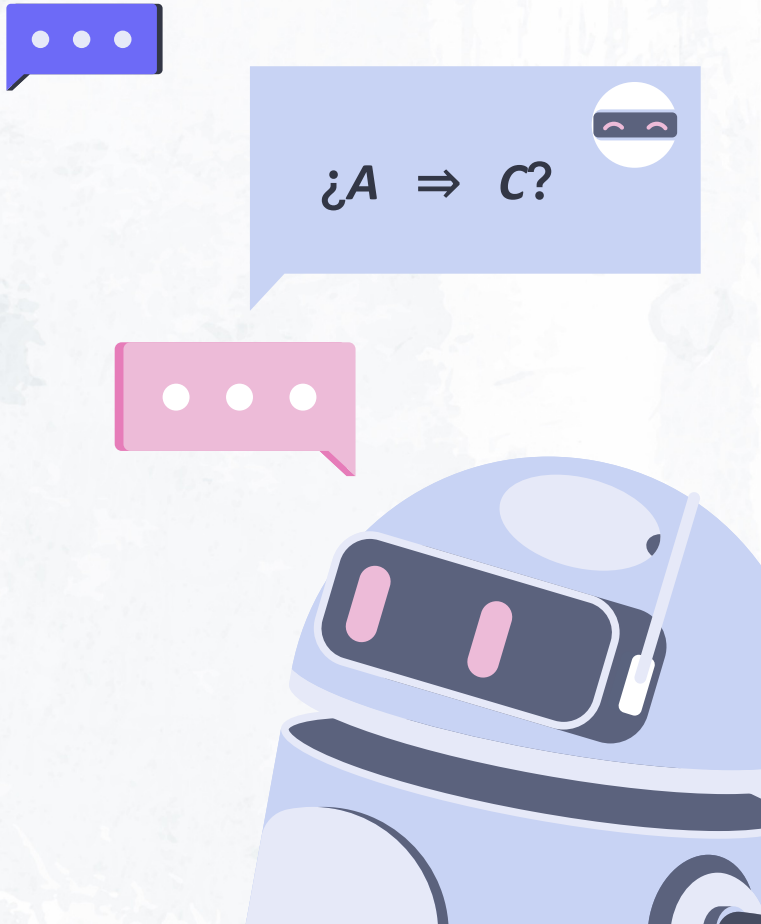
Reglas de asociación:

Metodología y Ejemplos



María Martínez Ballesteros
mariamartinez@us.es

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla

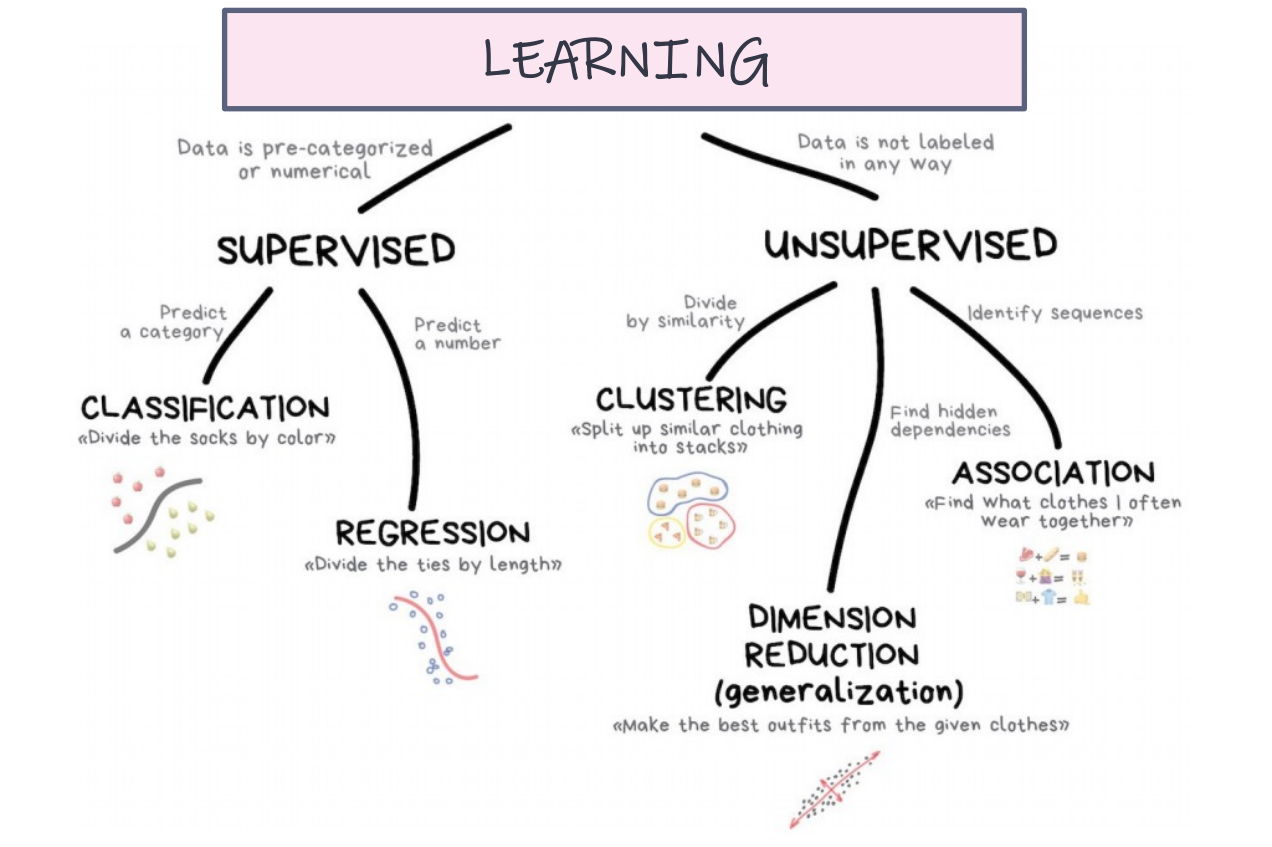
Índice

01 →

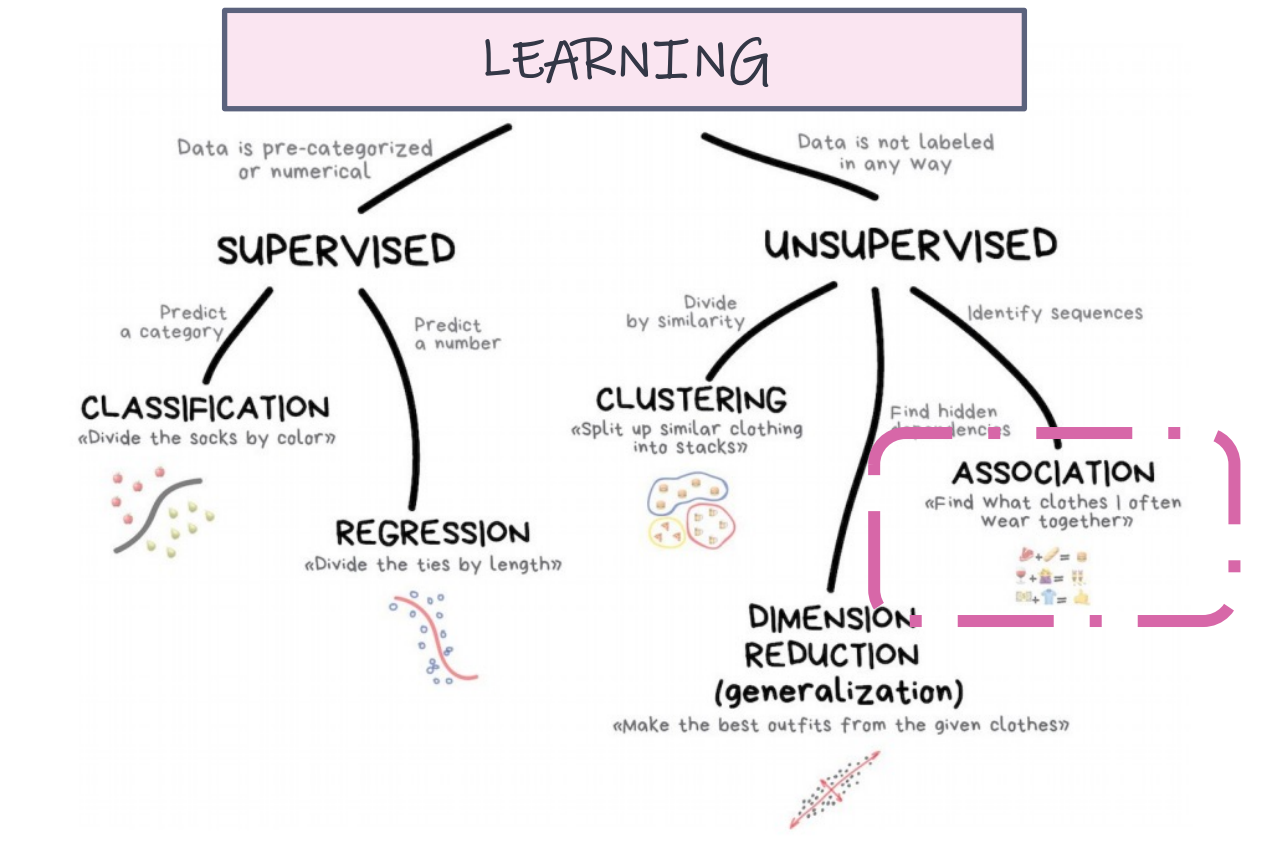
Introducción

(IA)

Vertientes en Machine Learning



Vertientes en Machine Learning



Supervisado vs No supervisado

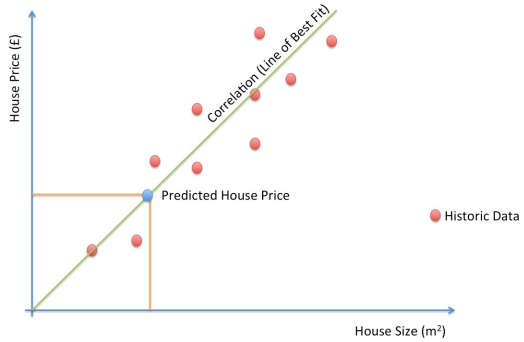
Supervisado	No Supervisado
Los datos de entrada están **etiquetados**	Los datos de entrada **NO están etiquetados**
Usa un conjunto de **entrenamiento** y otro de **validación**	Usa el conjunto de datos al **completo**
Usado para hacer **predicciones**	Usado para **análisis**
Número conocido de clases	Número de clases **desconocido**

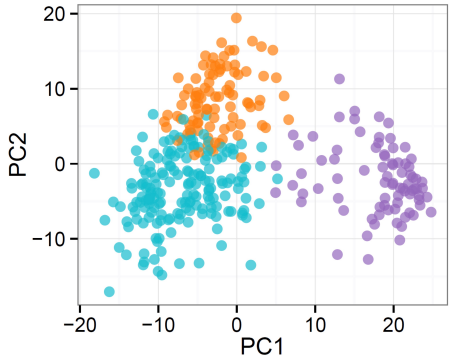
Supervisado vs No supervisado

Supervisado

No supervisado

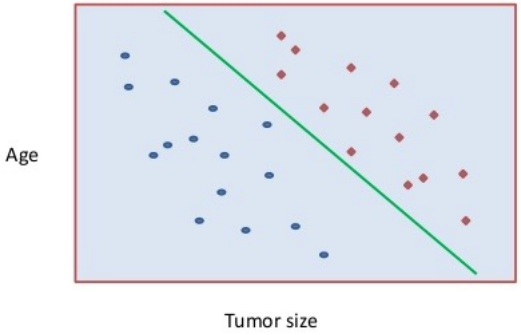
Regresión





Clustering

Clasificación



Asociaciones

02 →

Reglas de Asociación

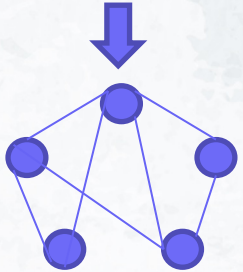
(IA)

Fenómenos naturales en los que algunas variables están relacionadas con otras

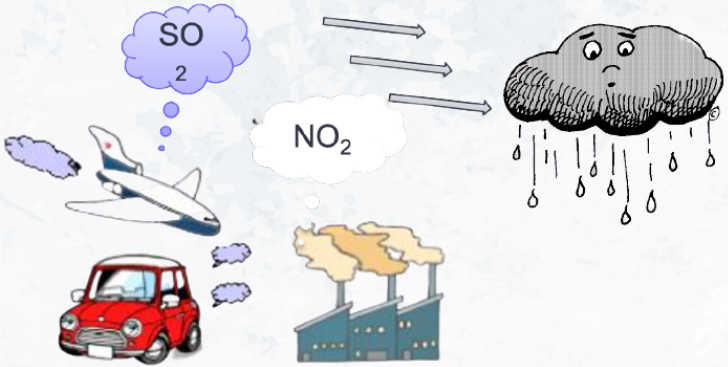
Bioinformática

¿Qué características tienen los genes en común?

