

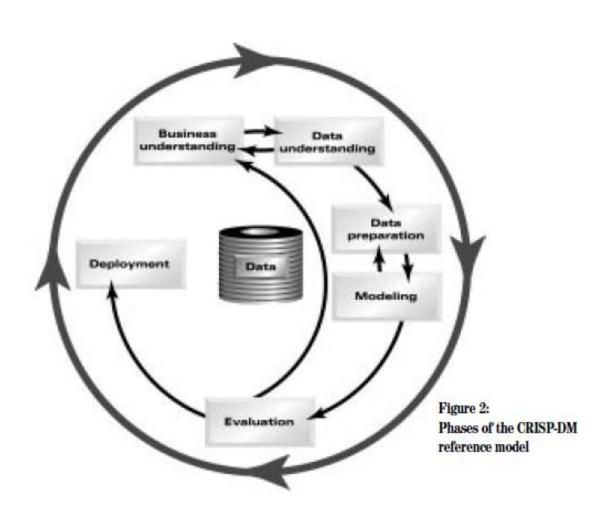


INTRODUCCIÓN AL APRENDIZAJE AUTOMÁTICO

CICLO EXPERIMENTAL EN CIENCIA DE DATOS

LAURA DIAZ DÁVILA – FRANCISCO TAMARIT

CRISP-DM



DOMINIO PROBLEMA

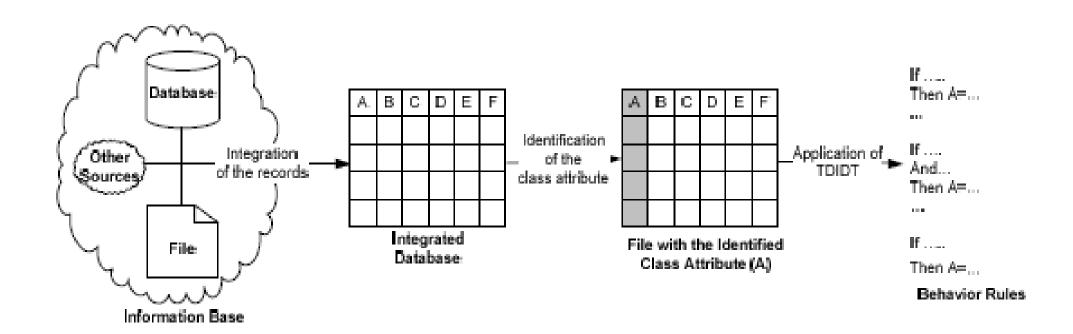
EXPERTOS DEL DOMINIO

EXPERTOS EN SOFTWARE

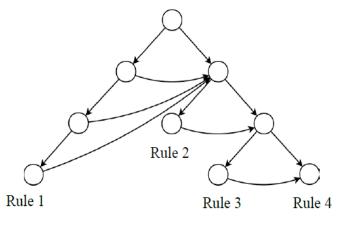
EXPERTOS EN SISTEMAS

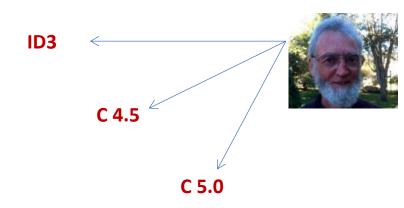
DATA MINNERS

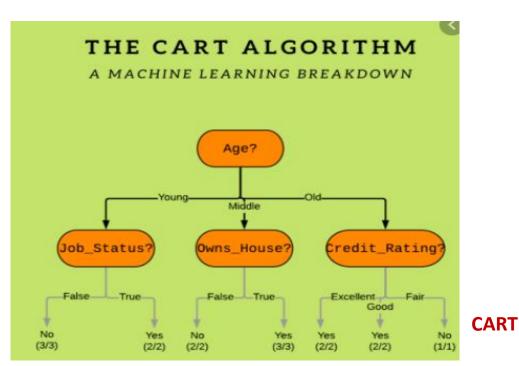
DISEÑO DE PROCESOS INTELIGENTES PROCESO: DESCUBRIMIENTO DE REGLAS DE COMPORTAMIENTO



