Pontificia Universidad JAVERIANA Bogotá

Técnicas Descriptivas para la Generación de Nuevo Conocimiento

ING. HERIBERTO FELIZZOLA JIMENEZ

Educación **Continua**

Generamos experiencias educativas

Pontificia Universidad JAVERIANA Bogotá

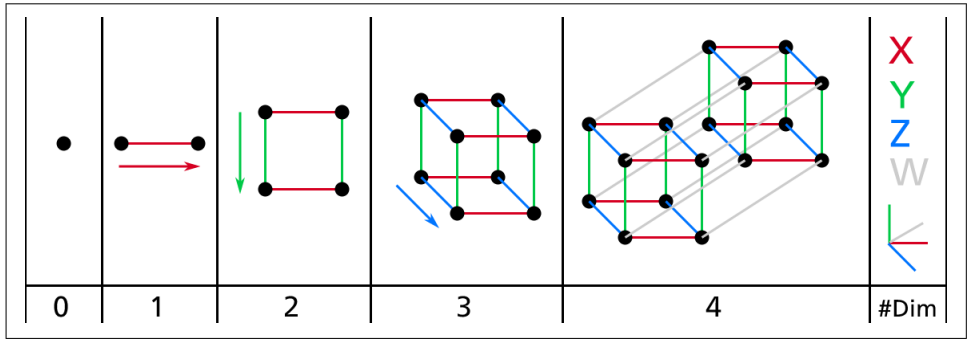
Reducción de Dimensiones

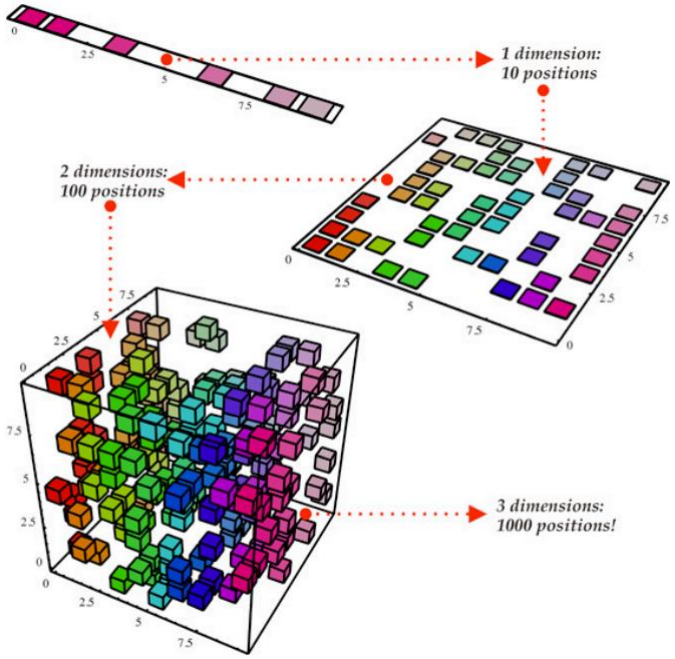
Educación **Continua**

Generamos experiencias educativas

Reducción de Dimensiones

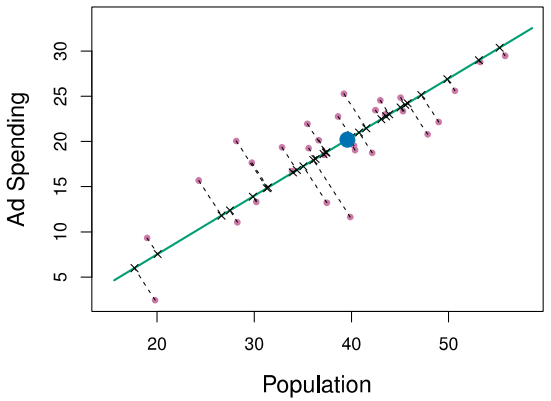
- La dimensionalidad se refiere a la cantidad de variables o atributos en conjunto de datos.
- La alta dimensionalidad es, muchas veces, un gran problema a la hora de analizar los datos y generar modelos de aprendizaje a partir de los datos.
- Los humanos pueden procesar 3, quizás 4 dimensiones a la vez.
- Si 10 dimensiones, te resultará difícil observar los tipos de patrones que saltan a la vista en 2 o 3 dimensiones.
- La reducción de dimensionalidad se utiliza a menudo para hacer que los datos sean más consumibles para los seres humanos y los algoritmos de aprendizaje.
- 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 ”generan nuevas variables” a partir de los atributos existentes a través de procesos de transformación.

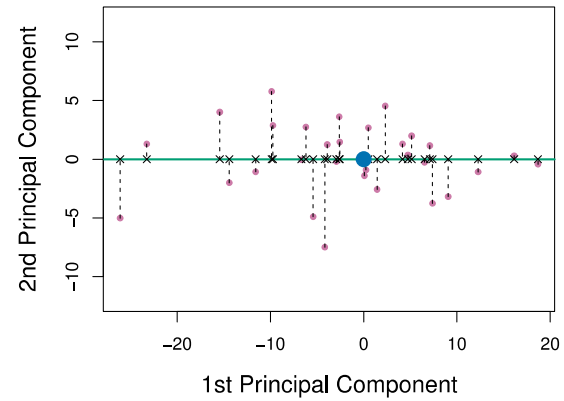


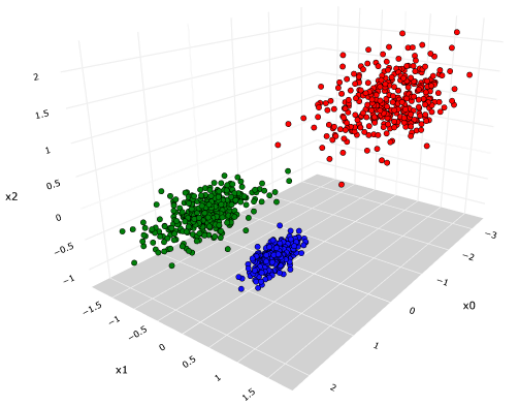


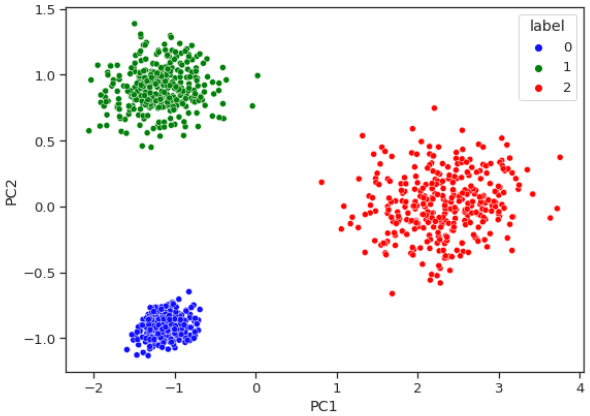
Componentes Principales (PCA)

- 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.