¿Cómo interactúan entre ellos?



Medio ambiente



Encontrar relaciones entre variables → Reglas de asociación

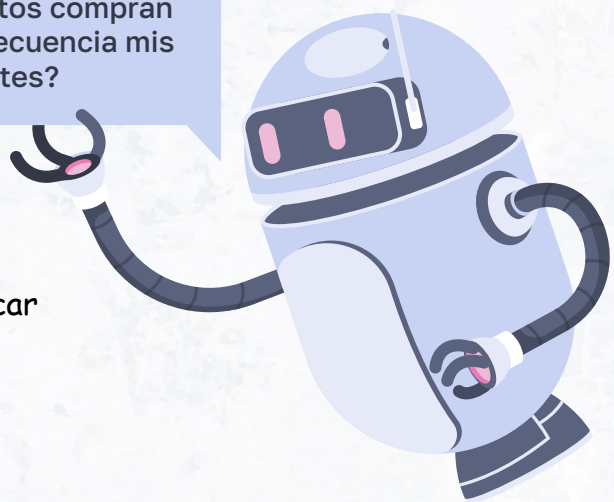
¿Qué son las reglas de asociación?

Se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos

Si **leche, huevos** entonces **¿pan?**

Si **azúcar** entonces **¿huevos?**

¿Qué productos compran juntos con frecuencia mis clientes?



Leche, huevos, azúcar, pan

Cliente 1

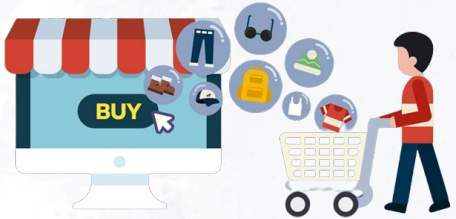
Leche, huevos, cereales, pan

Cliente 2

Huevos, azúcar

Cliente 3

¿Para qué se usan?



Supermercado:
identificar productos que se compran juntos con frecuencia
colocarlos para fomentar la venta cruzada.

Análisis comportamiento usuarios red social:
identificar patrones de interacción entre usuarios y predecir nuevas amistades o posibles conflictos.

Web de noticias: recomendar artículos relacionados con el tema que el usuario está leyendo, basándose en los patrones de lectura previos.

Sistema recomendación de películas:
identificar patrones en las valoraciones de los usuarios y recomendar películas similares.

Salud: identificar patrones en los datos de pacientes y predecir posibles enfermedades o problemas de salud.

Análisis de la cesta de la compra

PROBLEMA

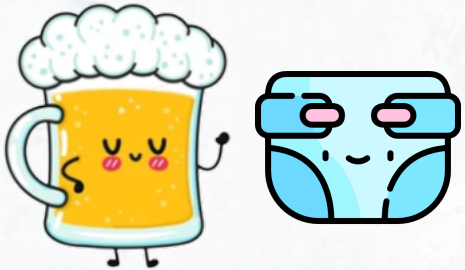
Dado un conjunto de transacciones, encontrar reglas que describen tendencias en los datos

Detectar cuándo la ocurrencia de un artículo está asociada a la ocurrencia de otros artículos en la misma compra

Análisis de la cesta de la compra

Colocación de productos en estanterías supermercado

- **Objetivo:** Identificar productos que muchos clientes compran juntos.
- **Solución:** Procesar los datos de los terminales de punto de venta proporcionados por los escáneres de código de barras.
- **Ejemplo:** Los viernes por la tarde, con frecuencia, quienes compran pañales, compran también cerveza.

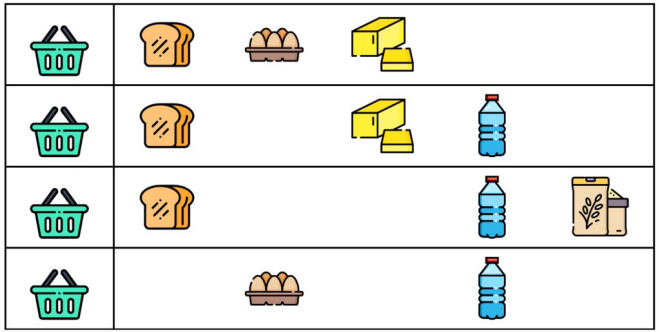


Análisis de la cesta de la compra

Promociones y ofertas

Regla del tipo: {café} → {leche}

- **Leche en consecuente** => Cómo incrementar las ventas
- **Café en antecedente** => Qué productos se verían afectados si se paran ventas de café.
- **Leche en el antecedente y café en el consecuente** => Qué productos deberían venderse con el café para promover las ventas de leche.



Análisis de la cesta de la compra

- En la terminología del "**análisis de la cesta de la compra**" los datos consisten en
  - Una serie de registros de transacciones
  - Cada registro contiene un conjunto de artículos comprados por un cliente

Cestas de la compra

Cliente	Compras
1	Leche, huevos, azúcar, pan
2	Leche, huevos, cereales, pan
3	Huevos, azúcar

Productos

- Leche
- Huevos
- Azúcar
- Pan
- Cereales

Análisis de la cesta de la compra

- En la terminología del "**análisis de la cesta de la compra**"
  - Los datos se pueden visualizar de forma tabular como una matriz

Cliente	Leche	Huevos	Azúcar	Pan	Cereales
1	1	1	1	1	0
2	1	1	0	1	1
3	0	1	1	0	0

Análisis de la cesta de la compra

En el ejemplo, S = {Leche, Huevos, Azúcar, Pan, Cereales}

Posibles preguntas:

- ¿Cuántos clientes compraron **leche** y **huevos**?
- ¿Qué otros productos compran generalmente los clientes que compran **leche**?

Cliente	Compras
1	Leche, huevos, azúcar, pan
2	Leche, huevos, cereales, pan
3	Huevos, azúcar

Posible regla:

- Si **leche** y **huevos** entonces **pan**
  - 66,6% de los clientes compran **leche**, **huevos** y **pan**
  - 100% de los clientes que compran **leche** y **huevos** también compran **pan**

Definiciones

Itemset:

- Conjunto de uno o más items, p.ej. {pan, leche}
- **K-itemset.** Itemset con k elementos

Soporte de un itemset (support):

- Fracción de las transacciones que contienen el itemset p.ej. sup({pan,leche}) = 2/3

Itemset frecuente:

- Itemset con soporte igual o superior a un umbral de soporte establecido por el usuario (MinSup).

Cliente	Compras
1	Leche, huevos, azúcar, pan
2	Leche, huevos, cereales, pan
3	Huevos, azúcar

Definiciones

Descubrimiento de Reglas de Asociación:

Aprendizaje no supervisado para descubrir relaciones entre atributos

Una regla de asociación es una **implicación** de la forma:

$$\boxed{A} \Rightarrow \boxed{C}$$

Antecedente **Consecuente**

A y C están formados por itemsets

Definiciones

Reglas de asociación booleanas:

- Asociaciones entre la presencia y ausencia de items. *E.g. compra A o no compra A*.

Leche y Huevos ⇒ Pan

Reglas de asociación nominales:

- Asociaciones entre las propiedades o valores de items.

Temperatura es FRÍA y Humedad es NORMAL ⇒ Jugar es SI

Reglas de asociación cuantitativas:

- Asociaciones entre items o atributos cuantitativos.

Temperatura $\in$ [38, 42] y Humedad $\in$ [25, 33] ⇒ Ozono Troposférico $\in$ [140, 206]

Definiciones

Regla de Asociación

- Implicación de la forma A -> C, donde A e C son itemsets

  {Leche, Huevos} -> {Pan}

Métricas de evaluación

- **Soporte**

  Fracción de ejemplos que contienen tanto A como C

$$Sup(A \rightarrow C) = P(A \wedge C) = \frac{\#(A \wedge C)}{N}$$

- **Confianza**

  Con qué frecuencia aparece C en los ejemplos que incluyen a A

$$Conf(A \rightarrow C) = P(A|C) = \frac{Sup(A \rightarrow C)}{Sup(A)}$$

Cliente	Compras
1	Leche, huevos, azúcar, pan
2	Leche, huevos, cereales, pan
3	Huevos, azúcar

Ejemplo: {Leche, Huevos} -> {Pan}

Sup(Leche & Huevos ⇒ Pan) = $\frac{2}{3} = 0.666 \rightarrow 66.6\%$

Conf(Leche & Huevos ⇒ Pan) =

$$= \frac{Sup(Leche\ \&\ Huevos \Rightarrow Pan)}{Sup(Leche\ \&\ Huevos)} = \frac{2}{2} = 1 \rightarrow 100\%$$

03 →

¿Cómo se evalúan?

(IA)

Otras medidas de evaluación

Instancia	F_1	F_2	F_3
t_1	35	183	88
t_2	42	154	47
t_3	37	186	93
t_4	30	199	112
t_5	33	173	83
t_6	24	178	75
t_7	63	177	91
t_8	22	167	60

A⇒B

R_1: $F_1 \in [30,38] \wedge F_2 \in [179,200] \Rightarrow F_3 \in [84,94]$

Sup(A) = 0.375 A se cumple en t_1, t_3 y t_4

Sup(B) = 0.375 B se cumple en t_1, t_3, t_7

Sup(A⇒B) = 0.25 A ⇒B se cumple en t_1 y t_3

Conf(A⇒B) = 0.25/0.375 = 0.666

A⇒C

R_2: $F_1 \in [30,38] \wedge F_2 \in [179,200] \Rightarrow F_3 \in [46,94]$

Sup(A) = 0.375 A se cumple en t_1, t_3 y t_4

Sup(C) = 0.875 C se cumple en todas excepto t_4

Sup(A⇒C) = 0.25 A ⇒C se cumplen en t_1 y t_3

Conf(A⇒C) = 0.25/0.375 = 0.666

¿Las 2 reglas tienen la misma calidad?

Otras medidas de evaluación

Interés: Lift

- Intervalo [0,∞)
- Valor < 1 Dependencia Negativa
- Valor > 1 Dependencia Positiva
- 1 Independencia

Factor de Certeza, Leverage

- Intervalo [-1,1]
- Valor < 0 Dependencia Negativa
- Valor > 0 Dependencia Positiva
- 0 Independencia

Implicación: Conviction

- Intervalo [0, ∞]

Lift: Cuándo una regla es mejor prediciendo el resultado que asumiendo el resultado de forma aleatoria

$$Lift(A \rightarrow C) = \frac{Conf(A \rightarrow C)}{Sup(C)}$$

Leverage: Proporción de ejemplos adicionales cubiertos por una regla (izquierda y derecha) sobre los cubiertos por cada parte si fueran independientes.

