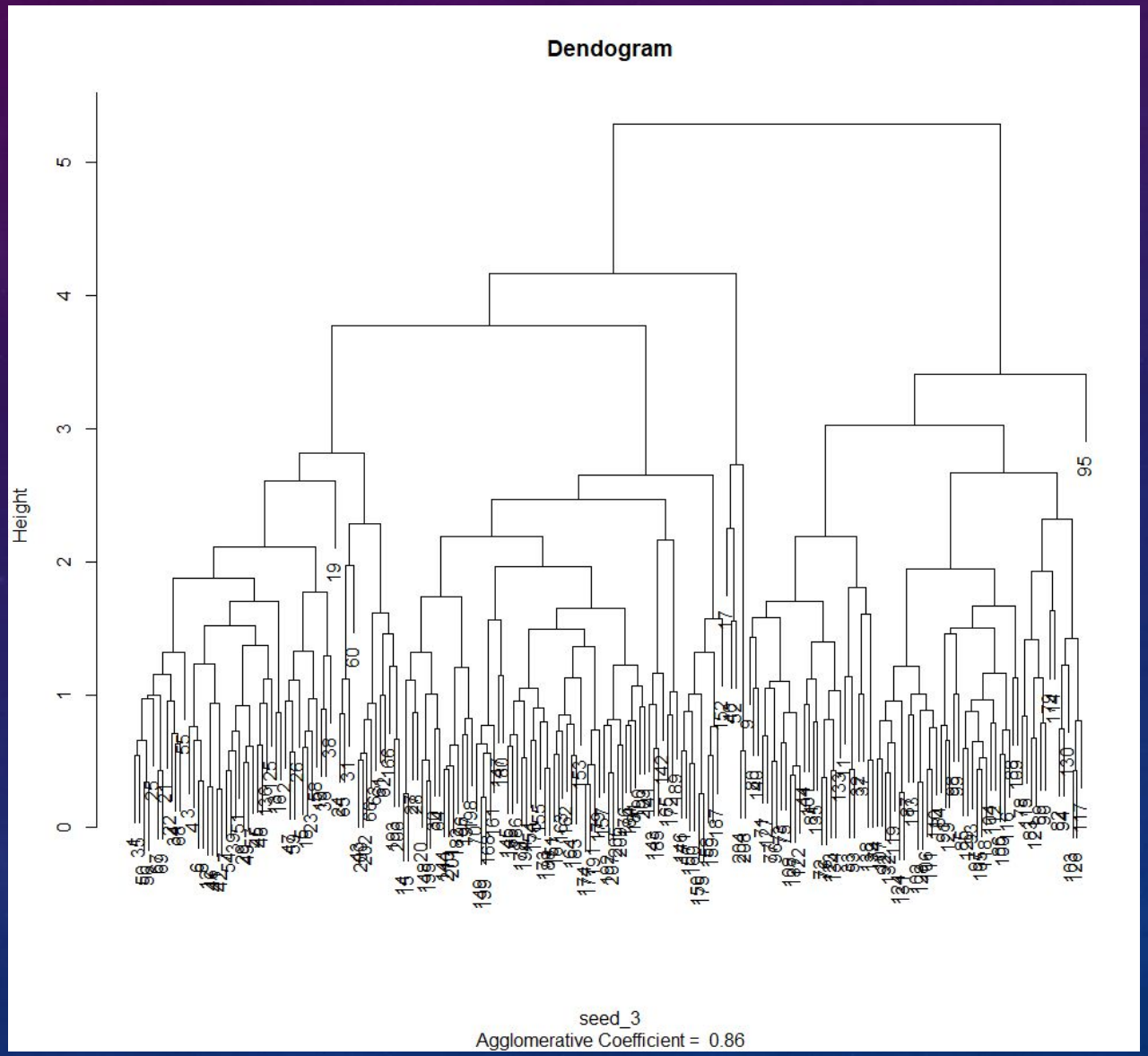
types			
	0	1	2
1	69	4	2
2	1	66	0
3	0	0	68

K = 3

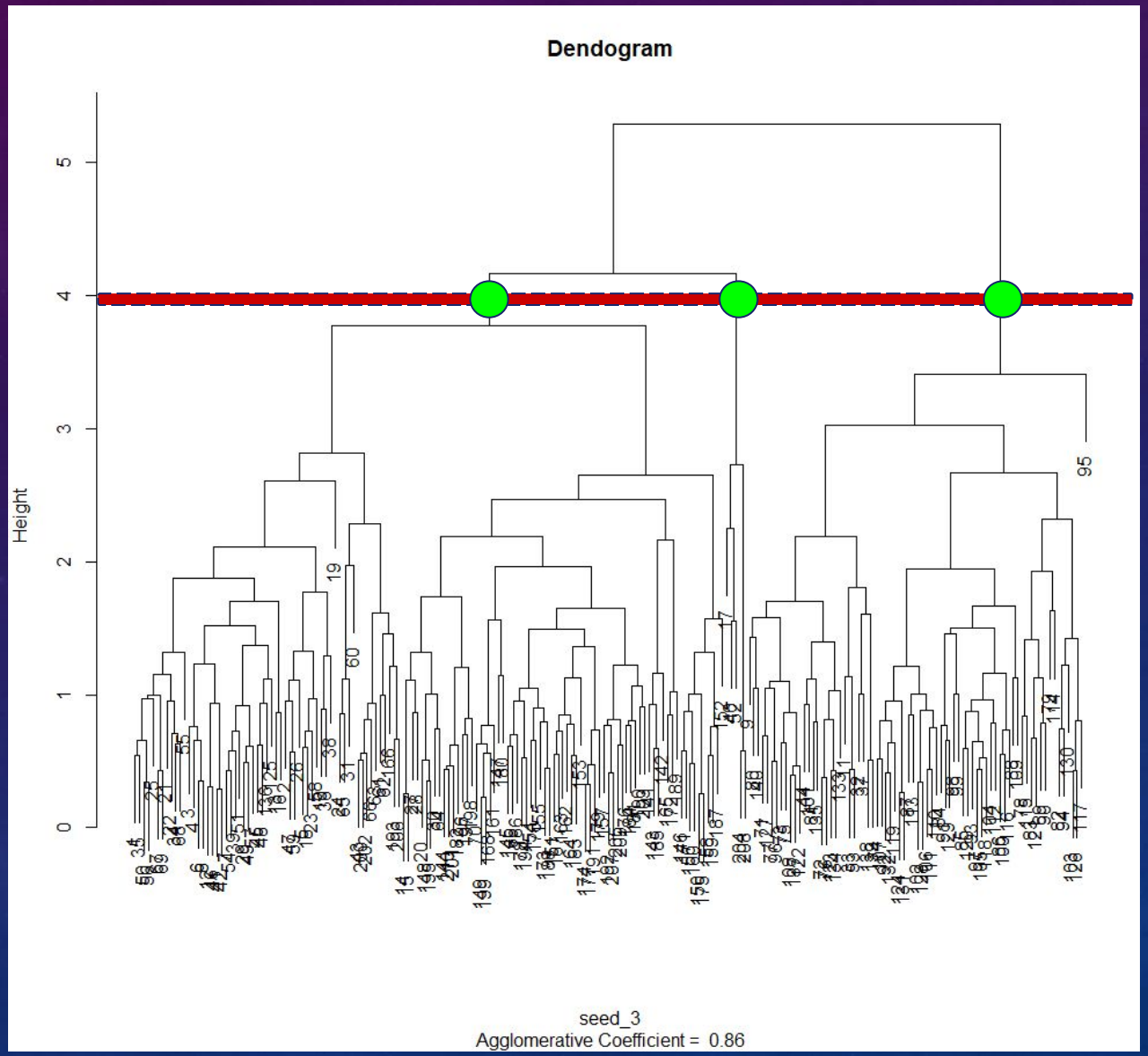
types			
	0	1	2
1	66	1	2
2	4	27	0
3	0	42	0
4	0	0	68