Determinación de las Componentes Principales

- Cada una de las dimensiones encontradas por PCA es una combinación lineal de las p variables.
- El primer componente principal PC_1 de un conjunto de variables X_1, X_2,. . . , X_p es la combinación lineal normalizada de características:

$$PC_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

- Esta primera componente recoge la mayor cantidad de información (varianza) de las variables originales.
- ϕ_{11} , $\phi_{21,}$ . . . , ϕ_{p1} como las cargas o los pesos de cada variable en el primer componente principal. La suma del cuadrado de estos pesos debe ser igual a 1.
- Dado un conjunto de datos X de n × p, ¿cómo calculamos el primer componente principal?

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2 \right\} \text{ subject to } \sum_{j=1}^{p}\phi_{j1}^2 = 1$$

El problema de optimización se puede resolver mediante una descomposición propia, una técnica estándar en álgebra lineal.

- El objetivo que estamos maximizando es simplemente la varianza muestral de los n valores de z_{i1}.
- Donde,

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip}$$

- El segundo componente principal es la combinación lineal de X_1,...,X_p que tiene la varianza máxima de todas las combinaciones lineales que no están correlacionadas con Z_1. De esta misma forma se determinan los demás componentes.

Ejemplo: Arrestos en Estados Unidos por Estado

- 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: Agresión, Asesinato y Violación.
- También se registra la Población como el porcentaje de la población de cada estado que vive en áreas urbanas.

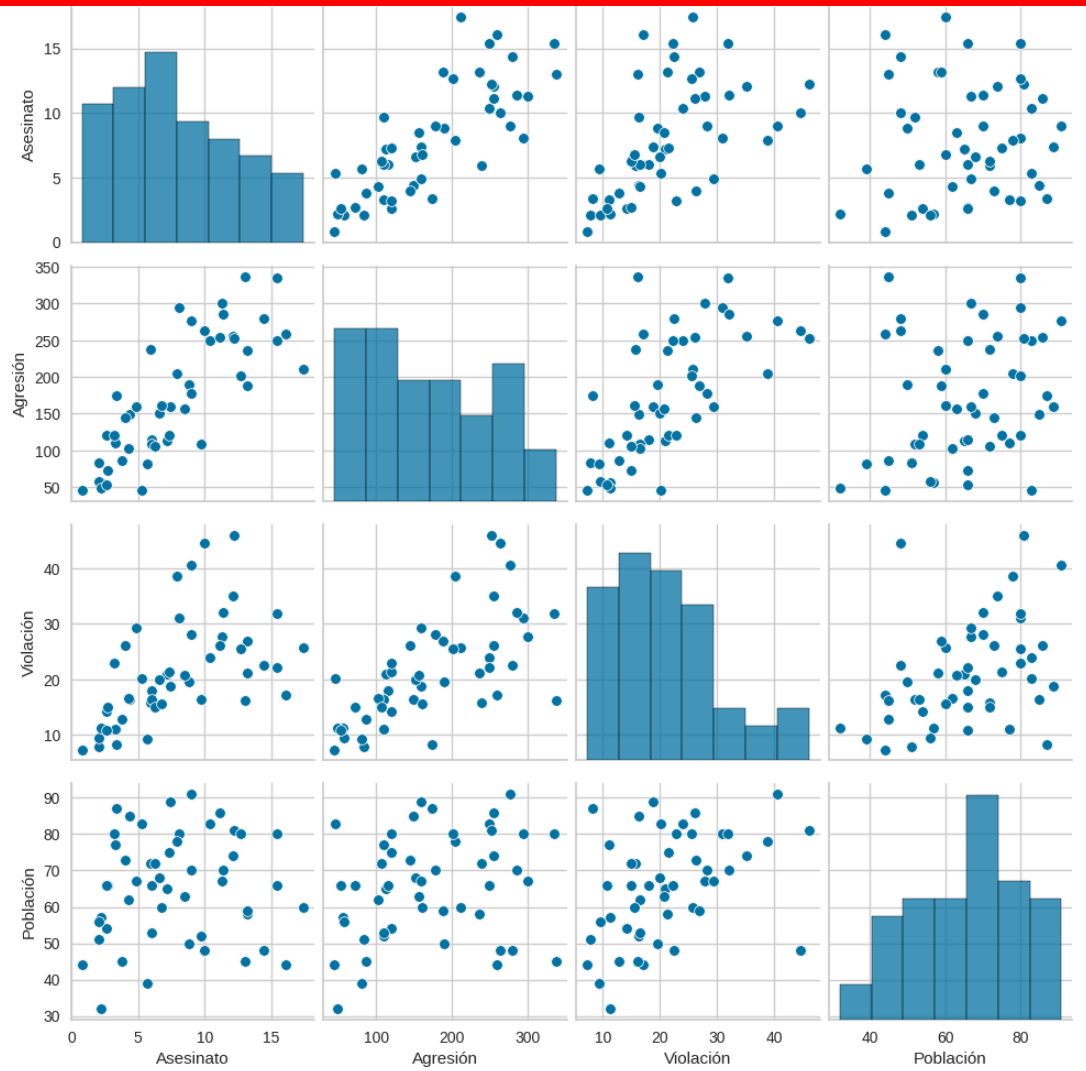
Tabla de Datos

Estado	Asesinato	Agresión	Violación	Población
Alabama	13.2	236	21.2	58
Alaska	10.0	263	44.5	48
Arizona	8.1	294	31.0	80
Arkansas	8.8	190	19.5	50
California	9.0	276	40.6	91
…	…	…	…	…

Resumen estadístico

	Asesinato	Agresión	Violación	Población
Promedio	7.79	170.76	21.23	65.54
Desviación	4.36	83.34	9.37	14.47
Mínimo	0.80	45.00	7.30	32.00
Máximo	17.40	337.00	46.00	91.00

¿Cómo caracterizar los estados en función del nivel de arresto y la población urbana?

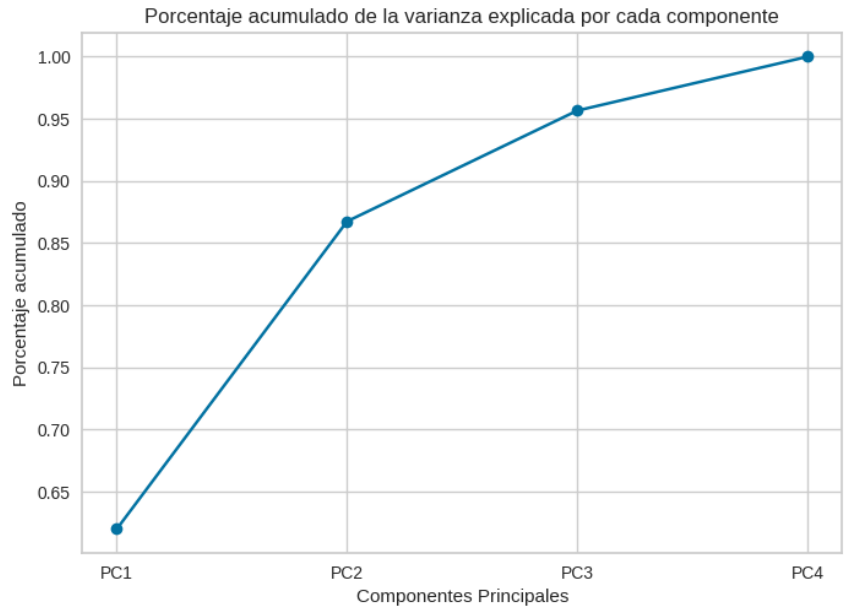


Vectores de peso

- Vemos que el primer vector de pesos otorga aproximadamente el mismo peso a Agresión, Asesinato y Violación, pero con mucho menos peso a Población.
- Por lo tanto, este componente corresponde aproximadamente a una medida de las tasas generales de delitos graves.
- El segundo vector de pesos coloca la mayor parte de su peso en Población y mucho menos peso en las otras tres características.
- Por tanto, este componente corresponde aproximadamente al nivel de urbanización del estado.
- En la gráfica se observa que los dos primeros componentes explican un poco mas de 85% de la información (varianza) contenida en las 4 cuatro variables originales.
- Por tanto, podemos analizar y caracterizar el nivel de arresto en cada estado haciendo uso de dos componentes principales:
  - PC1: tasa general de delitos graves
  - PC2: nivel de urbanización

