



Pontificia Universidad
JAVERIANA
Bogotá

Técnicas Predictivas y Descriptivas para la Generación de Nuevo Conocimiento

ING. HERIBERTO FELIZZOLA JIMENEZ

Educación **Continua**
Generamos experiencias educativas

Contenido

1. Introducción al Aprendizaje de Máquina
2. Aprendizaje Supervisado (Técnicas Predictivas)
 1. Modelo de Regresión Lineal
 2. Regresión Logística
 3. Arboles de Clasificación
 4. Métodos de Ensamble
 5. K Vecinos mas Cercanos
3. Aprendizaje No Supervisado (Técnicas Descriptivas y Exploratorias)
 1. Métodos de Clustering y Anomalías
 2. Reducción de dimensiones
 3. Reglas de asociación
4. Introducción a la visualización de datos
5. Introducción a Tecnologías Big Data



Pontificia Universidad
JAVERIANA
Bogotá

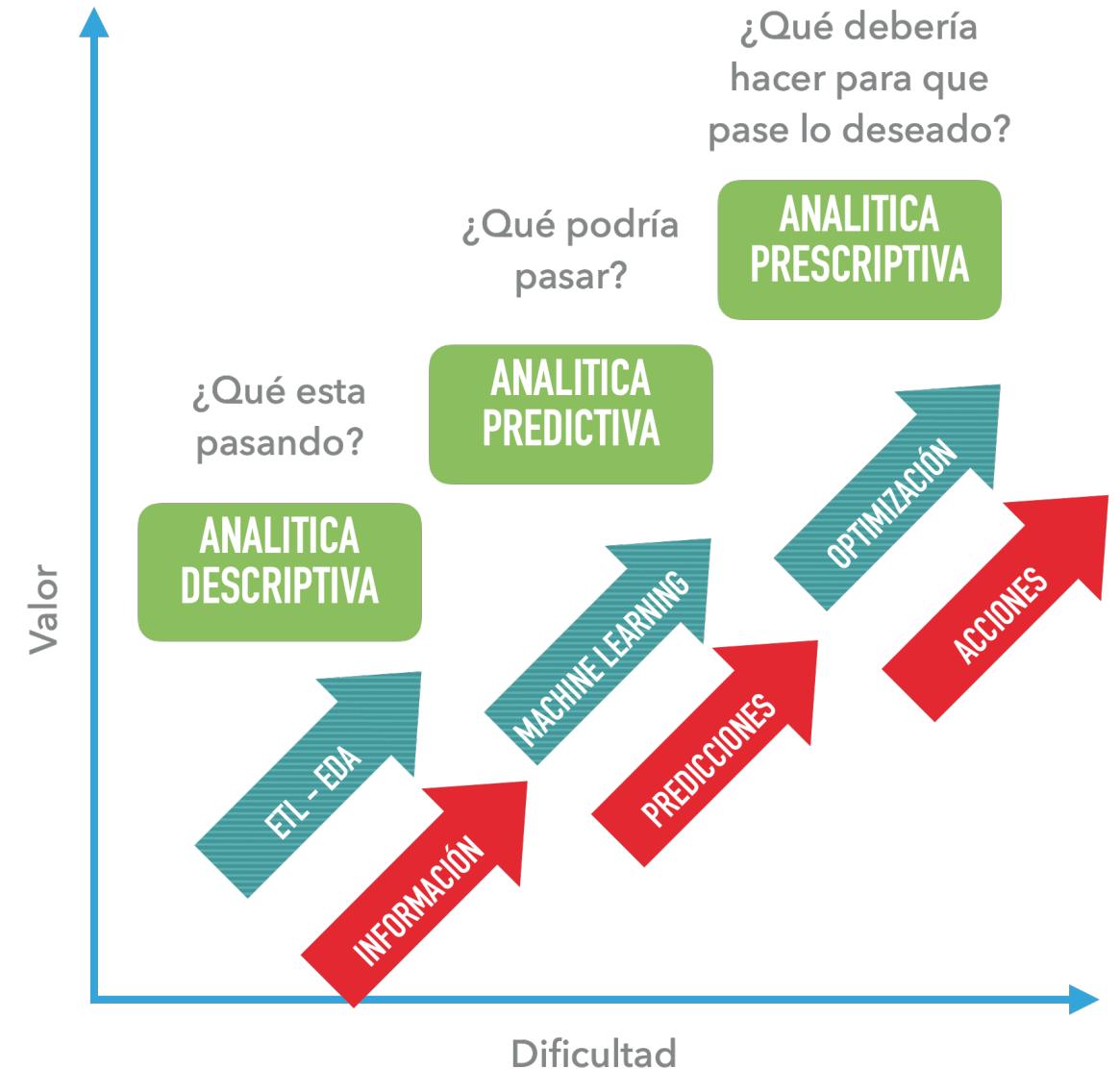
1. Introducción al Aprendizaje de Máquina

ING. HERIBERTO FELIZZOLA JIMENEZ

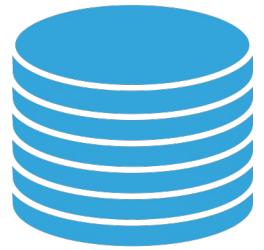
Educación **Continua**
Generamos experiencias educativas

Analítica Predictiva

- Es el arte de construir modelos que permiten realizar predicciones a partir de los patrones encontrados en los datos.
- Algunas aplicaciones de las predicciones son:
 - Precios de productos, servicios y mercado de valores
 - Recomendación de productos
 - Sistemas de recomendación
 - Riesgo en proyectos
 - Diagnóstico de enfermedades
 - Clasificación de imágenes, texto y video
 - Condiciones meteorológicas

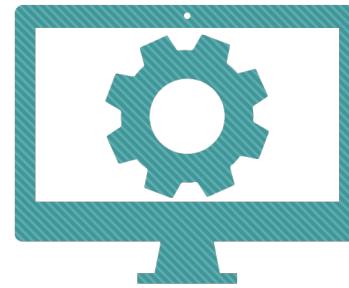
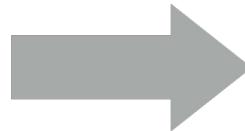


Aprendizaje de Máquina



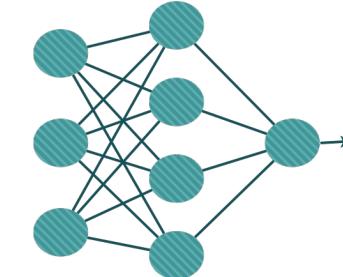
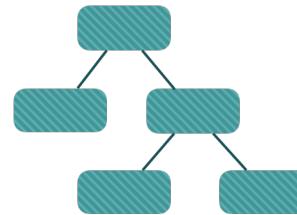
Datos de Entrenamiento

X ₁	X ₂	...	X _p	Y
X ₁₁	X ₁₂	...	X _{1p}	Y ₁
X ₂₁	X ₂₂	...	X _{2p}	Y ₃
...
X _{n1}	X _{n2}	...	X _{np}	Y _n

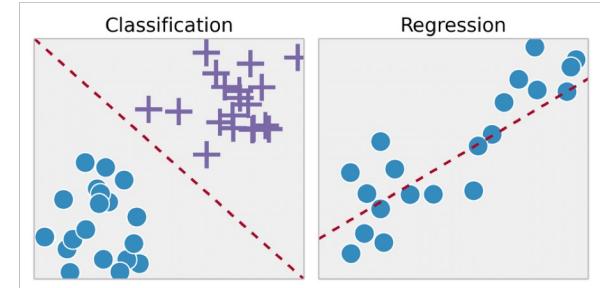


Algoritmos y Modelos de Aprendizaje de Maquina

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$



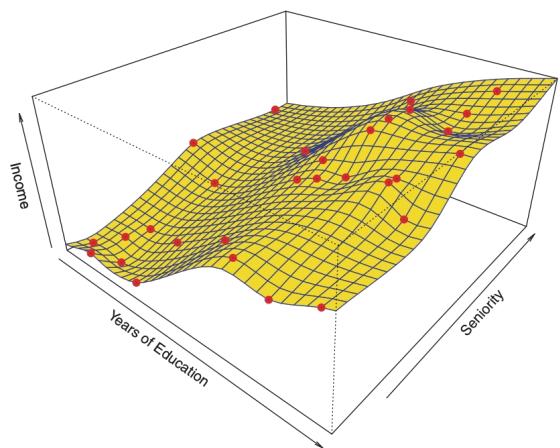
Predicciones



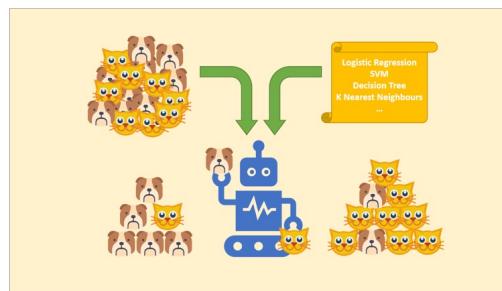
Tareas en Aprendizaje Automático

APRENDIZAJE SUPERVISADO

REGRESIÓN

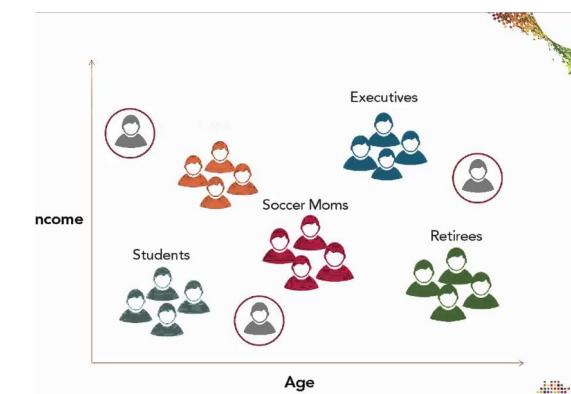


CLASIFICACIÓN

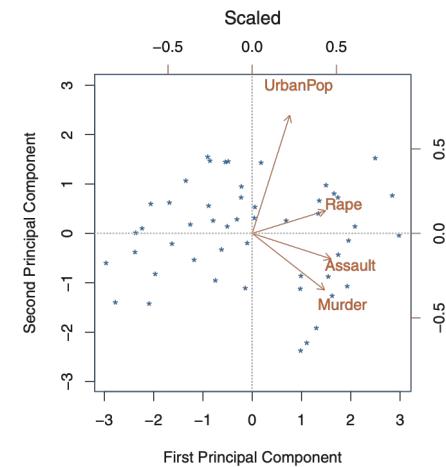


APRENDIZAJE NO SUPERVISADO

CLUSTERING



REDUCCIÓN DE DIMENSIONES



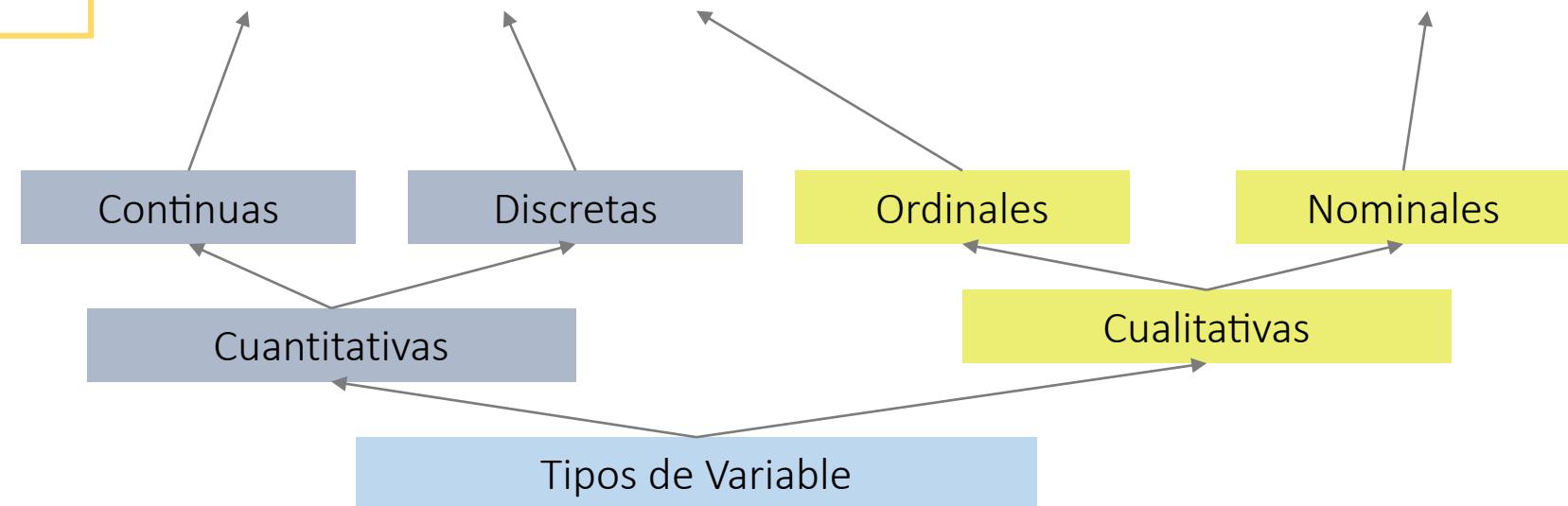
Tipos de Datos

Las columnas representan las características de cada uno de los préstamos, también denominadas variables.

Cada fila representa un único préstamo. Esto puede ser un caso/instancia/observación.

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
:	:	:	:	:	:	:	:
50	3000	7.96	36	A	CA	34000	rent

Cada celda es un valor único de la variable para una observación particular.

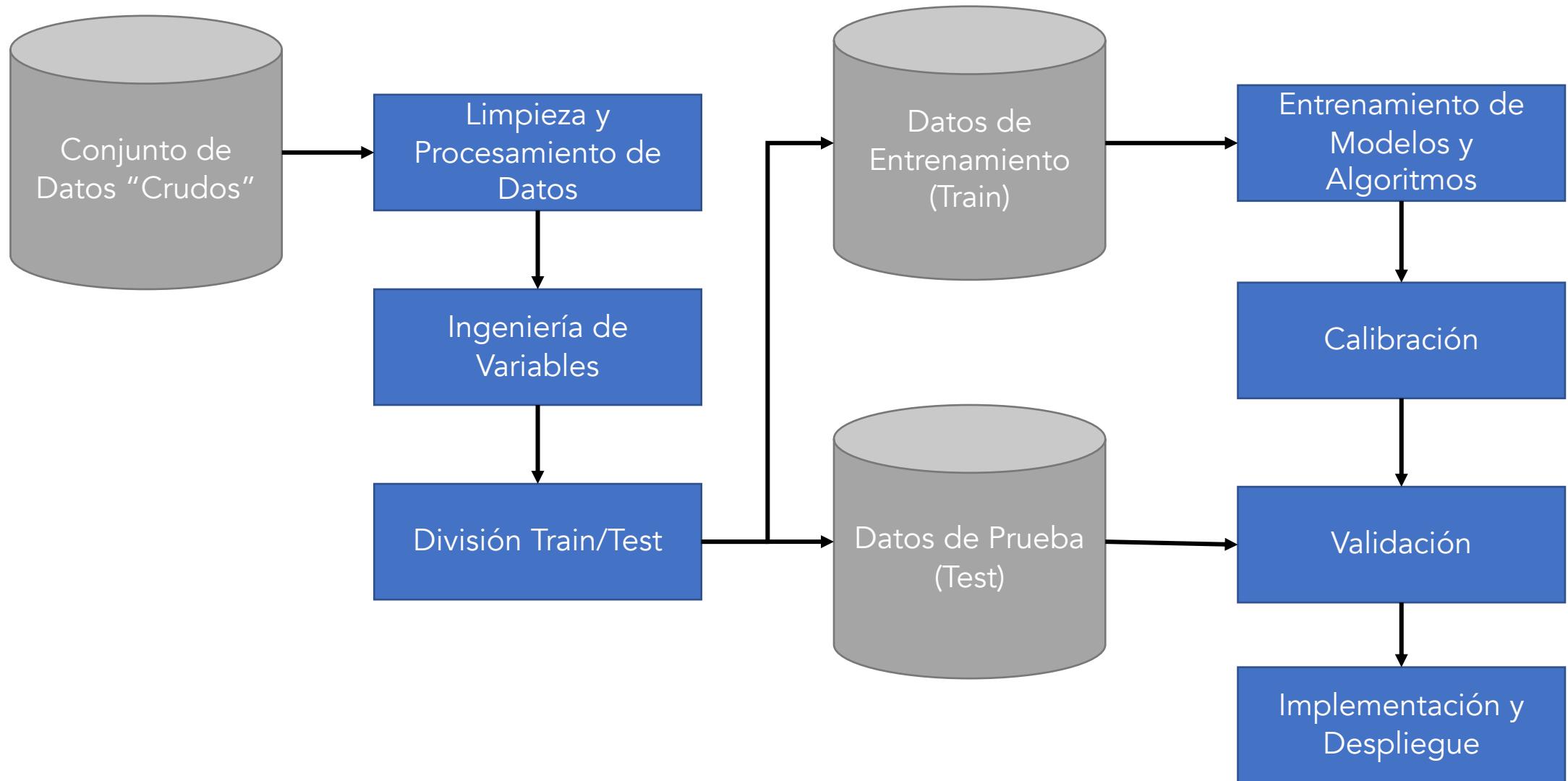


Instancia- Atributo- Salida

- La entrada a un esquema de aprendizaje automático es un conjunto de instancias.
- Estas instancias son las cosas que deben clasificarse, asociarse o agruparse.
- El conjunto de características de cada instancia es denominado atributos o variables de entrada (X).
- El resultado que se quiere predecir es denominado salida (Y).