K = 4

5.3. ANÁLISIS NO SUPERVISADO - AGNES



5.3. ANÁLISIS NO SUPERVISADO - AGNES



5.4. ANÁLISIS NO SUPERVISADO - CONCLUSIONES

¿DIFERENCIAS?

¿RENDIMIENTO?

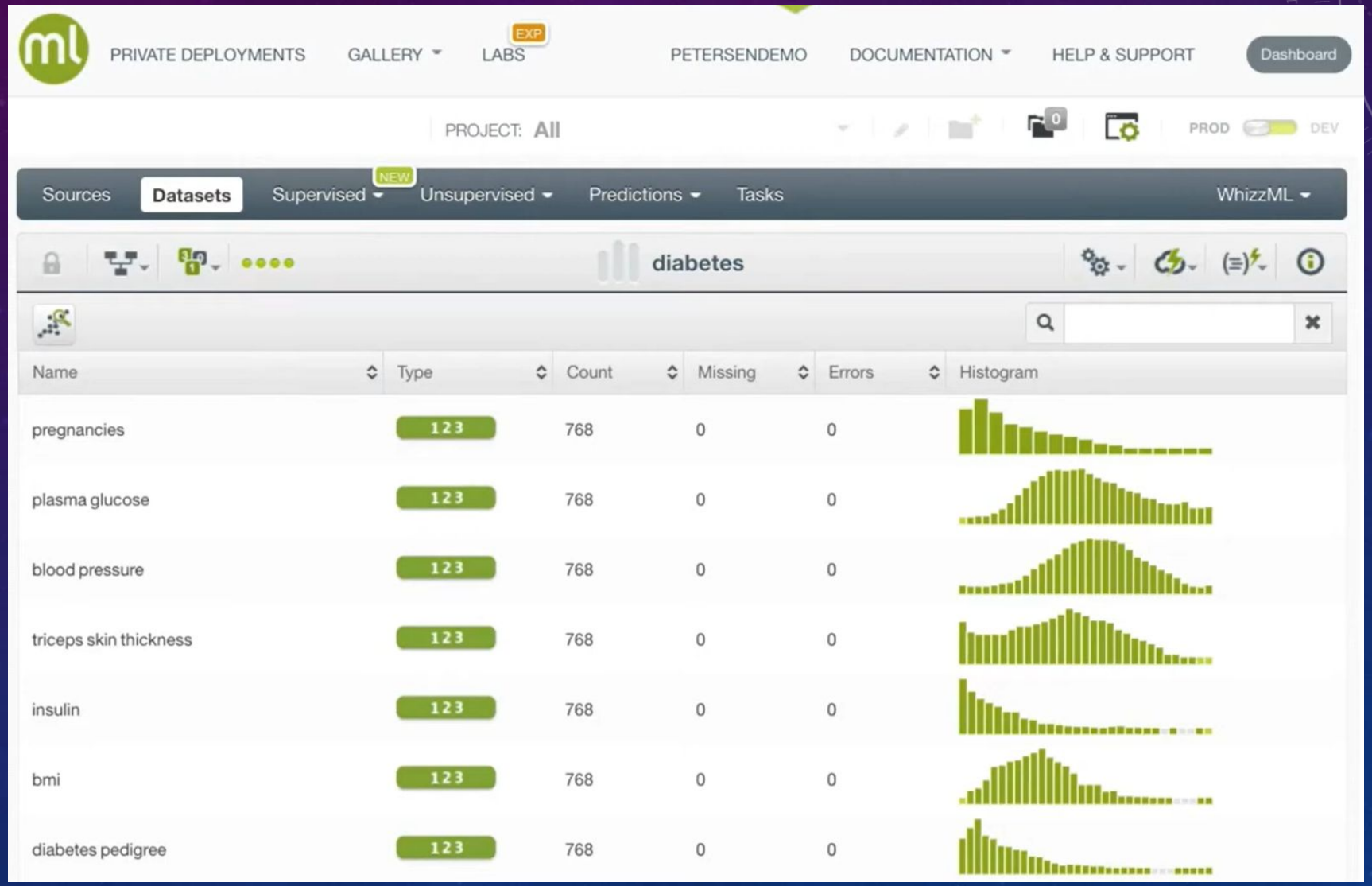
6. BIG-ML: CLUSTERS

INTRODUCCIÓN TEÓRICA:

- TÉCNICA DE APRENDIZAJE NO SUPERVISADO
  - NO ES NECESARIO ETIQUETAR LOS DATOS (CLASIFICARLOS)
- AGRUPA INSTANCIAS DE DATOS SIMILARES ENTRE SÍ
  - CONCEPTO DE SIMILITUD DEPENDE DEL CONTEXTO DEL PROBLEMA
- CADA GRUPO (CLUSTER) SE DEFINE POR SU CENTROIDE
  - CENTRO GEOMÉTRICO DEL GRUPO
  - REPRESENTA LA MEDIA DE LOS DATOS DEL GRUPO
  - EL NÚMERO DE CENTROIDES (K) PUEDE SER DETERMINADO A PRIORI
- EJEMPLO