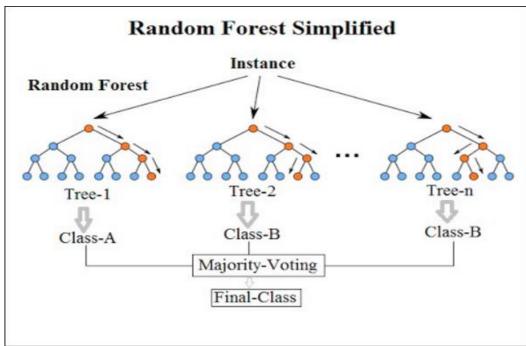
MODELO DE MACHINE LEARNING: TOP DOWN INDUCTION DECISION TREE



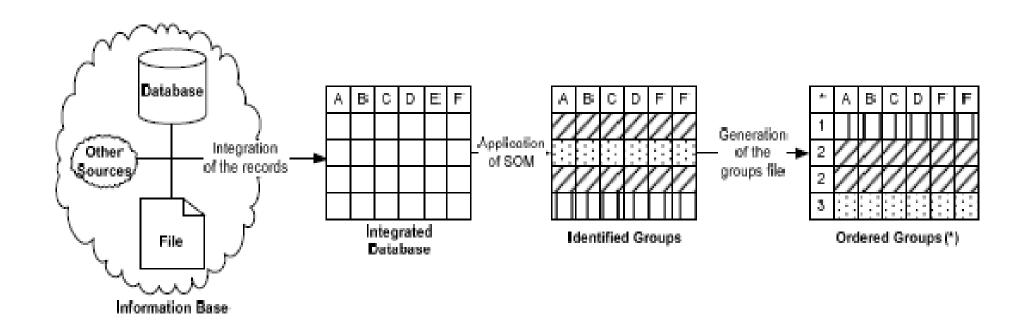




RANDOM FOREST



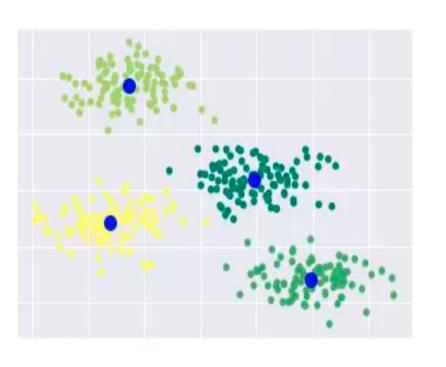
DISEÑO DE PROCESOS INTELIGENTES PROCESO: DESCUBRIMIENTO DE GRUPOS

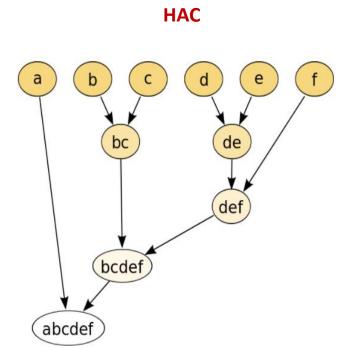


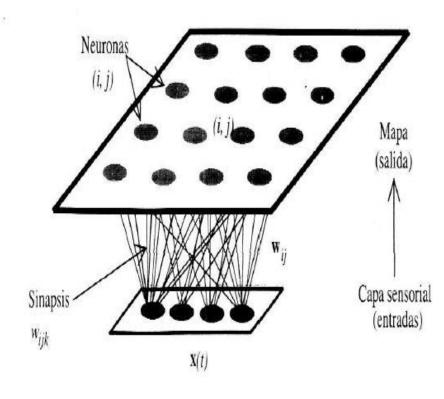
MODELO DE MACHINE LEARNING: MAPAS AUTOORGANIZADOS O CLUSTERING

K-MEANS

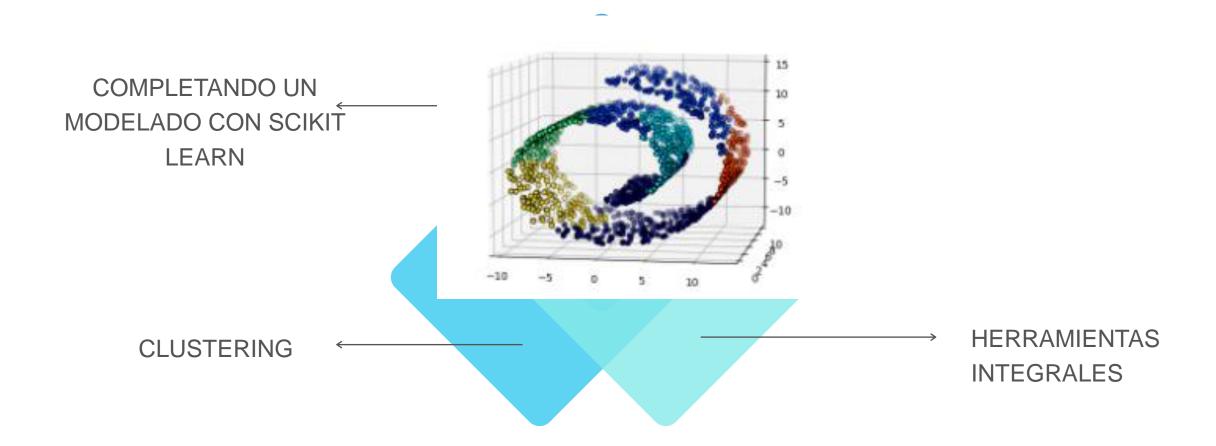
SOM - KOHONEN





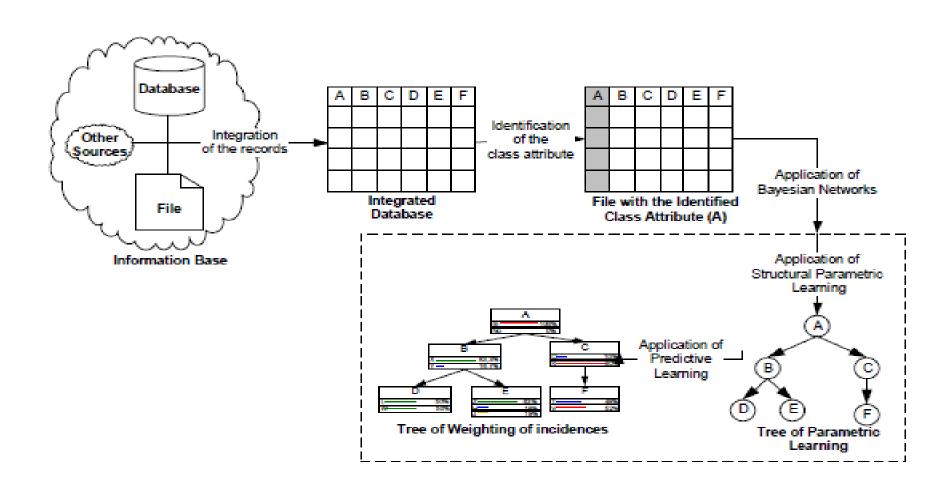


Procesos de descubrimientos de agrupamientos

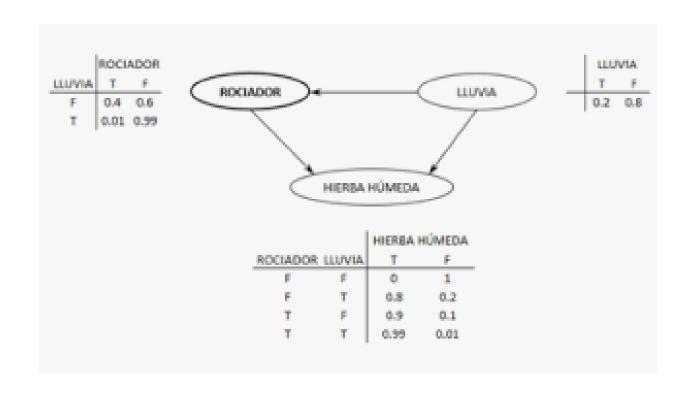


DISEÑO DE PROCESOS INTELIGENTES PROCESO:

DESCUBRIMIENTO DE ATRIBUTOS SIGNIFICATIVOS O INTERDEPENDENCIA



MODELO DE MACHINE LEARNING: REDES BAYESIANAS

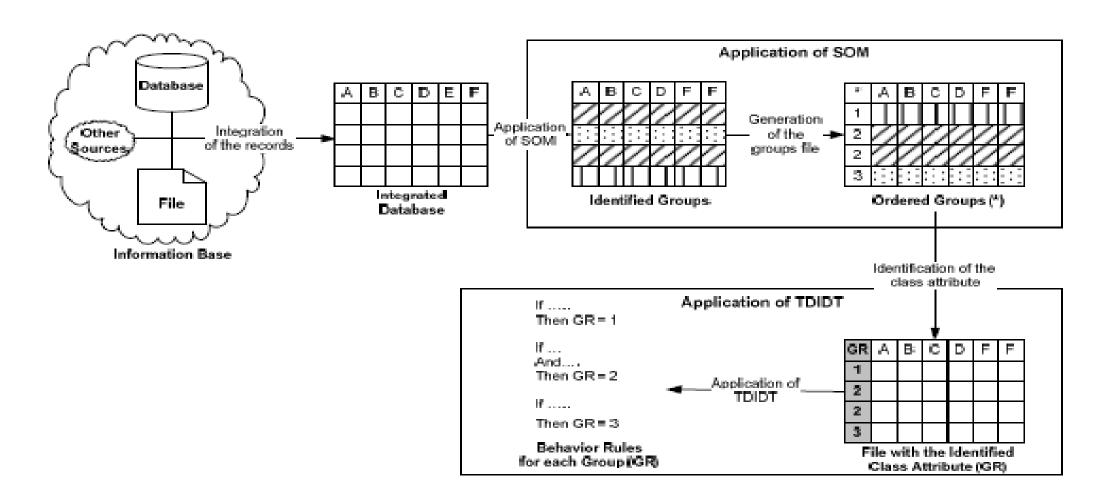




Naive bayes continuous