Instancias	Y: Salida					X: Atributos - Variables de Entrada				
	Outlook	Temperature	Humidity	Windy	Play Time	Outlook	Temperature	Humidity	Windy	Play Time
	Sunny	85	85	false	5	Rainy	70	96	false	40
	Sunny	80	90	true	0	Rainy	68	80	false	65
	Overcast	83	86	false	55	Rainy	65	70	true	45
	Rainy	70	96	false	40	Overcast	64	65	true	60
	Rainy	68	80	false	65	Sunny	72	95	false	0
	Rainy	65	70	true	45	Sunny	69	70	false	70
	Overcast	64	65	true	60	Rainy	75	80	false	45
	Sunny	72	95	false	0	Sunny	75	70	true	50
	Sunny	69	70	false	70	Overcast	72	90	true	55
	Rainy	75	80	false	45	Overcast	81	75	false	75
	Sunny	75	70	true	50	Rainy	71	91	true	10

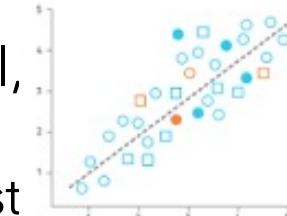
Proceso de Entrenamiento



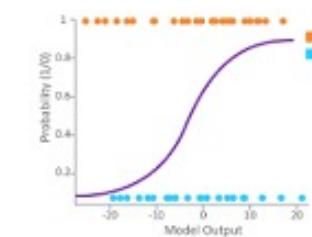
Modelos y Algoritmos

- Aprendizaje supervisado:
 - Modelos lineales: Mínimos Cuadrados, Análisis discriminante, Regresión Logística.
 - Modelos de penalización: Ridge, Lasso, ElasticNet
 - Modelos no lineales: Splines, Regresión Local, Modelos Generales Aditivos (GAM), Kernels.
 - Arboles: CART, C5.0, Bagging, Random Rorest Boosting
 - Maquinas de soporte vectorial SVM.
 - Naive-Bayes
 - Redes neuronales.
- Aprendizaje no supervisado:
 - Algoritmos de clustering: K-mean, Modelos Jerarquicos
 - Reducción de dimensiones: Componentes principales - PCA

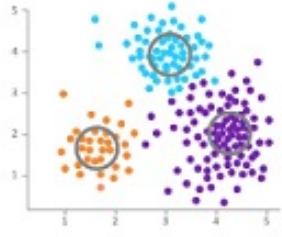
Linear Regression



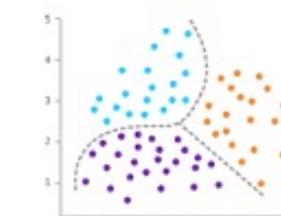
Logistic Regression



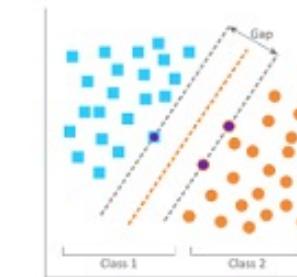
K-Means



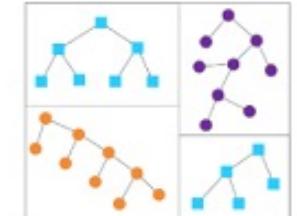
Naive Bayes



Support Vector Machines



Random Forest





Pontificia Universidad
JAVERIANA
Bogotá

2. Aprendizaje Supervisado Técnicas Predictivas

ING. HERIBERTO FELIZZOLA JIMENEZ

Educación **Continua**

Generamos experiencias educativas



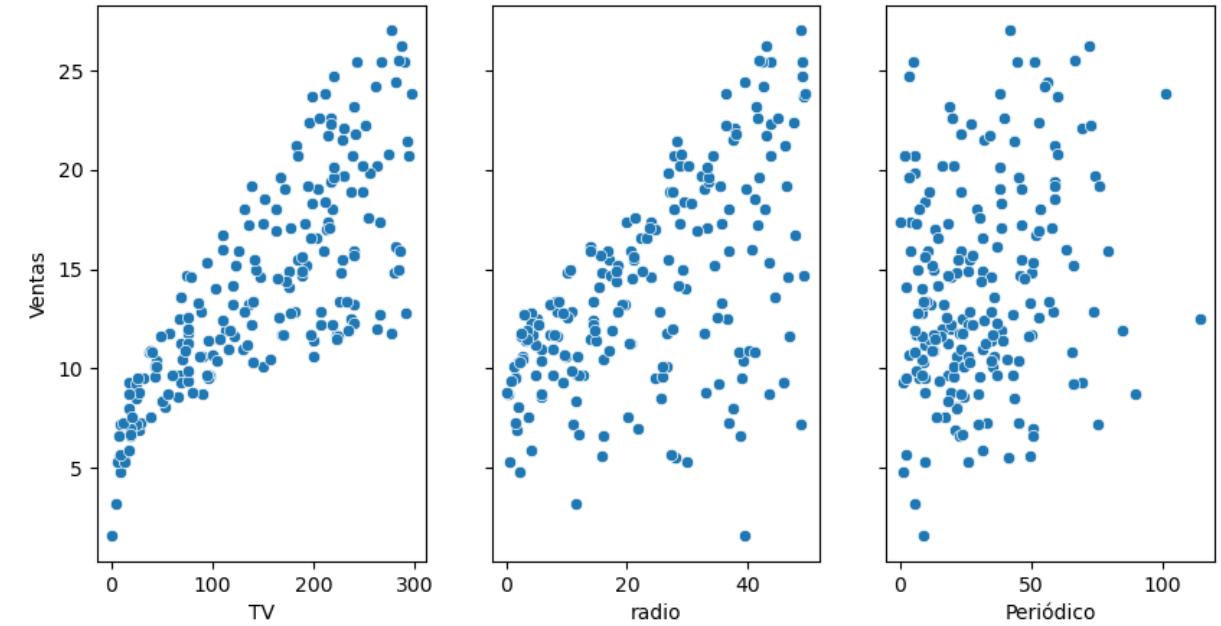
Pontificia Universidad
JAVERIANA
Bogotá

2.1 Regresión Lineal

Educación **Continua**
Generamos experiencias educativas

Ejemplo

- Supongamos que somos consultores estadísticos contratados por un cliente para investigar la asociación entre el gasto en publicidad y las ventas de un producto en particular.
- Nuestro cliente no puede incrementar directamente las ventas del producto pero puede controlar el gasto publicitario.
- Por lo tanto, si determinamos que existe una asociación entre publicidad y ventas, entonces podemos instruir a nuestro cliente para que ajuste los presupuestos publicitarios.



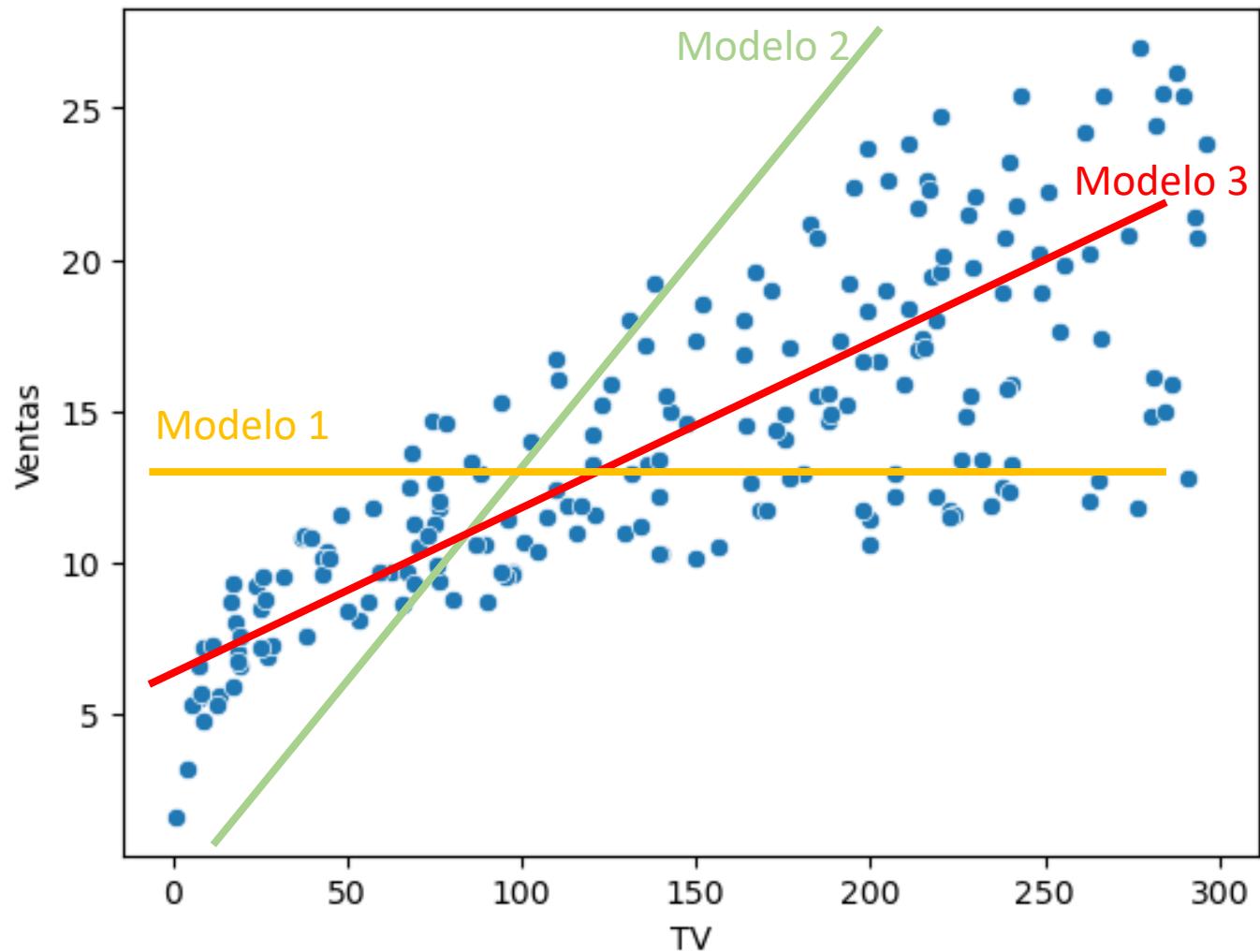
En otras palabras, nuestro objetivo es desarrollar un modelo preciso que se pueda utilizar para predecir las ventas sobre la base de los tres presupuestos de medios (TV, radio y periódico).

Con el modelo de regresión para predecir las ventas en función del presupuesto de publicidad en los diferentes medios, buscamos resolver las siguientes preguntas:

- ¿Existe una relación entre el presupuesto publicitario y las ventas?
- ¿Qué medios contribuyen a las ventas?
- ¿Qué contribución hace cada medio para la comportamiento de las ventas?
- ¿Con qué precisión podemos predecir las ventas futuras?
- ¿Existe correlación entre los medios publicitarios?

Regresión Lineal

- ¿Es posible trazar una línea que capture de forma perfecta todos los puntos?
- De forma intuitiva podemos decir que construir una modelo de predicción es encontrar la línea que nos permita capturar de la mejor manera la relación entre dos variables.
- En la gráfica podemos ver que el modelo 3, es el que mejor representa la relación entre las ventas y el presupuesto para TV.
- Observe que este modelo no captura de forma perfecta todos los puntos, por tanto la línea de tendencia central es el valor esperado de las ventas dado un presupuesto de TV.
- Debemos ser conscientes de que cualquier predicción de las ventas a partir del presupuesto para TV tendrá un error.
- No existe una predicción perfecta.



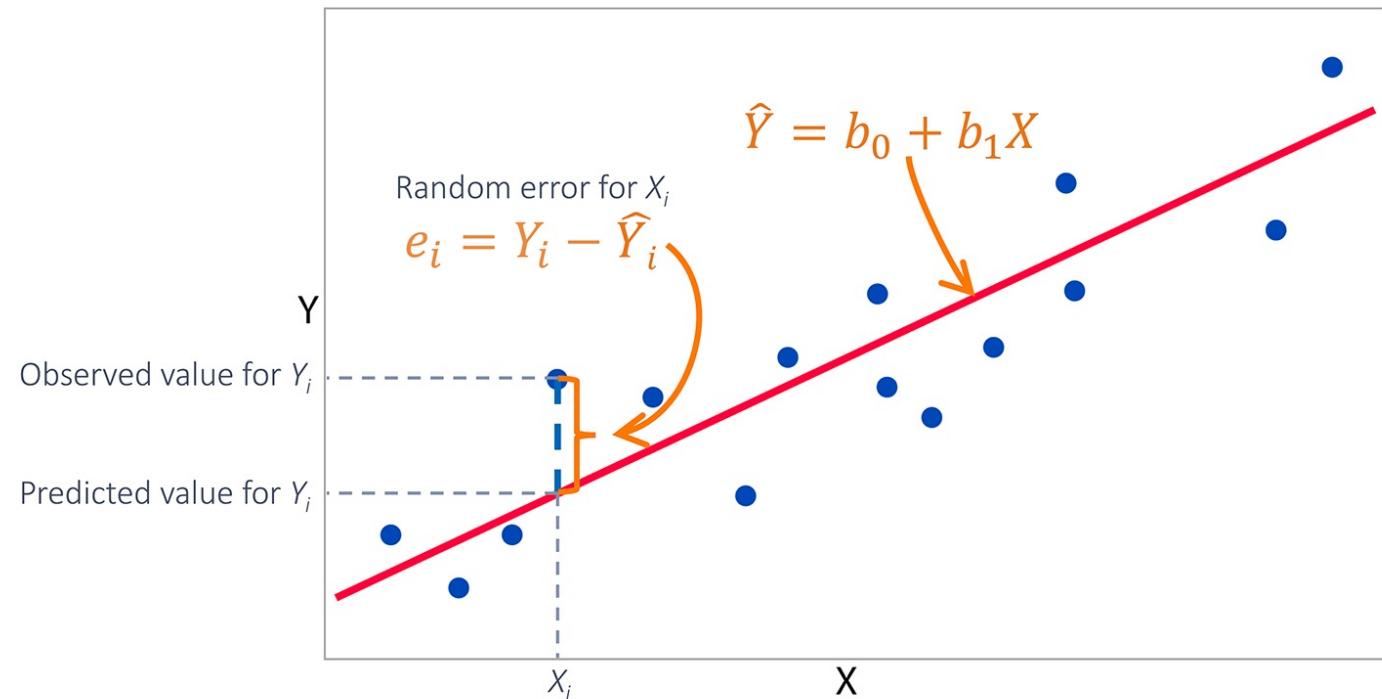
Conceptos Básicos

- **Objetivo o respuesta:** es la variable que deseamos predecir, en este caso las ventas. Generalmente nos referimos a la respuesta como la variable Y .
- **Entrada o predictor:** son las variables independientes o controlables que utilizamos para predecir la respuesta, nos referimos a estas como X 's, este caso los presupuestos de:
 - TV (X_1)
 - Radio (X_2)
 - Periódico (X_3)
- **Modelo de regresión:** es una función que relaciona la variable de respuesta con los predictores, generalmente se escribe como: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
- Donde, β_0 es el intercepto
- β_1, β_2 y β_3 son las contribuciones individuales (cuando las demás permanecen constantes) de cada variable de entrada a la predicción de la respuesta Y .
- **Error (ϵ):** es la diferencia entre el valor real de Y y su predicción \hat{Y} , por tanto
$$\rightarrow \epsilon = Y - \hat{Y}$$