6. BIG-ML: CLUSTERS

DATASET



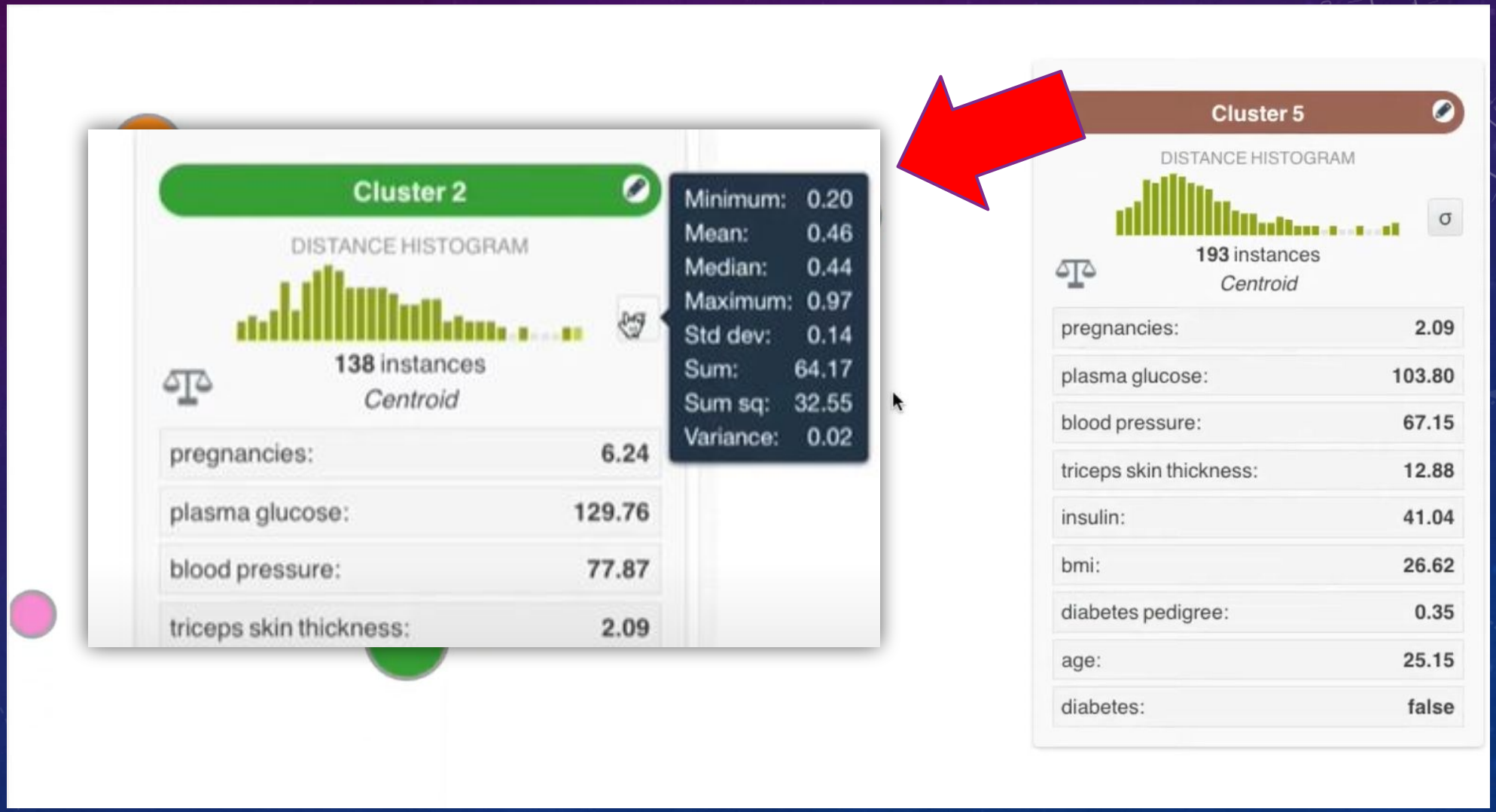
Name	Type	Count	Missing	Errors	Histogram
pregnancies	123	768	0	0	
plasma glucose	123	768	0	0	
blood pressure	123	768	0	0	
triceps skin thickness	123	768	0	0	
insulin	123	768	0	0	
bmi	123	768	0	0	
diabetes pedigree	123	768	0	0	

6. BIG-ML: CLUSTERS

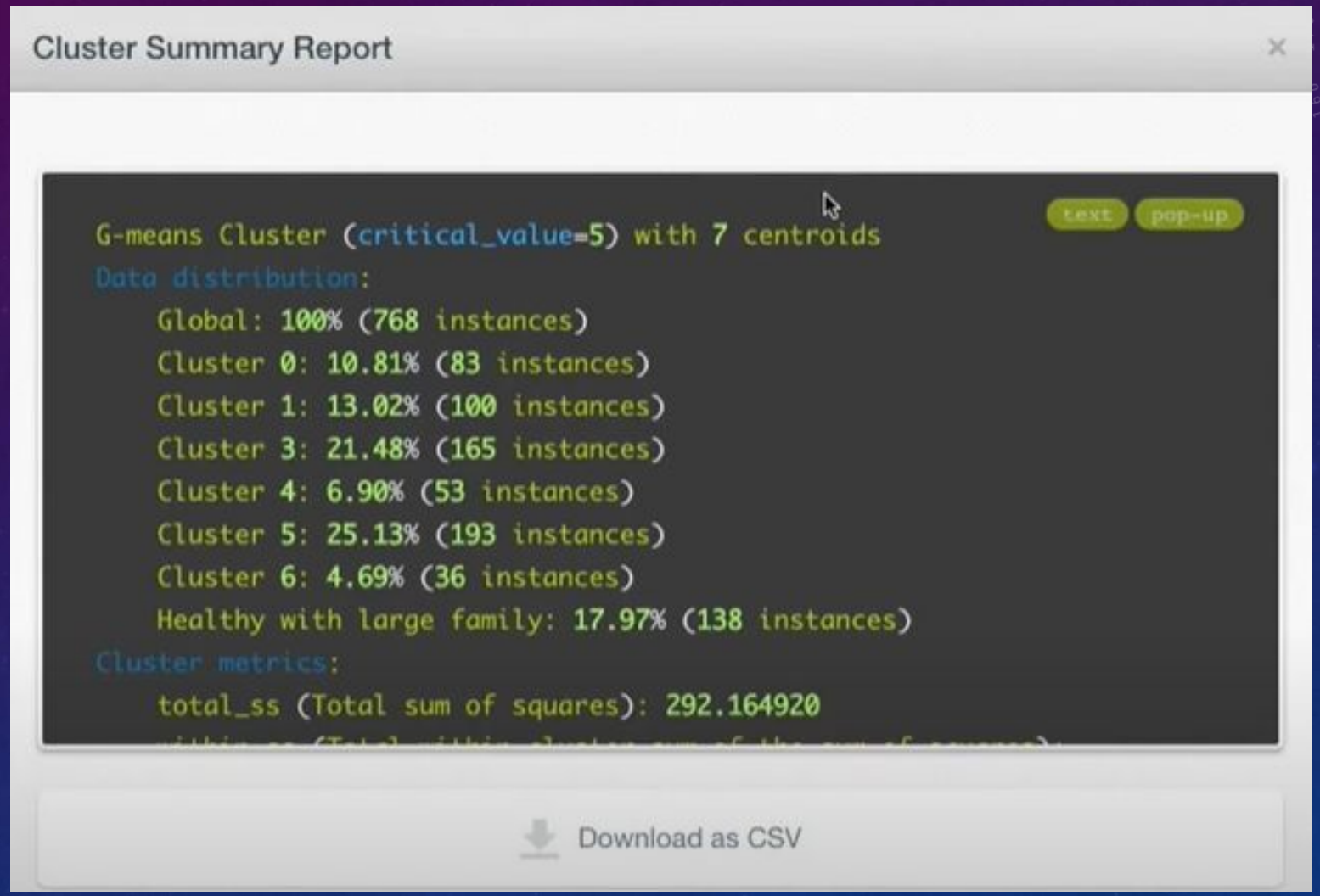
ONE CLICK CLUSTER (AUTOMATICO)