Peso de las variables en cada componente (ϕ)

Variable Original	PC1	PC2
Asesinato	0.5358995	−0.4181809
Agresión	0.5831836	−0.1879856
Violación	0.2781909	0.8728062
Población	0.5434321	0.1673186



Reducción de dimensiones con PCA

- Con los vectores de pesos ϕ de cada variable para cada componente se realiza la transformación de los datos aplicando la siguiente operación matricial:

$$Z = \phi^T X$$

- La tabla contiene los datos reducidos a 2 dimensiones.
- Los valores se podrían interpretar de la siguiente forma:
  - PC1: valor positivos altos indican mayor tasa de arrestos por delitos graves, como es el caso de florida y california.
  - PC2: valor positivos altos indican mayor nivel de urbanización como es el caso de Massachusetts.
- Cabe resaltar que antes de aplicar esta transformación se recomienda normalizar los datos de la matriz X de la siguiente manera:

$$X_j^{Scale} = \frac{X_j - \mu_j}{\sigma_j}$$

- De esta manera se evita el efecto de la escala de las variables en los resultados de los componentes principales.

Reducción de Dimensiones

Estado	PC1	PC2
Alabama	0.985566	-0.156267
Alaska	1.950138	0.438583
Arizona	1.763164	0.834653
Arkansas	-0.141420	0.182811
California	2.523980	0.341996
Colorado	1.514563	-0.001465
Connecticut	-1.358647	0.118469
Massachusetts	-0.486166	1.474496
Florida	3.013042	0.096285
…	…	…

En esta transformación la componente 2 se multiplicado por -1 para obtener la componente independiente y facilitar su interpretación

Biplot

¿Qué es un biplot?

Un biplot es una representación gráfica de los datos que combina dos componentes principales con las variables originales.

¿Para qué sirve un biplot?

- Visualizar la estructura de los datos
- Identificar grupos de datos
- Explorar la relación entre las variables

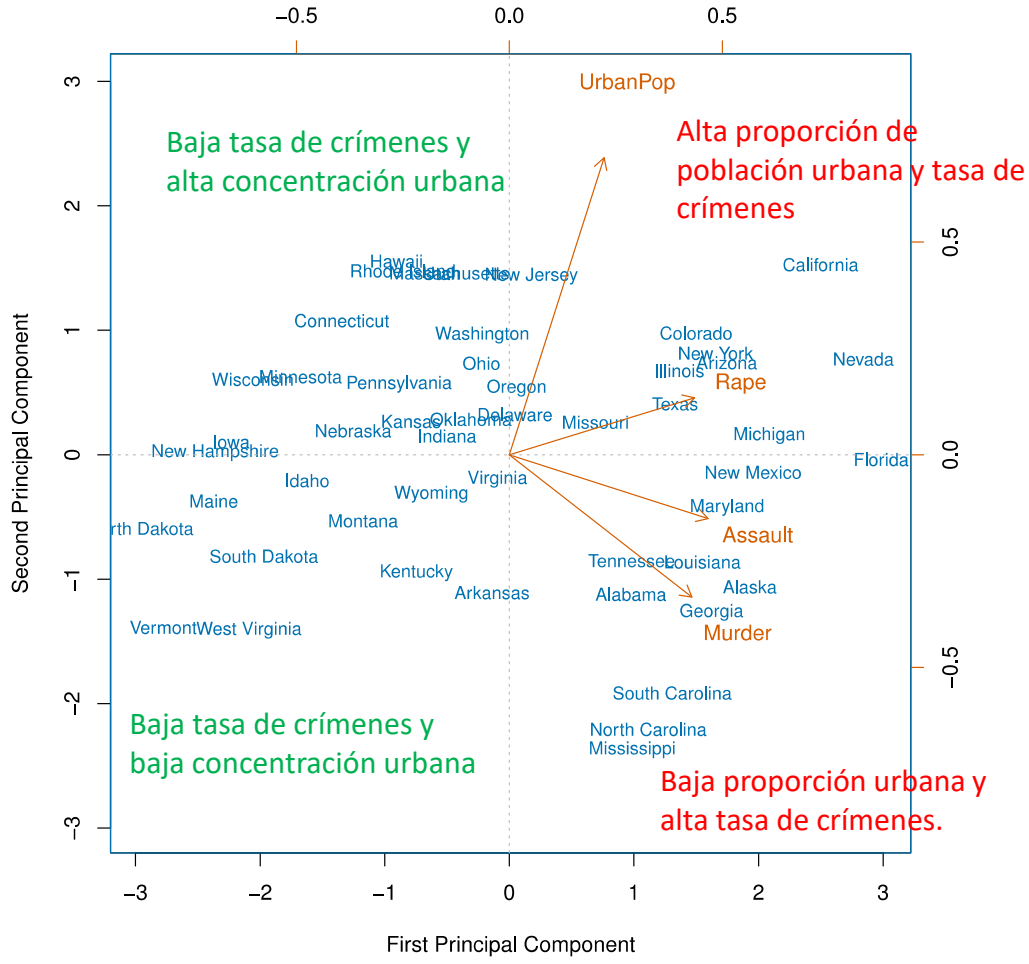
¿Cómo se interpreta un biplot?

En un biplot, los puntos representan los casos (observaciones) y las flechas representan las variables.

- La distancia entre un punto y el origen del biplot representa la magnitud de la contribución de ese caso a la CP representada por el eje vertical.
- La longitud de una flecha representa la importancia de la variable para la CP representada por el eje horizontal.

¿Cómo se interpreta el biplot para que el caso de los arrestos?

- Estas tres variables están correlacionadas en la misma dirección. Por tanto, la primera componente puede indicar el *"nivel de criminalidad"* de un estado.
- Cada cuadrante representa la combinación entre un nivel de arresto y un nivel de urbanización



Pontificia Universidad JAVERIANA Bogotá

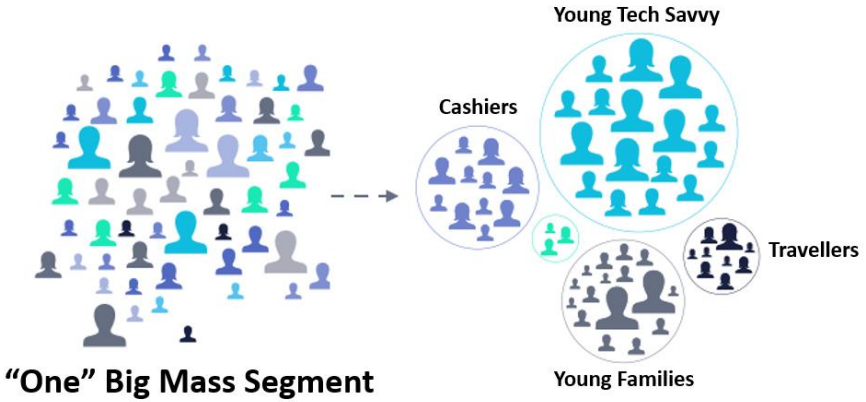
3.1 Métodos de Cluster y Anomalías

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.