DISEÑO DE PROCESOS INTELIGENTES PROCESO:

DESCUBRIMIENTO DE REGLAS DE PERTENENCIA A LOS GRUPOS



*Ejemplo FACULTAD DE CIENCIAS ECONÓMICAS*6500 REGISTROS

Attribute	Target	Input	Illustrative
A¤en la carrera	-	-	-
Aprobo informatica	-	yes	-
Demora en Cursarla	-	yes	-
Aprobo Infen Cursada	-	yes	-
CumplePlan	-	yes	-
Procedencia_Argentina	-	yes	-
Procedencia_Cordoba	-	yes	-
Procedencia Cordoba Capital	-	yes	-
CosteaEst_Trabajo	-	yes	-
CosteaEst_Familia	-	yes	-
CosteaEst_Beca	-	yes	-
Padre <u>Ultimos</u> estudios Grupos	-	yes	-
Madre <u>Ultimos</u> estudios grupos	-	yes	-
Ritmolnicial	-	yes	-
Cantidad deportes	-	-	-
Escala deportes	-	yes	-
pc casa	-	yes	-
PC universidad	-	yes	-
Internet casa	-	yes	-
Internetuniversidad	-	yes	-
internet para ocio	-	yes	-
Internet para capacitarse	-	yes	-
Ingles	-	yes	-
Portugues	-	yes	-
Cant idiomas	-	-	-
Escala idiomas	-	yes	-