6. BIG-ML: CLUSTERS

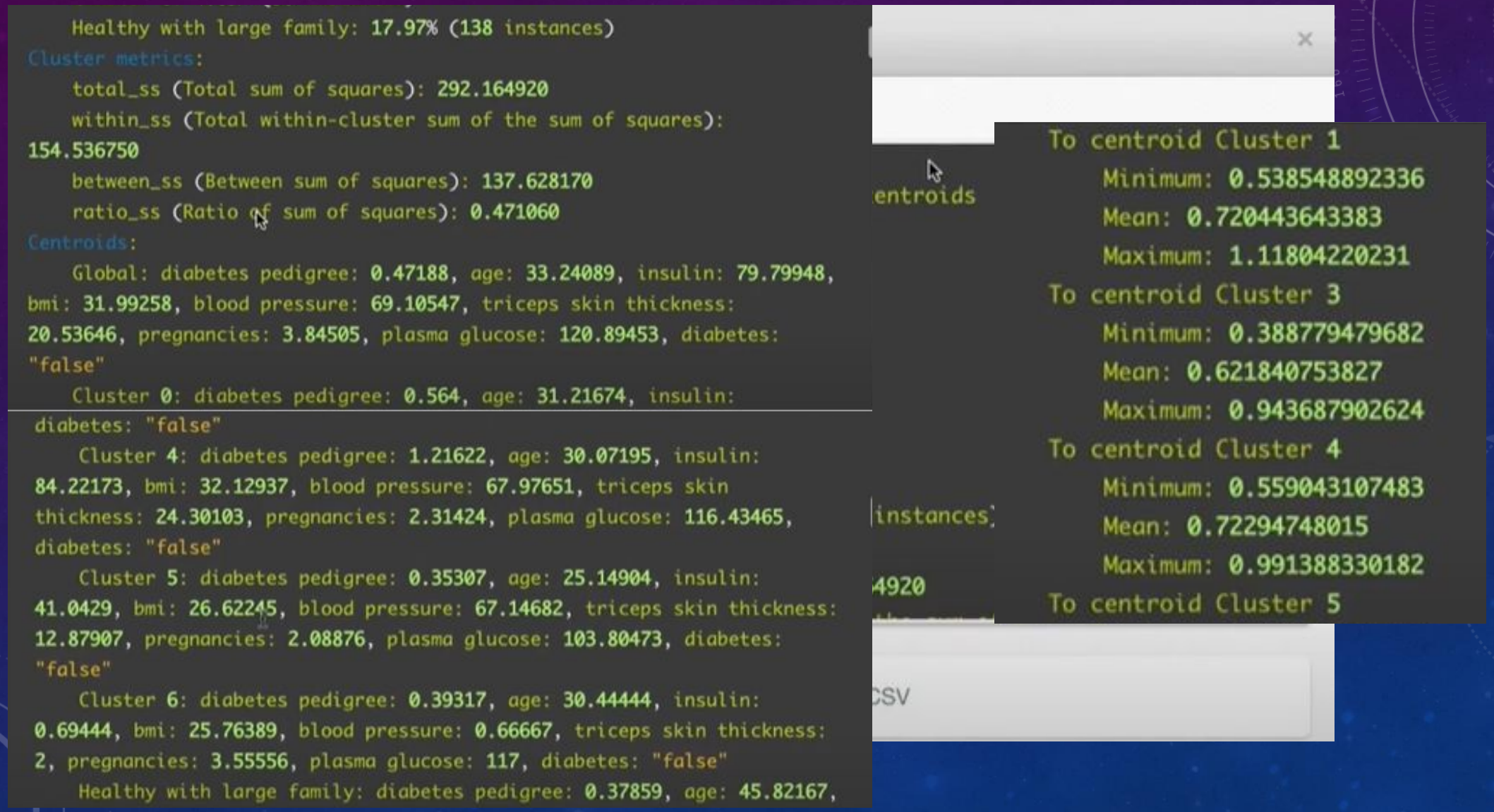
ONE CLICK CLUSTER (AUTOMATICO)