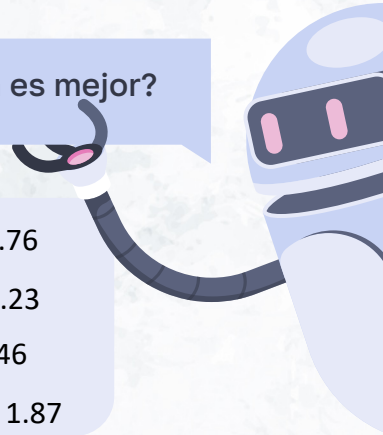
$$Lev(A \rightarrow C) = Sup(A \rightarrow C) - Sup(A) * Sup(C)$$

Conviction: Direccional y obtiene su máximo valor (infinito) si la implicación es perfecta, esto es, si siempre que A ocurre sucede también C.

$$Conv(A \rightarrow C) = \frac{Sup(A) * Sup(\neg C)}{Sup(A \rightarrow \neg C)} = \frac{1 - Sup(C)}{1 - Conf(A \rightarrow C)}$$

Otras medidas de evaluación

Instancia	F_1	F_2	F_3
t_1	35	183	88
t_2	42	154	47
t_3	37	186	93
t_4	30	199	112
t_5	33	173	83
t_6	24	178	75
t_7	63	177	91
t_8	22	167	60



A⇒B

R_1: $F_1 \in [30,38] \wedge F_2 \in [179,200] \Rightarrow F_3 \in [84,94]$

Sup(A) = 0.375 A se cumple en t_1, t_3 y t_4

Sup(B) = 0.375 B se cumple en t_1, t_3, t_7

Sup(A⇒B) = 0.25 A ⇒B se cumple en t_1 y t_3

Conf(A⇒B) = 0.25/0.375 = 0.666

Lift(A⇒B) = 1.76

Lev(A⇒B) = 0.23

CF(A⇒B) = 0.46

Conv(A⇒B) = 1.87

A⇒C

R_2: $F_1 \in [30,38] \wedge F_2 \in [179,200] \Rightarrow F_3 \in [46,94]$

Sup(A) = 0.375 A se cumple en t_1, t_3 y t_4

Sup(C) = 0.875 C se cumple en todas excepto t_4

Sup(A⇒C) = 0.25 A ⇒C se cumplen en t_1 y t_3

Conf(A⇒C) = 0.25/0.375 = 0.666

Lift(A⇒B) = 0.75

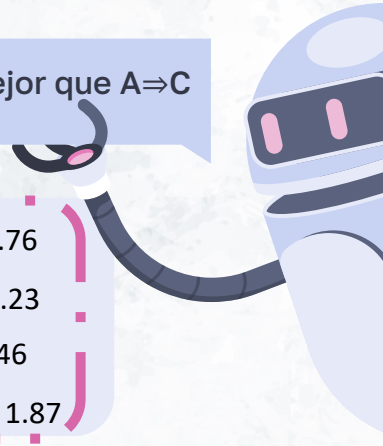
Lev(A⇒B) = -0.57

CF(A⇒B) = -0.24

Conv(A⇒B) = 0.37

Otras medidas de evaluación

Instancia	F_1	F_2	F_3
t_1	35	183	88
t_2	42	154	47
t_3	37	186	93
t_4	30	199	112
t_5	33	173	83
t_6	24	178	75
t_7	63	177	91
t_8	22	167	60



A⇒B es mejor que A⇒C

A⇒B

R_1: $F_1 \in [30,38] \wedge F_2 \in [179,200] \Rightarrow F_3 \in [84,94]$

Sup(A) = 0.375 A se cumple en t_1, t_3 y t_4

Sup(B) = 0.375 B se cumple en t_1, t_3, t_7

Sup(A⇒B) = 0.25 A ⇒B se cumple en t_1 y t_3

Conf(A⇒B) = 0.25/0.375 = 0.666

Lift(A⇒B) = 1.76

Lev(A⇒B) = 0.23

CF(A⇒B) = 0.46

Conv(A⇒B) = 1.87

A⇒C

R_2: $F_1 \in [30,38] \wedge F_2 \in [179,200] \Rightarrow F_3 \in [46,94]$

Sup(A) = 0.375 A se cumple en t_1, t_3 y t_4

Sup(C) = 0.875 C se cumple en todas excepto t_4

Sup(A⇒C) = 0.25 A ⇒C se cumplen en t_1 y t_3

Conf(A⇒C) = 0.25/0.375 = 0.666

Lift(A⇒B) = 0.75

Lev(A⇒B) = -0.57

CF(A⇒B) = -0.24

Conv(A⇒B) = 0.37

04 →

¿Cómo extraemos reglas de asociación?

(IA)

Extracción de reglas de asociación

- Dado un conjunto de transacciones, encontrar todas las reglas con:
  - **Soporte:** sup(X -> Y) >= umbral *minsup*
  - **Confianza:** conf (X-> Y) >= umbral *minconf*
- Solución por fuerza bruta
  - Buscar todas las reglas de asociación posibles
  - Calcular el soporte y la confianza para cada regla
  - Eliminar las reglas que no superen los umbrales mínimos de soporte y confianza

¡¡¡Computacionalmente muy costoso!!!

Extracción de reglas de asociación

Cliente	Compras
1	Leche, huevos, azúcar, pan
2	Leche, huevos, cereales, pan
3	Huevos, azúcar

Ejemplo de reglas:

{Leche, Huevos} -> {Pan} (sup = 0.66, conf = 1)
{Pan, Huevos} -> {Leche} (sup = 0.66, conf = 1)
{Leche} -> {Huevos, Pan} (sup = 0.66, conf = 0.66)
{Huevos} -> {Leche, Pan} (sup = 0.66, conf = 0.66)
{Pan} -> {Leche, Huevos} (sup = 0.66, conf = 1)

Observaciones:

- Todas las reglas son particiones binarias del mismo itemset: {Leche, Huevos, Pan}
- Las reglas originadas del mismo itemset tienen el **mismo soporte** pero la **confianza** puede ser **diferente**

Extracción de reglas de asociación

Enfoque basado en dos pasos:

1. **Generación de Itemset frecuentes**

   Generar todos los itemsets cuyo soporte ≥ minsup

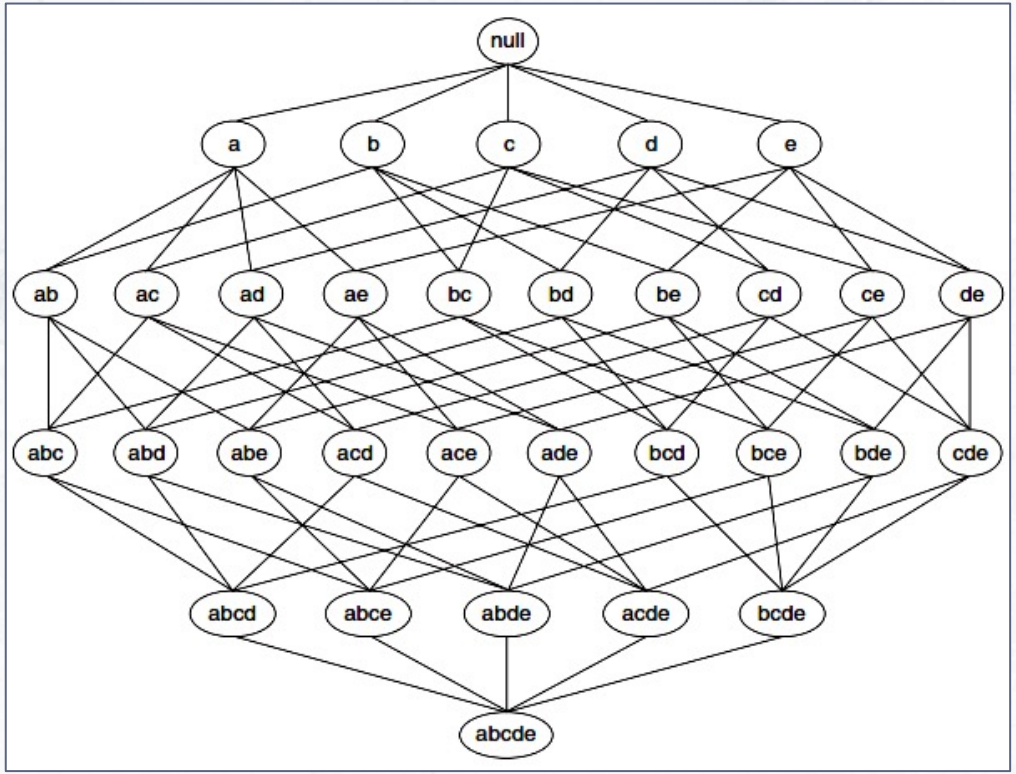
2. **Generación de reglas**

   Generar reglas con alta confianza (cuyo confianza ≥ minconf) a partir de cada itemset frecuente, donde cada regla es una partición binaria de un itemset frecuente

Generación de itemsets sigue siendo muy costosa computacionalmente

Extracción de reglas de asociación

- Generación de todos los itemsets posibles para 5 candidatos
- Dados *d* items, hay 2^d posibles itemsets candidatos



Extracción de reglas de asociación

Estrategias

- Reducir el numero de candidatos (M)
  - Uso de técnicas de poda
  - Ejemplo: algoritmos Apriori y DHP (Direct Hashing and Pruning)
- Reducir el numero de transacciones (N)
  - Reducir N conforme aumenta el tamaño del itemset
  - Ejemplo: algoritmo AprioriTID
- Reducir el numero de comparaciones (NM)
  - Uso de estructuras de datos eficientes para almacenar los candidatos o las entradas, de forma que no haya que comparar cada candidato con todas las transacciones.

Algoritmo clásico: Apriori

- Fue propuesto por Agrawal y Srikant en 1994.
- **Idea**: usar un conjunto de items (*itemset*) L con “k” atributos para generar uno nuevo con “k+1” atributos.

Si {A,B} es un *itemset* frecuente entonces {A} y {B} son también *itemsets* frecuentes

Objetivo:
Encontrar los itemsets L con mayor frecuencia.

Algoritmo clásico: Apriori

- **Principio**: Si un itemset es frecuente, entonces todos los subconjuntos deben ser frecuentes también.
- El soporte de un itemset nunca puede ser mayor que el de cualquiera de sus subconjuntos

$$\forall X, Y : (X \subseteq Y) => s(X) \geq s(Y)$$

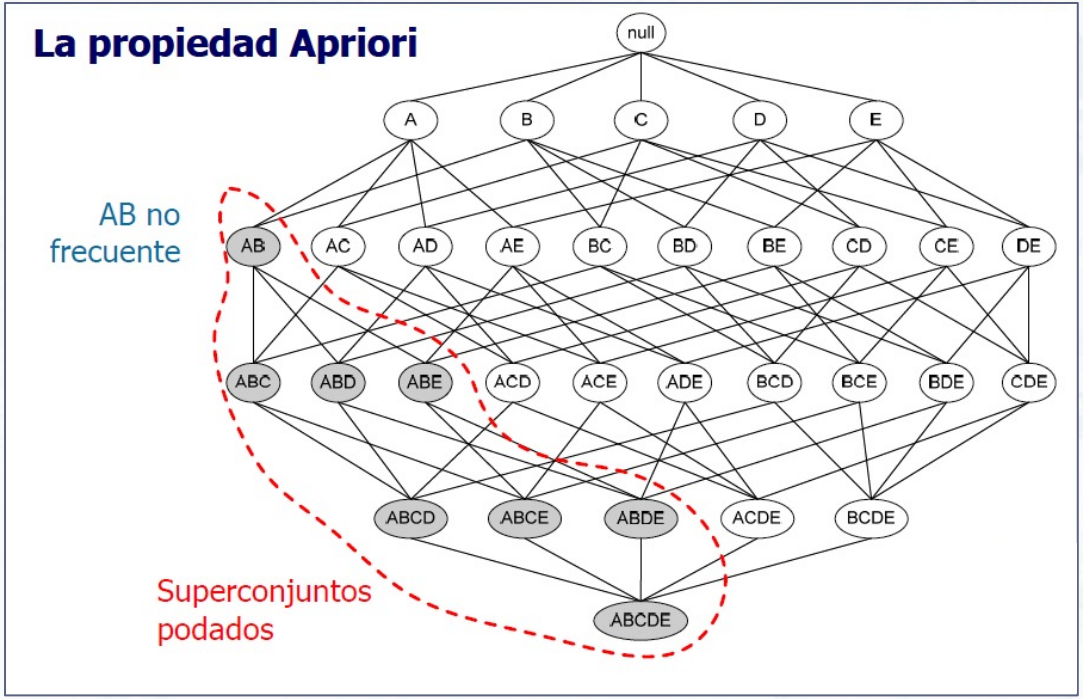
- Esta propiedad se conoce con el nombre de antimonotonía del soporte

Algoritmo clásico: Apriori

Algoritmo de descubrimiento de reglas de asociación basado en dos fases:

1. Búsqueda de itemsets frecuentes, es decir, aquellos que tengan un soporte mínimo predeterminado ($sup\ (A \rightarrow C) \geq MinSup$)
2. Generación de reglas a partir de dichos itemsets que superen una confianza mínima predeterminada ($conf\ (A \rightarrow C) \geq MinConf$)