SOM

- Aprobo Inf en Cursada < 0,5000
 - Ingles < 0,5000 then Cluster_SOM_1 = c_som_2_2 (100,00 % of 381 examples)
 - Ingles >= 0,5000
 - Procedencia_Cordoba < 0,5000
 - Demora en Cursarla < 0,5000 then Cluster_SOM_1 = c_som_3_2 (53,68 % of 190 examples)
 - Demora en Cursarla >= 0,5000 then Cluster_SOM_1 = c_som_2_1 (99,64 % of 842 examples)
 - Procedencia_Cordoba >= 0,5000
 - Portugues < 0,5000
 - Demora en Cursarla < 0,5000 then Cluster_SOM_1 = c_som_3_2 (84,82 % of 494 examples)
 - Demora en Cursarla >= 0,5000 then Cluster_SOM_1 = c_som_1_2 (91,99 % of 1811 examples)
 - Portugues >= 0,5000
 - CumplePlan < 2,5000 then Cluster_SOM_1 = c_som_1_1 (96,90 % of 2127 examples)
 - CumplePlan >= 2,5000 then Cluster_SOM_1 = c_som_3_2 (65,32 % of 222 examples)
- Aprobo Inf en Cursada >= 0,5000 then Cluster_SOM_1 = c_som_3_1 (98,97 % of 584 examples)

Subgro	oup = Local				
Examples		[33,0 %] 219			
Att - Desc	Test value	Group	Overral		
Continuous attributes : Mean (StdDev)				
Portugues	57,82	0,97 (0,16)	0,47 (0,50)		
Escala idiomas	46,85	2,20 (0,40)	1,60 (0,73)		
Procedencia_Cordoba	27,17	1,00 (0,00)	0,82 (0,39)		
Demora en Cursaria	19,30	1,49 (0,64)	1,21 (0,81)		
CosteaEst_Trabajo	16,60	0,71 (0,45)	0,57 (0,50)		
Procedencia_Cordoba_Capital	15,01	0,63 (0,48)	0,50 (0,50)		
Ingles	14,31	1,00 (0,00)	0,94 (0,24)		
Internet para capacitarse	10,64	0,32 (0,47)	0,24 (0,43)		
Aprobo informatica	4,44	0,36 (0,48)	0,33 (0,47)		
Procedencia_Argentina	3,59	1,00 (0,00)	1,00 (0,06)		
pc casa	3,15	0,95 (0,22)	0,94 (0,24)		
Internet casa	1,89	0,90 (0,30)	0,89 (0,31)		
PC universidad	1,67	0,23 (0,42)	0,22 (0,42)		
CosteaEst_Familia	0,56	0,02 (0,15)	0,02 (0,14)		
Internet universidad	0,01	0,20 (0,40)	0,20 (0,40)		
CosteaEst_Beca	-2,06	0,00 (0,02)	0,00 (0,05)		
Escala deportes	-2,18	0,95 (0,82)	0,99 (0,87)		
internet para ocio	-5,23	0,39 (0,49)	0,43 (0,50)		
Padre Ultimos estudios Grupos	-5,54	1,92 (0,83)	2,00 (0,85)		
Madre Ultimos estudios grupos	-6,51	2,12 (0,82)	2,22 (0,86)		
Aprobo Inf en Cursada	-17,76	0,00 (0,00)	0,09 (0,28)		
Ritmolnicial	-21,29	0,16 (0,54)	0,47 (0,84)		
CumplePlan	-26,00	0,22 (0,71)	0,91 (1,51)		