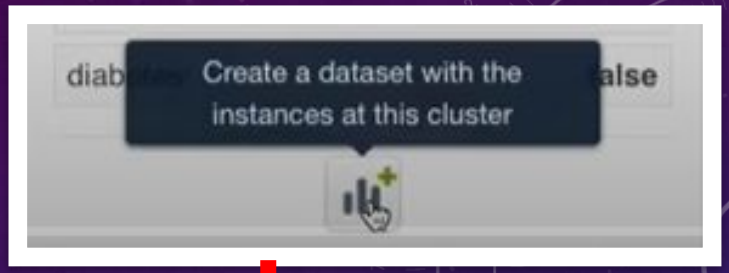
6. BIG-ML: CLUSTERS



6. BIG-ML: CLUSTERS



6. BIG-ML: CLUSTERS



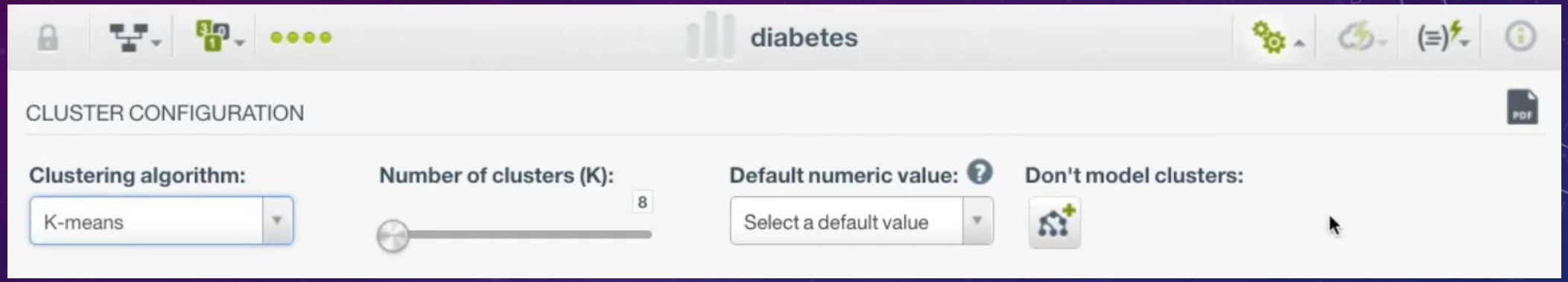
AISLAR DATOS DE UN CLUSTER

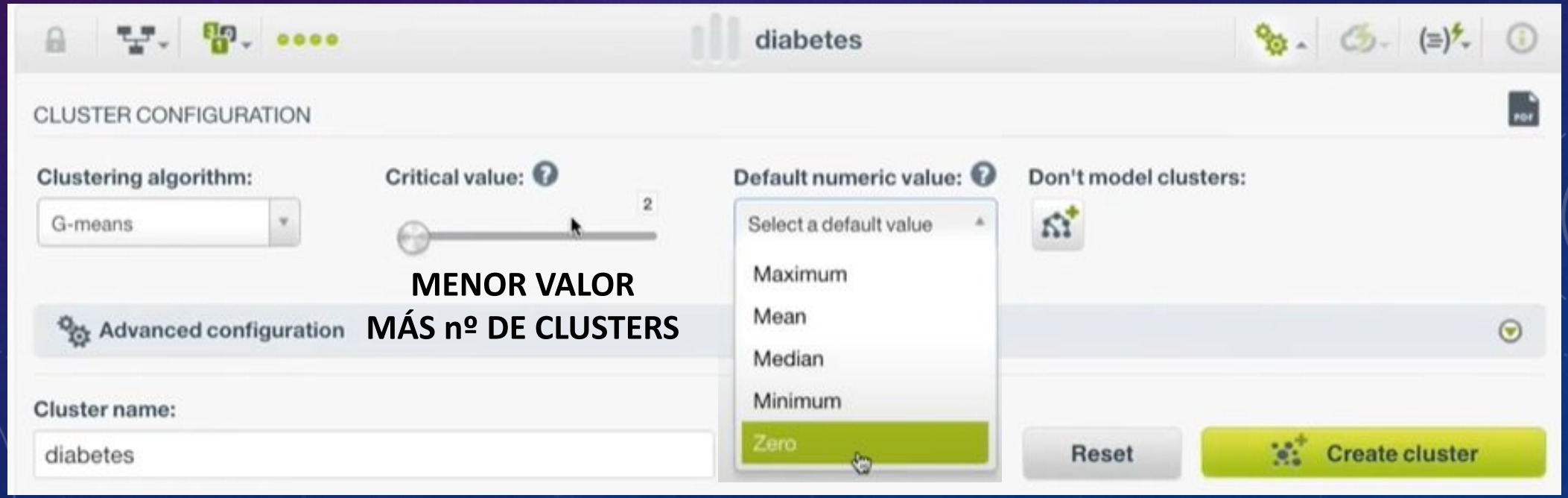
Healthy with large family - diabetes

Name	Type	Count	Missing	Errors	Histogram
pregnancies	123	138	0	0	
plasma glucose	123	138	0	0	
blood pressure	123	138	0	0	
triceps skin thickness	123	138	0	0	
insulin	123	138	0	0	
bmi	123	138	0	0	

6. BIG-ML: CLUSTERS

CONFIGURACIÓN BÁSICA (EN VEZ DE ONE CLICK OPTION)