Algoritmo clásico: Apriori



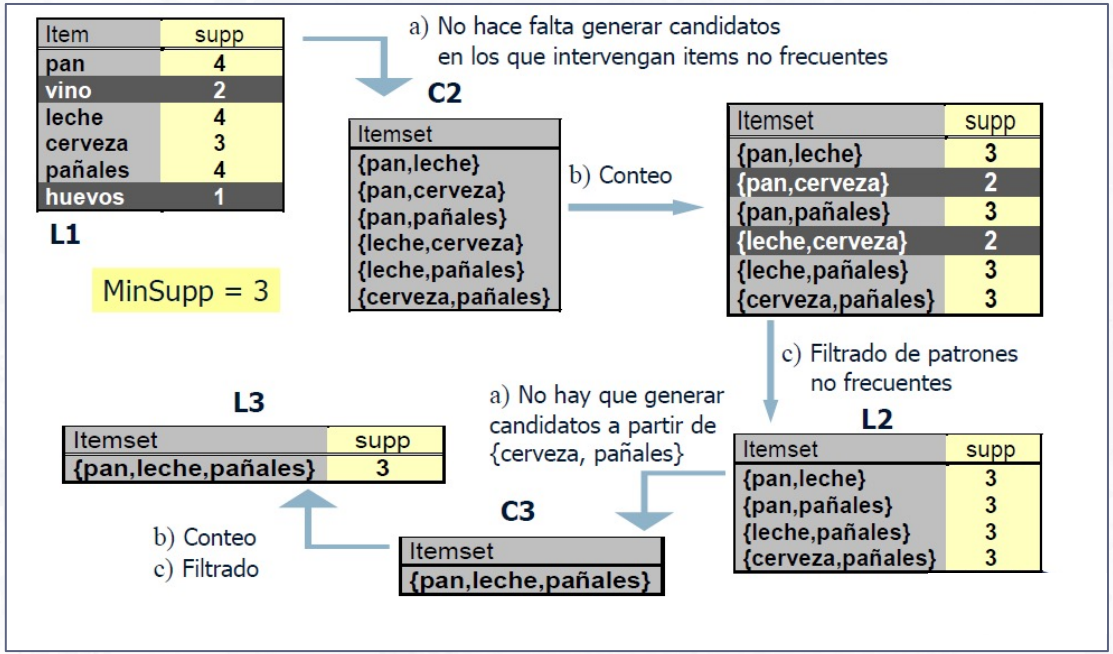
Algoritmo clásico: Apriori

Fase 1: Reducción del número de candidatos

- Generar todos los *itemsets* L con un único elemento. Usarlos para generar los *itemsets* con 2 elementos y así sucesivamente.
- Se toman todos los **posibles pares** cuyo soporte sea mayor o igual a ***minsup*** (lo cual permite ir eliminando algunas combinaciones).

Algoritmo clásico: Apriori

Ejemplo con soporte mínimo de 50% (tuplas)



Si se considerara cada subconjunto:

$$^{6}C_1 + {}^{6}C_2 + {}^{6}C_3 = 41$$

Con el filtrado basado en soporte:

$$6 + 6 + 1 = 13$$

Algoritmo clásico: Apriori

Fase 2: Generación de reglas

- Dado un *itemset* frecuente L, se encuentran todos los subconjuntos no vacíos que satisfacen el umbral de confianza mínima ***minconf***.
- Ejemplo: A partir del *itemset* frecuente {A,B,C,D} se generan las siguientes reglas candidatas y posteriormente se filtran con ***minconf***:

ABC-> D, ABD->C, ACD->B, BCD->A, A->BCD, B->ACD, C->ABD, D->ABC, AB->CD,

AC->BD, AD->BC, BC->AD, BD->AC, CD->AB

- Si |L| = k, entonces hay $2^k - 2$ reglas de asociación candidatas

Algoritmo clásico: Apriori

¿Cómo generar las reglas de forma eficiente?

- ¿Es la confianza anti-monótona como el soporte?

  NO: La confianza de ABC → D puede ser mayor o menor que la confianza de AB → D

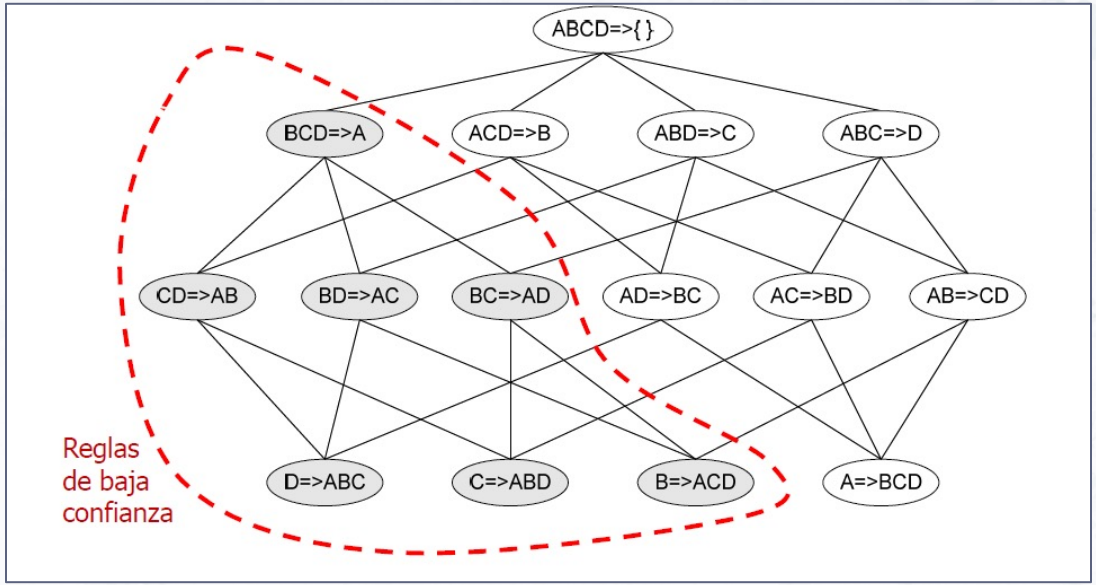
- Pero la confianza de las reglas generadas de un mismo itemset tienen una propiedad antimonótona:

  Para L = {A,B,C,D}: c(ABC → D) ≥ c(AB → CD) ≥ c(A → BCD)

- La confianza es antimonótona con respecto al número de items en la parte derecha de la regla

Algoritmo clásico: Apriori

¿Cómo generar las reglas de forma eficiente?



Algoritmo clásico: Apriori

Cliente	Compras
1	Leche, huevos, azúcar, pan
2	Leche, huevos, cereales, pan
3	Huevos, azúcar

Se comprueba que se cumple la regla anterior:

{Leche, Huevos} -> {Pan} (sup = 0.66, conf = 1)
{Leche} -> {Huevos, Pan} (sup = 0.66, conf = 0.66)

{Pan,Huevos} -> {Leche} (sup = 0.66, conf = 1)
{Huevos} -> {Leche, Pan} (sup = 0.66, conf = 0.66)

Algoritmo clásico: Apriori

Problemas de eficiencia

- **Elección del umbral:** umbrales soporte mínimo demasiado bajos -> muchos itemsets e incremento complejidad.
- **Número de ítems:** afecta al rendimiento del algoritmo.
- **Tamaño base de datos:** incremento tiempo de ejecución (múltiples pasadas a todos los datos).
- **Longitud transacciones:** aumento longitud itemsets frecuentes (más almacenamiento).

Otros algoritmos clásicos

AIS (*Agrawal* et al., 1993)

APRIORI (*Agrawal* et al., 1994)

SETM (*Houtsma and Swami*, 1995)

ECLAT (*Zaki, 2000*)

FP-GROWTH (*Han et al., 2004)*

Limitaciones

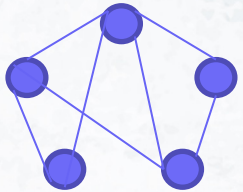
- No trabajan con datos reales o continuos
- Requieren discretización previa
- Umbrales mínimos de soporte y confianza
- Formato reglas: 1 atributo en el consecuente

Fenómenos naturales en los que algunas variables están relacionadas con otras

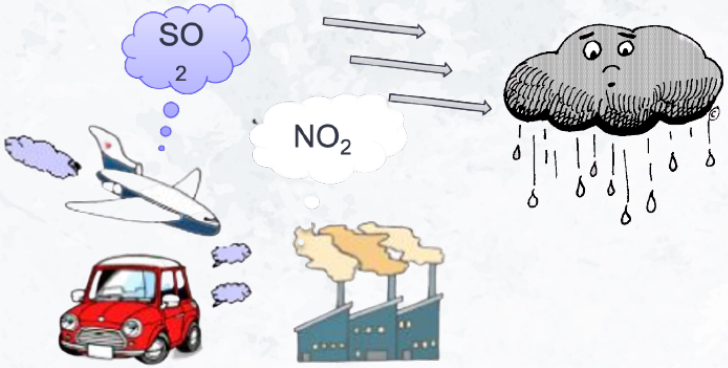
Bioinformática

¿Qué características tienen los genes en común?

¿Cómo interactúan entre ellos?



Medio ambiente



Encontrar relaciones entre variables → Reglas de asociación

Definiciones

Reglas de asociación booleanas:

- Asociaciones entre la presencia y ausencia de items. *E.g. compra A o no compra A*.

Leche y Huevos ⇒ Pan

Reglas de asociación nominales:

- Asociaciones entre las propiedades o valores de items.

Temperatura es FRÍA y Humedad es NORMAL ⇒ Jugar es SI

Reglas de asociación cuantitativas:

- Asociaciones entre items o atributos cuantitativos.

Temperatura ∈ [38, 42] y Humedad ∈ [25, 33] ⇒ Ozono Troposférico ∈ [140, 206]

Algoritmos evolutivos para RA Cuantitativas

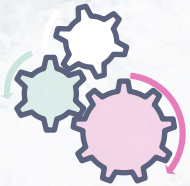
$A \Rightarrow C$

Temperatura $\in$ [38, 42] y
Humedad $\in$ [25, 33] $\Rightarrow$ Llover NO

RA en datos continuos, booleanos y categóricos:

- Atributos continuos: Intervalos adaptativos sin discretización
- Consecuente: puede ser fijo o no fijo

Esquema de aprendizaje:



- Basado en Computación Evolutiva
- Aprendizaje iterativo para cubrir todas las instancias
- Aprendizaje incremental basado en ventanas: conjuntos de datos de gran escala[3]

Objetivos seleccionados usando análisis estadístico

Parámetros para definir tipos de reglas deseados por usuario: reglas específicas, reglas genéricas, etc.

Algoritmos evolutivos para RA Cuantitativas

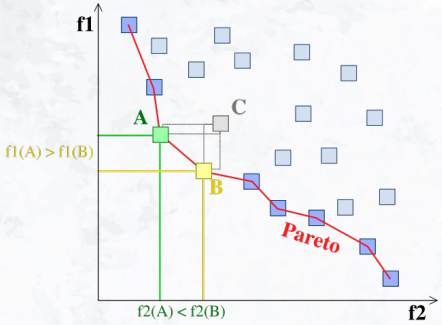
Algoritmo QARGA[1]:

Quantitative Association Rules by Genetic Algorithm

- Función fitness de pesos ponderados para guiar la búsqueda

Algoritmo MOQAR[2]:

Multi-Objective evolutionary algorithm to discover Quantitative Association Rule



- Aprendizaje basado en Pareto: Conjunto de soluciones óptimas
- Mejor equilibrio entre dos o más objetivos en conflicto.

[1]*M. Martínez Ballesteros, F. Martínez-Álvarez, A. Troncoso Lora, J.C. Riquelme Santos. "An Evolutionary Algorithm to Discover Quantitative Association Rules in Multidimensional Time Series". Soft Computing (SOCO). Vol. 15, No. 10, pp. 2065-2084, 2011.*

[2]*M. Martínez Ballesteros, F. Martínez-Álvarez, A. Troncoso Lora, J.C. Riquelme Santos. "Improving a multi-objective evolutionary algorithm to discover quantitative association rules". Knowledge and Information Systems 49*, pp 481–509, 2016.