Rule: Cluster_SOM_1 = c_som_1_2

Subgrou	up = Local			
Examples	[2	6,2 %] 1741		
Att - Desc	Test value	Group	Overral	
Continuous attributes : Mean (S	tdDev)			
Procedencia_Cordoba	23,05	1,00 (0,00)	0,82 (0,39)	
Demora en Cursarla	20,10	1,55 (0,58)	1,21 (0,81)	
Ingles	12,13	1,00 (0,00)	0,94 (0,24)	
Procedencia_Cordoba_Capital	11,06	0,61 (0,49)	0,50 (0,50)	
CosteaEst_Trabajo	7,16	0,64 (0,48)	0,57 (0,50)	
Procedencia_Argentina	3,04	1,00 (0,00)	1,00 (0,06)	
CosteaEst_Familia	2,86	0,03 (0,17)	0,02 (0,14)	
CosteaEst_Beca	-0,40	0,00 (0,04)	0,00 (0,05)	
Internet casa	-2,18	0,88 (0,33)	0,89 (0,31)	
Internet para capacitarse	-2,60	0,22 (0,41)	0,24 (0,43)	
pc casa	-2,91	0,92 (0,27)	0,94 (0,24)	
internet para ocio	-6,19	0,37 (0,48)	0,43 (0,50)	
Escala deportes	-7,37	0,85 (0,85)	0,99 (0,87)	
Madre Ultimos estudios grupos	-8,01	2,08 (0,88)	2,22 (0,86)	
Internet universidad	-8,10	0,13 (0,34)	0,20 (0,40)	
Padre Ultimos estudios Grupos	-9,46	1,84 (0,80)	2,00 (0,85)	
PC universidad	-9,85	0,14 (0,35)	0,22 (0,42)	
CumplePlan	-10,09	0,60 (1,12)	0,91 (1,51)	
Aprobo Inf en Cursada	-15,06	0,00 (0,00)	0,09 (0,28)	
Aprobo informatica	-15,48	0,18 (0,38)	0,33 (0,47)	
Ritmolnicial	-22,49	0,08 (0,40)	0,47 (0,84)	
Escala idiomas	-29,47	1,15 (0,37)	1,60 (0,73)	
Portugues	-45,68	0,00 (0,00)	0,47 (0,50)	

K-means

```
    Aprobo Inf en Cursada < 0,5000</li>

    Portugues < 0,5000</li>

    Ingles < 0,5000 then Cluster KMeans 1 = c kmeans 3 (97,76 % of 357 examples)</li>

              Ingles >= 0,5000

    pc casa < 0,5000 then Cluster_KMeans_1 = c_kmeans_3 (100,00 % of 185 examples)</li>

                     pc casa >= 0,5000

    Ritmolnicial < 0,5000</li>

    CumplePlan < 1,5000 then Cluster KMeans 1 = c kmeans 6 (95,39 % of 1650 examples)</li>

                                   o CumplePlan >= 1,5000
                                          o internet para ocio < 0,5000 then Cluster_KMeans_1 = c_kmeans_6 (62,86 % of 140 examples)</li>
                                          o internet para ocio >= 0,5000 then Cluster KMeans 1 = c kmeans 4 (85,92 % of 277 examples)

    Ritmolnicial >= 0,5000 then Cluster KMeans 1 = c kmeans 4 (95,70 % of 582 examples)

       Portugues >= 0,5000

    pc casa < 0,5000 then Cluster_KMeans_1 = c_kmeans_3 (95,36 % of 151 examples)</li>

              pc casa >= 0.5000
                     CumplePlan < 2,5000</li>

    Demora en Cursarla < 0,5000</li>

    internet para ocio < 0,5000 then Cluster KMeans 1 = c kmeans 1 (73,66 % of 186 examples)</li>

                                   o internet para ocio >= 0,5000 then Cluster KMeans 1 = c kmeans 4 (74,77 % of 111 examples)

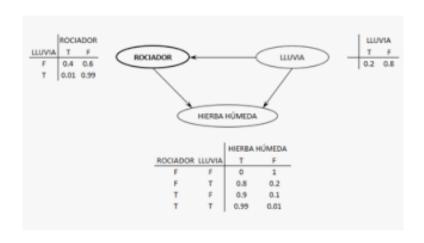
    Demora en Cursarla >= 0,5000 then Cluster KMeans 1 = c kmeans 1 (98,17 % of 2182 examples)

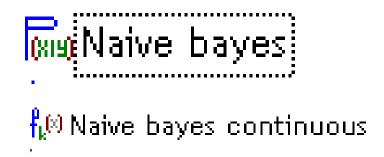
    CumplePlan >= 2,5000 then Cluster KMeans 1 = c kmeans 4 (80,49 % of 246 examples)

    Aprobo Inf en Cursada >= 0,5000 then Cluster_KMeans_1 = c_kmeans_2 (99,49 % of 584 examples)
```