Mínimos Cuadrados

- Construir un modelo de regresión implica encontrar las contribuciones de las variables TV, Radio y Periódico (β_1, β_2 y β_3) que permiten predecir las ventas con el mínimo error.
- Para estimar las constantes de la regresión se utiliza el método de mínimos cuadrados (Least Square).
- El método busca elegir los valores para $\beta_0, \beta_1, \beta_2$ y β_3 que minimicen la suma de cuadrados de los errores ϵ .
- En otras palabras:

$$\text{Min RSS} = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$



OLS Regression Results

```
=====
Dep. Variable:           ventas    R-squared:                 0.897
Model:                  OLS        Adj. R-squared:            0.896
Method:                 Least Squares   F-statistic:              570.3
Date:                   Sat, 23 Oct 2021   Prob (F-statistic):       1.58e-96
Time:                   19:37:49      Log-Likelihood:            -386.18
No. Observations:        200        AIC:                      780.4
Df Residuals:             196        BIC:                      793.6
Df Model:                      3
Covariance Type:          nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
TV	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
periodico	-0.0010	0.006	-0.177	0.860	-0.013	0.011
constante	2.9389	0.312	9.422	0.000	2.324	3.554

<=====

Omnibus:	60.414	Durbin-Watson:	2.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241
Skew:	-1.327	Prob(JB):	1.44e-33
Kurtosis:	6.332	Cond. No.	454.

<=====

Significancia del Modelo

¿Existe una relación entre el presupuesto publicitario y las ventas?

- Esta pregunta se responde evaluando la significancia de la regresión.
- Esta prueba permite determinar si existe una relación lineal entre la variable de respuesta y las variables de entrada.

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$Ha: \beta_j \neq 0, \text{ para al menos una } j$$

- Para esto se revisa el valor P de la regresión, en la tabla el **Prob (F-statistic)**
- Cuando este valor es muy bajo, por lo general menor a 0.05, entonces podemos inferir que al menos una variable de respuesta genera un efecto significativo sobre la variable de respuesta.

OLS Regression Results							
Dep. Variable:	ventas	R-squared:	0.897				
Model:	OLS	Adj. R-squared:	0.896				
Method:	Least Squares	F-statistic:	570.3				
Date:	Sat, 23 Oct 2021	Prob (F-statistic):	1.58e-96				
Time:	19:37:49	Log-Likelihood:	-386.18				
No. Observations:	200	AIC:	780.4				
Df Residuals:	196	BIC:	793.6				
Df Model:	3						
Covariance Type:	nonrobust						
	coef	std err	t	P> t	[0.025	0.975]	
IV	0.0458	0.001	32.809	0.000	0.043	0.049	
radio	0.1885	0.009	21.893	0.000	0.172	0.206	
periodico	-0.0010	0.006	-0.177	0.860	-0.013	0.011	
constante	2.9389	0.312	9.422	0.000	2.324	3.554	
Omnibus:		60.414	Durbin-Watson:				2.084
Prob(Omnibus):		0.000	Jarque-Bera (JB):				151.241
Skew:		-1.327	Prob(JB):				1.44e-33
Kurtosis:		6.332	Cond. No.				454.

En este caso, el valor p correspondiente al estadístico F en la tabla es muy bajo, lo que indica una clara evidencia de una relación entre publicidad y ventas.

Inferencia de los Parámetros

¿Qué medios contribuyen a las ventas?

- Para responder a esta pregunta, podemos examinar los valores p ($P > |t|$) asociados con el estadístico t de cada predictor.
- Para cada variable del modelo se plantea la siguiente hipótesis:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

- Los valores p para televisión y radio son bajos ≈ 0 , pero el valor p para periódicos (0.86) no lo es.

OLS Regression Results						
Dep. Variable:	ventas	R-squared:	0.897			
Model:	OLS	Adj. R-squared:	0.896			
Method:	Least Squares	F-statistic:	570.3			
Date:	Sat, 23 Oct 2021	Prob (F-statistic):	1.58e-96			
Time:	19:37:49	Log-Likelihood:	-386.18			
No. Observations:	200	AIC:	780.4			
Df Residuals:	196	BIC:	793.6			
Df Model:	3					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
TV	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
periodico	-0.0010	0.006	-0.177	0.860	-0.013	0.011
constante	2.9389	0.312	9.422	0.000	2.324	3.554
Omnibus:		60.414	Durbin-Watson:			2.084
Prob(Omnibus):		0.000	Jarque-Bera (JB):			151.241
Skew:		-1.327	Prob(JB):			1.44e-33
Kurtosis:		6.332	Cond. No.			454.

Esto sugiere que solo la televisión y la radio están relacionadas con las ventas. En cambio, el presupuesto para periódico no afecta significativamente en las ventas.

Interpretación de los Parámetros

¿Qué contribución hace cada medio para la comportamiento de las ventas?

- Las contribuciones de cada medio están expresadas en sus coeficientes (*coef*):
 - TV:** por cada peso que se gasta en TV las ventas aumentan 0.0458.
 - Radio:** por cada peso que se gasta en Radio las ventas aumentan 0.1885.
 - Periódico:** por cada peso que se gasta en periódicos las ventas disminuyen 0.0010
- Para cada contribución también se presentan intervalos de confianza del 95%.

OLS Regression Results						
Dep. Variable:	ventas	R-squared:	0.897			
Model:	OLS	Adj. R-squared:	0.896			
Method:	Least Squares	F-statistic:	570.3			
Date:	Sat, 23 Oct 2021	Prob (F-statistic):	1.58e-96			
Time:	19:37:49	Log-Likelihood:	-386.18			
No. Observations:	200	AIC:	780.4			
Df Residuals:	196	BIC:	793.6			
Df Model:	3					
Covariance Type:	nonrobust					
Intervalos de Confianza						
	coef	std err	t	P> t	[0.025	0.975]
TV	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
periodico	-0.0010	0.006	-0.177	0.860	-0.013	0.011
constante	2.9389	0.312	9.422	0.000	2.324	3.554
Omnibus:		60.414	Durbin-Watson:			2.084
Prob(Omnibus):		0.000	Jarque-Bera (JB):			151.241
Skew:		-1.327	Prob(JB):			1.44e-33