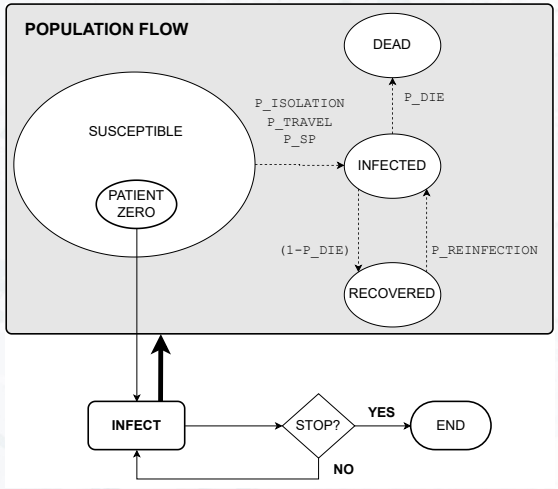
[3]*M. Martínez Ballesteros, J. Bacardit, A. Troncoso Lora, J.C. Riquelme Santos. "Enhancing the scalability of a genetic algorithm to discover quantitative association rules in large-scale datasets". Integrated Computer-Aided Engineering 22*, pp 21–39, 2015.

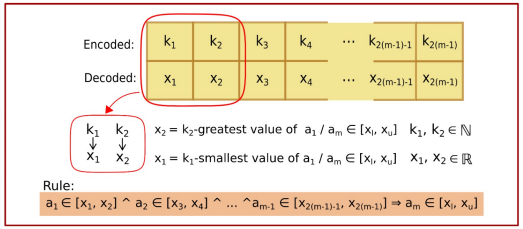
Evolutivo CVOA para RA Cuantitativas

Intervalos basados en k mayores y menores valores

Búsqueda iterativa basada en el modelo de propagación de COVID-19 (algoritmo de optimización de coronavirus – CVOA)

Optimización basada en el interés de las reglas (lift) para caracterizar outliers



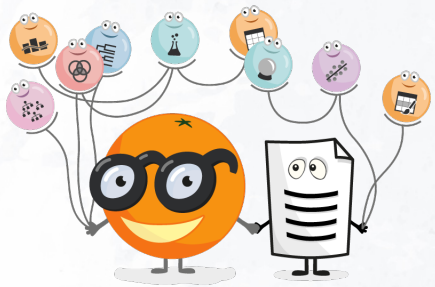


05 →

¿Qué herramientas existen para RA?

(IA)

Herramientas



Orange
(FP-Growth)

Weka
(Apriori, FP-Growth)

R
- ARules: Apriori, FpGrowth, Eclat
- ARulesViz

- Mlxtend: Apriori, FpGrowth
- Apyori, Pycaret

Open for Innovation KNIME

(Apriori)

GitHub

https://orangedatamining.com/widget-catalog/associate/associationrules/

https://cran.r-project.org/web/packages/arules/arules.pdf

06 →

Aplicaciones y ejemplos

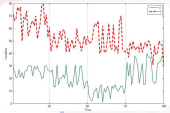
(IA)

Aplicaciones

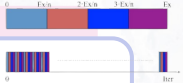
Reglas de asociación

$A \Rightarrow C$

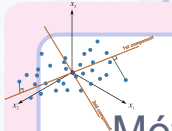
Datos climatológicos

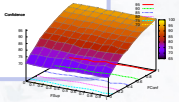


Datos gran escala



Métricas





Educación

Datos biomédicos

Explicabilidad

06 →

Aplicaciones y ejemplos:

(a) Climatológicos y medioambientales

(IA)

Contaminación Atmosférica: O_3, NO, SO_2

RAs para condiciones climatológicas con reespecto a O_3 (mg / m^3) y SO_2 (mg / m^3)

Rule	Conf. (%)	Lift
Temp. $\in$ [38, 42] and Hum. $\in$ [25, 33] and Hour $\in$ [15, 18] $\Rightarrow$ O_3 $\in$ [140, 206]	90	6.61
Temp. $\in$ [16, 22] and Hum. $\in$ [75, 90] $\Rightarrow$ O_3 $\in$ [22, 110]	100	1.43
Temp. $\in$ [42.9, 45.0] $\Rightarrow$ SO_2 $\in$ [3.7, 7.5]	100	1.72

Reglas de asociación encontradas por Apriori

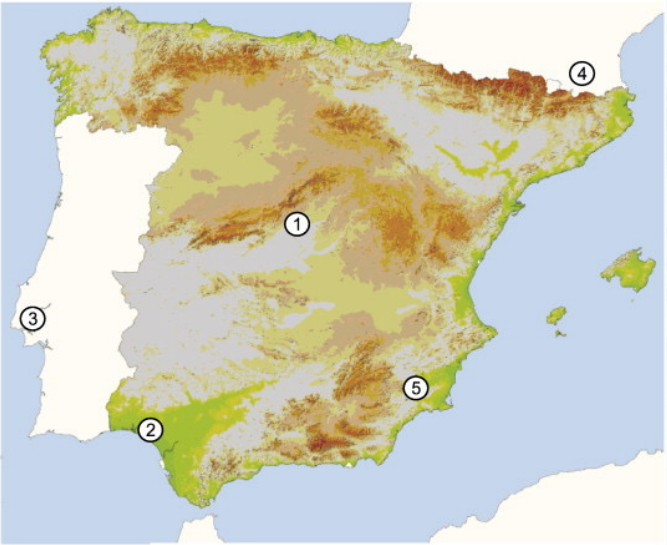
Rule	Conf. (%)	Lift	#intervals discretization
Temp. $\in$ [24, 27] $\Rightarrow$ O_3 $\in$ [90, 115]	33	1.43	10
Hum. $\in$ [14, 40] and Dir. $\in$ [120, 240] $\Rightarrow$ O_3 $\in$ [99, 183]	73	1.80	3

*** Todas las variables fueron proporcionadas por la estación meteorológica de Sevilla**

***M. Martínez Ballesteros**, F. Martínez-Álvarez, A. Troncoso Lora, J.C. Riquelme Santos. "An Evolutionary Algorithm to Discover Quantitative Association Rules in Multidimensional Time Series". Soft Computing. Vol. 15, No. 10, pp. 2065-2084. 2011. ISSN: 1432-7643. DOI:10.1007/S00500-011-0705-4.*

Modelado de Contenido Ozono Total (TOC)

Localización de diferentes estaciones de observación TOC consideradas:



Training	Test
1. Madrid	2. Arenosillo
3. Lisboa	4. Montlouis
5. Murcia	

Medidas

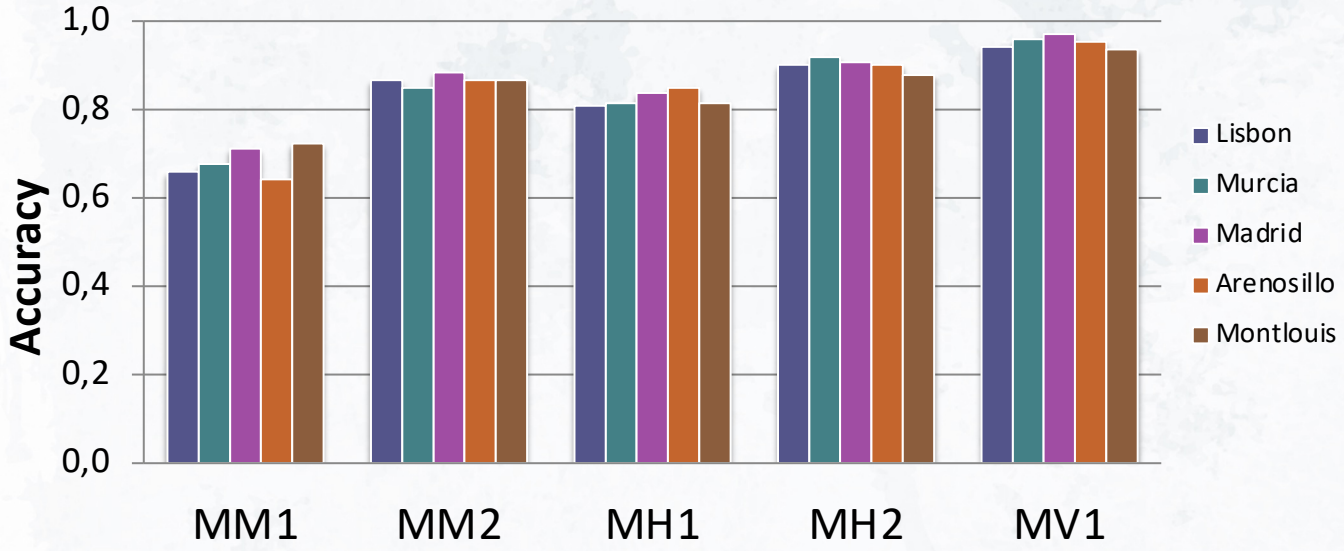
- Altura tropopausa
- Longitud de radiación de onda saliente
- Velocidad del viento vertical
- Temperatura 50 hpa

- Todas las variables han sido proporcionadas por la Agencia Estatal de Meteorología (AEMET)
- Mediciones de datos de TOMS del satélite NASA Nimbus7

M. Martínez-Ballesteros, S. Salcedo-Sanz, J. C. Riquelme, C. Casanova-Mateo, J. L. Camacho. "Evolutionary Association Rules for Total Ozone Content Modeling from Satellite Observations". Chemometrics and Intelligent Laboratory Systems. Vol. 109, No. 2, pp. 217- 227, 2011. ISSN: 0169-7439. DOI:10.1016/J.CHEMOLAB.2011.09.011.

Modelado de Contenido Ozono Total (TOC)

Accuracy de las RAC para la concentración de TOC entrenadas con datos de Madrid y testeadas con datos de Lisboa, Murcia, Madrid, Arenosillo y Montlouis



Modelado de Contenido Ozono Total (TOC)

ID	Association rules for TOC concentration at Madrid	TOC (DU)
M_{m1}	TP_G [14.1,12.5]	[285.6, 327.8]
M_{m2}	TP_G [12.4, 11.9] & OLR [262.9, 270.3] & t_{50} [215.4, 217.2]	[329, 347.5]
M_{h1}	TP_C [12.0, 11.3] & t_{50} [215.1, 216.4]	[334.1, 361]
M_{h2}	TP_C [10.8, 11.5] & t_{50} [214.4, 216.1]	[356.3, 392.5]
M_{v1}	TP_G [10.6, 11.2] & OLR [224.4, 245.4] & t_{50} [214.5, 216.8]	[368.1, 402.8]

06 →

Aplicaciones y ejemplos:

(b) Biomédicos

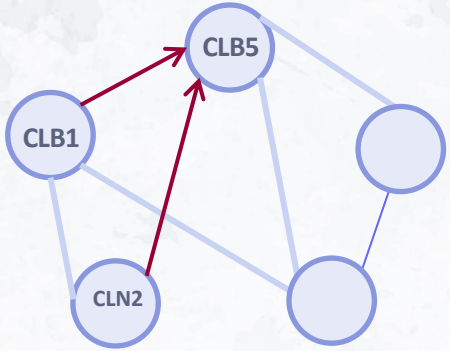
(IA)

Construcción de Redes de Genes

Asociaciones gen-gen a partir de RAC

$$\textbf{\textit{CLB1}} \in [-0.68, 0.05] \wedge \textbf{\textit{CLN2}} \in [-0.85, 0.1] \Longrightarrow \textbf{\textit{CLB5}} \in [-1.11, 0]$$

CLB1 $\Rightarrow$ **CLB5**
CLN2 $\Rightarrow$ **CLB5**



*Microarray of Spellman and Cho for the budding yeast (*Saccharomyces cerevisiae*) cell-cycle Time fold differential expressions

M. Martínez-Ballesteros, I. Nepomuceno-Chamorro, J. C. Riquelme. "Discovering gene association networks by multi-objective evolutionary quantitative association rules". Journal of Computer and System Sciences. Vol. 180, No.1 pp. 118-136, 2014

Construcción de Redes de Genes

Asociaciones gen-gen a partir de RAC

Dataset 1

$$\textit{CLB1} \in [-0.68, 0.05] \wedge \textit{CLN2} \in [-0.85, 0.1] \Rightarrow \textit{CLB5} \in [-1.11, 0]$$
$$\textit{CLN1} \in [0.1, 0.45] \Rightarrow \textit{CLB5} \in [0.05, 0.5]$$
$$\textit{CLB1} \in [-0.68, 0.05] \Rightarrow \textit{SWI5} \in [-0.5, -0.01]$$
$$\textit{CLN1} \in [0.2, 0.41] \Rightarrow \textit{CLB2} \in [0.1, 0.56] \wedge \textit{CLN2} \in [0.03, 0.4]$$