Subgro	oup = Local	l.		1020000	weeks and		
Examples		[36,0 %] 2396		Subgroup = Local			
Att - Desc	Test value	Group	Overral	Examples			[8,7 %] 581
Continuous attributes : Mean (StdDev)			Att - Desc	Test value	Group	Overral
Portugues	63,19	0,98 (0,12)	0.47 (0.50)	Continuous attributes : Mean (StdDev)		
Escala idiomas	CHATCON !	2,21 (0,41)	SUBSTITUTE OF THE STREET	Aprobo Inf en Cursada	81,31	1,00 (0,00)	0,09 (0,28)
Demora en Cursaria	10000000	1,51 (0,62)	September September	Ritmolnicial	38,91	1,77 (0,61)	0,47 (0,84)
	0.0000000000000000000000000000000000000	INTERPRETATIONS	PODUROS POR DESIGNATION OF THE PERSON OF THE	Aprobo informatica	36,08	1,00 (0,00)	0,33 (0,47)
Ingles	VV52500	1,00 (0,00)	SCHOOLS NOT HOSE	CumplePlan	29,92	2,70 (1,71)	0,91 (1,51)
pc casa	110000000	1,00 (0,05)	ACCOMPANIAN STATES	Madre Ultimos estudios grupos	9,93	2,56 (0,76)	2,22 (0,86)
CosteaEst_Trabajo	14,19	0,68 (0,47)	0,57 (0,50)	internet para ocio	9,50	0,62 (0,49)	0,43 (0,50)
Internet para capacitarse	11,03	0,32 (0,47)	0,24 (0,43)	Padre Ultimos estudios Grupos	9,30	2,32 (0,82)	2,00 (0,85)
Internet casa	8,50	0,94 (0,24)	0,89 (0,31)	Internet casa	6,33	0,97 (0,17)	0,89 (0,31)
Aprobo informatica	7,63	0,39 (0,49)	0,33 (0,47)	PC universidad	5,69	0,32 (0,47)	0,22 (0,42)
Procedencia_Cordoba_Capital	5,02	0,54 (0,50)	0,50 (0,50)	pc casa	5,65	0.99 (0.09)	0.94 (0.24)
Procedencia_Argentina	3,83	1,00 (0,00)	1,00 (0,06)	Ingles	1,500	INTERNATION OF THE PARTY OF THE	0,94 (0,24)
Procedencia_Cordoba	2,87	0,83 (0,37)	0,82 (0,39)	Internet universidad	200	0,27 (0,44)	Promise Control of the Control of th
Escala deportes	1,43	1,01 (0,84)	0,99 (0,87)	Escala deportes	1,7,2 (1)	1,14 (0,87)	PASSAGRAPHICAL PROPERTY AND ADDRESS OF THE PASSAGRAPHICAL PROPERTY A
CosteaEst_Familia	-0,91	0,02 (0,14)	0,02 (0,14)	Procedencia Argentina		1,00 (0,00)	SECONDOCUSES
PC universidad	-1,34	0,21 (0,41)	0,22 (0,42)	CosteaEst_Beca	UTCOM	Contract Con	0,00 (0,05)
Padre Ultimos estudios Grupos	-1,43	1,98 (0,83)	2,00 (0,85)	Escala idiomas	0,20	1,61 (0,68)	1,60 (0,73)
Madre Ultimos estudios grupos	-1,62	2,20 (0,83)	2,22 (0,86)	Procedencia_Cordoba	-0,02	0,82 (0,39)	0,82 (0,39)
CosteaEst_Beca	-2,25	0,00 (0,02)	0,00 (0,05)	CosteaEst_Familia	-1,02	0,02 (0,12)	0,02 (0,14)
Internet universidad	-2,84	0,18 (0,38)	0,20 (0,40)	Portugues	-2,33	0,42 (0,49)	0,47 (0,50)
internet para ocio	-5,60	0,39 (0,49)	0,43 (0,50)	Procedencia_Cordoba_Capital	-2,97	0,44 (0,50)	0,50 (0,50)
Aprobo Inf en Cursada	-18,98	0,00 (0,00)	0,09 (0,28)	Internet para capacitarse	-4,71	0,16 (0,37)	0,24 (0,43)
Ritmolnicial	-24,85	0,13 (0,49)	0,47 (0,84)	CosteaEst_Trabajo	-13,84	0,29 (0,46)	0,57 (0,50)
CumplePlan	-30,01	0,17 (0,59)	0,91 (1,51)	Demora en Cursarla	-34,78	0,10 (0,30)	1,21 (0,81)
Discrete attributes : [Recall] Accuracy			Discrete attributes : [Recall] A	ccuracy			