Kurtosis:		6.332	Cond. No.			454.

El modelo para predecir las ventas tiene la forma:

$$Ventas = 2.939 + 0.046TV + 0.189Radio - 0.0010Periodico$$

Coeficiente de Determinación

¿Con qué precisión podemos predecir las ventas futuras?

- La precisión se puede evaluar por la cantidad de variabilidad observada que es explicada por las variables de entrada, algunas de los indicadores utilizados son:
 - R^2 (R-squared): Indica que proporción de la variabilidad total es absorbida por las variables regresoras.
 - $R^2_{Ajustado}$ (Adj. R-squared): No siempre aumenta con la inclusión de nuevas variables, pero si la diferencia entre R^2 y $R^2_{Ajustado}$ es considerable, entonces existe el riesgo de haber agregado términos no significativos.
- Estos indicadores se encuentran entre 0 – 1, entre más cercano mejor la precisión del modelo.

OLS Regression Results						
Dep. Variable:	ventas					
Model:		OLS				
Method:			Least Squares			
Date:			Sat, 23 Oct 2021			
Time:			19:37:49			
No. Observations:			200			
Df Residuals:			196			
Df Model:			3			
Covariance Type:			nonrobust			
coef	std err	t	P> t	[0.025	0.975]	
TV	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
periodico	-0.0010	0.006	-0.177	0.860	-0.013	0.011
constante	2.9389	0.312	9.422	0.000	2.324	3.554
Omnibus:		60.414	Durbin-Watson:			2.084
Prob(Omnibus):		0.000	Jarque-Bera (JB):			151.241
Skew:		-1.327	Prob(JB):			1.44e-33
Kurtosis:		6.332	Cond. No.			454.

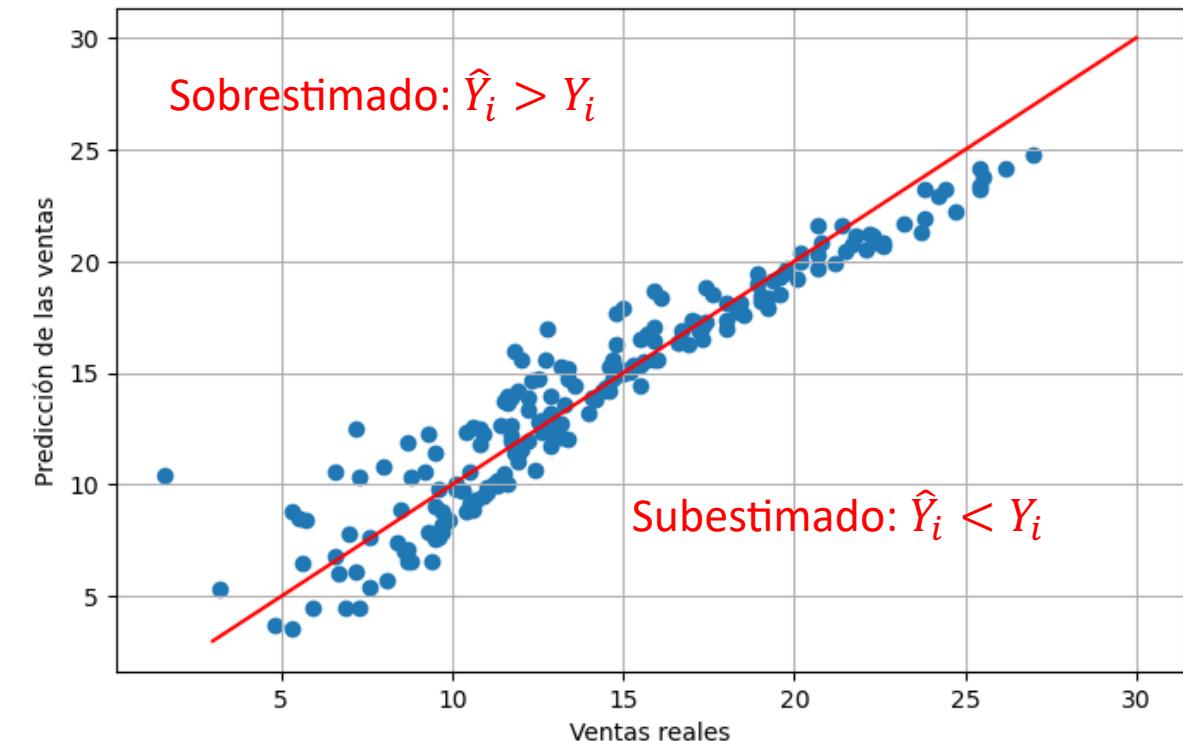
Los valores R^2 (0.897) y $R^2_{Ajustado}$ (0.896) indican una buena capacidad predictiva del modelo

¿Con qué precisión podemos predecir las ventas futuras?

- Otra forma de evaluar la precisión del modelo es comparar el valor real de Y con la predicción.
- Además, se pueden utilizar los siguientes indicadores:
 - Error cuadrático medio MSE:**
 - $MSE = \frac{1}{n} \sum (Y_i - \hat{Y}_i)^2$
 - Raíz del error cuadrático medio RMSE:**
 - $RMSE = \sqrt{MSE}$
 - Porcentaje de error:**
 - $CV = \frac{RMSE}{|\bar{Y}|} = \frac{1.67}{|14.02|}$
 - $CV = 0.12$
- El CV indica un porcentaje de error de aproximadamente el 12%

Y Real	Predicción
22.1	20.523974
10.4	12.337855
9.3	12.307671
18.5	17.597830
12.9	13.188672
...	...
7.6	5.370342
9.7	8.165312
12.8	12.785921
25.5	23.767321
13.4	15.173196

$$MSE = 2.78$$
$$RMSE = 1.67$$
$$\bar{Y} = 14.02$$



La gráfica y el CV indican una buena capacidad predictiva del modelo.

1. Falta de normalidad de los errores
2. Varianza no constante de los términos de error.
3. Valores atípicos.
4. Puntos de alto apalancamiento.
5. Multicolinealidad.
6. No linealidad de las relaciones respuesta-predictor.

Variables Categóricas

- Hasta el momento hemos asumido que todas las variables en nuestro modelo de regresión lineal son cuantitativas.
- Pero en la práctica algunos predictores son cualitativos.
- Si un predictor cualitativo (también conocido como factor), entonces incorporarlo a un modelo de regresión es muy simple, simplemente creamos un indicador o variable “dummie”

Factor					Variables Dummies			
TV	radio	periodico	ventas	zona	zona_norte	zona_occidente	zona_oriente	zona_sur
230.1	37.8	69.2	22.1	sur	0	0	0	1
44.5	39.3	45.1	10.4	occidente	0	1	0	0
17.2	45.9	69.3	9.3	occidente	0	1	0	0
151.5	41.3	58.5	18.5	occidente	0	1	0	0
180.8	10.8	58.4	12.9	oriente	0	0	1	0
...
38.2	3.7	13.8	7.6	occidente	0	1	0	0
94.2	4.9	8.1	9.7	sur	0	0	0	1
177.0	9.3	6.4	12.8	occidente	0	1	0	0
283.6	42.0	66.2	25.5	norte	1	0	0	0
232.1	8.6	8.7	13.4	norte	1	0	0	0



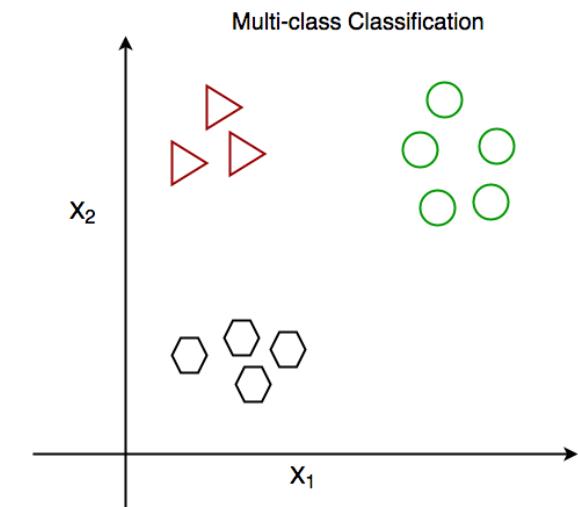
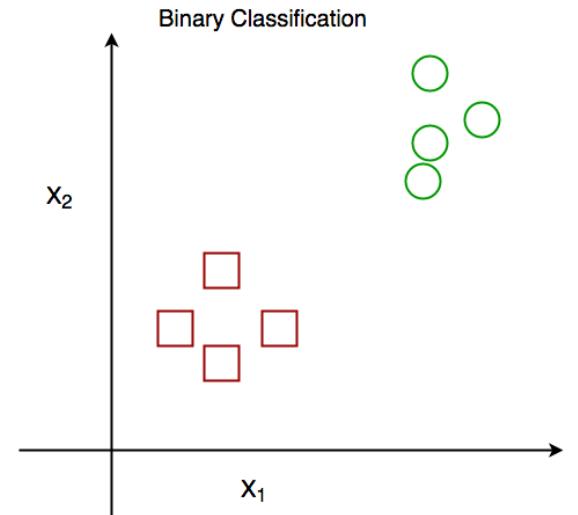
Pontificia Universidad
JAVERIANA
Bogotá

2.2 Regresión Logística

Educación **Continua**
Generamos experiencias educativas

Problema de Clasificación

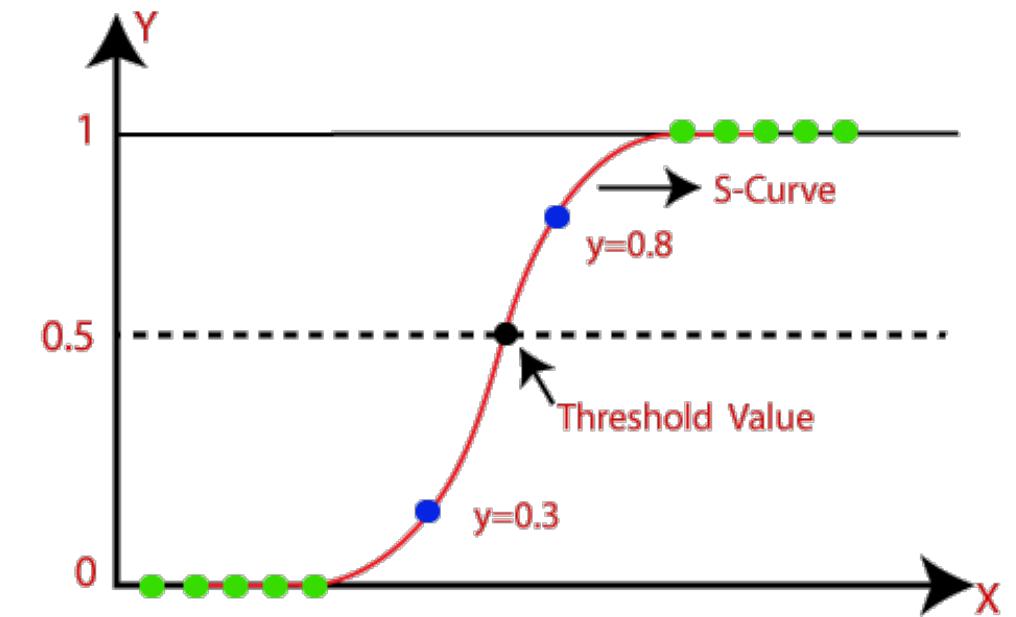
- El problema de clasificación consiste en predecir una clase o categoría
- Se dividen en:
 - *Problema Clasificación binaria* -> la variable a predecir tiene dos posibles resultados.
 - *Problema Multi-Clase* -> la variable a predecir tiene más de dos posibles resultados.
- Los modelos y algoritmos para predecir son llamados clasificadores.
- Algunos ejemplos de problemas de clasificación:
 - Determinar si un proyecto tendrá retrasos o sobrecostos
 - Predecir del impago (default) de un crédito
 - Determinar si un correo es spam
 - Determinar la condición médica de una persona de acuerdo a sus síntomas
 - Detectar transacciones fraudulentas
 - Clasificar imágenes de objetos



Regresión Logística - RL

- Se utiliza para modelar y predecir una variable discreta o binaria, esto es $y = \{0, 1\}$
- Los valores 0 y 1 representan categorías binarias, ejemplo:
 - Compra (1) vs No Compra (0)
 - Fraude (1) vs No Fraude (0)
 - Cumple (1) vs No Cumple (0)
 - Enfermo (1) vs Sano (0)
 - Spam (1) vs No Spam (0)
- Usualmente la categoría que interesa predecir se codifica con 1.
- La RL predice la probabilidad de $y = 1$ a partir de la siguiente ecuación:

$$E(y) = P(y = 1|x) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}}$$



Estrategia de Promoción

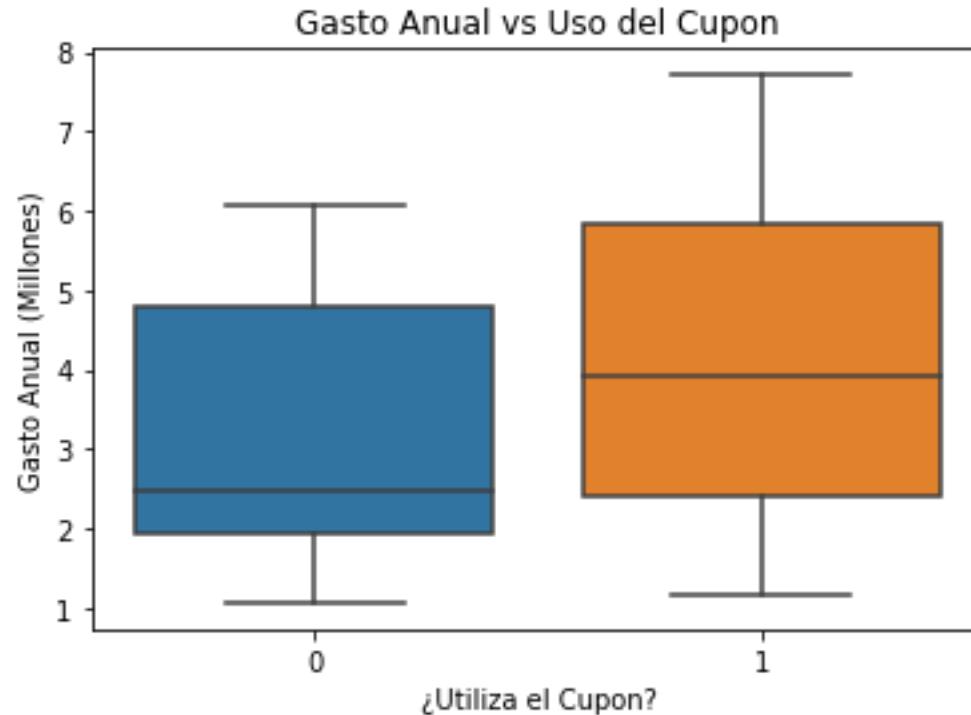
- Una cadena nacional de ropa realizará una promoción por correo.
- Ha ordenado imprimir copias de su catálogo y en cada uno incluye un cupón de \$50 de descuento en compras por \$200 o más.
- La cadena desea enviarlo sólo a aquellos clientes que tengan mayor probabilidad de usar el cupón.
- La gerencia considera que:
 1. La cantidad gastada anualmente por el cliente (X_1)
 2. Si posee o no una tarjeta de crédito de la tienda (X_2)
- son dos variables útiles para predecir si ese cliente usará el cupón.