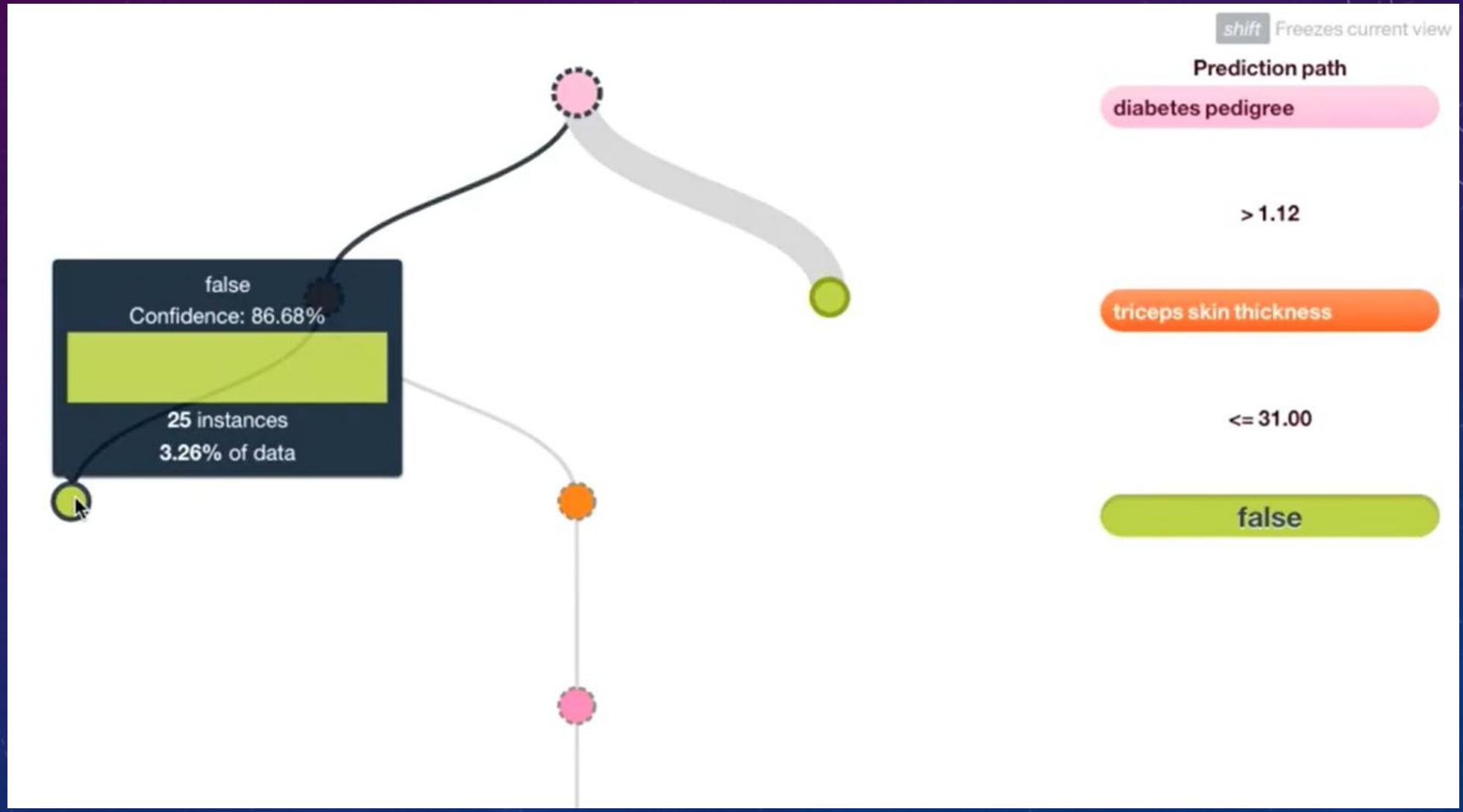




MENOR VALOR
MÁS nº DE CLUSTERS

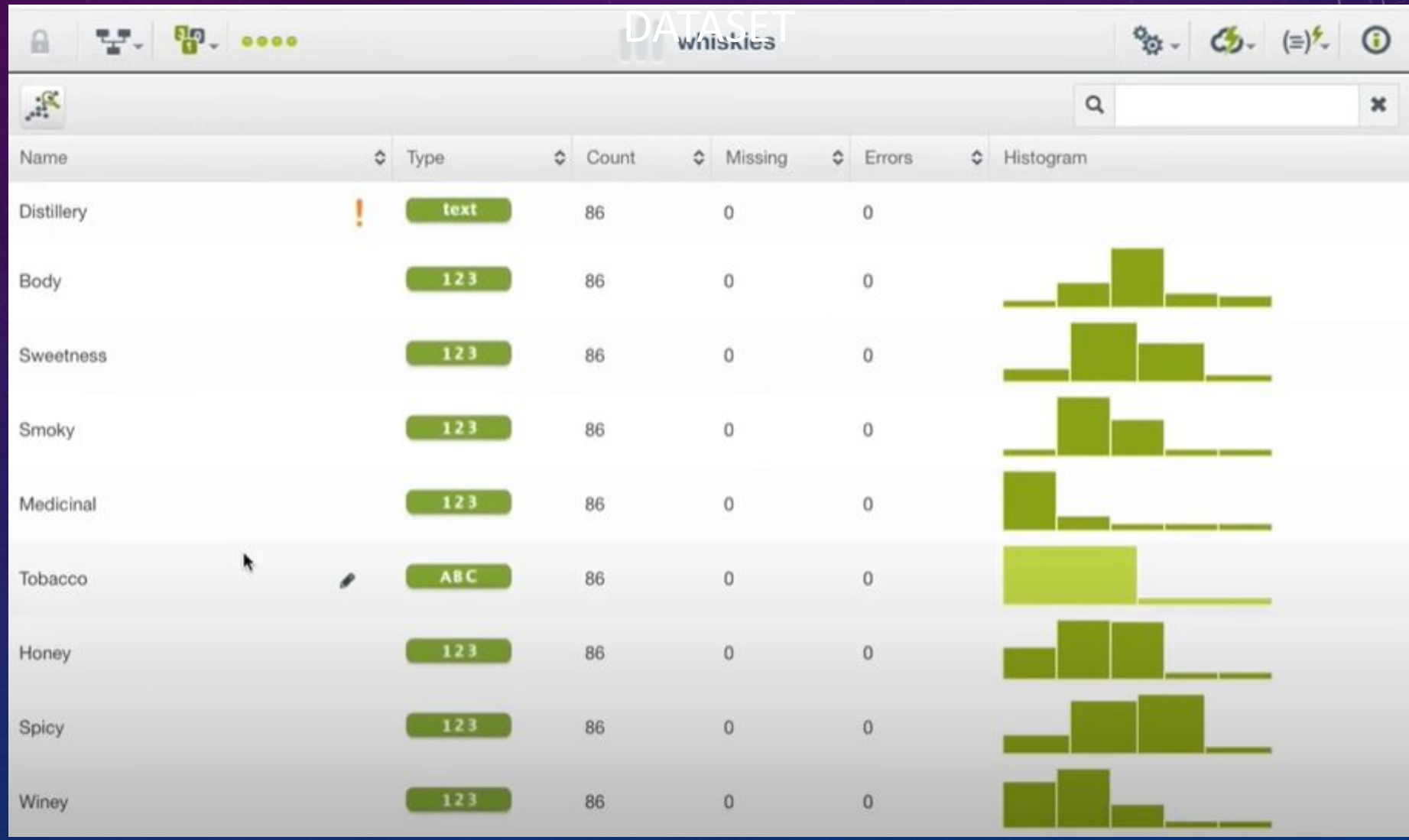
6. BIG-ML: CLUSTERS

CLUSTER MODEL



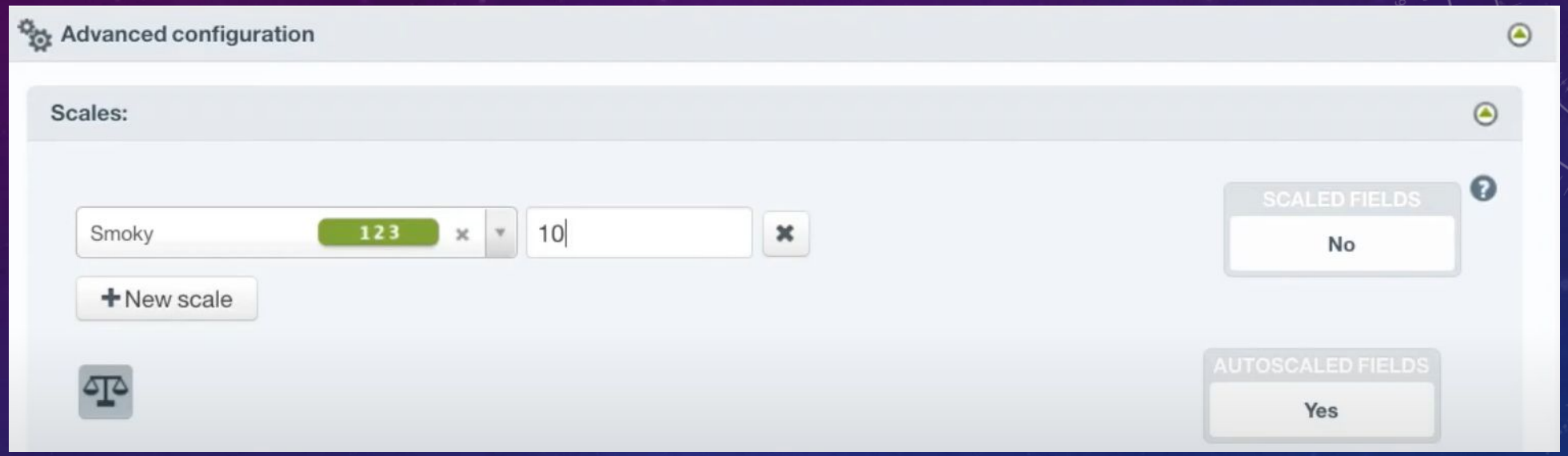
6. BIG-ML: CLUSTERS

NUEVO DATASET



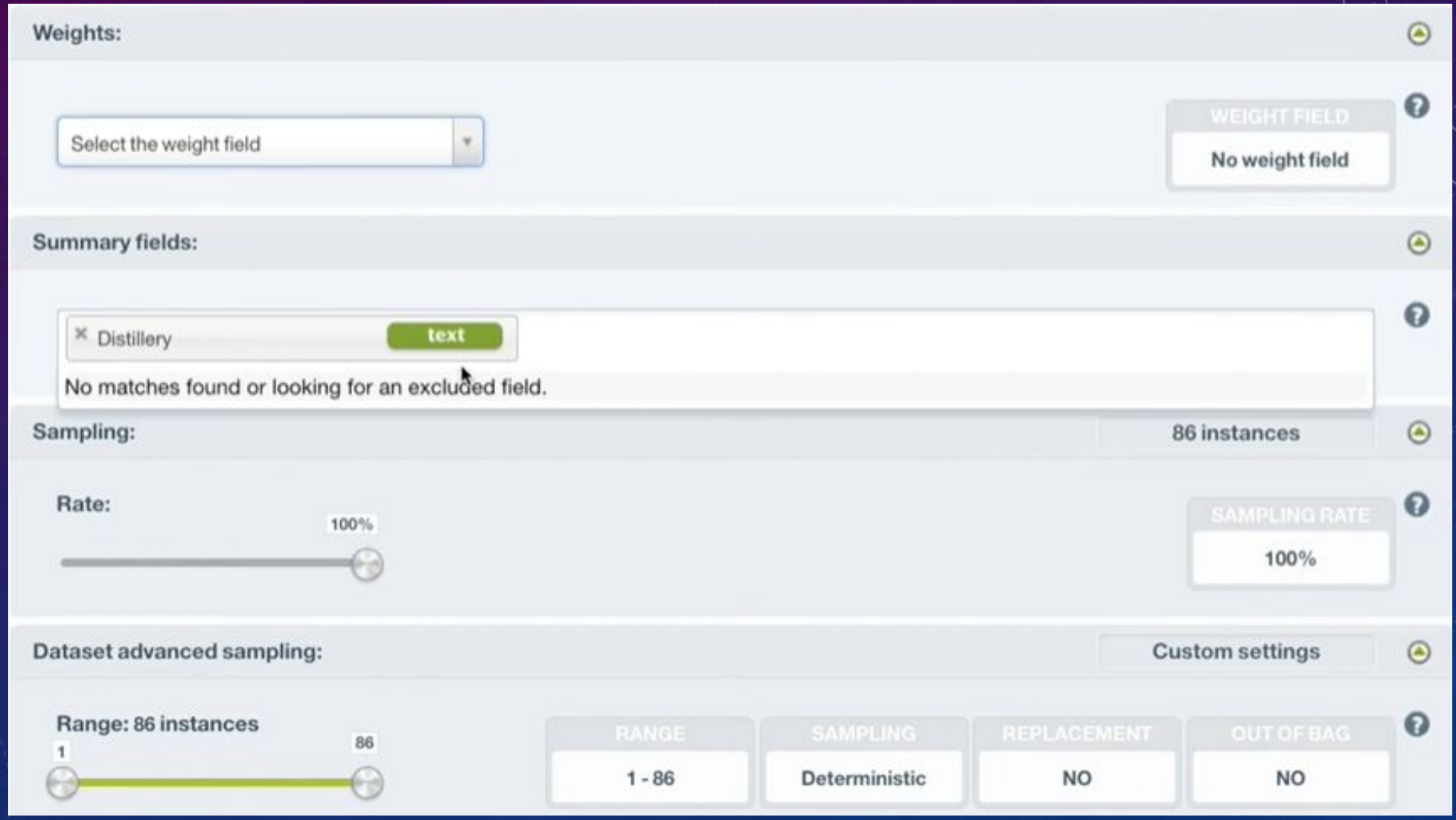
6. BIG-ML: CLUSTERS

CLUSTER CONFIGURATION



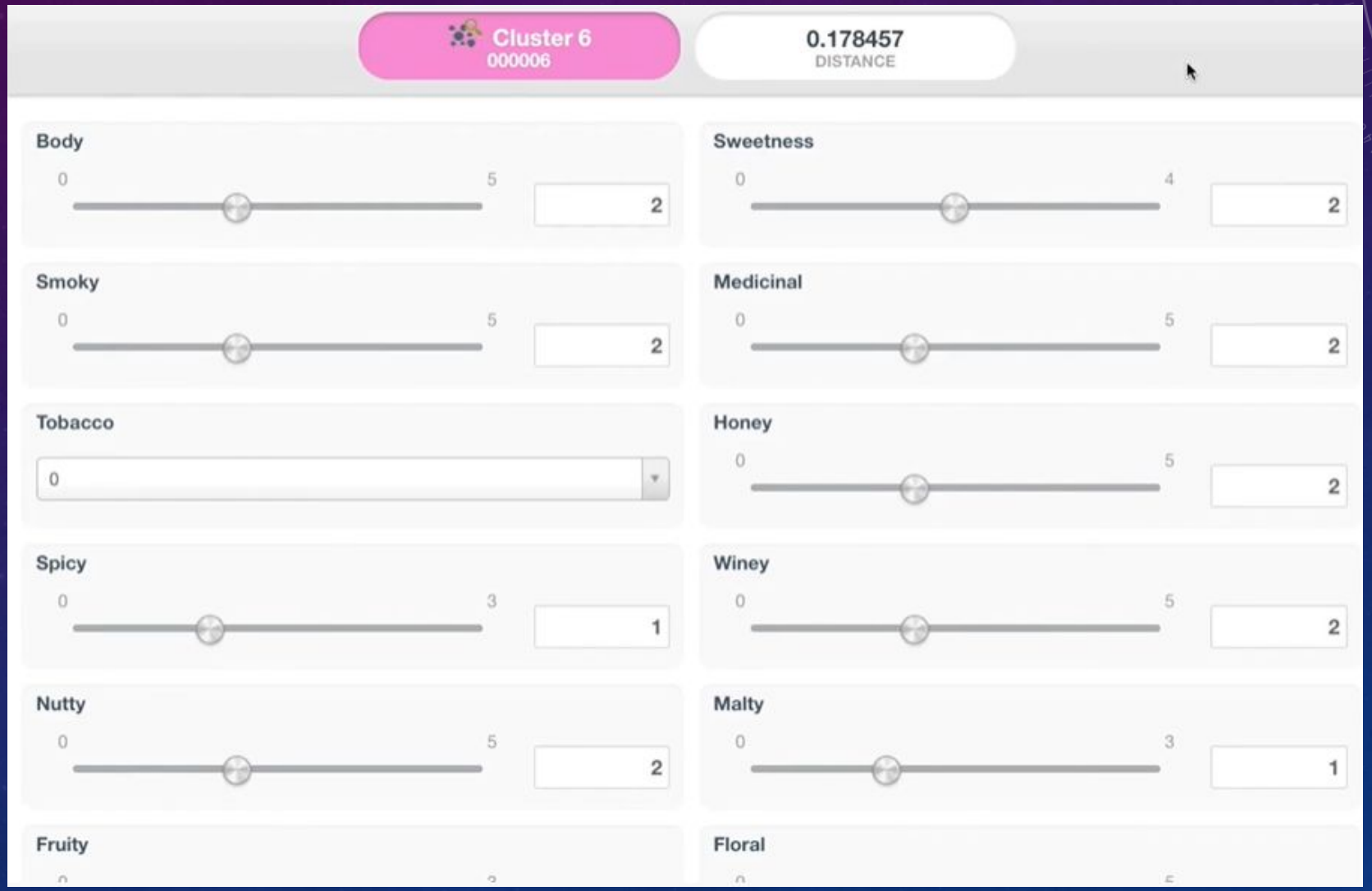
6. BIG-ML: CLUSTERS

CLUSTER CONFIGURATION



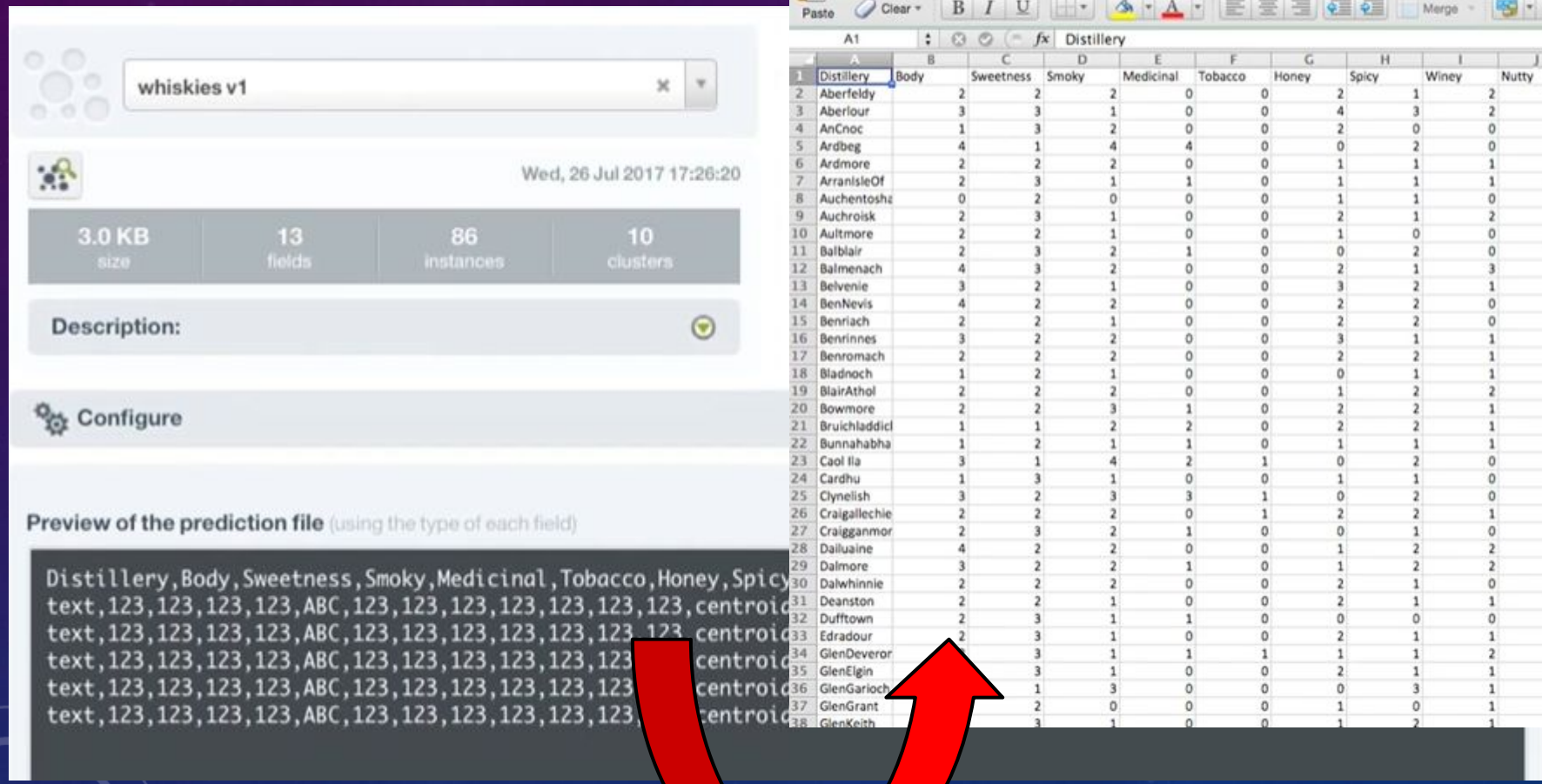
6. BIG-ML: CLUSTERS

BATCH - CENTROIDES



6. BIG-ML: CLUSTERS

BATCH-CENTROIDES



Distillery	Body	Sweetness	Smoky	Medicinal	Tobacco	Honey	Spicy	Winey	Nutty	Malty	Fruity	Floral	cluster
Aberfeldy	2	2	2	0	0	2	1	2	2	2	2	2	Cluster 1
Aberlour	3	3	1	0	0	4	3	2	2	3	3	2	Cluster 1
AnCnoc	1	3	2	0	0	2	0	0	2	2	3	2	Cluster 3