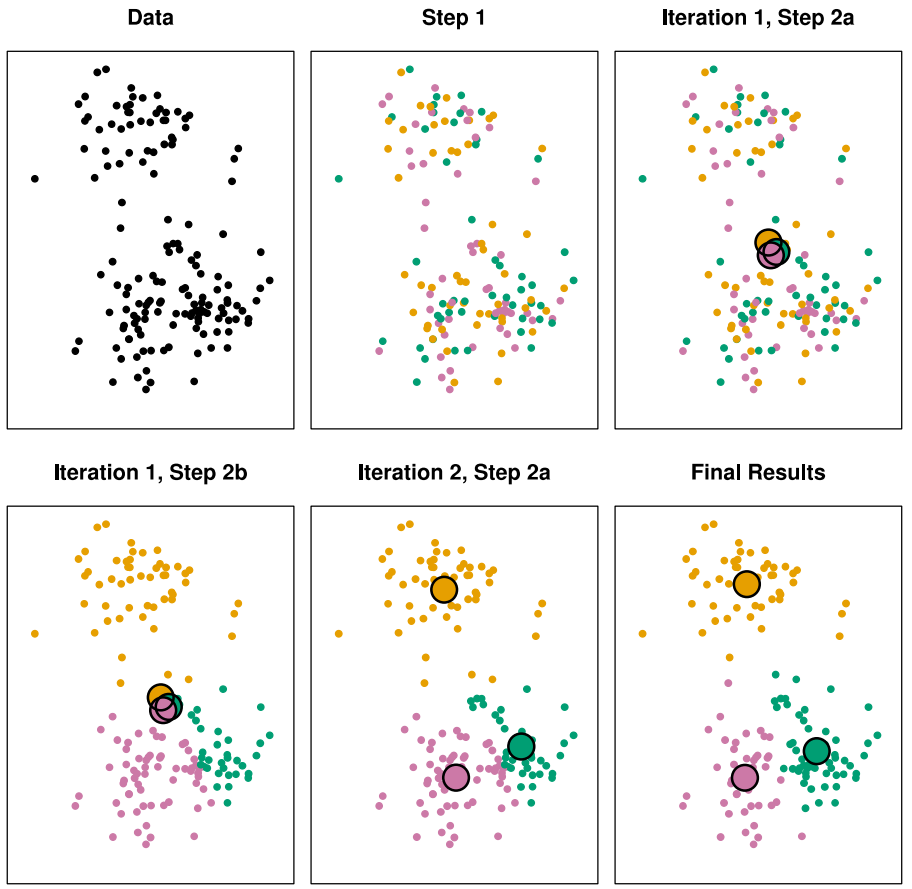
Método de K-medias (Kmeans)

¿En que consiste?

- El método KMeans es un método de agrupación en clústeres no supervisado.
- Se utiliza para dividir un conjunto de datos en k grupos, de modo que los datos dentro de cada grupo sean lo más similares posible y los datos entre grupos sean lo más diferentes posible.

Pasos del algoritmo:

1. El usuario debe especificar el número de grupos deseados, k.
2. Se seleccionan k puntos aleatorios como centroides iniciales.
3. Cada punto de datos se asigna al centroide más cercano.
4. Los centroides se actualizan para que sean el promedio de los puntos de datos asignados a ellos.
5. Los pasos 3 y 4 se repiten hasta que los centroides no cambien o hasta que se alcance un límite de iteraciones.



Para determinar la distancia entre los puntos y de los puntos a los centroides se utilizan las medidas de distancias vistas con los algoritmos de los k vecinos

Número de Clusters

El método de codo es una forma de elegir el número óptimo de clústeres en un conjunto de datos.

¿Cómo funciona?

- El método de codo se basa en la idea de que, a medida que aumenta el número de clústeres, la suma de cuadrados de las distancias entre los puntos y sus centroides disminuye.
- Sin embargo, a un cierto punto, la disminución de la suma de cuadrados se ralentiza. Este punto se conoce como el codo.

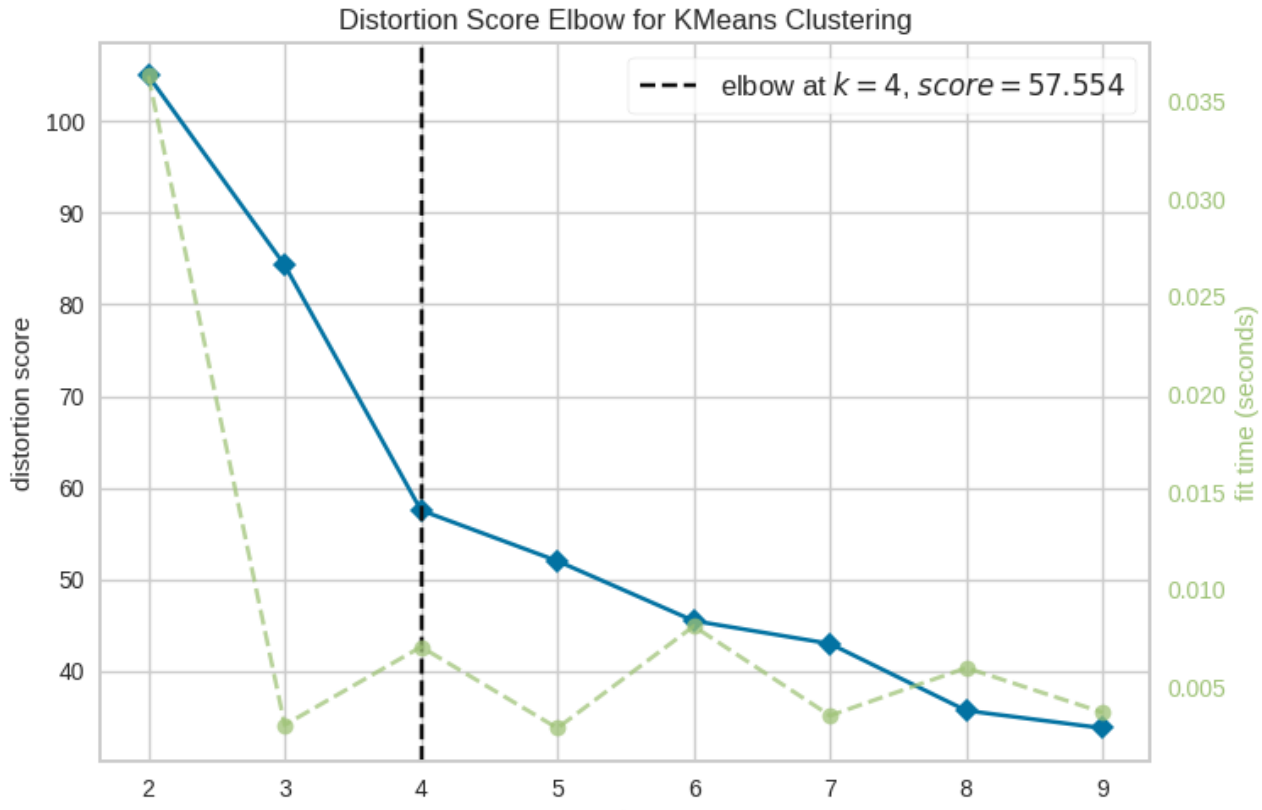
¿Cómo se usa?