- Realiza un testeo con 32 clientes para obtener datos y construir un modelo de RL que le permita predecir si un cliente utilizará o no un cupón.

Cliente	Gasto Anual Millones	Tarjeta	Cupón
1	2,291	1	1
2	3,215	1	0
3	2,135	1	1
4	3,924	0	1
5	2,528	1	1
6	2,473	0	0
7	2,384	0	0
8	7,708	0	1
9	1,182	1	1
10	3,345	0	1
11	5,547	0	0
12	1,917	0	0
...

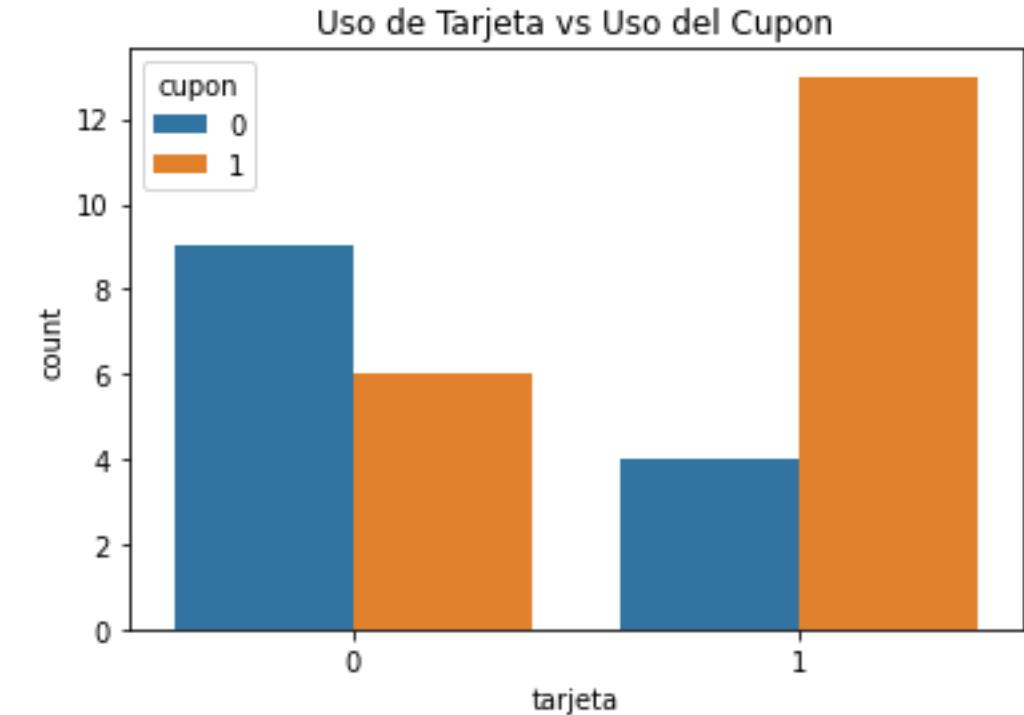
$Y = 0 \rightarrow$ No utiliza el cupon

$Y = 1 \rightarrow$ Utiliza el cupon

Análisis exploratorio



Los clientes que utilizan el cupón (1) tienen un gasto medio superior respecto a los que no lo utilizan (0)



Clientes que tienen tarjeta son mas propensos a utilizar el cupón en comparación a los que no tienen tarjeta

Estimación de Parámetros

- Para entrenar el modelo utilizamos el 70% de los datos (20 datos), a esto le llamamos conjunto de entrenamiento y el resto (10 datos) lo vamos a utilizar para evaluar.
- Con estos datos la ecuación de regresión logística tiene la siguiente forma:

$$P(y = 1|x_1, x_2) = \frac{e^{-4.2587+0.6863X_1+3.4334X_2}}{1 + e^{-4.2587+0.6863X_1+3.4334X_2}}$$

Esta ecuación se obtiene con el *método de máxima verosimilitud*.

- Si se quiere predecir la probabilidad de que un cliente use el cupón, entonces tendremos que reemplazar en la ecuación el gasto anual y si tiene o no tarjeta de crédito.
- Considere un cliente con gasto anual $X_1 = 2.64$ millones y Si tiene tarjeta de crédito $X_2 = 1$. La probabilidad de que use el cupón es:

$$P(y = 1|x_1, x_2) = \frac{e^{-4.2587+0.6863(2.64)+3.4334(1)}}{1 + e^{-4.2587+0.6863(2.64)+3.4334(1)}} = 0.7284$$

- Esto indica que el cliente tiene una probabilidad de 0.7284 de utilizar el cupón por tanto.
- Cuando la función da como resultado una probabilidad igual o superior a 0.5, entonces se asigna la categoría $y = 1$.

Cociente de Posibilidad (Odds Ratio)



Pontificia Universidad
JAVERIANA
Bogotá

- Otra forma de interpretar el resultado es haciendo uso del cociente de posibilidades (odds ratio):

$$odds = \frac{P(X = 1)}{1 - P(X = 1)}$$

- Para la predicción que se realizó anteriormente el odds:

$$odds = \frac{0.7284}{1 - 0.7284} = 2.68$$

- Esto indica que el cliente tiene 2.68 veces más de posibilidades de usar el cupón que de no usarlo.
- Esto también se puede calcular como:

$$odds = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p} = e^{-4.2587 + 0.6863(2.64) + 3.4334(1)} = 2.68$$

- Al aplicar logaritmo natural a la ecuación para el odds, obtenemos:

$$\log(odds) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = -2.5813 + 0.4288X_1 + 2.6628X_2$$

- Por tanto, los parámetros de la regresión logística β_1 y β_2 se interpretan como el cambio que se produce en el $\log(\text{odds})$ cuando X_1 o X_2 cambian en una unidad, siendo que las demás variables permanecen constantes.
- Observe que el uso de la tarjeta de crédito tiene un mayor efecto en el $\log(\text{odds})$ para uso del cupón, en comparación a un incremento en 1 millón de pesos en el gasto anual.

Matriz de Confusión

- Al utilizar el modelo surge la pregunta:

¿Qué tanto acierta el clasificador en las predicciones?

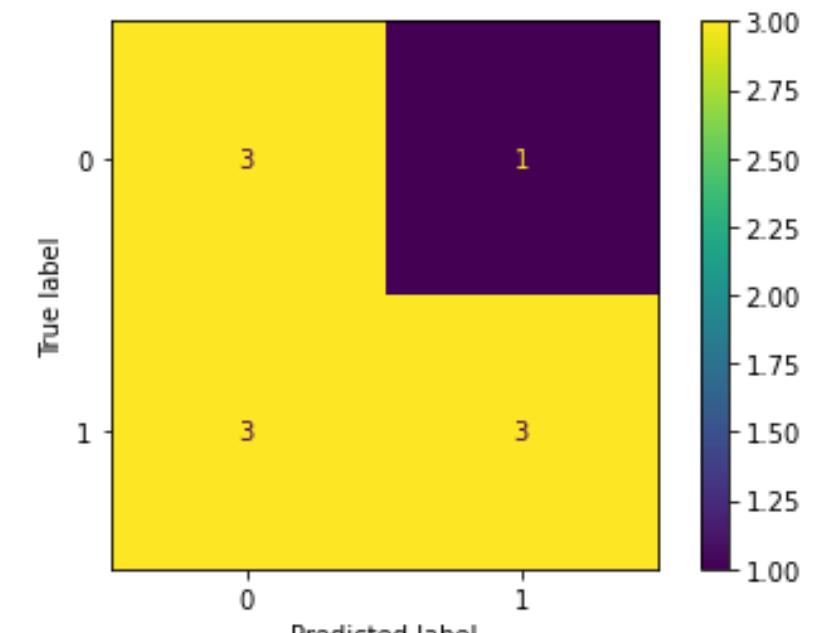
- Para responder a esta pregunta se utiliza la **matriz de confusión**, la cual resume los resultados de las predicciones.
- Con esta matriz se generan las siguientes métricas:
 - Accuracy**: proporción de predicciones acertadas.
 - Sensibilidad (Recall)**: proporción predicciones positivas correctas, con respecto al total de clases reales positivas.
 - Precisión**: proporción de predicciones positivas correctas, con respecto al total de predicciones.
 - F-1 Score**: se calcula como la media harmónica entre el Recall y la Precisión.

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

$$F1 - score = 2 \frac{Precision \times Recall}{Precision + Recall}$$

Predicciones para el Conjunto de Prueba

Cliente	Gasto Anual	Tarjeta	Utiliza Cupón	Probabilidad	Utiliza Cupón Predicción
12	1.92	0	0	0.0502	0
11	5.55	0	0	0.3895	0
4	3.92	0	1	0.1725	0
9	1.18	1	1	0.4961	0
27	1.27	1	0	0.5116	1
25	2.64	1	1	0.7284	1
20	6.28	0	1	0.5129	1
29	4.80	0	0	0.2760	0
10	3.35	0	1	0.1235	0
31	5.53	1	1	0.9512	1



$$Accuracy = \frac{3 + 3}{3 + 1 + 3 + 3} = 0.6 \quad \text{En solo en 6 de las 10 predicciones el modelo acertó (60%)}$$

$$Sensibilidad = \frac{3}{3 + 3} = 0.5 \quad \text{De los 6 clientes que usaron el cupón, el modelo detectó 3 (50%)}$$

$$Precisión = \frac{3}{3 + 1} = 0.75 \quad \text{El modelo predijo que 4 clientes usarían el cupón, pero solo 3 lo hicieron (75%)}$$



Pontificia Universidad
JAVERIANA
Bogotá

2.3 Arboles de Decisión

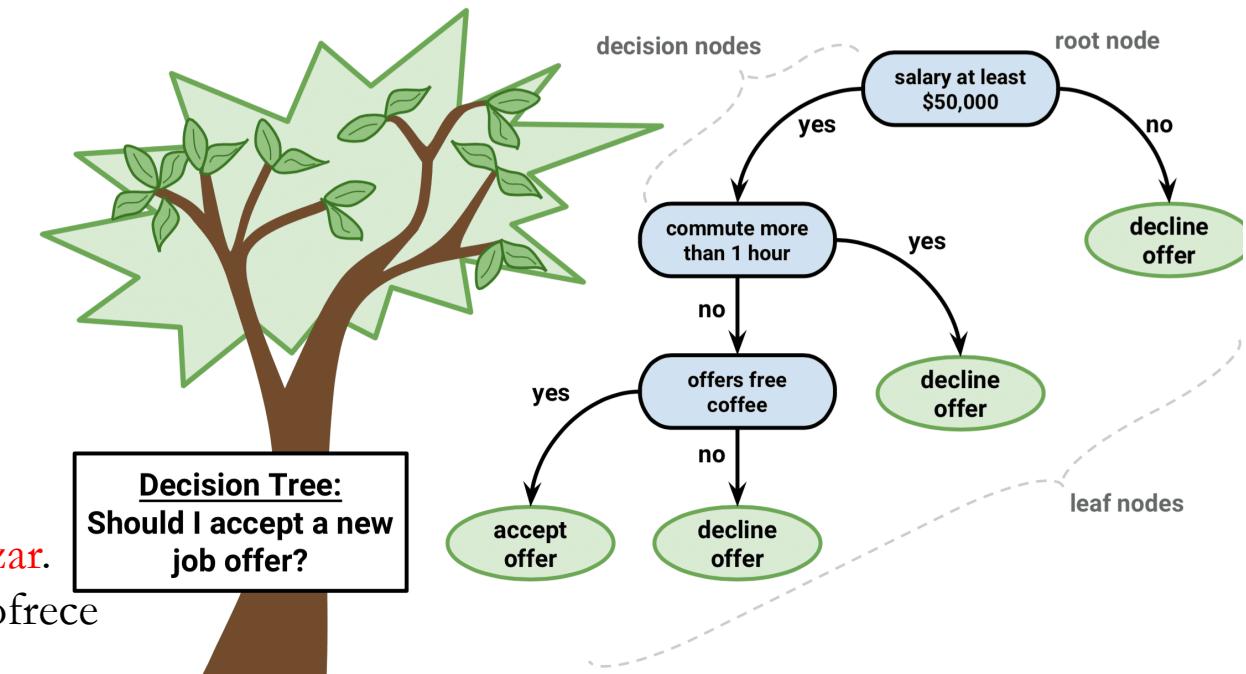
Educación **Continua**
Generamos experiencias educativas

Arboles de Decisión

Un árbol de decisión es un **conjunto de condiciones** organizadas en una **estructura jerárquica**, de tal manera que la **decisión final** a tomar se puede determinar siguiendo las **condiciones** que se cumplen desde la **raíz** del árbol hasta sus **hojas**.

De este árbol se pueden extraer las siguientes reglas (caminos)

1. Salario < \$50.000 -> **Rechazar**.
2. Salario \geq \$50.000 + Transporte mas de 1 hora -> **Rechazar**.
3. Salario \geq \$50.000 + Transporte menos de 1 hora + NO ofrece café gratis -> **Rechazar**.
4. Salario \geq \$50.000 + Transporte menos de 1 hora + ofrece café gratis -> **Aceptar**.



Usted puede construir un árbol de decisión a partir de sus propias reglas.
Sin embargo esto puede que aplique solo para usted, y no es útil para generalizar en un grupo de interés.
En este curso nos enfocaremos en arboles de decisión que “aprende” reglas a partir de datos.

Ejemplo