Ardbeg	4	1	4	4	0	0	2	0	1	2	1	0	Medicinal or Smokey
Ardmore	2	2	2	0	0	1	1	1	2	3	1	1	Cluster 6
ArranIsleOf	2	3	1	1	0	1	1	1	0	1	1	2	Cluster 8
Auchentoshz	0	2	0	0	0	1	1	0	2	2	3	3	Cluster 3
Auchroisk	2	3	1	0	0	2	1	2	2	2	2	1	Cluster 1
Aultmore	2	2	1	0	0	1	0	0	2	2	2	2	Cluster 3
Balblair	2	3	2	1	0	0	2	0	2	1	2	1	Cluster 8
Balmenach	4	3	2	0	0	2	1	3	3	0	1	2	Cluster 0
Belvenie	3	2	1	0	0	3	2	1	0	2	2	2	Cluster 9
BenNevis	4	2	2	0	0	2	2	0	2	2	2	2	Cluster 0
Benriach	2	2	1	0	0	2	2	0	0	2	3	2	Cluster 9
Benrinnes	3	2	2	0	0	3	1	1	2	3	2	2	Cluster 1
Benromach	2	2	2	0	0	2	2	1	2	2	2	2	Cluster 1
Bladnoch	1	2	1	0	0	0	1	1	0	2	2	3	Cluster 8
BlairAthol	2	2	2	0	0	1	2	2	2	2	2	2	Cluster 1
Bowmore	2	2	3	1	0	2	2	1	1	1	1	2	Cluster 5
Bruichladdicl	1	1	2	2	0	2	2	1	2	2	2	2	Cluster 6