Dataset 2

$$\textit{CLB2} \in [0.05, 0.5] \wedge \textit{CLB6} \in [0.1, 0.62] \Rightarrow \textit{CLN1} \in [0.06, 0.33]$$
$$\textit{CLN1} \in [-0.43, 0.03] \Rightarrow \textit{CLN2} \in [-0.85, 0.1]$$
$$\textit{CLN1} \in [0, 0.43] \Rightarrow \textit{CLB5} \in [0.02, 0.45]$$
$$\textit{CLB1} \in [0.1, 0.45] \Rightarrow \textit{CLB5} \in [0, 0.51]$$
$$\textit{CLB5} \in [-1.11, 0] \Rightarrow \textit{CLN2} \in [-0.85, 0.1]$$

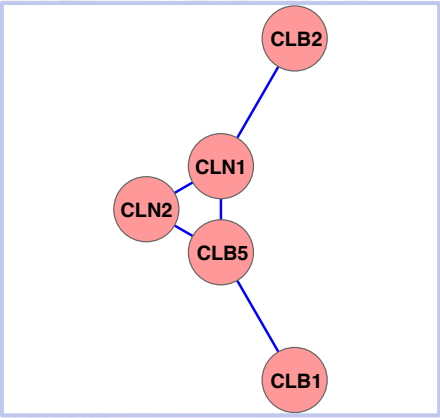
$\prod_1$

CLB1 – CLB5
CLN2 – CLB5
CLN1 – CLB5
CLB1 – SWI5
CLN1 – CLB2
CLN1 – CLN2

$\prod_2$

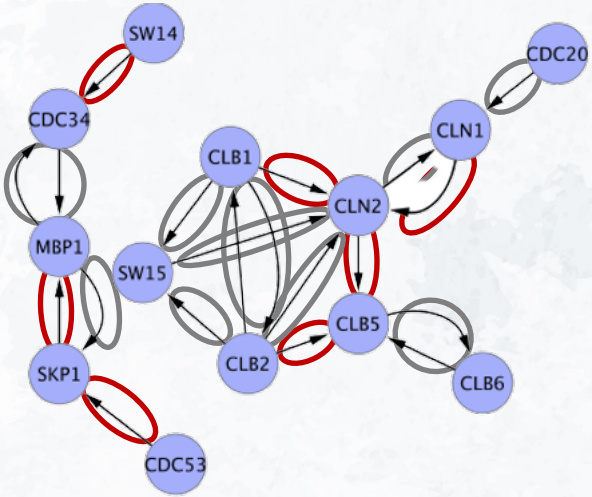
CLB2 – CLB6
CLB2 – CLN1
CLN1 – CLN2
CLN1 – CLB5
CLB1 – CLB5
CLB5 – CLN2

$$\prod = \prod_1 \cap \prod_2$$

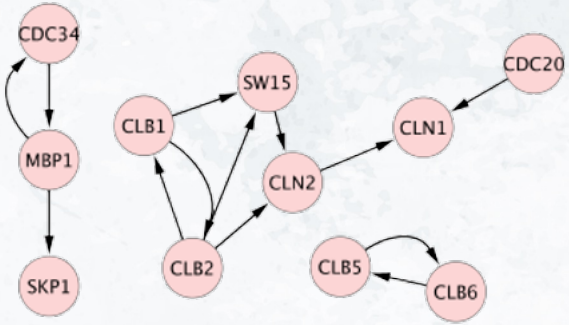


*Microarray of Spellman and Cho for the budding yeast (*Saccharomyces cerevisiae*) cell-cycle

Construcción de Redes de regulación de Genes



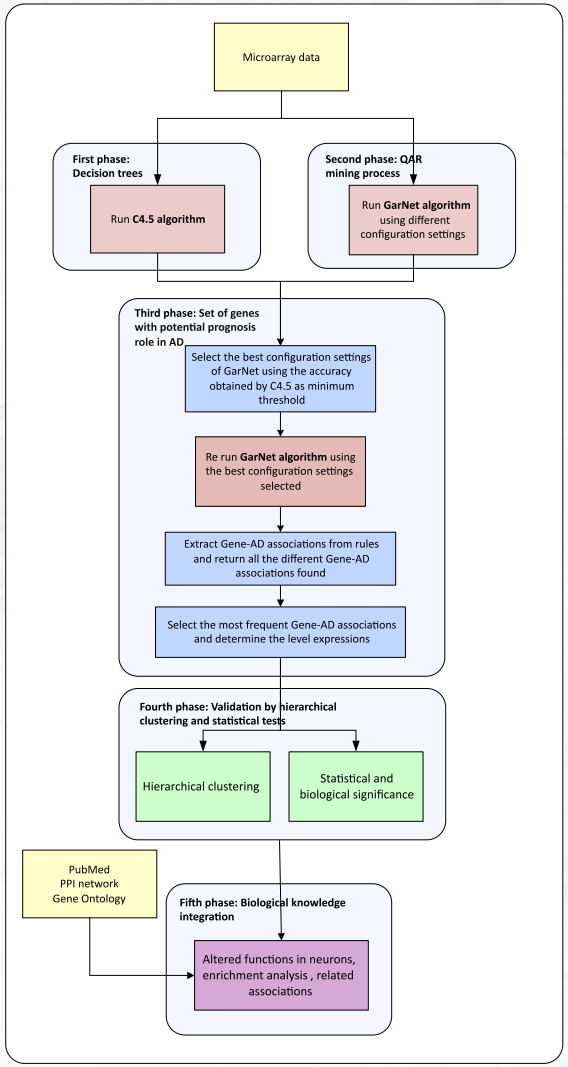
Grafo dirigido obtenido por MOQAR



Grafo dirigido obtenido por SOINOV

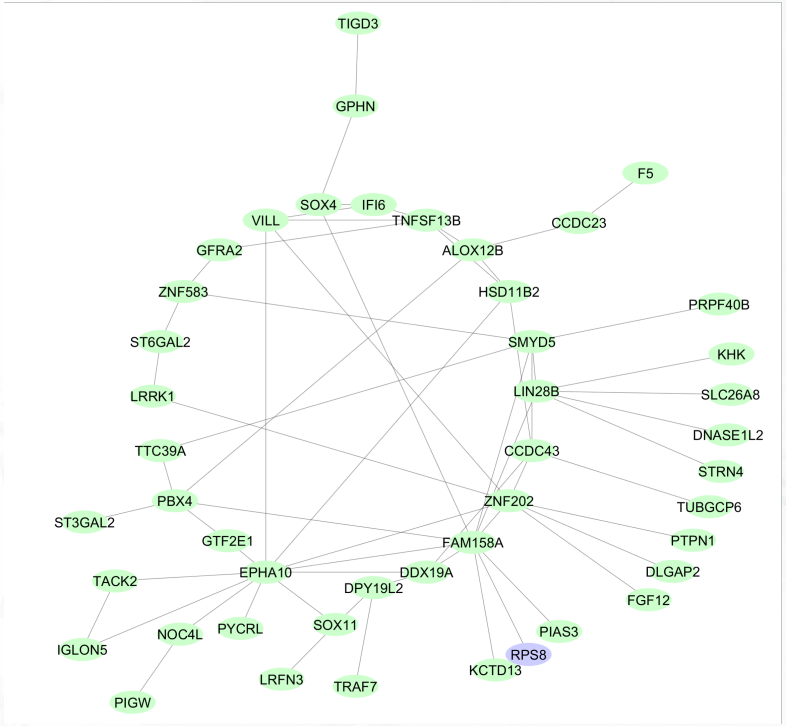
Datos Alzheimer

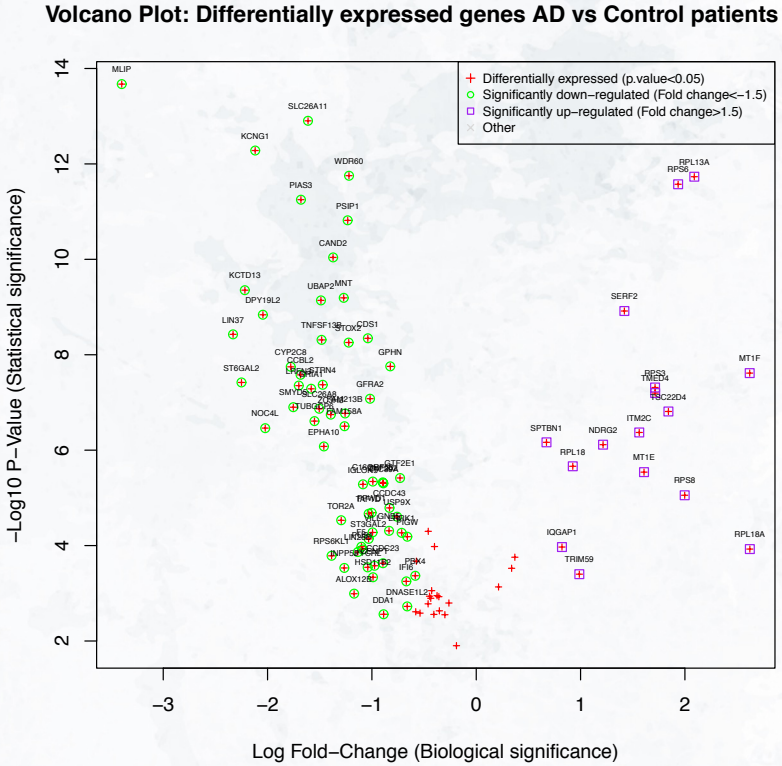
- Reglas de asociación cuantitativas entre pacientes de control y de Alzheimer
- Genes asociados a pacientes con Alzheimer
- Validación:
  - Test estadísticos
  - Fold-Change
  - Otros datasets



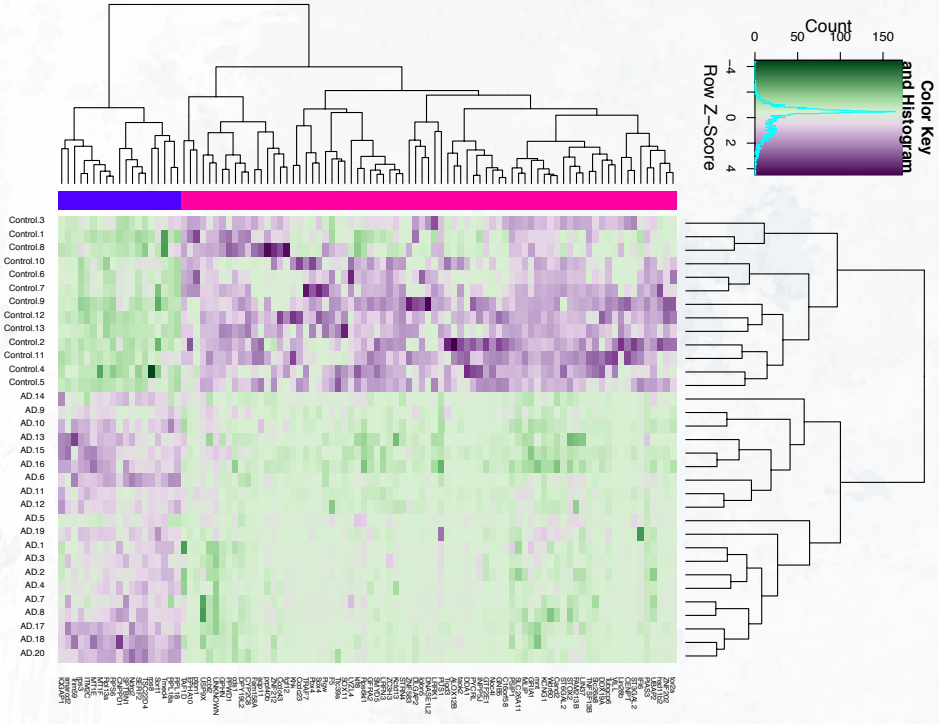
M. Martínez Ballesteros, J.M. García-Heredia, I. Nepomuceno-Chamorro, J.C. Riquelme Santos. "Machine learning techniques to discover genes with potential prognosis role in Alzheimer's disease using different biological sources", Information Fusion, Vol. 36, pp. 114 – 129, 2017. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2016.11.005