Suponga que se desea construir un Árbol de Decisión para generar un sistema de recomendación de géneros de películas.

Se ha tomado una muestra de 10 personas a las que se registró:

1. Sexo
2. Edad
3. Genero de película de su preferencia

Sexo	Edad	Genero de Película
F	20	romantica
F	24	accion
M	25	accion
F	35	romantica
M	35	accion
M	40	romantica
F	40	romantica
F	50	documental
F	55	romantica
M	60	documental

¿Cómo se Construye un Árbol de Decisión?

Divide y reinaras: Para todas las variables se generan divisiones de las clases y se calcula la impureza del nodo.



Al parecer la edad genera un división mas “pura” (homogénea) de los géneros de películas.

Pero, ¿Cómo medimos la pureza?

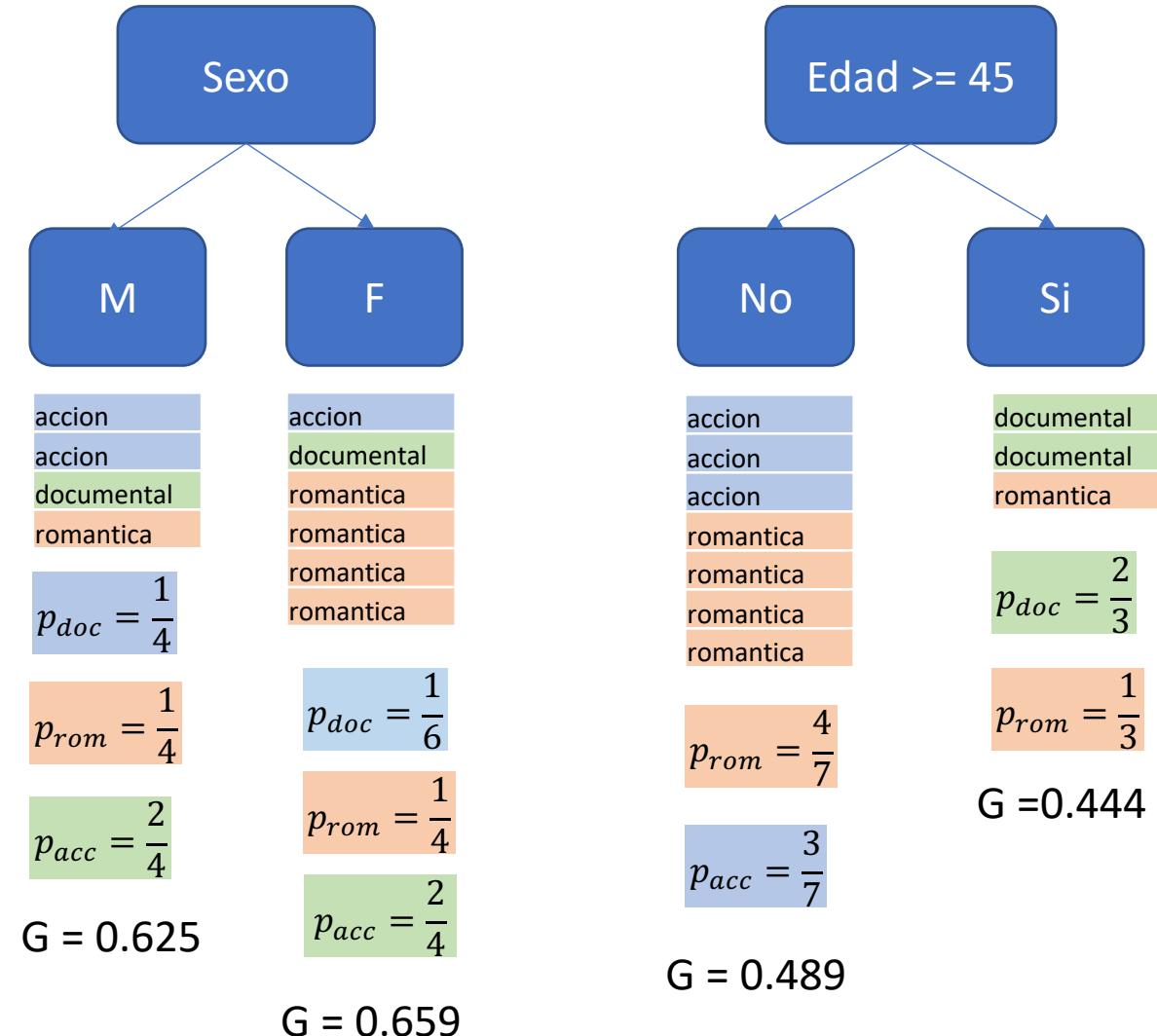
- El índice de Gini se conoce como una medida de la pureza del nodo: **un valor pequeño indica que un nodo contiene predominantemente observaciones de una sola clase.**

$$G_{nodo} = 1 - \sum_{k=1}^K p_k^2$$

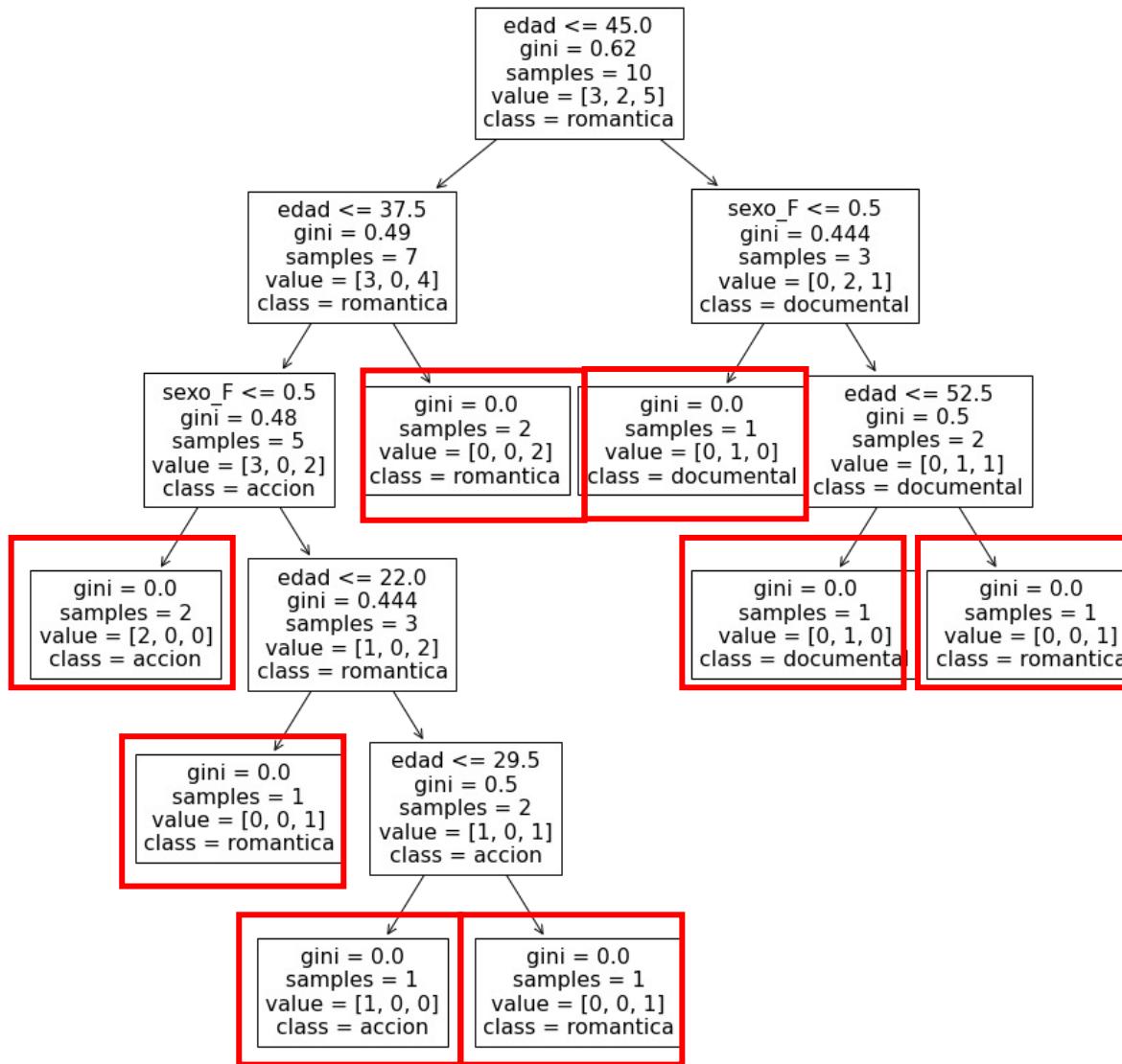
- Donde, k son las clases en cada nodo y p_k la proporción la clase k dentro del nodo.
- Para cada partición se saca la suma ponderada de cada uno de los nodos que genera.
- De esta manera:

$$G(\text{Sexo}) = (4/10)*0.625 + (6/10)*0.659 = 0.645$$

$$G(\text{Edad} \geq 45) = (7/10)*0.489 + (3/10)*0.444 = 0.476$$
- Con este resultado la mejor partición (menor impureza) se genera con la partición edad ≥ 45 .
- A partir de los nodos generados se pueden generar mas particiones teniendo en cuenta el mismo criterio.



¿Cuando detener la Partición?



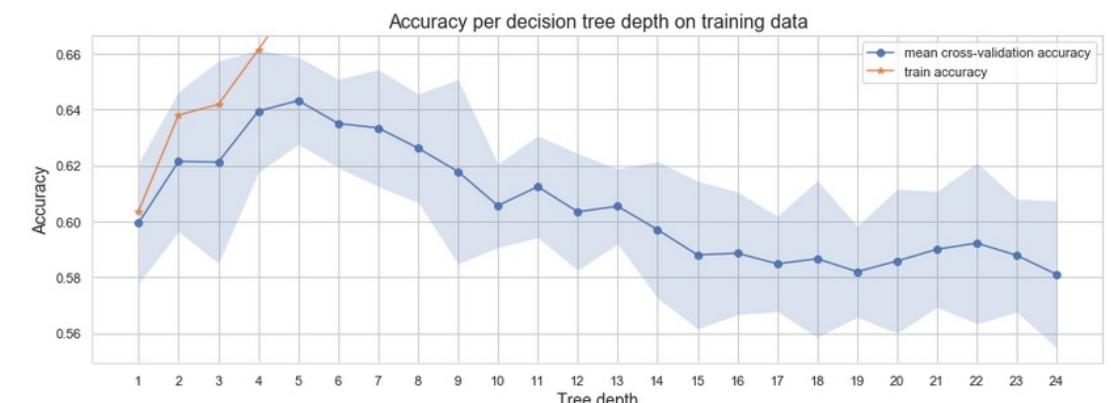
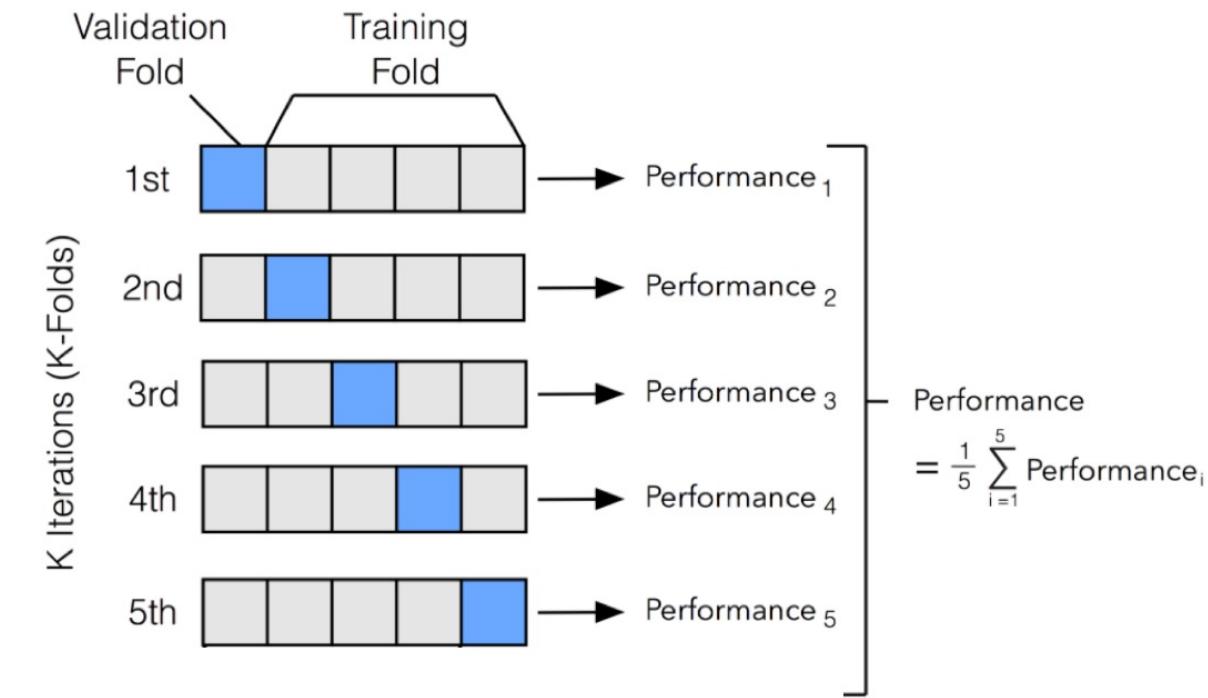
Observe que total hay 8 nodos terminales para 10 instancias y están compuestos por una misma clase (ósea error 0). Si bien, esto parece funcionar perfectamente, genera un problema que se denomina sobreajuste, lo que hace que el árbol no sea efectivo para predecir nuevas observaciones.

La profundidad del árbol es un parámetro que se debe calibrar para evitar el sobreajuste (overfitting).

El procedimiento para calibrar parámetros en un árbol o cualquier otro algoritmo se denomina validación cruzada. Este procedimiento lo veremos con mayor detalle mas adelante.

Optimización de Parámetros

- En general se utiliza una técnica de remuestreo para estimar el rendimiento de los modelos y calibrar parámetros.
- Se utiliza un subconjunto de muestras para ajustar un modelo y las muestras restantes se utilizan para estimar la eficacia del modelo.
- Este proceso se repite varias veces y los resultados se agregan y resumen.
- En la validación cruzada las muestras se dividen aleatoriamente en k conjuntos de tamaño aproximadamente igual tamaño.
- Se ajusta un modelo utilizando todas las muestras excepto un subconjunto, el cual se utiliza para validación.





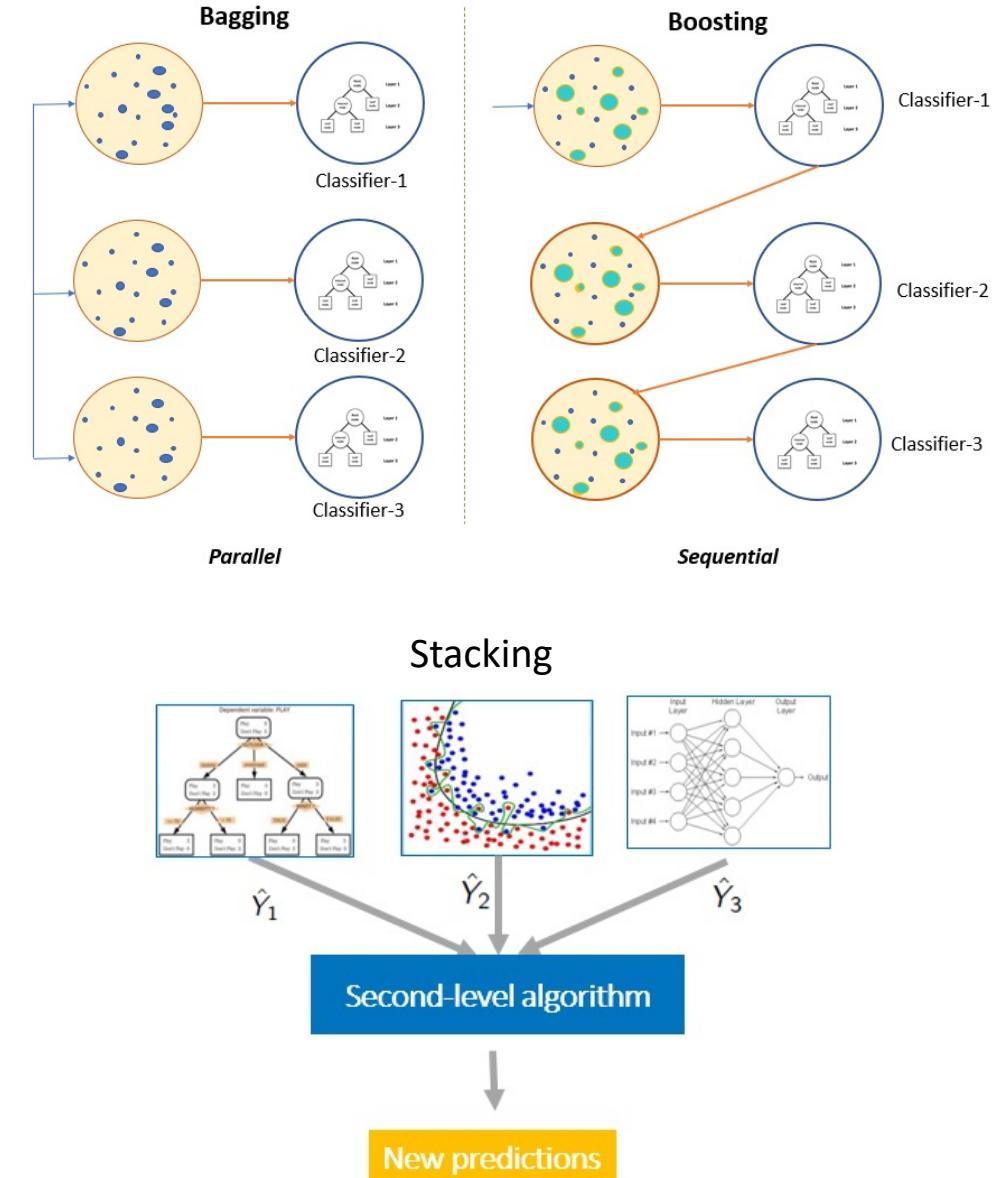
Pontificia Universidad
JAVERIANA
Bogotá

2.4 Métodos de Ensamble

Educación **Continua**
Generamos experiencias educativas

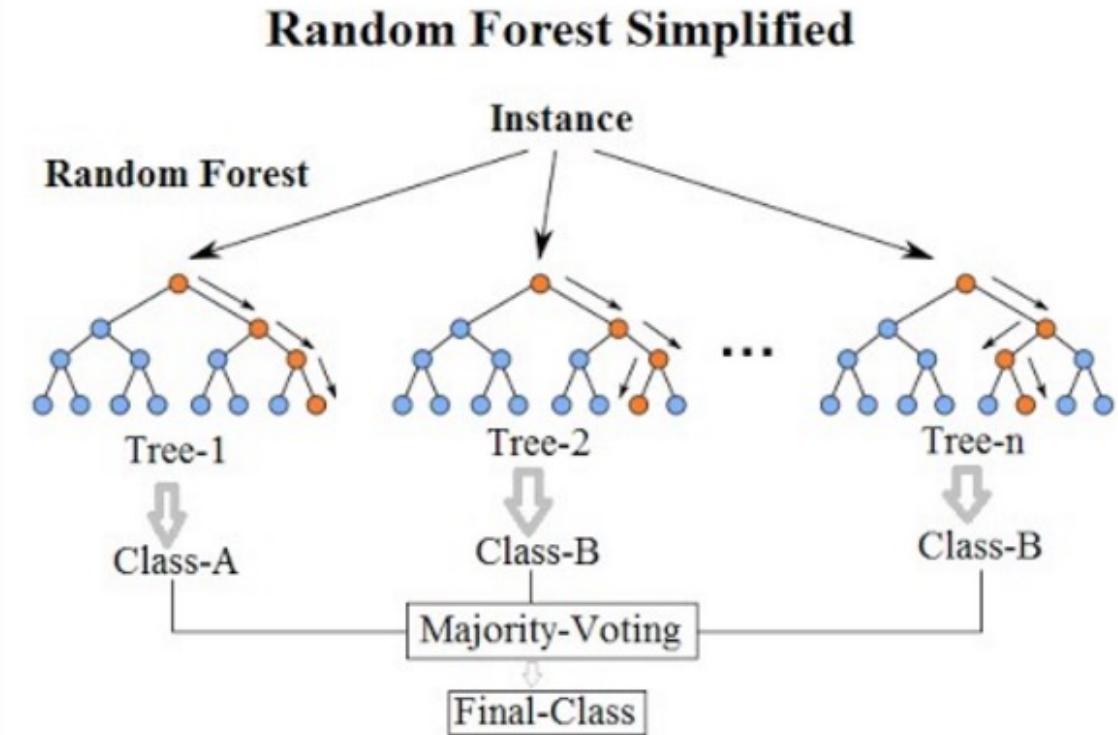
Métodos de Ensamble

- Un método de ensamble es un enfoque que combina muchos modelos simples.
- La idea es tener "bloques de construcción" para obtener un modelo único y potencialmente muy poderoso.
- Estos modelos de bloques de construcción simples a veces se conocen como aprendices débiles, ya que pueden conducir a predicciones mediocres por sí mismos.
- Estos son métodos de conjunto para los cuales el bloque de construcción puede ser una simple regresión o un árbol de clasificación.
- Los enfoques de ensamble mas utilizados son:
 - Bagging (Bootstrap Aggregation)
 - Boosting
 - Stacking



Bagging – Random Forest

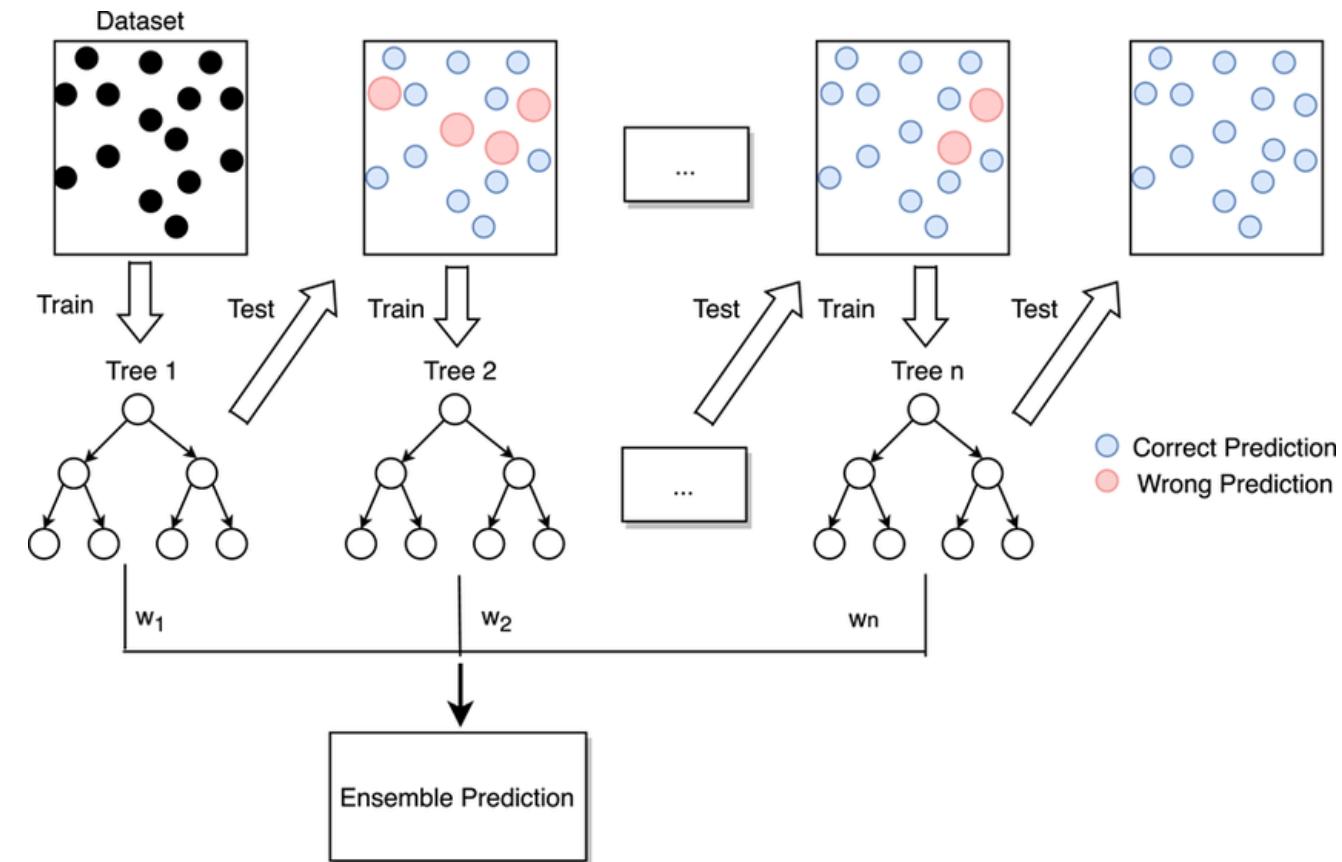
- Los árboles de decisión sufren de una alta variabilidad.
- Esto significa que si dividimos los datos de entrenamiento en dos partes y ajustamos un árbol de decisión a ambas mitades, los resultados que obtengamos podrían ser bastante diferentes.
- Por el contrario, un procedimiento con poca varianza arrojará resultados similares si se aplica repetidamente a distintos conjuntos de datos.
- El Bagging es un procedimiento de propósito general para reducir la varianza de un método de aprendizaje estadístico.
- El Random Forest es el algoritmo mas conocido que funciona bajo el método Bagging.
- Con el Random Forest se generan una cantidad determinada arboles a partir de una selección aleatoria de variables e instancias.
- EL numero de arboles, la cantidad de variables e instancias son parámetros de calibración.



La predicción final se da por promedio para la regresión o por votación mayoritaria para un problema de clasificación.

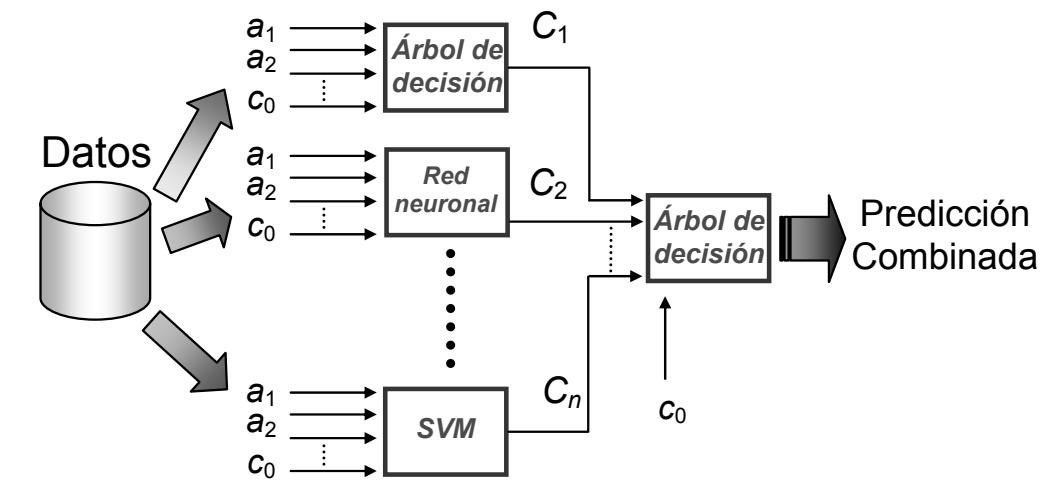
Boosting

- Se basa de la asignación de un peso a cada instancia del conjunto de entrenamiento.
- En cada iteración, *boosting* aprende un modelo que minimiza la suma de los pesos de los ejemplos clasificados erróneamente.
- Los errores de cada iteración se utilizan para actualizar los pesos de los ejemplos del conjunto de entrenamiento.
- De esta forma se incrementa el peso de los ejemplos errados y se reduce el peso de los ejemplos acertados.
- Se busca que el modelo en la siguiente iteración otorgue más relevancia a los ejemplos que los modelos anteriores habían clasificado erróneamente.
- Algunos de los métodos mas conocidos:
 - Adaboost
 - Gradiente Boosting Machine
 - Extreme Gradient Boosting



Stacking

- Es un método multiclásificador bastante simple basado en la combinación de modelos generados por diferentes algoritmos de aprendizaje.
- Dado que cada modelo se aprende con un mecanismo de aprendizaje distinto se consigue que los modelos del conjunto sean diferentes.