Bunnahabha	1	2	1	1	0	1	1	1	1	2	2	3	Cluster 3
Caol Ila	3	1	4	2	1	0	2	0	2	1	1	1	Medicinal or Smokey
Cardhu	1	3	1	0	0	1	1	0	2	2	2	2	Cluster 3
Clynelish	3	2	3	3	1	0	2	0	1	1	2	0	Medicinal or Smokey
Craigallechie	2	2	2	0	1	2	2	1	2	2	1	4	Cluster 2
Craigganmor	2	3	2	1	0	0	1	0	2	2	2	2	Cluster 8
Dailuaine	4	2	2	0	0	1	2	2	2	2	2	1	Cluster 0
Dalmore	3	2	2	1	0	1	2	2	1	2	3	1	Cluster 0
Dalwhinnie	2	2	2	0	0	2	1	0	1	2	2	2	Cluster 3
Deanston	2	2	1	0	0	2	1	1	1	3	2	1	Cluster 1
Dufftown	2	3	1	1	0	0	0	0	1	2	2	2	Cluster 8
Edradour	2	3	1	0	0	2	1	1	4	2	2	2	Cluster 1
GlenDeveror	2	3	1	1	1	1	1	2	0	2	0	1	Cluster 7
GlenElgin	2	3	1	0	0	2	1	1	1	1	2	3	Cluster 3
GlenGarioch	1	3	0	0	0	0	3	1	0	2	2	2	Cluster 5
GlenGrant	2	0	0	0	1	0	1	2	1	2	2	1	Cluster 3
GlenKeith	2	1	0	0	1	2	1	2	1	2	2	1	Cluster 8

GRACIAS

¿PREGUNTAS?