- Para utilizar el método de codo, primero debemos calcular la suma de cuadrados de las distancias entre los puntos y sus centroides para un rango de valores de k.
- Luego, podemos trazar un gráfico de la suma de cuadrados en función de k. El codo del gráfico nos indicará el número óptimo de clústeres.

¿Qué significa el codo?

- El codo representa el punto en el que la disminución de la suma de cuadrados se ralentiza.

Gráfico de Codo para el caso de Arrestos en US

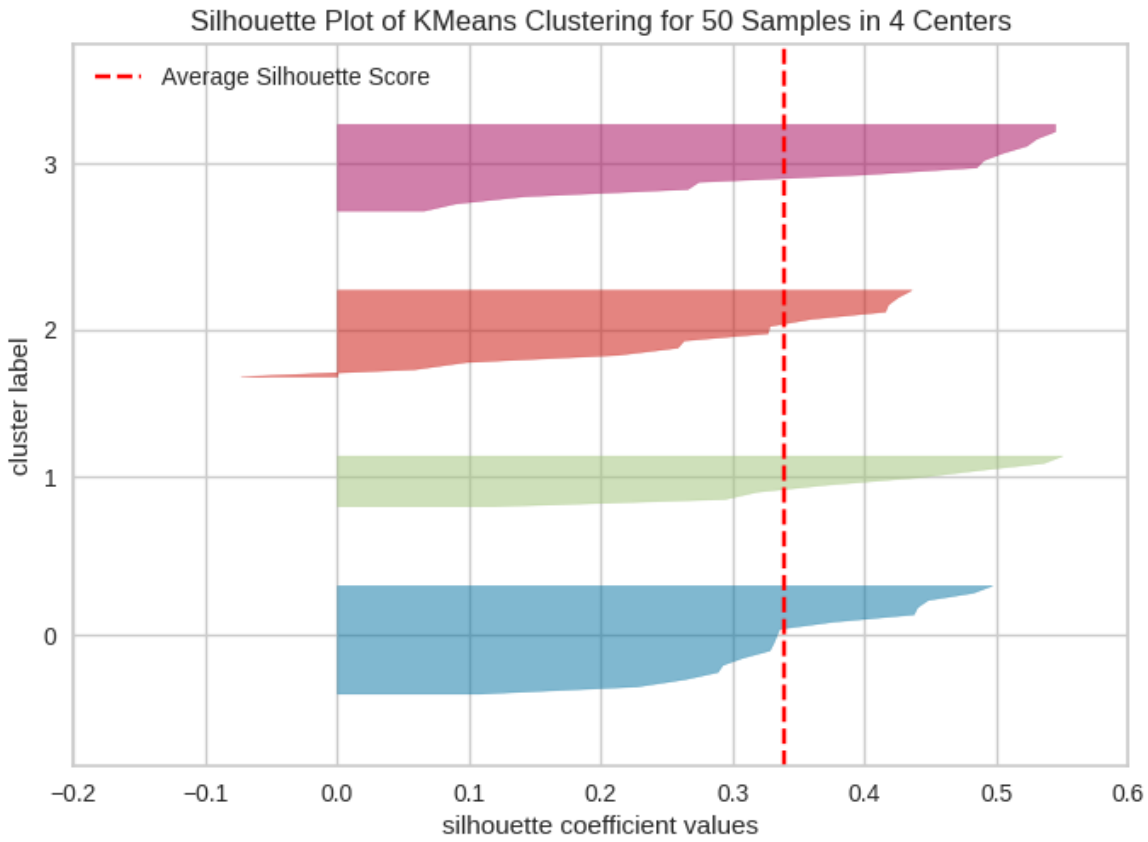


Observe que en 4 clusters la métrica comienza de disminuir de forma mas lenta

Gráfico de Silueta

- El gráfico de silueta se utiliza para evaluar la calidad de los clústeres generados por un algoritmo de clustering.
- El coeficiente de silueta se calcula como la diferencia entre la distancia promedio dentro del grupo y la distancia media del grupo más cercano para cada muestra.
- Un coeficiente de silueta de 1 indica que una muestra está perfectamente asignada a su clúster.
- Los clústeres con coeficientes de silueta más altos son los más densos y bien definidos.
- Para el caso de los arrestos se puede observar que la mayoría de estados están bien representados en los clusters.
- Solo en el cluster 2 existen algunos estados que que no están lo suficientemente cerca del cluster.

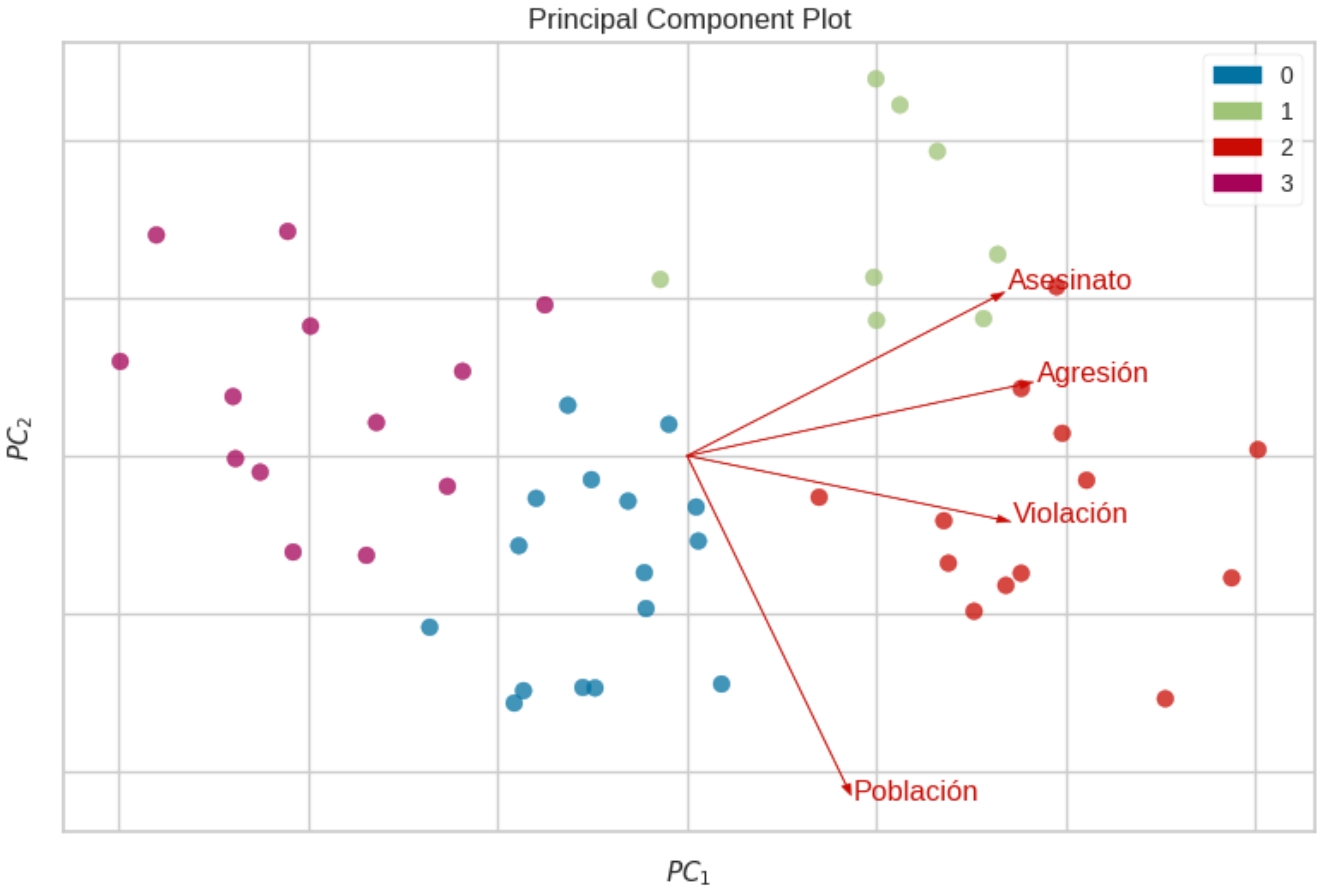
Gráfica de Silueta para el Caso de Arrestos en US