- Una primera y simple versión del *stacking* consiste en aprender un modelo con uno de los algoritmos que deseemos utilizar.
- Si, por ejemplo, queremos utilizar árboles de decisión, regresión logística y redes neuronales, generaríamos un modelo con cada técnica y para un mismo conjunto de datos.
- La clasificación se podría realizar mediante una votación mayoritaria.





Pontificia Universidad
JAVERIANA
Bogotá

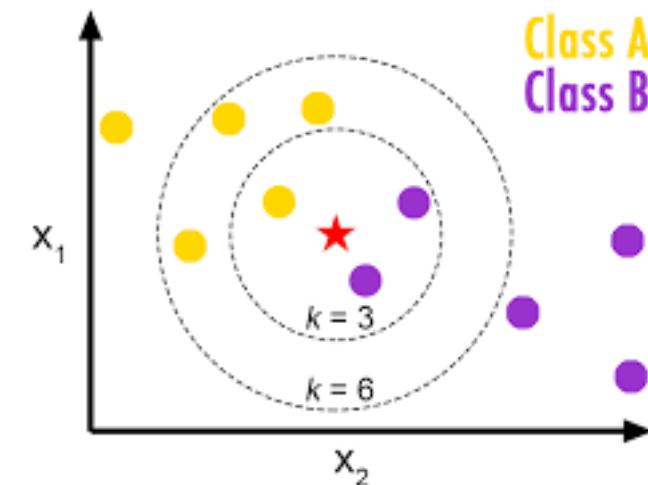
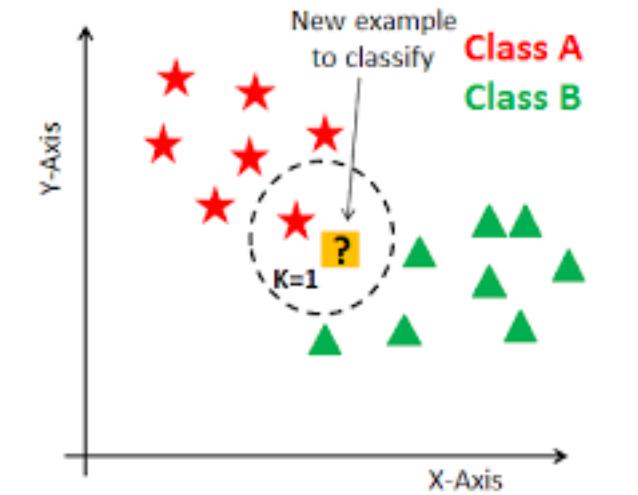
2.5 K Vecinos mas Cercanos

Educación **Continua**

Generamos experiencias educativas

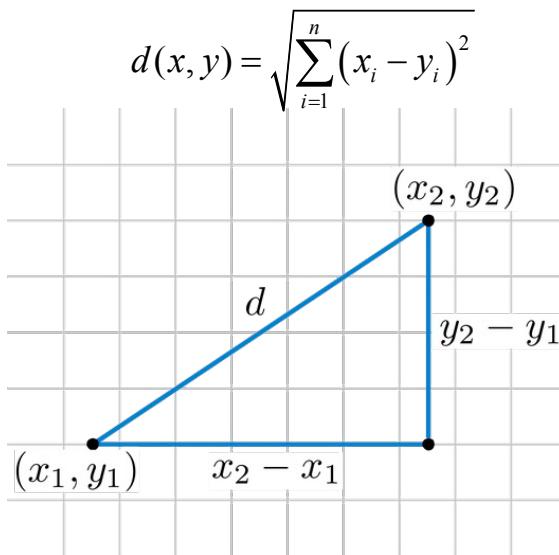
La Regla del Vecino mas Próximo

- La regla del vecino más próximo simplemente asigna la clase del ejemplo más próximo, utilizando una función de distancia.
- Esta regla, conocida como 1-NN (one nearest neighbor), tiene bastantes problemas, ya que ignora la densidad o la región donde se encuentra el ejemplo.
- Una variante de este método son los k vecinos más próximos (kNN, k-nearest neighbors) en el que se asigna la clase mayoritaria entre los k vecinos más próximos.
- En un problema de regresión se asigna el promedio o mediana de los vecinos mas próximos.

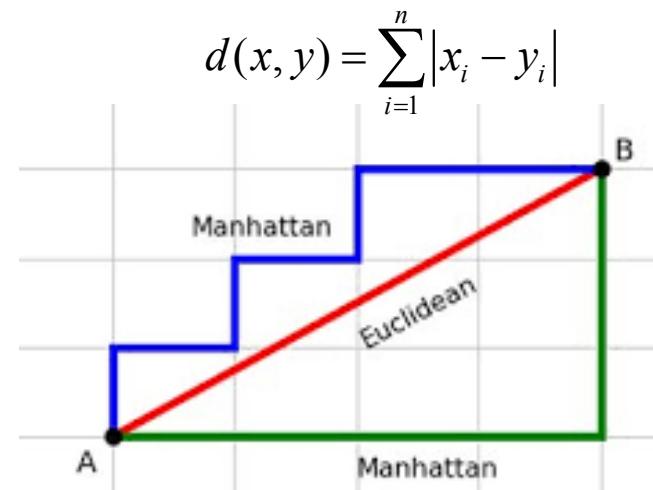


Las medidas de distancia se utilizan para dimensionar la separación de dos instancias u observaciones.

Distancia Euclídea. Es la distancia clásica, como la longitud de la recta que une dos puntos en el espacio euclídeo:

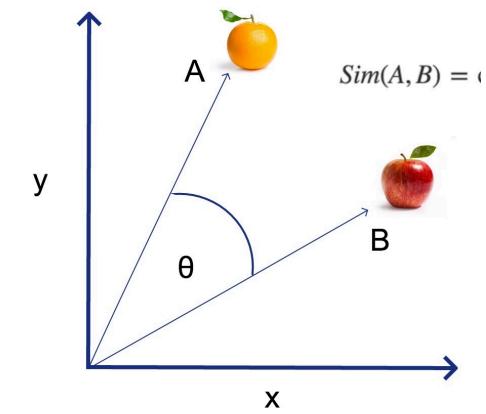


Distancia de Manhattan: hace referencia a recorrer un camino no en diagonal o en zigzag.



Distancia del coseno. Si se considera que cada ejemplo es un vector, la distancia sería el coseno del ángulo que forman

$$d(x, y) = \arccos\left(\frac{x^T y}{\|x\| \cdot \|y\|}\right)$$





Pontificia Universidad
JAVERIANA
Bogotá

3. Aprendizaje No Supervisado

Técnicas Descriptivas y Exploratorias

Educación **Continua**

Generamos experiencias educativas



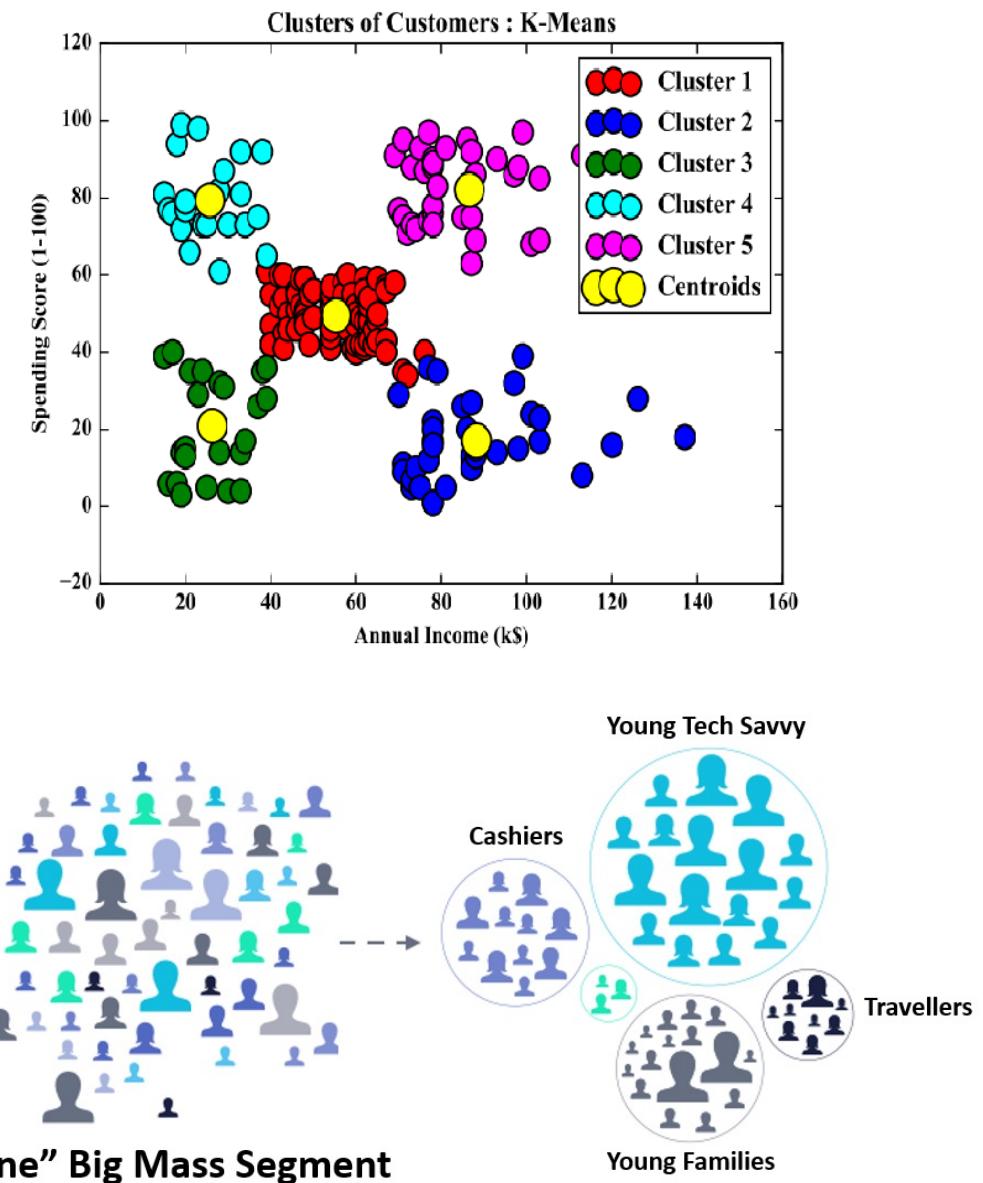
Pontificia Universidad
JAVERIANA
Bogotá

3.1 Métodos de Cluster y Anomalías

Educación **Continua**
Generamos experiencias educativas

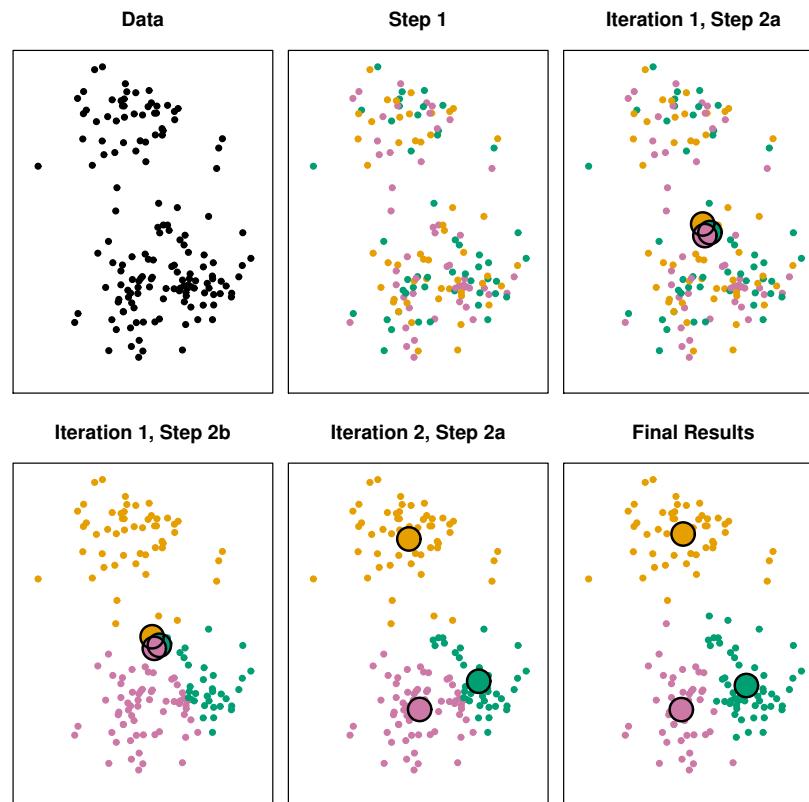
Métodos de Cluster

- Permiten encontrar grupos compuestos de observaciones o instancias con características similares o cercanas.
- La cercanía entre las entidades u observaciones que conforman el grupo o cluster, se determina a partir de las medidas de distancia que se abordaron con el métodos de los vecinos mas cercanos.
- Una de las aplicas mas comunes es en mercadeo.
- Es posible que tengamos acceso a un gran número de variables de clientes :
 - ingresos familiares medios,
 - ocupación,
 - distancia desde el área urbana más cercana
- Con esto se puede identificar grupos de personas que podrían ser más receptivas a una forma particular de publicidad, o más propensas a comprar un producto en particular.
- Los grupos son una abstracción conceptual a partir de sus características.
- Esto significa, que no tenemos etiquetas previas que los identifiquen. Normalmente, estas se determinan posterior al análisis del clúster.



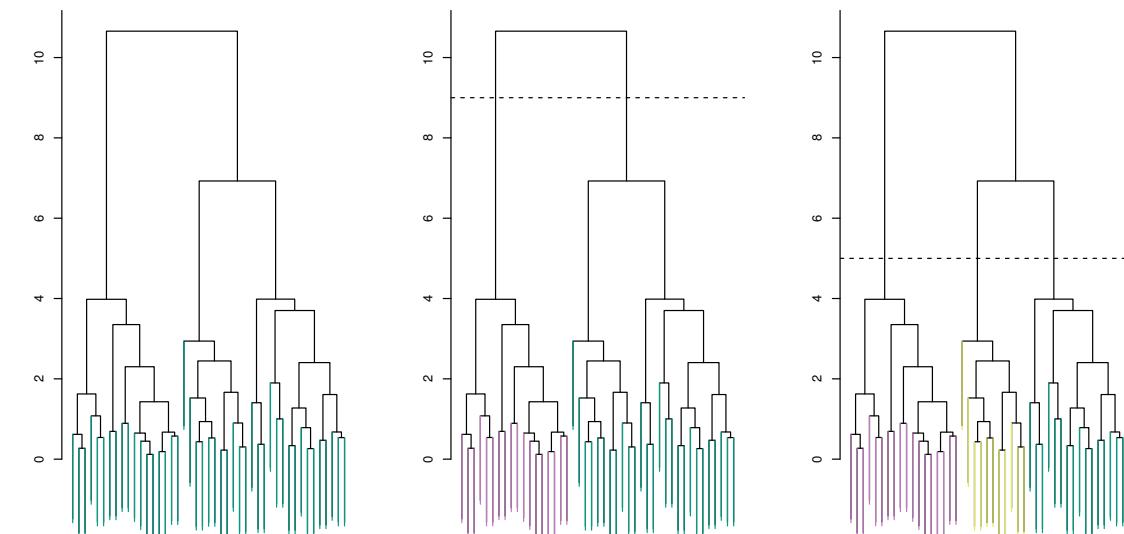
K-medias (Kmeans)

Para realizar el agrupamiento de K-medias, primero debemos especificar el número deseado de agrupamientos K; entonces el algoritmo de K-medias asignará cada observación exactamente a uno de los K conglomerados.



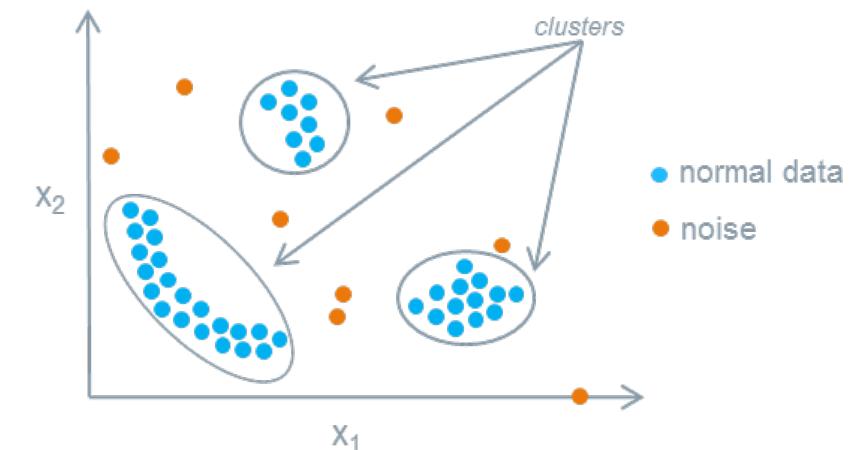
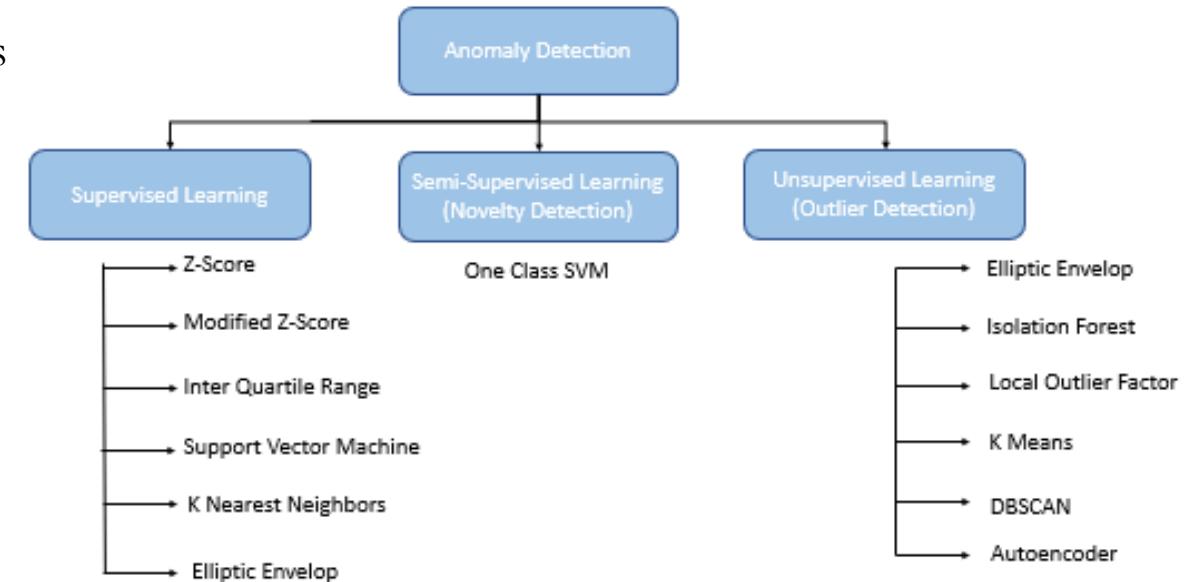
Agrupamiento Jerárquico (Hierarchical Clustering)

El agrupamiento jerárquico es un enfoque alternativo que no requiere que nos comprometamos con una elección particular de K. El agrupamiento jerárquico da como resultado una atractiva representación basada en árboles de las observaciones, llamada dendrograma.



Detección de Anomalías

- La detección de anomalías consiste en identificar observaciones que presentan un patrón atípico.
- Estas observaciones se denominan: atípicos (outliers) o novedades (novelty).
- Hay una diferencia sutil entre estos dos conceptos, pero podríamos intentar:
 - Outlier: valor anormal contenido en los datos.
 - Novelty: nuevo valor en los datos que podría tener un comportamiento anormal y desconocido.
- Existen tres tipos de técnicas para detectar anomalías:
 - **Aprendizaje supervisado:** se conoce de antemano los valores atípicos.
 - **Aprendizaje semisupervisado:** se conocen algunas observaciones atípicas.
 - **Aprendizaje no supervisado:** no se conocen los datos atípicos. En este caso se identifican como observaciones alejadas de grupos o clusters con alta densidad de observaciones.





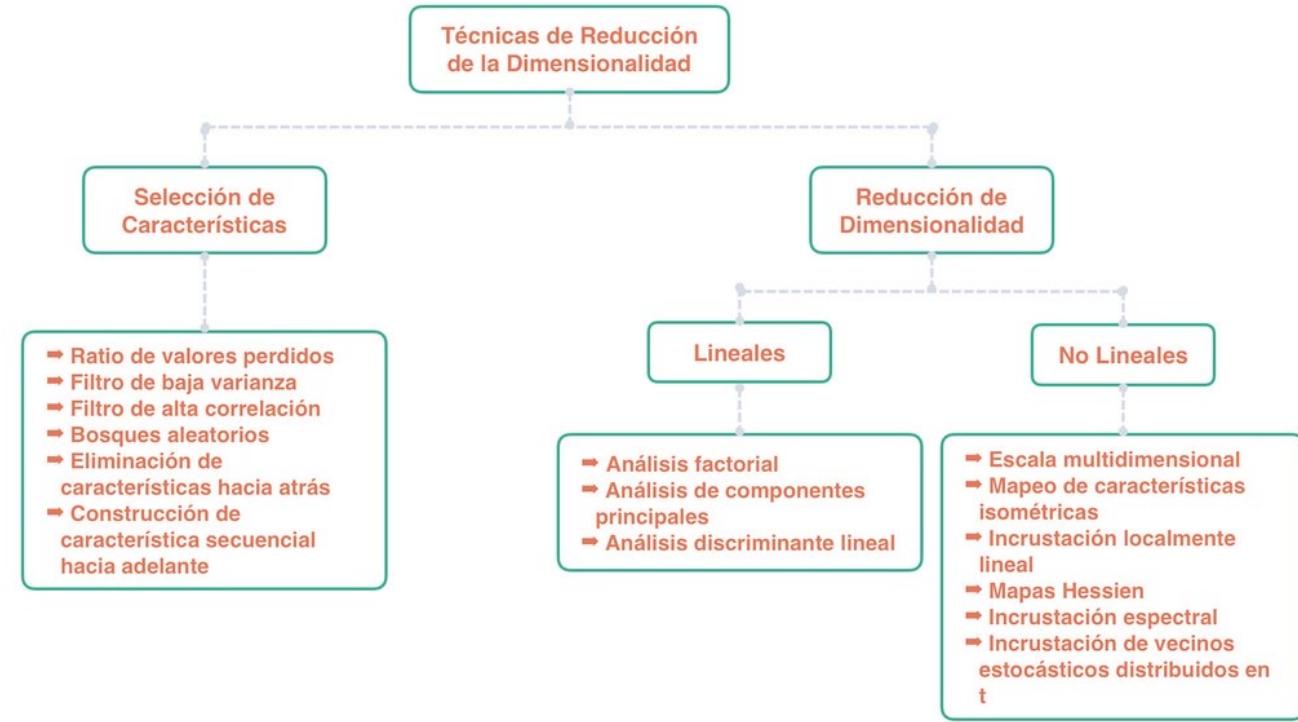
Pontificia Universidad
JAVERIANA
Bogotá

3.2 Reducción de Dimensiones

Educación **Continua**
Generamos experiencias educativas

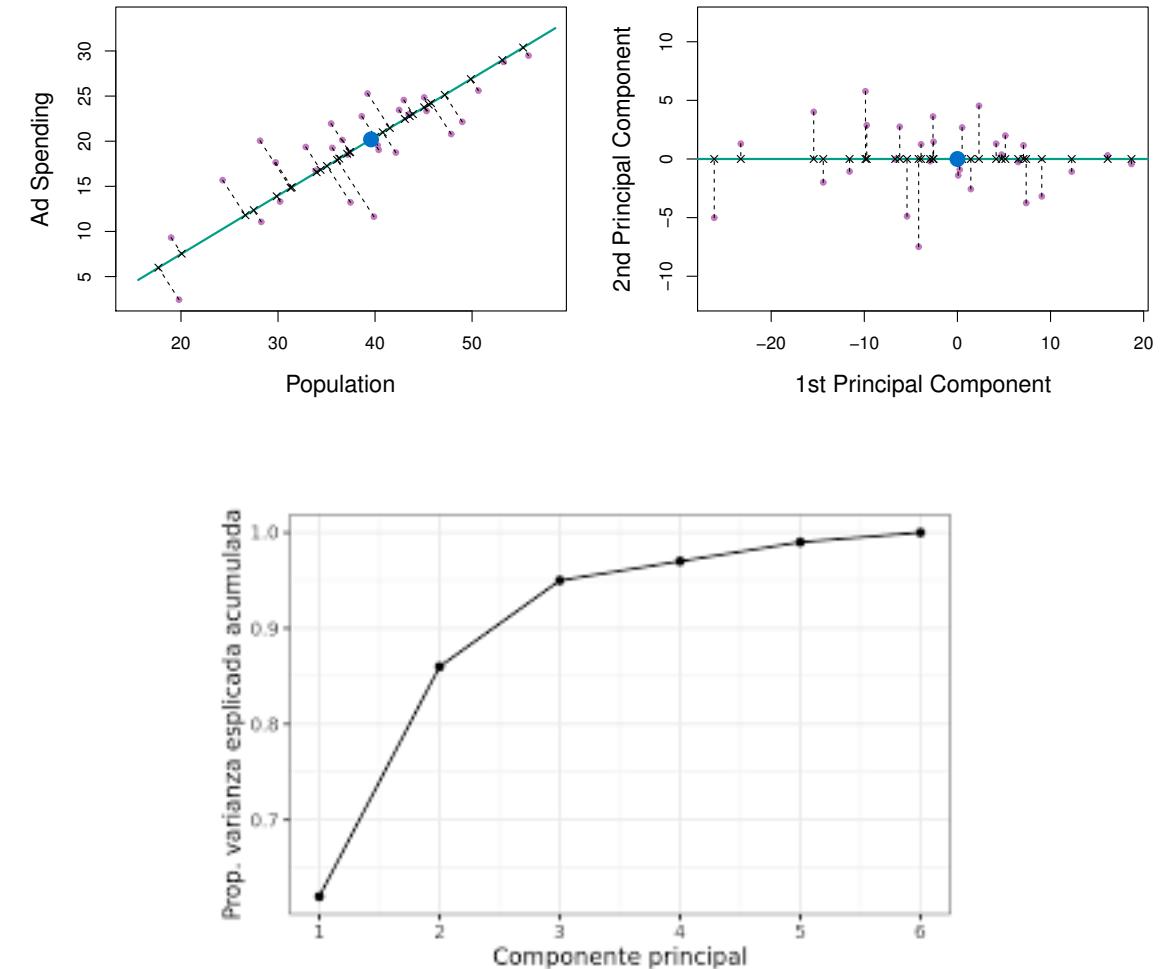
Reducción de Dimensiones

- La dimensionalidad se refiere a la cantidad de variables o atributos en conjunto de datos.
- La alta dimensionalidad es, muchas veces, un gran problema a la hora de aprender de los datos.
- Si tenemos muchas dimensiones (atributos) respecto a la cantidad de instancias o ejemplos, nos encontraremos con una situación poco deseable para el análisis de los datos.
- Lo que se busca es reducir el numero de atributos a partir de la selección o transformación de atributos.
- En la selección se descartan atributos que no bridan información relevante para el aprendizaje de patrones.
- En la transformación se ”construyen nuevas variables” a partir de los atributos existentes a través de procesos de transformación.



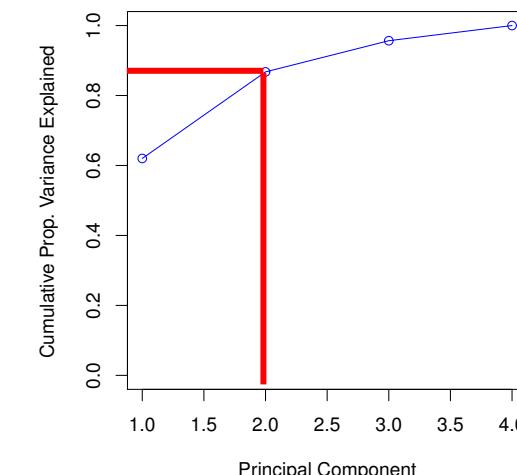