MODELO DE MACHINE LEARNING: REDES BAYESIANAS







1.9.1. Gaussian Naive Bayes

1.9.2. Multinomial Naive Bayes

1.9.3. Complement Naive Bayes

1.9.4. Bernoulli Naive Bayes

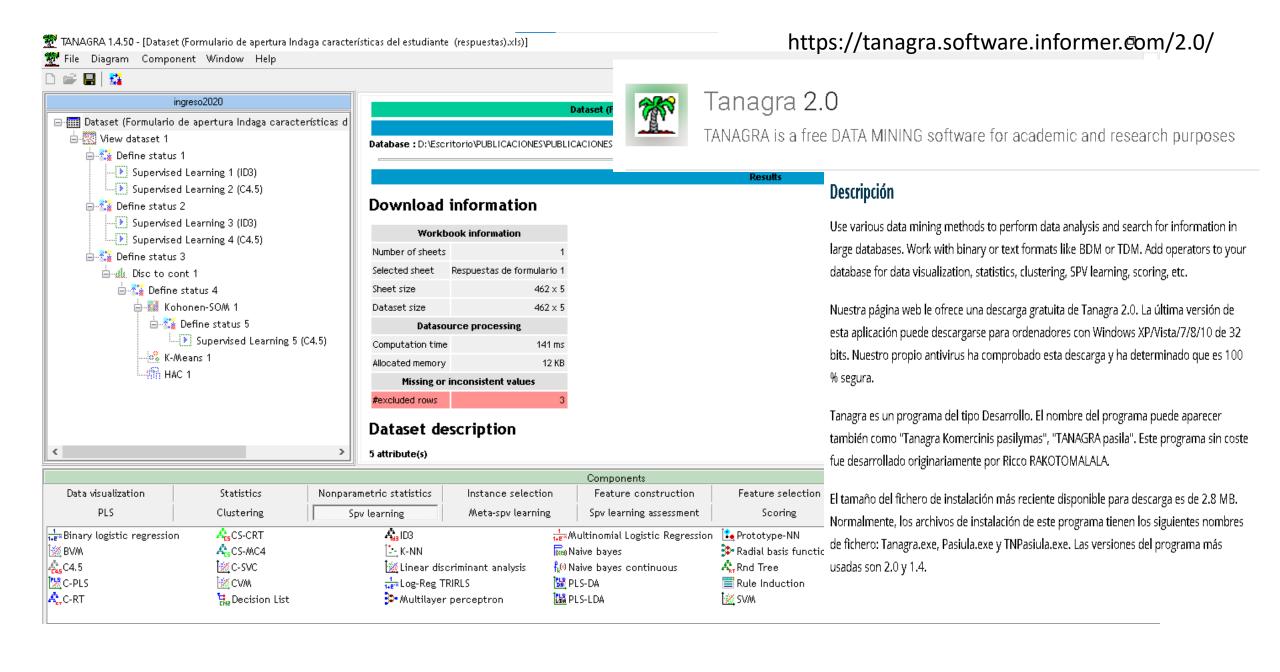
1.9.5. Categorical Naive Bayes

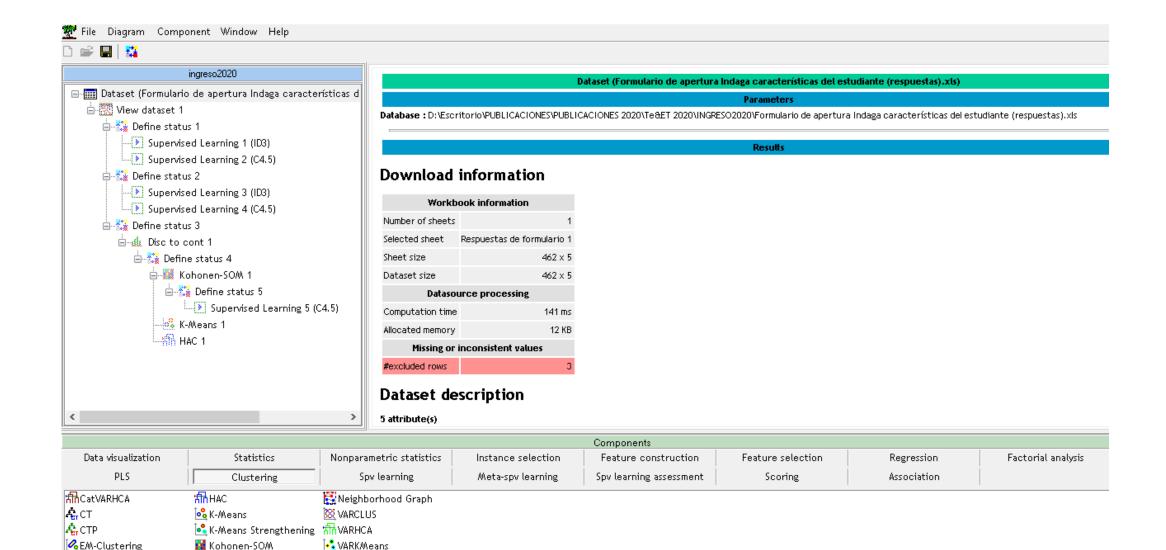
1.9.6. Out-of-core naive Bayes

model fitting

https://scikit-learn.org/stable/modules/naive_bayes.html







Kohonen-SOM

🏡 LVQ

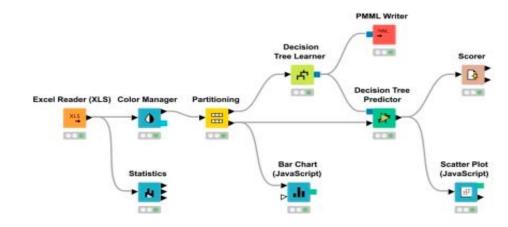
₹EM-Selection

VARKMeans



End to End Data Science

At KNIME, we build software to create and productionize data science using one easy and intuitive environment, enabling every stakeholder in the data science process to focus on what they do best.







Lightning-fast unified analytics engine

Download

Libraries -

Documentation -

Examples

Community -

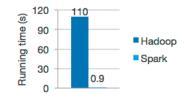
Developers -

Apache Spark™ is a unified analytics engine for large-scale data processing.

Speed

Run workloads 100x faster.

Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.