PCA y Cluster

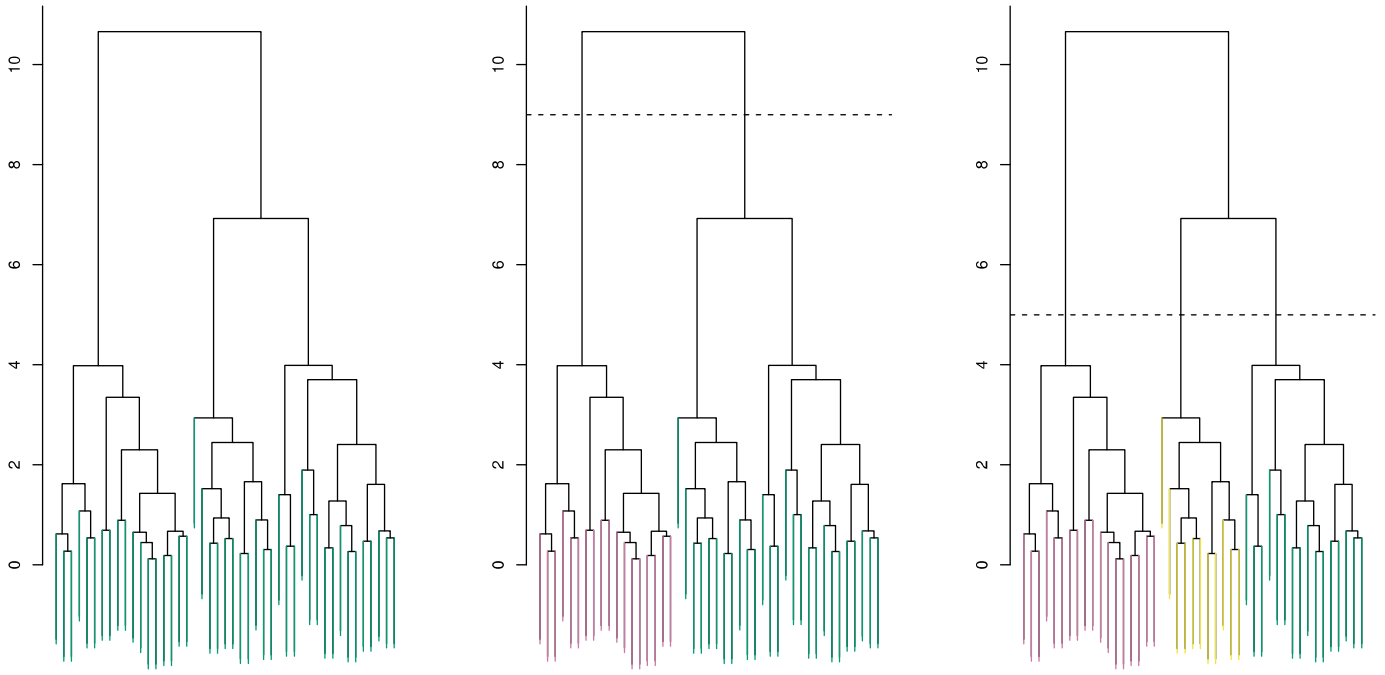
- La gráfica de biplot es una herramienta útil para analizar la composición de los cluster.
- Permite visualizar la relación entre las variables y los casos, y puede ayudar a identificar grupos de casos que son similares entre sí.
- Al combinar el análisis de cluster con PCA podemos analizar como las variables aportan en la agrupación de las instancias.

Visualización de los Cluster del Caso de Arrestos en Estado de US



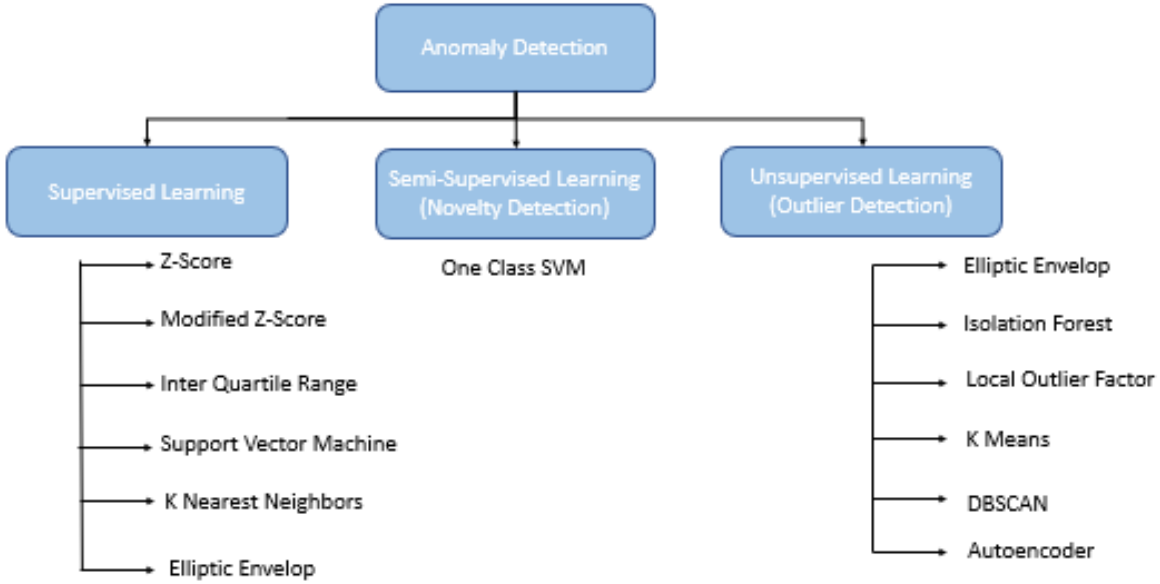
Agrupamiento Jerárquico (Hierarchical Clustering)