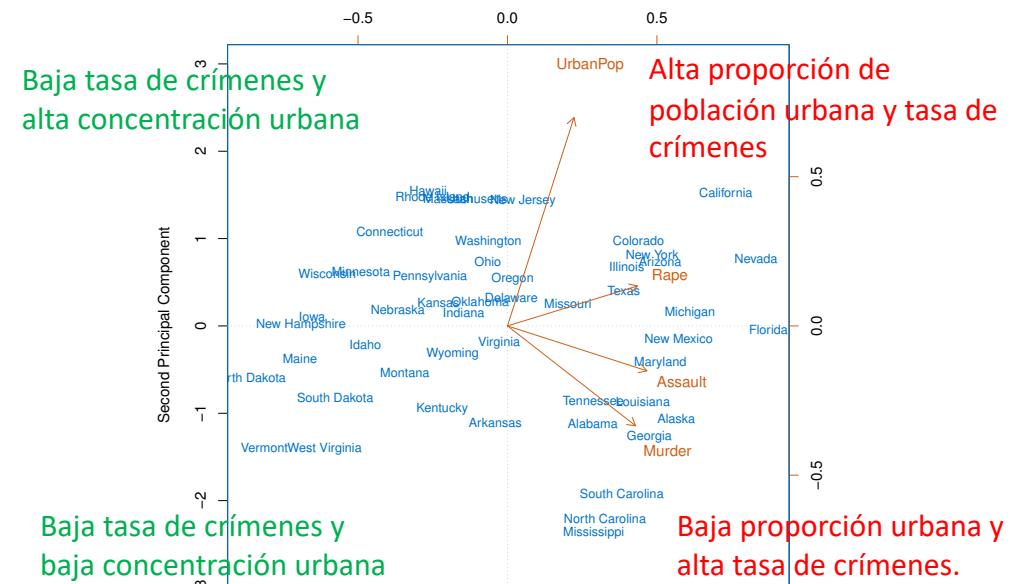
Análisis de Componentes Principales (PCA)

- PCA es una técnica para reducir la dimensión de una matriz de datos con n observaciones y p variables.
- La dirección del primer componente principal de los datos es aquella a lo largo de la cual las observaciones varían más.
- Observe la que muestra el tamaño de la población (población) en decenas de miles de personas y el gasto publicitario de una empresa en particular (anuncio) en miles de dólares, para 100 ciudades.
- La línea continua verde representa la dirección del primer componente principal de los datos.
- Podemos ver a simple vista que esta es la dirección a lo largo de la cual existe la mayor variabilidad en los datos, entre aprox. -20 y 20.
- La segunda componente concentra una menor variabilidad, entre aprox. -5 y 5.
- El objetivo es encontrar un numero de componentes que concentren la mayor la variabilidad (cantidad de información) de las variables originales.



Análisis de Componentes Principales (PCA)

- Ilustramos una aplicación con datos estadísticos arrestos para cada uno de los 50 estados de los Estados Unidos.
- El conjunto de datos contiene el número de arrestos por cada 100.000 residentes por cada uno de los tres delitos: Asalto, Asesinato y Violación.
- También registramos UrbanPop (el porcentaje de la población de cada estado que vive en áreas urbanas).
- En la figura 1 podemos ver que las variables Asalto, Asesinato y Violación apuntan en la dirección positiva de la primera componente. Esto revela lo siguiente:
 - Estas tres variables están correlacionadas en la misma dirección. Por tanto, la primera componente puede indicar el *"nivel de criminalidad"* de un estado.
 - Los estados con valores positivos en esa componente tienen alto nivel de criminalidad.
- La segunda componente mide la concentración en zonas urbanas. En la dirección positiva están los estados con mayor concentración.
- La figura de abajo muestra que un poco mas del 80% de la información de los 4 atributos originales es explicado por 2 componentes, una reducción de dimensiones del 50%.





Pontificia Universidad
JAVERIANA
Bogotá

3.3 Reglas de Asociación

Educación **Continua**
Generamos experiencias educativas

Reglas de Asociación

- Es un método de aprendizaje de patrones basado en reglas
- Su objetivo es identificar reglas sólidas en bases de datos
- Su uso se relaciona con el análisis de cesta de mercado (Market Basket Analysis)
- Con esto se busca identificar patrones de compra o consumo de los clientes, basado históricos de compra
- A partir de esto, se pueden:
 - Generar recomendaciones de compra
 - Diseñar promociones
 - Desarrollar estrategias de localización de productos
- Algunas aplicaciones pueden ser:
 - Web Mining
 - Sistemas de recomendación
 - Detección de fraude
 - Mercadeo y ventas
 - Seguridad informática
 - Análisis de sistemas de producción

ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
...	...

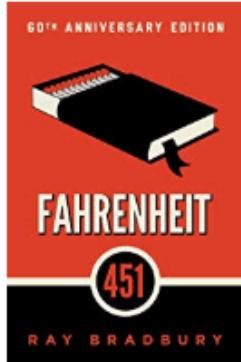
market
basket
transactions

{Diapers, Beer} Example of a frequent itemset

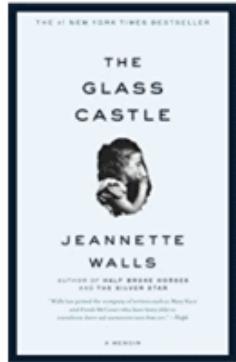
{Diapers} → {Beer} Example of an association rule

Algunas Aplicaciones

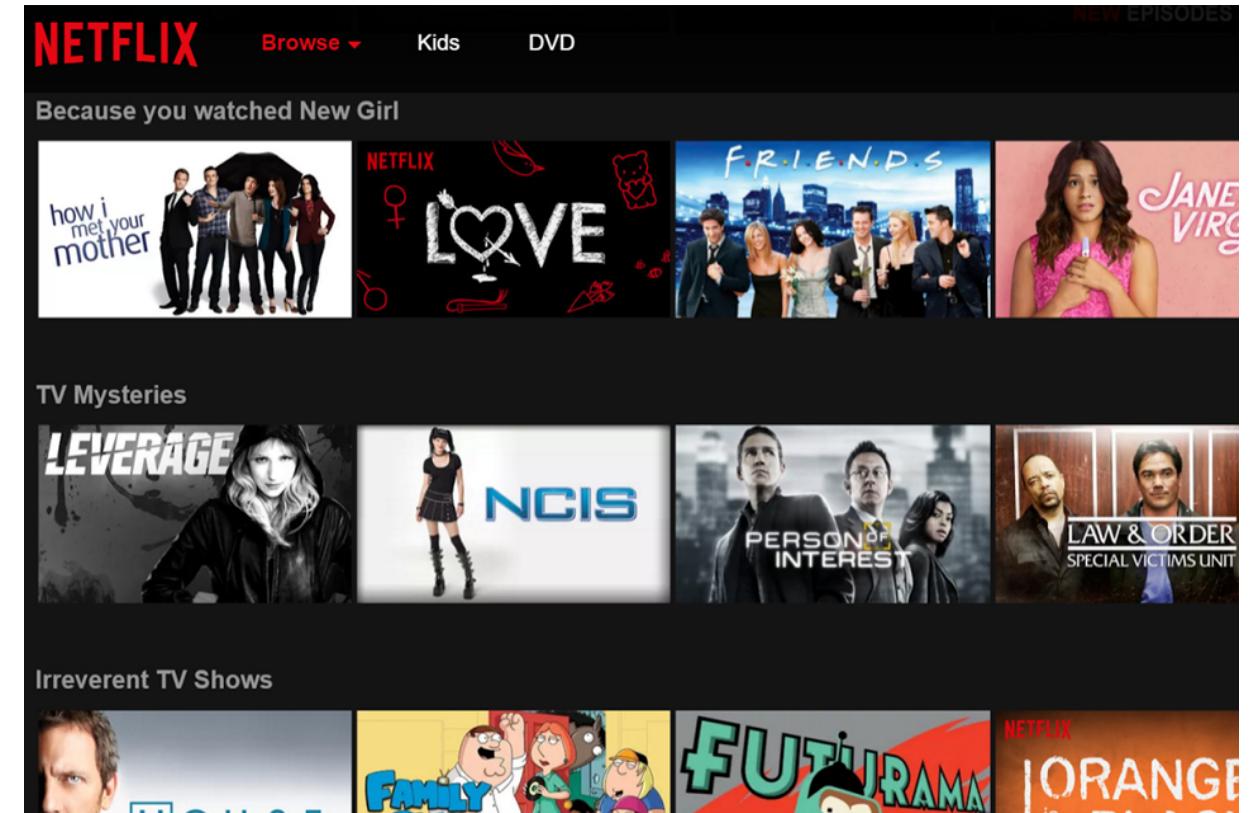
Customers who bought this item also bought



Fahrenheit 451
› Ray Bradbury
★★★★★ 3,502
#1 Best Seller in
Censorship & Politics
Paperback
\$8.99



The Glass Castle: A Memoir
› Jeannette Walls
★★★★★ 7,651
#1 Best Seller in Journalist
Biographies
Paperback
\$9.79



The screenshot shows the Netflix homepage with the title "Because you watched New Girl". It features promotional banners for various shows: "how i met your mother", "LOVE", "F.R.I.E.N.D.S.", "Jane the Virgin", "LEVERAGE", "NCIS", "PERSON OF INTEREST", "LAW & ORDER SPECIAL VICTIMS UNIT", "HOUSE", "FAMILY GUY", "FUTURAMA", and "ORANGE IS THE NEW BLACK". Below these are sections for "TV Mysteries" and "Irreverent TV Shows".

Importancia de una Regla de Asociación

La importancia de las Reglas de Asociación ($X \rightarrow Y$) se puede determinar mediante 3 parámetros que se utilizan para identificar la fuerza del algoritmo:

Soporte (support): es una indicación de la frecuencia con la que aparece el conjunto de elementos en el conjunto de datos.

$$\text{Soporte}(X, Y) = \frac{\text{frecuencia}(X, Y)}{N}$$

Confianza (confidence): es el porcentaje de todas las transacciones que satisfacen X que también satisfacen Y.

$$\text{Confianza}(X \rightarrow Y) = \frac{\text{Soporte}(X, Y)}{\text{Soporte}(X)}$$

Elevación (lift): indica el nivel de independencia entre dos eventos. Si la elevación es mayor 1, existe un alto nivel de dependencia entre los dos ítems, por tanto la regla es potencialmente útil

$$\text{Levantamiento}(X \rightarrow Y) = \frac{\text{Soporte}(X, Y)}{\text{Soporte}(X) \times \text{Soporte}(Y)}$$

Para determinar la importancia de la regla {Diapers, Beer}:

$\text{Soporte}(\{\text{Diapers}, \text{Beer}\}) = \frac{3}{5} = 0.6$. En el 60% de las transacciones aparece esta regla o patron.

$\text{Confianza}(\{\text{Diapers}, \text{Beer}\}) = \frac{3/5}{4/5} = 0.75$. Con un 75% de confianza si alguien compra pañales también comprará cerveza.

$\text{Levantamiento}(\{\text{Diapers}, \text{Beer}\}) = \frac{3/5}{(\frac{4}{5})(\frac{3}{5})} = 1.25$. Esto nos indica que esta relación es mas frecuente, que la aparición de cada ítem por separado.

ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
...	...

$$N = 5$$

$$\text{Frecuencia}(\text{Diapers}) = 4$$

$$\text{Frecuencia}(\text{Beer}) = 3$$

$$\text{Frecuencia}(\{\text{Diapers}, \text{Beer}\}) = 3$$