Logistic regression in Hadoop and Spark

Ease of Use

Write applications quickly in Java, Scala, Python, R. and SQL.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala, Python, R, and SQL shells.

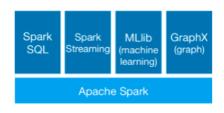
df = spark.read.json("logs.json") df.where("age > 21") .select("name.first").show()

Spark's Python DataFrame API Read JSON files with automatic schema inference

Generality

Combine SQL, streaming, and complex analytics.

Spark powers a stack of libraries including SQL and DataFrames, MLlib for machine learning, GraphX, and Spark Streaming. You can combine these libraries seamlessly in the same application.



Runs Everywhere

Spark runs on Hadoop, Apache Mesos, Kubernetes, standalone, or in the cloud. It can access diverse data sources.

You can run Spark using its standalone cluster mode, on EC2, on Hadoop YARN, on Mesos, or on Kubernetes. Access data in HDFS, Alluxio, Apache Cassandra, Apache HBase, Apache Hive, and hundreds of other data sources.



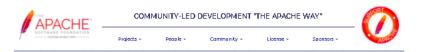












PLATINUM SPONSORS:



our Spansors. If you are interested in spansoring the ASF, please read our spansorship page











Tencent



Latest News

New repository service for sparkpackages (Apr 28, 2021)

Spark 3.1.1 released (Mar 02, 2021)

Spark 3.0.2 released (Feb 19, 2021)

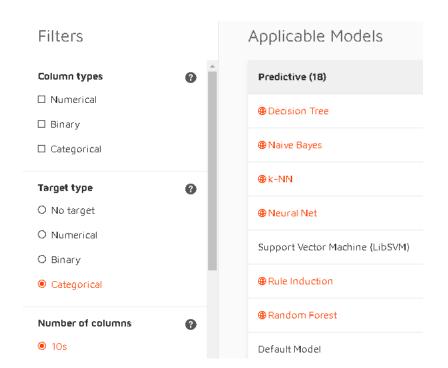
Next official release: Spark 3.1.1 (Jan 07, 2021)



In the news

Depth for Data Scient Simplified for Everyo

Join the 40,000+ global organizations in every indus RapidMiner data science platform to drive revenue, risk.



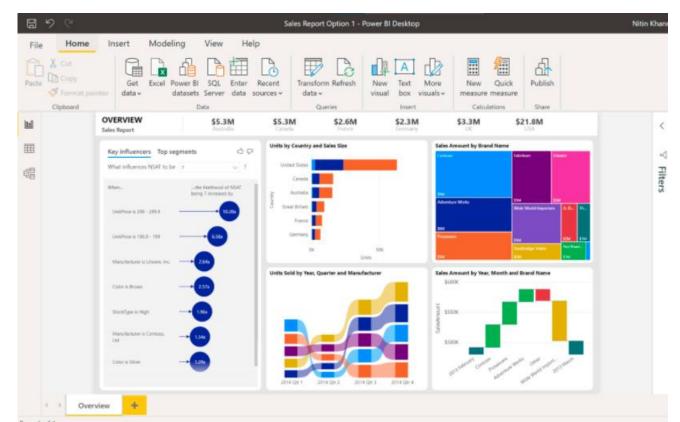
RapidMiner Named a Visionary in 2021 Gartner Magic Quadrant for Data Science and Machine Learning Platforms

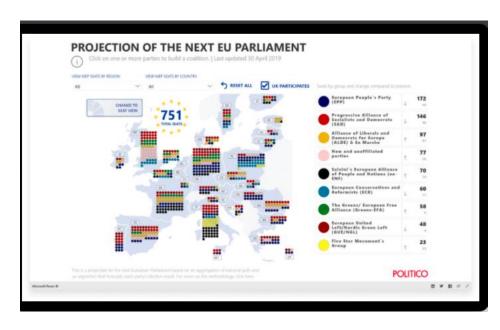
March 9, 2021

RapidMiner, a leading enterprise AI platform for people of all skill levels, today announced that it has been recognized as a Visionary in the 2021 Gartner Magic Quadrant for Data Science and Machine Learning



Filters	Applicable Models
— -···-, ☐ Categorical	Predictive (6)
La Categorical	
Target type	Decision Tree
O No target	⊕ Naive Bayes
O Numerical	⊕ k-NN
O Binary	
Categorical	Default Model
Number of columns (2)	Naive Bayes (Kernel)
O 10s	
O 100s	
O 1000s	
● 10,000s	Segmentations (1)





Vea qué dicen las empresas de investigación independientes

Gartner. Forrester

Por decimocuarto año consecutivo, Gartner otorga a Microsoft el reconocimiento de líder en el segmento de plataformas de análisis y business intelligence en el informe Gartner Magic Quadrant de 2021.1

Leer el informe >

Explore con el informe sobre el impacto económico total de Microsoft Power Bl cómo las organizaciones obtuvieron una rentabilidad de la inversión del 366 por ciento.

Leer el estudio > Ver el seminario web >

¿PREGUNTAS?

¡GRACIAS!