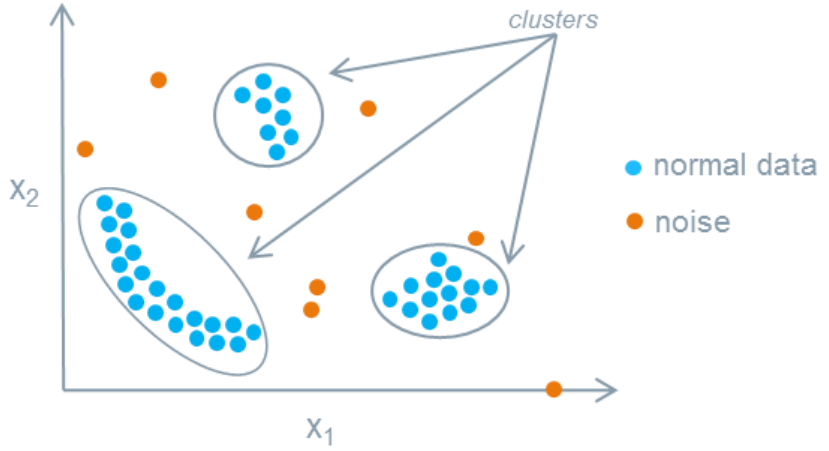
El agrupamiento jerárquico es un enfoque alternativo que no requiere que nos comprometamos con una elección particular de K. El agrupamiento jerárquico da como resultado una atractiva representación basada en árboles de las observaciones, llamada dendrograma.



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 observaciones.





Pontificia Universidad JAVERIANA Bogotá

3.3 Reglas de Asociación

Educación **Continua**

Generamos experiencias educativas

Reglas de Asociación

- Se utiliza para descubrir relaciones entre atributos en un conjunto de datos.
- Estas relaciones se pueden utilizar para identificar patrones en los datos, tomar decisiones y mejorar la toma de decisiones.
- 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 o consumo
  - Diseñar promociones
  - Desarrollar estrategias de localización de productos

ID	Elementos
1	{Pan, Leche}
2	{Pan, **Pañales**, **Cerveza**, Huevos}
3	{Leche, **Pañales**, **Cerveza**, Gaseosa}
4	{Pan, Leche, **Pañales**, **Cerveza**}
5	{Pan, Leche, Pañales, Gaseosa}
…	…

Transacciones

{Pañales, Cerveza} Ejemplo de un conjunto frecuente

{Pañales} -> {Cerveza} Ejemplo de una regla de asociación

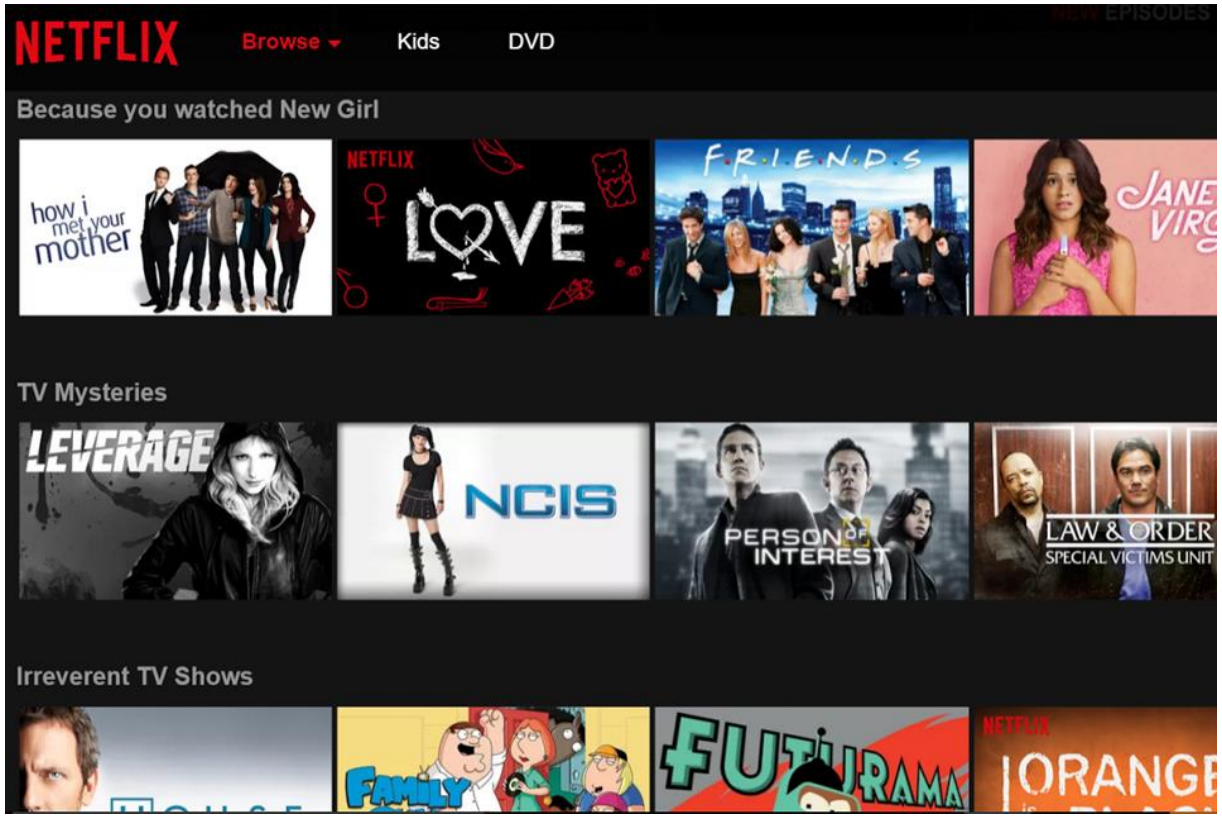
Algunas Aplicaciones

Pontificia Universidad JAVERIANA Bogotá

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



Aplicaciones

- Análisis de Mercado y Compra Conjunta: Analizar patrones de compra. Los minoristas pueden identificar productos que a menudo se compran juntos, lo que les permite mejorar la disposición de la tienda, el cross-selling y la comercialización dirigida.
- Recomendaciones en E-commerce: pproporcionar recomendaciones de productos a los usuarios.
- Salud: identificar posibles condiciones médicas, resultados de tratamientos e interacciones de medicamentos.
- Detección de Fraudes: Ayuda a identificar patrones de comportamiento inusuales y posibles casos de fraude.
- Minería de Uso de Sitios Web: comprender cómo los usuarios navegan por sus sitios. Estos datos se pueden utilizar para mejorar la disposición del sitio web, la experiencia del usuario y las recomendaciones de contenido.
- Manufactura y Control de Calidad: identificar factores que llevan a defectos o problemas de calidad del producto

Aplicaciones

- Telecomunicaciones: Analizar registros de llamadas y comprender patrones de llamadas, detectar comportamientos inusuales (como llamadas fraudulentas) y optimizar el tráfico de la red.
- Gestión de la Cadena de Suministro: identificar relaciones y dependencias en la cadena de suministro, lo que puede llevar a una mejor gestión de inventarios y reducción de costos.
- Minería de Texto: analizar y descubrir patrones en grandes corpus de texto, lo que ayuda en la recuperación de información, el análisis de sentimientos y las recomendaciones de contenido.
- Análisis de Redes Sociales: sugerir amigos, conexiones o grupos basados en los intereses, actividades y conexiones de los usuarios.
- Análisis Ambiental: analizar la relación en los datos ambientales, como patrones climáticos, niveles de contaminantes y su impacto en los ecosistemas.
- Farmacovigilancia: detectar posibles efectos secundarios e interacciones medicamentos mediante el análisis de registros de pacientes y reportes de eventos adversos.