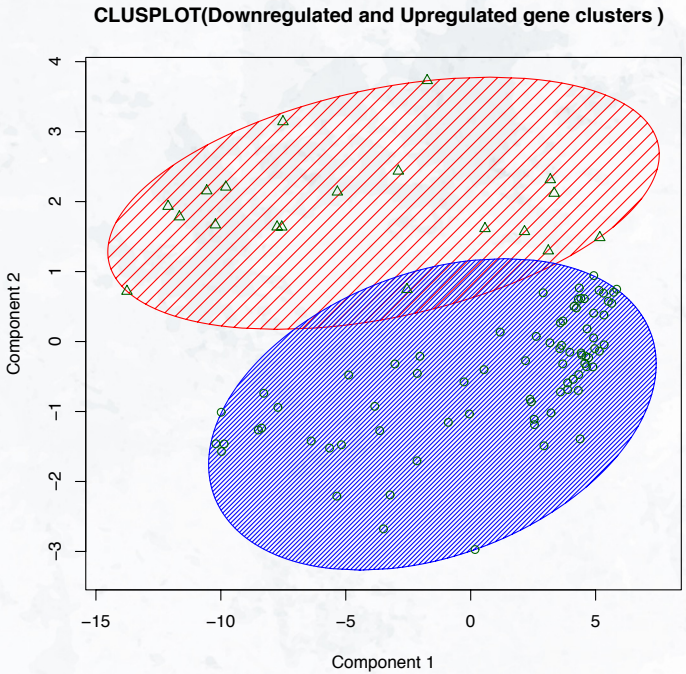
Datos Alzheimer





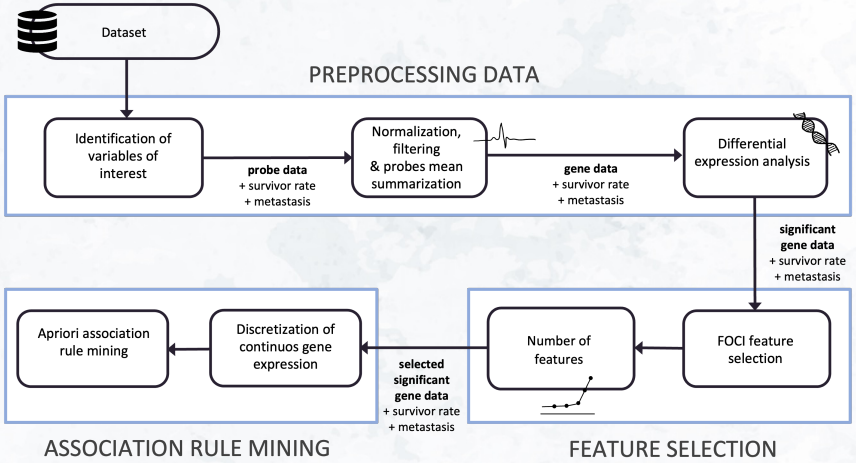
Datos Alzheimer





Datos Sarcoma

- Reglas de asociación cuantitativas entre pacientes con y sin metastasis
- Detección de patrones entre genes asociados a metastasis y baja supervivencia (menos de 5 años)

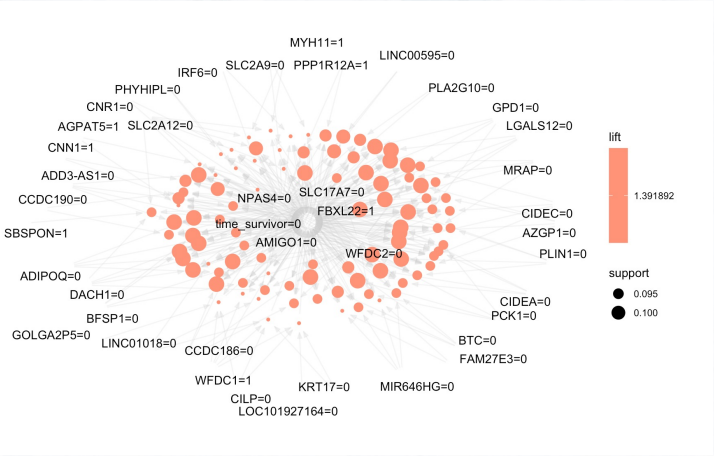


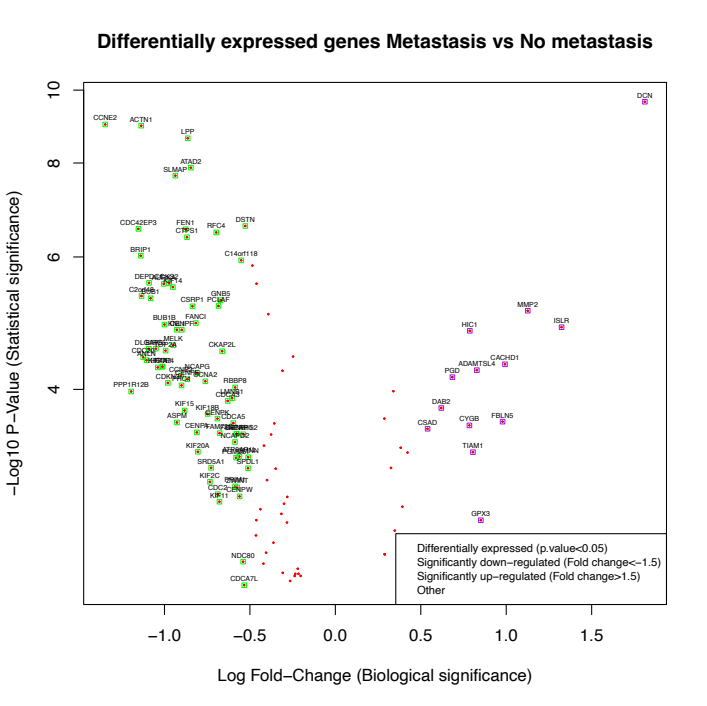
Antecedent	Consequent	Sup	Lift	Conf
CKAP2=UP ∧ GART=DOWN ∧ H2AZ1=DOWN	t_survivor < 5 years ∧ metastasis	0.055	1.64	1.00
ADAMTSL4=UP ∧ CDC42EP2=DOWN ∧ DSTN=DOWN	t_survivor < 5 years ∧ metastasis	0.058	1.64	1.00
ADAMTSL4=UP ∧ DSTN=DOWN ∧ GMNN=UP	t_survivor < 5 years ∧ metastasis	0.055	1.64	1.00

Linares-Barrera M.L., Martínez-Ballesteros M., García-Heredia J.M., Riquelme J.C. A Feature Selection and Association Rule Approach to Identify Genes Associated with Metastasis and Low Survival in Sarcoma. 18th International Conference on Hybrid Artificial Intelligence Systems (HAIS 2023).

Datos Sarcoma

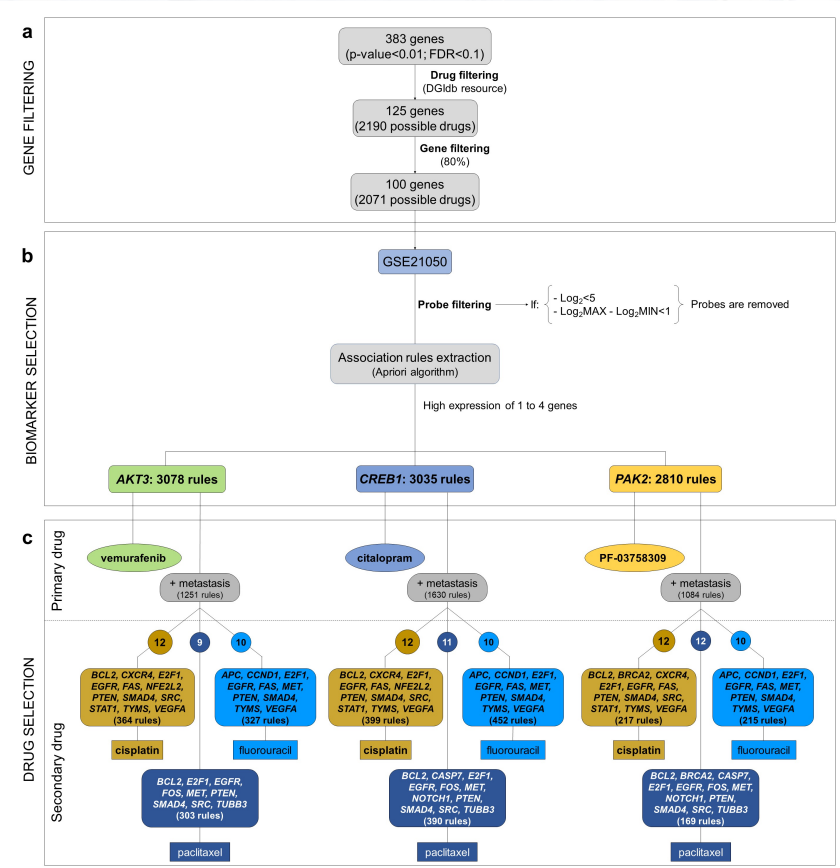
- Reglas de asociación cuantitativas entre pacientes con y sin metastasis
- Detección de patrones entre genes asociados a metastasis y baja supervivencia (menos de 5 años)





Datos Sarcoma

- Reglas para encontrar relaciones entre genes asociados al sarcoma mestastásico y que tuvieran fármacos asociados



Carnero A., García-Heredia J.M., Pérez M., Verdugo-Sivianes E., Martínez-Ballesteros M., Ortega-Campos S. A new treatment for sarcoma extracted from combination of miRNA deregulation and gene association rules. Signal Transduction and Targeted Therapy. 2023. JCR (2021): 35.130 (Q1)

06 →

Aplicaciones y ejemplos:

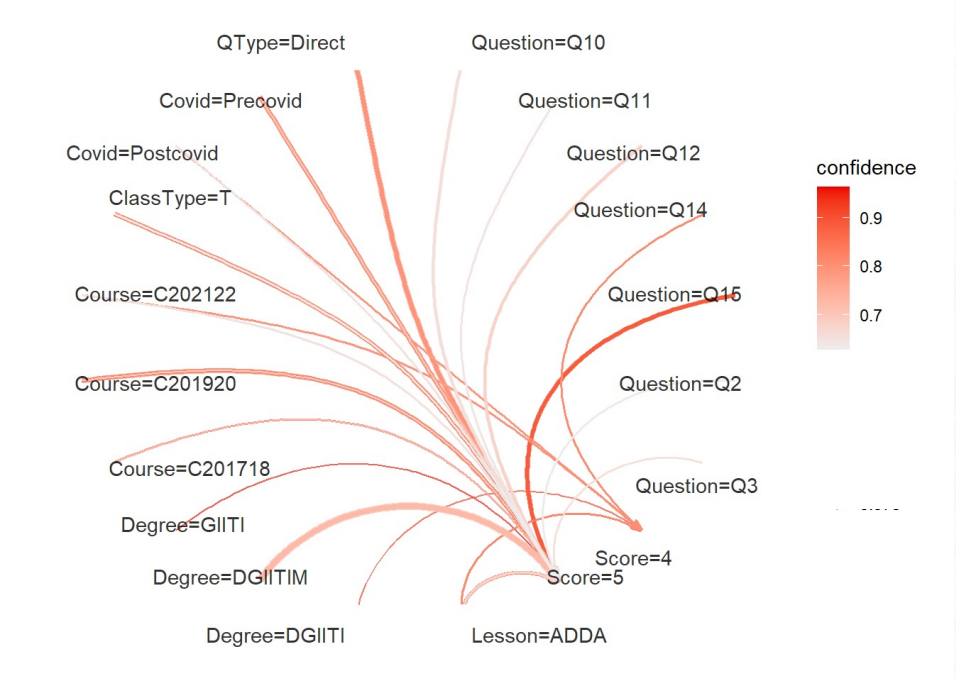
(c) Educación

(IA)

Encuestas de satisfacción de estudiantes

Items analizados:

- Preguntas:
  - Planificación y organización educativa
  - Apoyo estudiantil
  - Evaluación
  - Satisfacción general
- Asignatura
- Grado
- Curso
- Covid, Precovid, Postcovid
- Tipo de clase: teoría o prácticas
- Puntuación



Jiménez Navarro M., Vega Márquez B., Luna-Romera J.M., Carranza- García M., Martínez-Ballesteros M. Association Rule Analysis of Student Satisfaction Surveys for Teaching Quality Evaluation. 14th International Conference on EUropean Transnational Educational (ICEUTE 2023). 2023.