Conceptos Básicos

Definiciones:

- **Conjunto de transacciones:** Una colección de registros que describen las compras de los clientes.
- **Conjunto de atributos:** Una lista de todos los productos que se pueden comprar.
- **Regla de asociación:** Una implicación de la forma A ⇒ B, donde A y B son conjuntos de atributos.

Ejemplos:

- **Conjunto de transacciones:**

  [pan, mantequilla, leche]

  [pan, huevos, leche]

  [carne, queso]

- **Conjunto de atributos:**

  [pan, mantequilla, leche, huevos, carne, queso]

- **Regla de asociación:**

  [pan, mantequilla] ⇒ [leche]

Soporte (Support)

- El soporte es una medida que indica la frecuencia con la que un conjunto de elementos (o ítems) aparece en el conjunto de datos.
- Se calcula dividiendo el número de transacciones en las que aparece el conjunto de elementos por el número total de transacciones en el conjunto de datos.

$$Support(A) = \frac{frecuencia(A)}{N}$$

- Donde, A es el conjunto de ítem que se desea analizar y N el total de registros o transacciones.

ID	Elementos
1	{Pan, Leche}
2	{Pan, **Pañales**, **Cerveza**, Huevos}
3	{Leche, **Pañales**, **Cerveza**, Gaseosa}
4	{Pan, Leche, **Pañales**, **Cerveza**}
5	{Pan, Leche, Pañales, Gaseosa}

- Un alto valor de soporte indica que el conjunto de elementos es común en el conjunto de datos, mientras que un valor bajo sugiere que es raro.
- Ejemplo: para las transacciones que se muestran en la tabla los pañales tienen un soporte igual a:

$$Frecuencia(Pañales) = 4$$
$$N = 5$$
$$Support(Pañales) = \frac{4}{5} = 0.8$$

- Lo que indica que en el 80% de las transacciones se compran pañales.
- Ahora calcule el soporte para: X = Gaseosa, X = Leche y Pan

Confianza (Confidence)

- La confianza mide la probabilidad de que un conjunto de elementos B ocurra en una transacción dado que otro conjunto de elementos A ya ha ocurrido en la misma transacción.
- Se calcula dividiendo el número de transacciones en las que aparecen tanto A como B por el número de transacciones en las que aparece A.

$$Confidence(A,B) = \frac{Support(A,B)}{Support(A)}$$

ID	Elementos
1	{Pan, Leche}
2	{Pan, **Pañales**, **Cerveza**, Huevos}
3	{Leche, **Pañales**, **Cerveza**, Gaseosa}
4	{Pan, Leche, **Pañales**, **Cerveza**}
5	{Pan, Leche, Pañales, Gaseosa}

- Una alta confianza indica que la ocurrencia de A es un buen indicador de la ocurrencia de B.
- Ejemplo: Calcule la confianza de que se compre Cerveza dado que se han comprado pañales:

$$\text{Support}(Pañales) = \frac{4}{5} = 0.8$$

$$\text{Support}(Pañales, Cerveza) = \frac{3}{5} = 0.6$$

$$Support(Pañales) = \frac{0.6}{0.8} = 0.75$$

- Lo que indica que en el 75% de las transacciones se espera que se compre cerveza dado que se compran pañales.
- Ahora calcule la confianza para las reglas: (A: Pan, B: Leche), y (A: Pan;Leche, B:Huevo)

Elevación (lift)

- Es una medida del desempeño de una regla de asociación para predecir o clasificar casos que tienen una respuesta mejorada en comparación con la población en su conjunto, en contraste con una regla de asociación aleatoria
- Se calcula dividiendo el soporte de la regla (A,B) entre el soporte de cada ítem por separado.

$$Lift(A,B) = \frac{Support(A,B)}{Support(A) \times Support(B)}$$

ID	Elementos
1	{Pan, Leche}
2	{Pan, **Pañales**, **Cerveza**, Huevos}
3	{Leche, **Pañales**, **Cerveza**, Gaseosa}
4	{Pan, Leche, **Pañales**, **Cerveza**}
5	{Pan, Leche, Pañales, Gaseosa}

- Se interpreta de la siguiente manera:
  - Lift de 1: Independencia entre A y B. No se extraen reglas significativas.
  - Lift > 1: Dependencia entre A y B. Reglas útiles para asociación.
  - Lift < 1: A y B son elementos sustitutivos. La presencia de uno afecta negativamente al otro.

Ahora calcule el lift para las reglas: (A: Pan, B: Leche), y (A: Pan;Leche, B:Huevo)

- Calcule el lift de la regla A: pañales, B: cerveza:

$$\text{Support}(Pañales) = \frac{4}{5} = 0.8$$

$$\text{Support}(\text{Cerverza}) = \frac{3}{5} = 0.6$$

$$\text{Support}(Pañales, Cerveza) = \frac{3}{5} = 0.6$$

$$Support(Pañales) = \frac{0.6}{(0.8)(0.6)} = 1.25$$

Lift > 1, por tanto, la regla de asociación es útil.

Generación de Reglas – Algoritmo Apriori

El algoritmo Apriori es un algoritmo ampliamente utilizado en la minería de datos para encontrar reglas de asociación en un conjunto de datos.

Fases del algoritmo Apriori:

1. Fase de Generación de Conjuntos de Items Candidatos: comienza identificando todos los elementos individuales (ítems) que aparecen con suficiente frecuencia en el conjunto de datos. Luego, genera conjuntos candidatos de ítems más grandes a partir de estos elementos individuales.
2. Fase de Podado de Conjuntos Candidatos: filtra los conjuntos candidatos para eliminar aquellos que no cumplen con un nivel mínimo de soporte predefinido. Los conjuntos que no cumplen con el nivel mínimo de soporte se eliminan.
3. Fase de Generación de Reglas de Asociación: Con los conjuntos candidatos que cumplen con el nivel de soporte mínimo, el algoritmo Apriori genera reglas de asociación.
4. Repetición del Proceso: El algoritmo Apriori repite las fases anteriores, incrementando gradualmente el tamaño de los conjuntos candidatos hasta que ya no se pueden generar reglas adicionales que cumplan con el nivel de soporte mínimo.