06 →

Aplicaciones y ejemplos:

(d) Explicabilidad

(IA)

Explicabilidad con RA: Predicción Series Temporales

- Posthoc XAI: Independiente del modelo
- Calidad de las reglas: confianza, cobertura, soporte
- Reglas sobre conjunto de entrenamiento para explicar cómo se ha entrenado el modelo DL
- Valores cubiertos en el consecuente
- Elementos que se repiten en el antecedente

Troncoso-García A.R., Martínez-Ballesteros M., Martínez-Álvarez F., Troncoso A. A new approach based on association rules to add explainability to time series forecasting models. Information Fusion. 2023. 10.1016/j.inffus.2023.01.021. JCR (2021): 17.564.

Troncoso-García A. R., Martínez-Ballesteros M., Martínez-Álvarez F., Troncoso A. Evolutionary Computation to Explain Deep Learning Models for Time Series forecasting. . The 38th ACM/SIGAPP Symposium on Applied Computing, pp. 433-436. 2023.

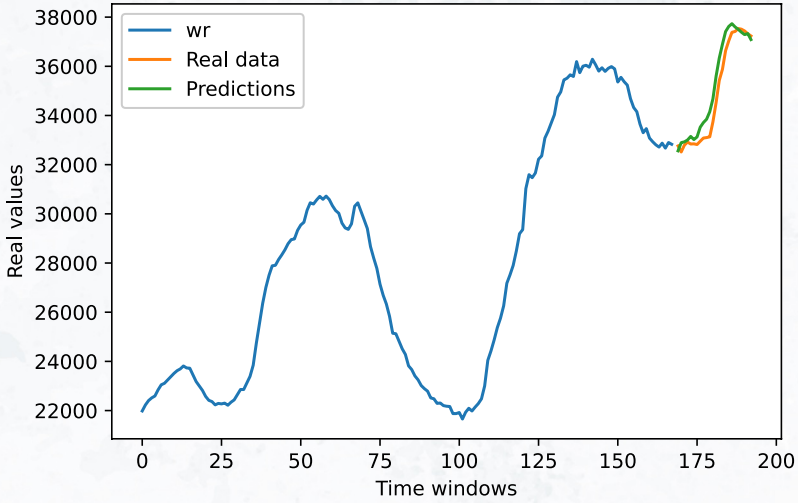
Explicabilidad con RA: Predicción Series Temporales

Datos

- Consumo eléctrico en España
- Frecuencia: 10 min
- Ventana temporal:
  1 día y 4 horas (168 elementos)
- Horizonte de predicción:
  4 horas (24 elementos)

Deep Learning

- WkNN
- BigPSF
- LSTM

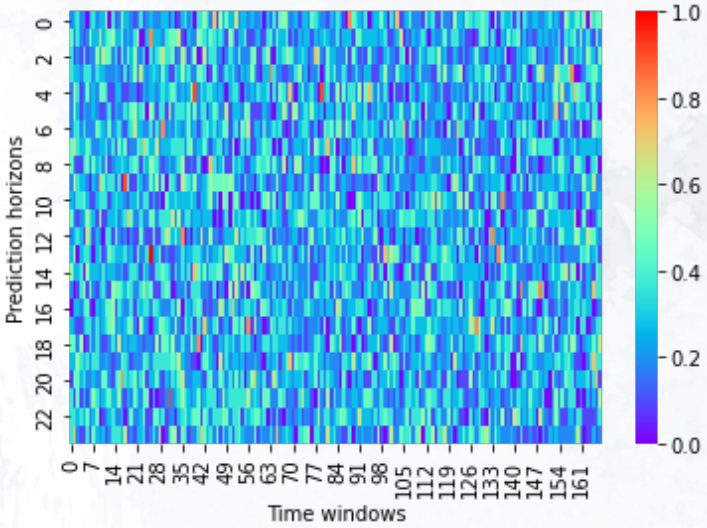


Explicabilidad con RA: Predicción Series Temporales

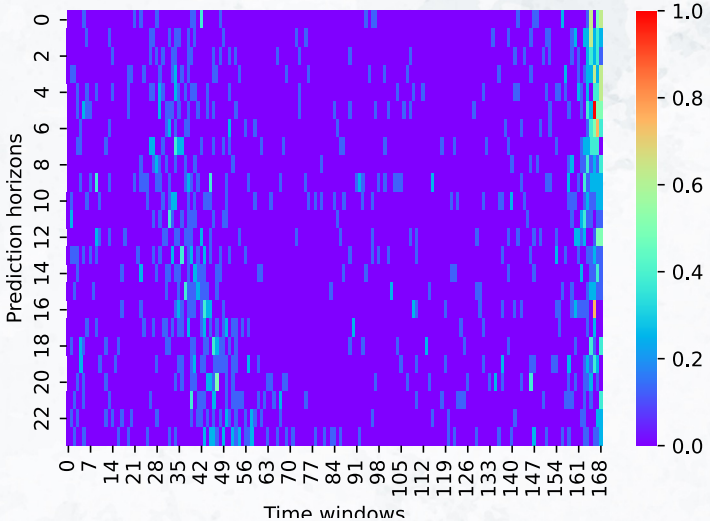
$$\text{IF } X_{t-1} \in [19144,\ 28171] \Longrightarrow \hat{X}_{t+1} \in [19017.01,\ 28913.1]$$
$$\text{IF } X_{t-2} \in [24681,\ 31386] \Longrightarrow \hat{X}_{t+2} \in [23969.40,\ 32495{,}10]$$
$$\text{IF } X_{t-1} \in [23712,\ 31206\ ] \Longrightarrow \hat{X}_{t+3} \in [23061.80,\ 31713.90]$$
$$\text{IF } X_{t} \in [24278,\ 29659] \Longrightarrow \hat{X}_{t+4} \in [22952.34,\ 31432.35]$$
$$\text{IF } X_{t-7} \in [21853,\ 34799] \text{ AND } X_{t-3} \in [24217,\ 31889] \Longrightarrow \hat{X}_{t+5} \in [23578.02,\ 33340.73]$$
$$\text{IF } X_{t-137} \in [24123,\ 34045] \text{ AND } X_{t} \in [24228,\ 33258] \Longrightarrow \hat{X}_{t+6} \in [23522.44,\ 34127.99]$$
$$\text{IF } X_{t-1} \in [19479,\ 29894\ ] \Longrightarrow \hat{X}_{t+7} \in [20182.95,\ 319619.64]$$

* Se obtuvieron reglas para los 24 horizontes de predicción

Explicabilidad con RA: Predicción Series Temporales

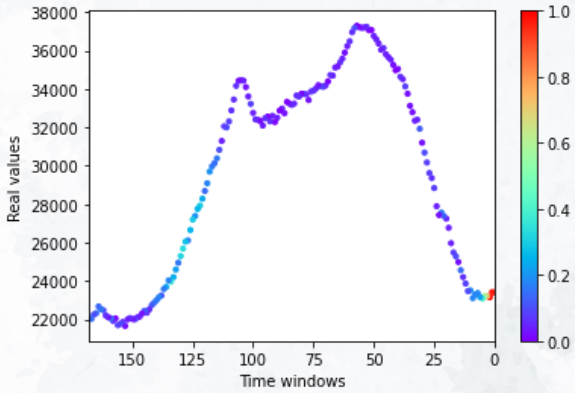


LIME

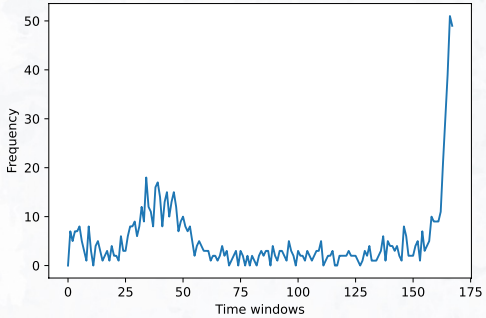


Reglas de asociación en LSTM

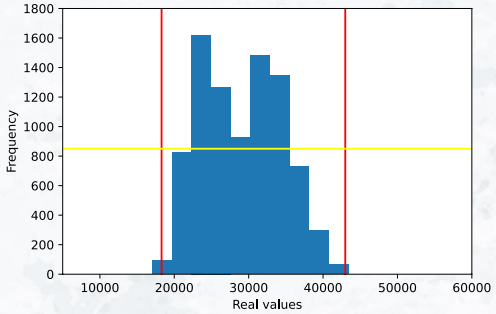
Explicabilidad con RA: Predicción Series Temporales



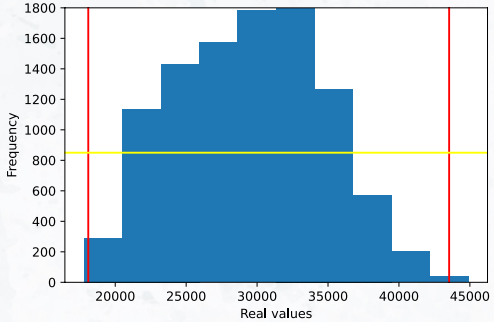
Una serie temporal en concreto



Importancia de cada ventana temporal para todas las predicciones



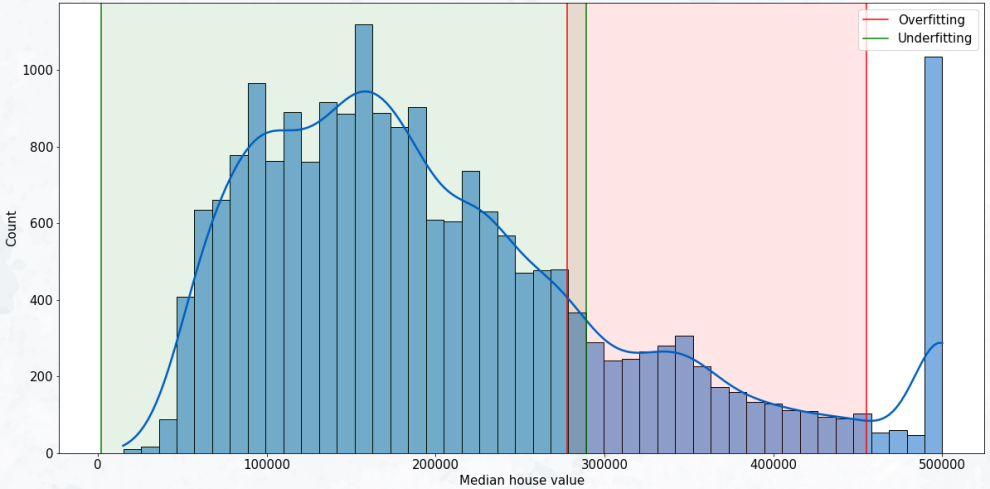
BigPSF



LSTM

Otro caso: Overfitting vs Underfitting

- Independiente del modelo
- Metodología para analizar el comportamiento de modelos de caja negra.
- Taxonomía de reglas según métricas y error
- Identificar posibles escenarios



Jiménez-Navarro M.J., Martínez-Ballesteros M., Martínez-Álvarez F., Asencio-Cortés G. Explaining learned patterns in deep learning by association rules mining. 18th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2023).

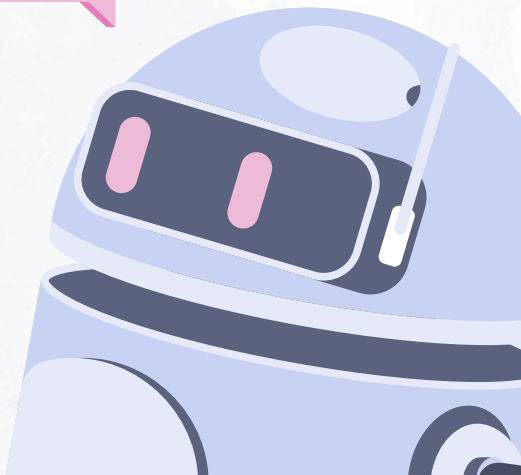
Problemas abiertos

- Métricas para evaluar la calidad de un grupo de reglas
- Interpretabilidad en reglas de asociación
- Reglas de asociación temporales
- Reglas de asociación raras
- Reglas de asociación en streaming
- Analizar problemas de otros dominios

Reglas de asociación:

Metodología y Ejemplos

MUCHAS GRACIAS por vuestra atención



María Martínez Ballesteros
mariamartinez@us.es

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla