R-Ladies Lima

Clustering

MBA Jacquelin Flor

Presentación

- Profesional de Ingeniería Estadística de la Universidad Nacional de Ingeniería, con especializaciones en Marketing Relacional en la UPC y Finanzas Corporativas en la UP y una maestría en Administración de Negocios en el IE Business School (Madrid, España). Candidata a Máster en Inteligencia Artificial en la Universidad de la Rioja. Asimismo, cuento con más de trece años de experiencia en temas relacionados con gestión de información, desarrollo e implementación de modelos predictivos, segmentación y desarrollo de estrategias orientadas a marketing, recursos humanos y riesgos en sectores como telecomunicaciones, banca y micro-finanzas

- Docente en la Universidad de Piura (pre-grado, post-grado) , DMC Perú.

- Actualmente me desempeño como Gerente de Smart Data en Valtx.

Agenda

1. Introducción
2. Clustering
3. Casos de uso

Introducción

Introducción

1. R es un lenguaje de programación que permite realizar comandos e implementar técnicas estadísticas en un entorno interactivo para el análisis estadístico y gráfico.

2. Es un lenguaje de programación con funciones orientadas a objetos.

3. R fue inicialmente diseñado por ***Robert Gentleman y Ross Ihaka (1993)***, miembros del Departamento de Estadística de la Universidad de Auckland, en Nueva Zelanda.

¿Por qué usar R?

¿Por qué usar R?

- La sintaxis es simple e intuitiva.

```
#Creando vectores.

> x<-c(10.4,5.6,3.1,6.4,21.7)
> x
[1] 10.4  5.6  3.1  6.4 21.7

> w<-c("rojo","verde","azul")
> w
[1] "rojo"  "verde" "azul"
```

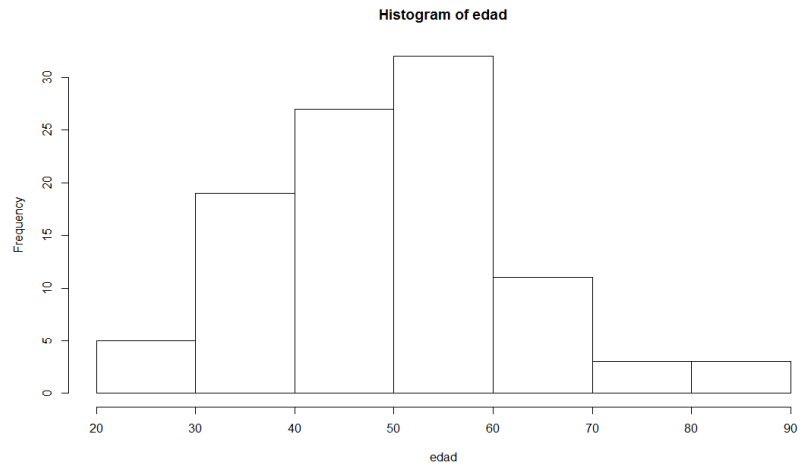
```
#Creando objetos y realizando cálculos aritméticos.

> a=10
> b=25
> c=25
> y=a+b+c
> y
[1] 60
```

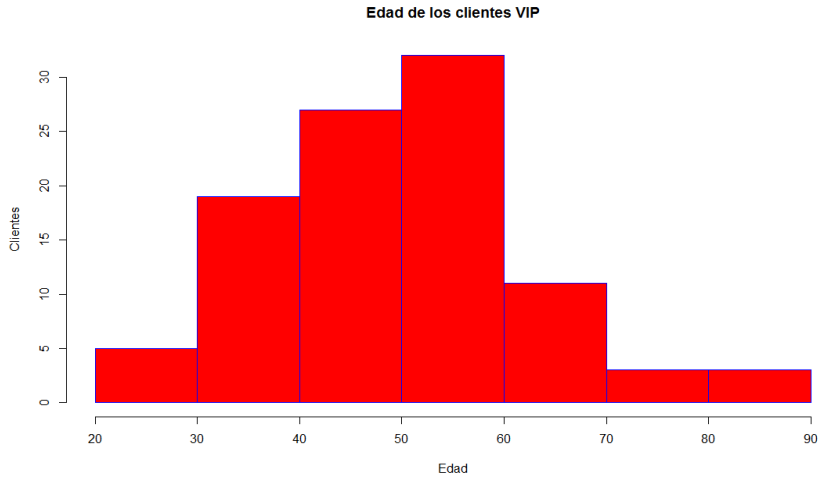
¿Por qué usar R?

- La estructura y facilidad de uso de R nos permite implementar nuestras propias funciones y rutinas a medida que aparecen nuestras necesidades.

```
>hist(edad)
```



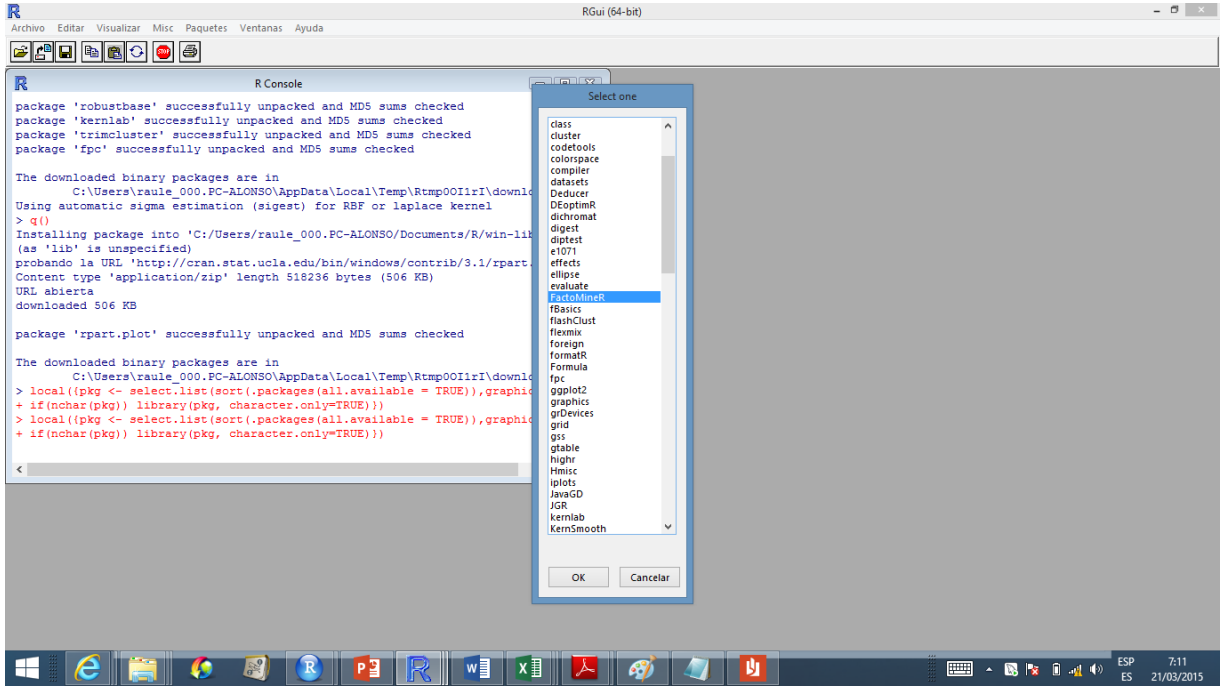

```
>hist(edad,col="red", xlab="Edad",ylab="Clientes",
main="Edad de los clientes VIP",border="blue")
```



¿Por qué usar R?

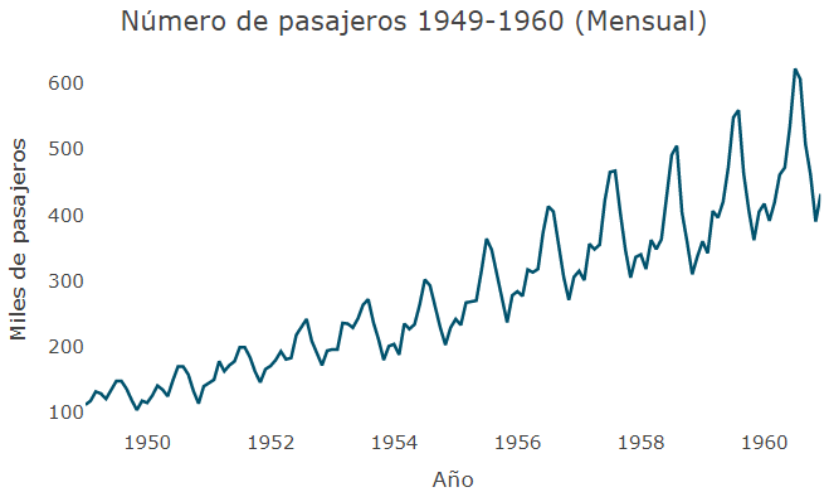
- La distribución de R viene acompañada de un numeroso conjunto de librerías base.
- Asimismo, es posible añadir librerías adicionales.

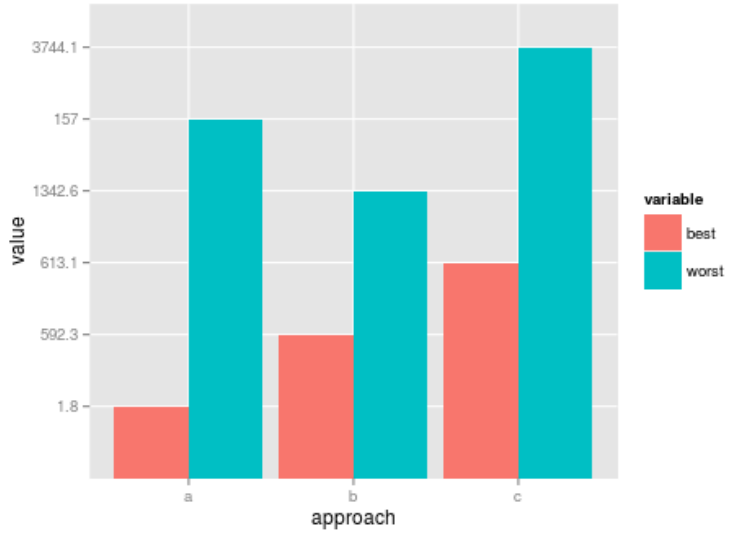
Librerías propias y librerías adicionales como el FactoMineR

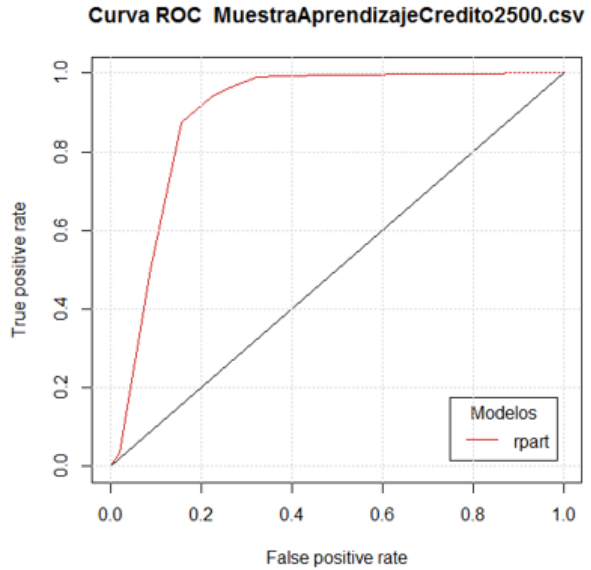


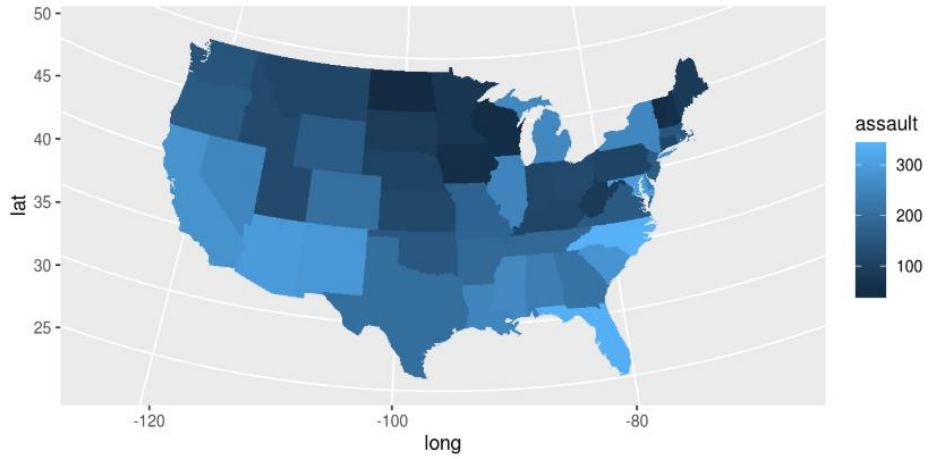
¿Por qué usar R?

- Gran variedad de librerías gráficas.



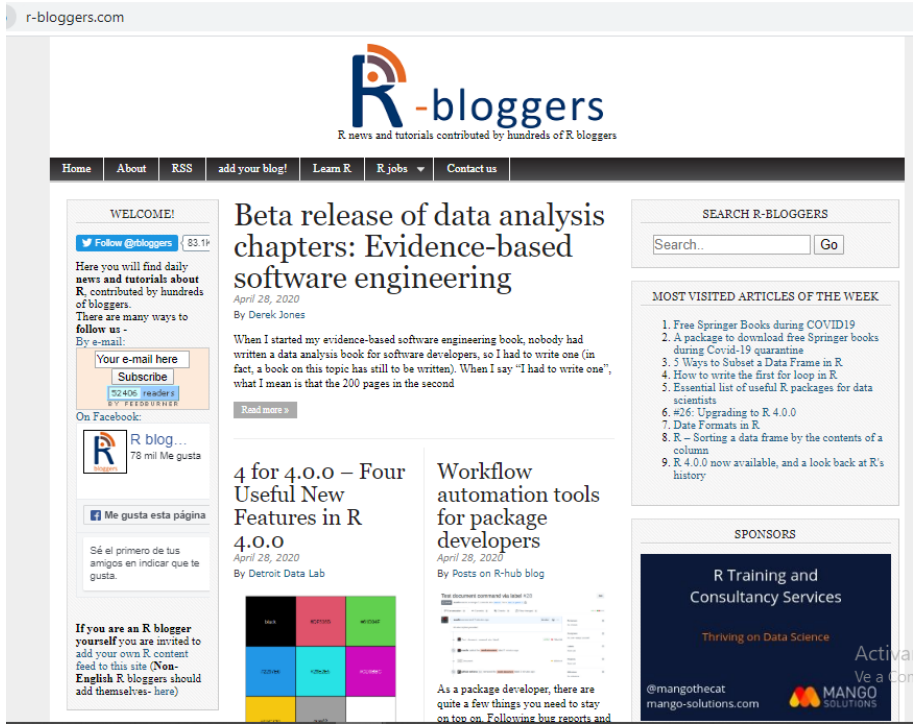






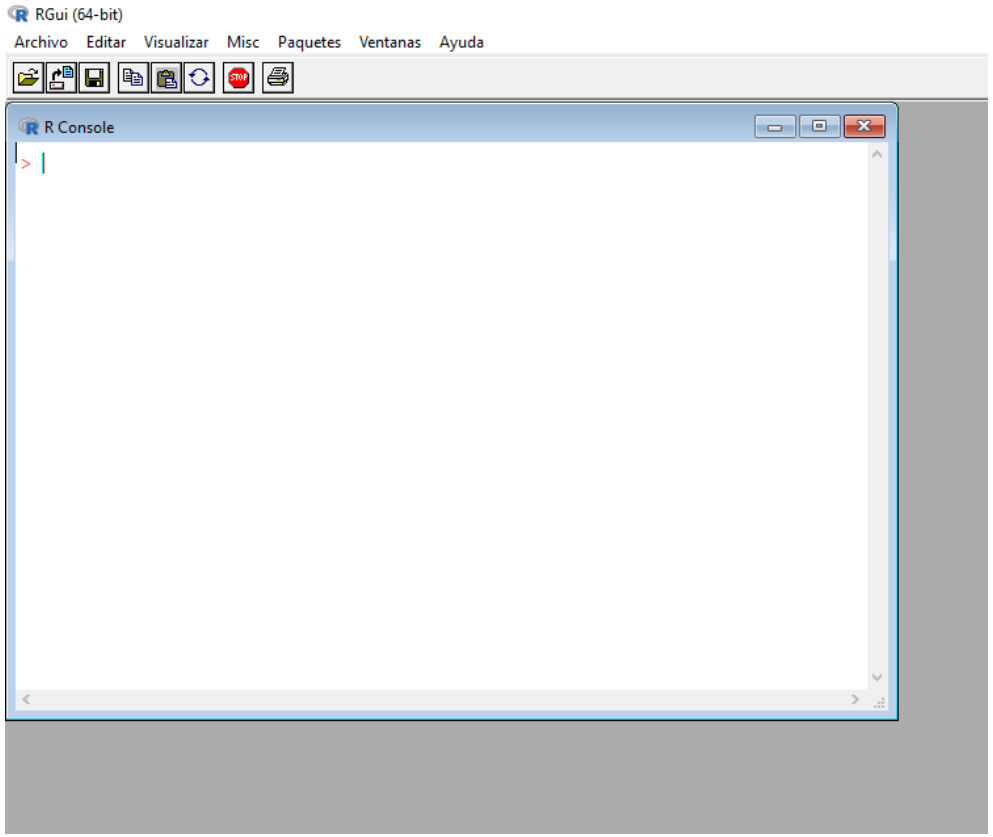
¿Por qué usar R?

- Gran red de apoyo y soporte disponible en foros, blogs, Facebook, etc.
- Por ejemplo: https://www.r-bloggers.com/



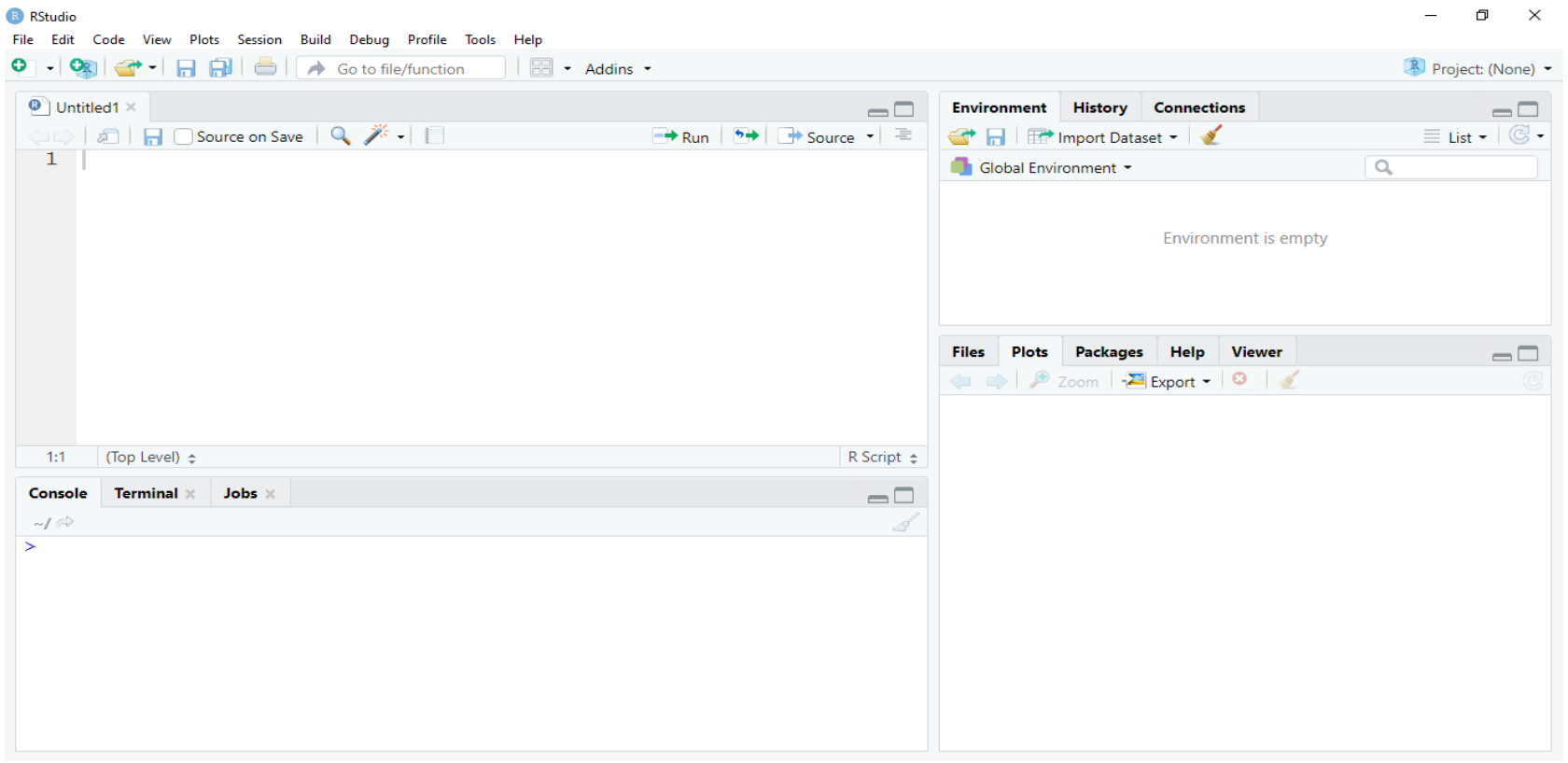
Reseñas básicas

- Consola de R.



Reseñas básicas

- Consola de Rstudio.



Reseñas básicas

- El lenguaje R distingue las mayúsculas de las minúsculas.
- El símbolo " . " es el que separa la parte entera de la decimal en R y no el símbolo " , " .
- Se utiliza el símbolo # para hacer comentarios dentro de un script.
- Las teclas **Ctrl + l** limpian la consola.
- Una orden consiste en una expresión que se evalúa, imprime y su valor se pierde.
- Una asignación, por el contrario, evalúa una expresión, no la imprime y guarda su valor en una variable.

```
> 2+3
[1] 5

> g<-c(1,2,3,4)
> g
[1] 1 2 3 4
```

Reseñas básicas

- Generación de sucesiones.

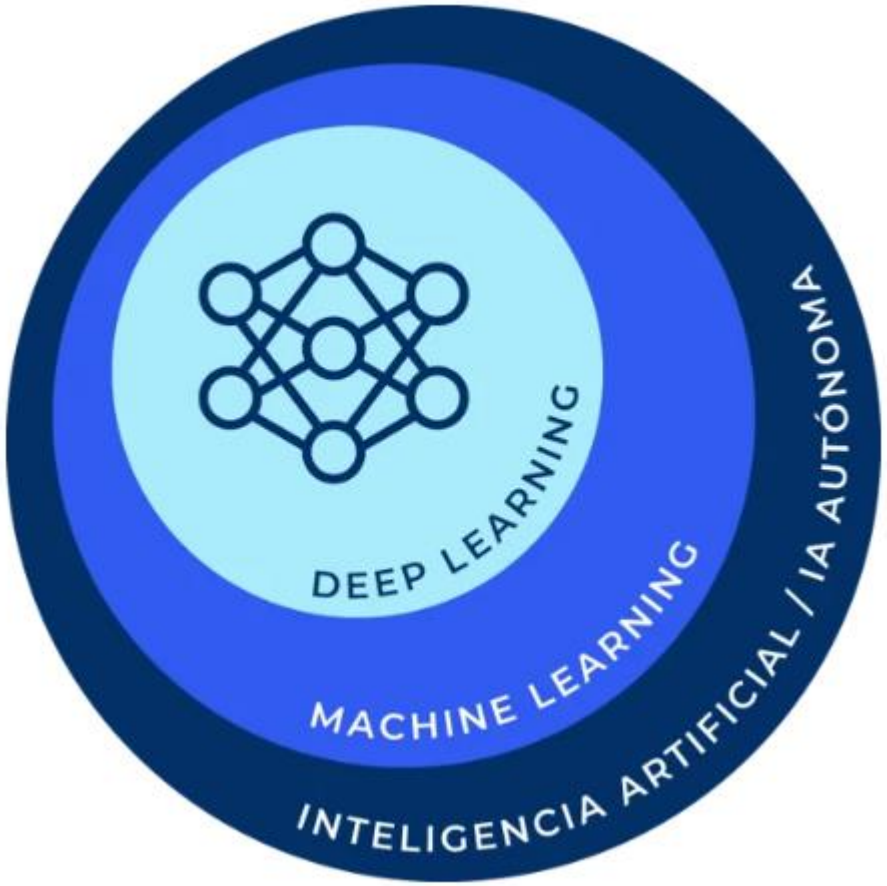
```
#Generación de una sucesión del 1 al 20.
> 1:20
[1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20

#Se realiza la sucesión primero y luego las operaciones aritméticas.
> 1:3*5+2
[1]  7 12 17

#Genera una secuencia que inicia en 1 y termina en 10 con elementos que van de 2 en 2.
> seq(1,10,by=2)
[1] 1 3 5 7 9
```

- Operaciones matemáticas +, -, *, /, ^
- Operadores de comparación <, ==, >, <=, >=, !=
- Operadores lógicos (and, or, not) &, |, !

Machine Learning



- ML es una rama de la inteligencia artificial que se ocupa de desarrollar algoritmos y técnicas que permiten a las computadoras aprender y tomar decisiones sin ser explícitamente programadas.
- En lugar de seguir instrucciones específicas, las máquinas aprenden a través de la experiencia y el análisis de datos.

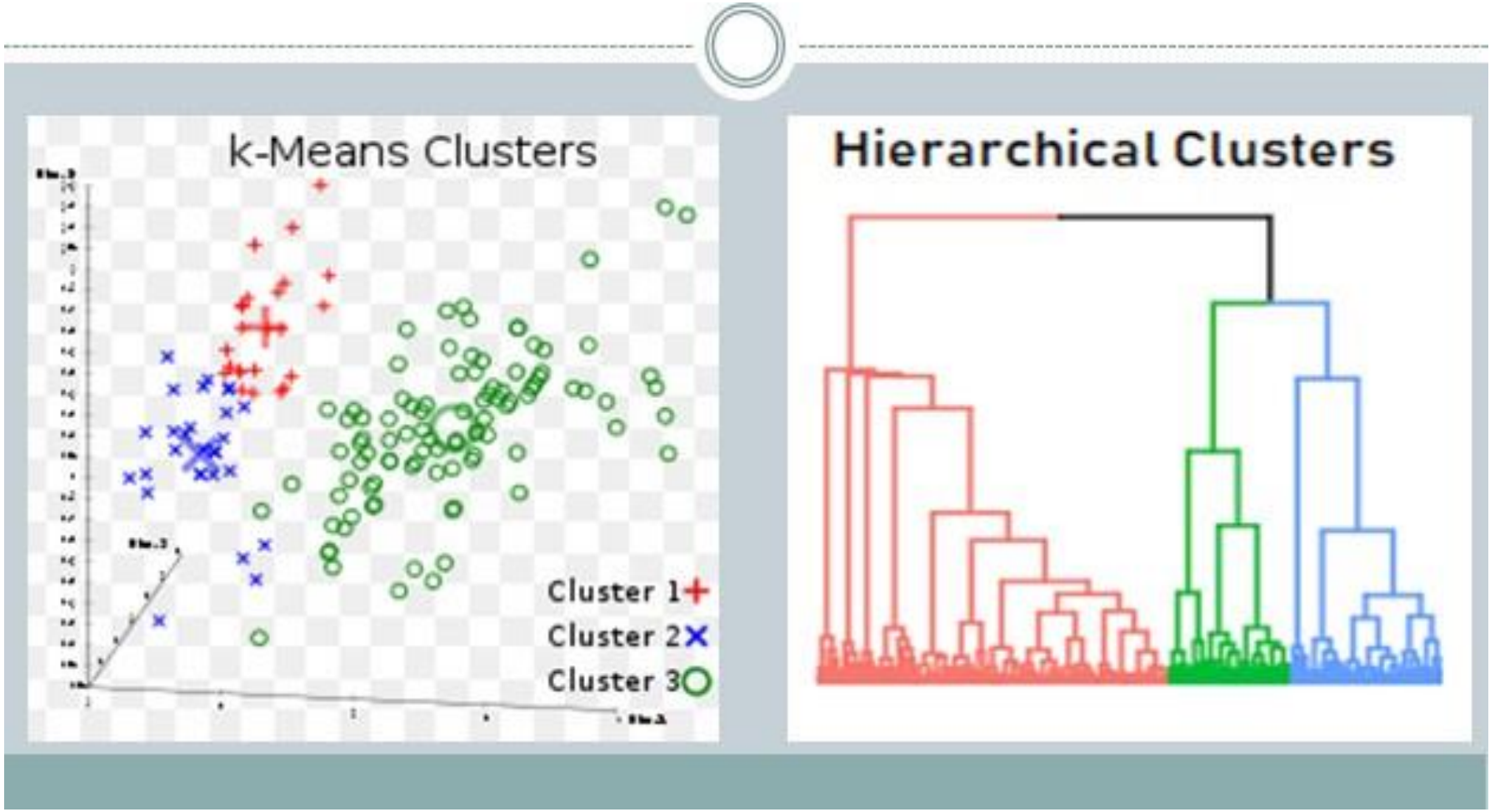
Machine Learning

Clustering

Introducción

- Clustering: (clasificación no supervisada, aprendizaje no supervizado), el objetivo es particionar o segmentar un conjunto de datos o individuos en grupos que pueden ser disjuntos o no. Los grupos se forman basados en la similitud de los datos o individuos en ciertas variables. Como los grupos no son dados a priori (por ejemplo, en una clasificación jerárquica) el experto debe dar una interpretación de los grupos que se forman.
- Clustering: también llamado Análisis de Conglomerados
- Algunos métodos:
  - Clasificación Jerárquica
  - K-means

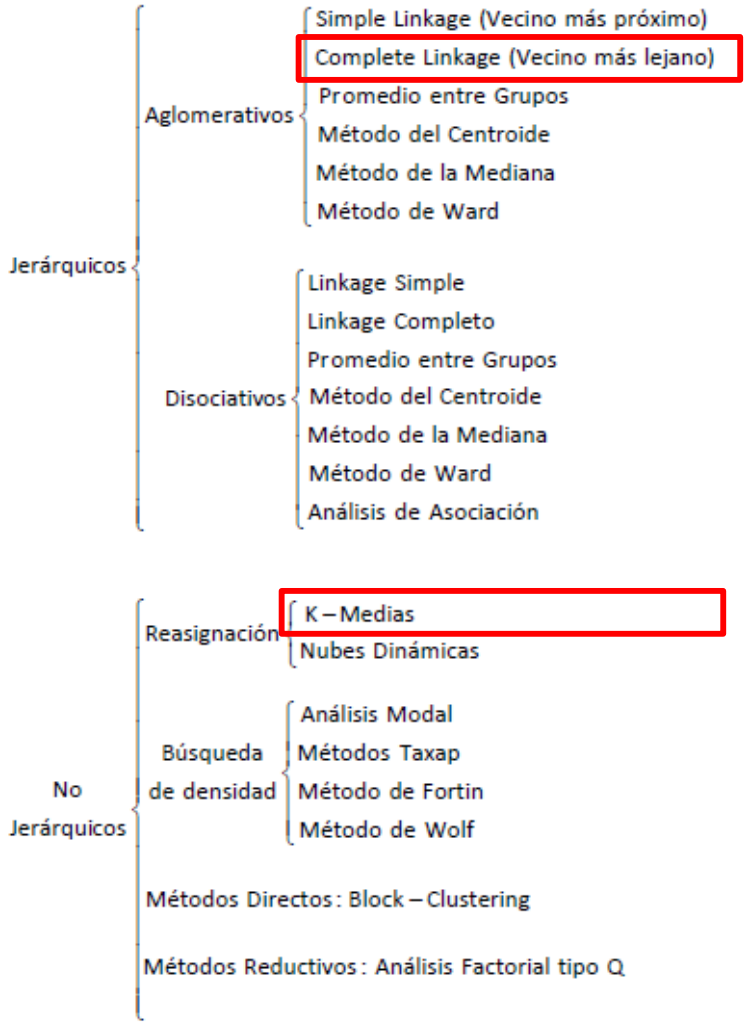
Introducción



Etapas del análisis de conglomerados

- Planteamiento del problema
- Elección de variables
- Análisis exploratorio
- Elección de la medida de asociación
- Elección de la técnica cluster
- Validación de resultados

Tipos de análisis de conglomerados



Método Jerárquico

- Los algoritmos jerárquicos son métodos que entregan una jerarquía de divisiones del conjunto de elementos en conglomerados.
- Se consideran dos tipos de métodos jerárquicos y son:

1. Método jerárquico aglomerativo

Parte con una situación en que cada observación forma un conglomerado y en sucesivos pasos se van uniendo, hasta que finalmente todas las situaciones están en un único conglomerado.

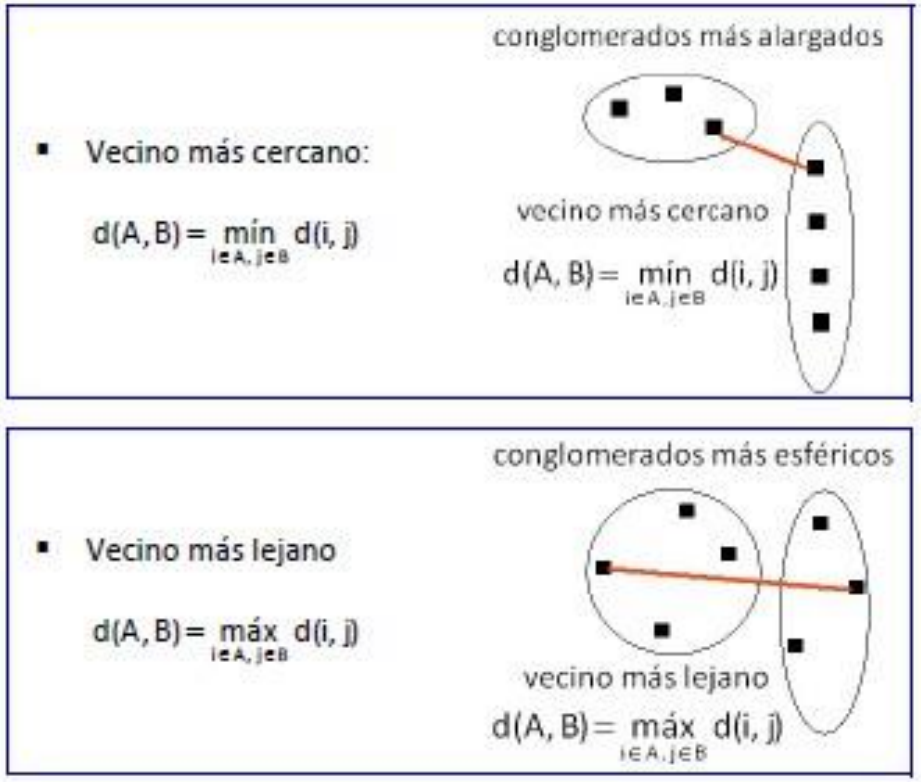
2. Método jerárquico disociativo

Sigue el sentido inverso, parte de un gran conglomerado y en pasos sucesivos se va dividiendo hasta que cada observación queda en un conglomerado diferente.

Distancia entre conglomerados

- Las distancias entre los conglomerados son funciones de las distancias entre observaciones, hay varias formas de definirlas:

A y B son dos conglomerados



Método Linkage Aglomerativo

- Conocidas las distancias o similaridades entre dos individuos, se observa cuáles son los más próximos (menor distancia o mayor similaridad); éstos dos individuos formarán un grupo que no vuelve a separase durante el proceso.
- Se repite el proceso, volviendo a medir la distancia o similaridad entre todos los individuos de la siguiente forma:

Cuando se mide la distancia entre el grupo formado y el individuo, se toma la distancia **máxima** de los individuos del grupo al nuevo individuo.

Cuando se mide la similitud entre el grupo formado y el individuo, se toma la distancia **mínima** de los individuos del grupo al nuevo individuo.

Ejemplo

- Se tienen las siguientes similaridades (coeficiente de correlación entre variables):

Distancia	A	B	C	D	E
A	1				
B	0,39	1			
C	0,75	0,24	1		
D	0,56	0,63	0,42	1	
E	0,81	0,72	0,12	0,93	1

Tabla simétrica debido a que:
d(A,B)=d(B,A)

- Similaridad máxima: s(D,E)=**0,93 → Por lo tanto, D y E forman un grupo.**
- Se miden las similaridades de nuevo:

Distancia	A	B	C	D-E
A	1			
B	0,39	1		
C	0,75	0,24	1	
D-E	0,56	0,63	0,42	1

- Similaridad máxima: s(C,A)=**0,75 → Por lo tanto A y C forman un grupo.**

Ejemplo

- Se miden las similaridades de nuevo:

Distancia	A-C	B	D-E
A-C	1		
B	0,24	1	
D-E	0,12	0,63	1

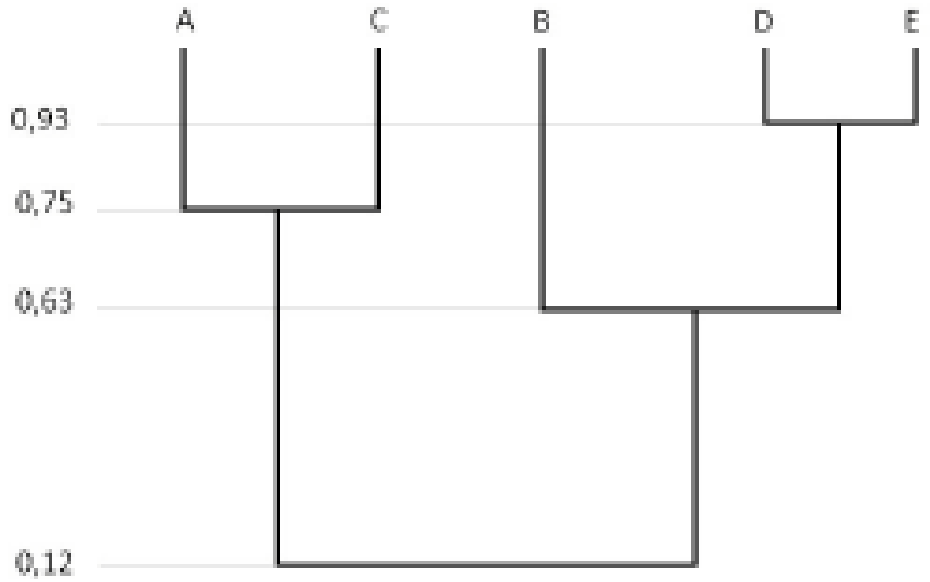
- Similaridad máxima: s(B,D-E)=**0,63 → Por lo tanto, B, D-E forman un grupo.**
- Se miden las similaridades de nuevo:

Distancia	A-C	B-D-E
A-C	1	
B-D-E	0,12	1

- Similaridad máxima: s(A-C,B-D-E)=**0,12 → Por lo tanto A-C-B-D-E forman un grupo.**

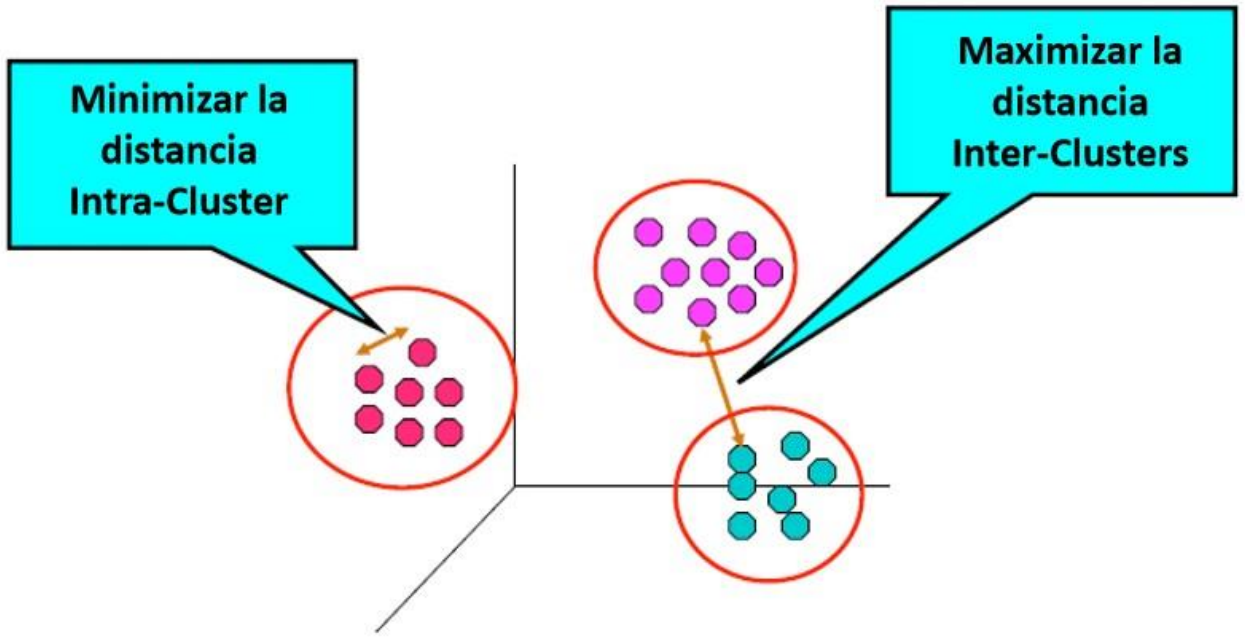
Dendograma

- Es una representación gráfica en forma de árbol que resume el proceso de agrupación en un análisis de conglomerados



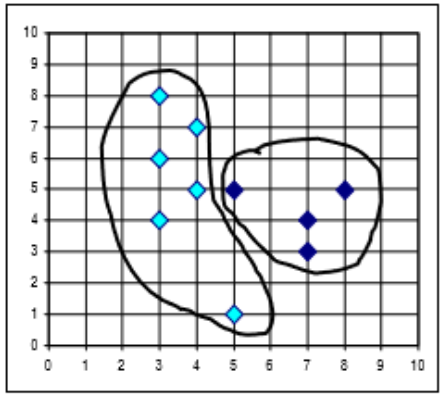
Método de K-medias

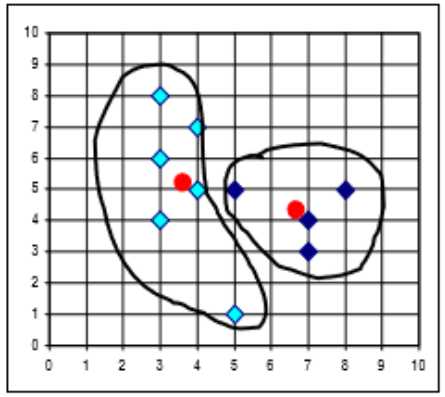
- Objetivo: Obtener la homogeneidad dentro de los grupos y la heterogeneidad entre grupos.



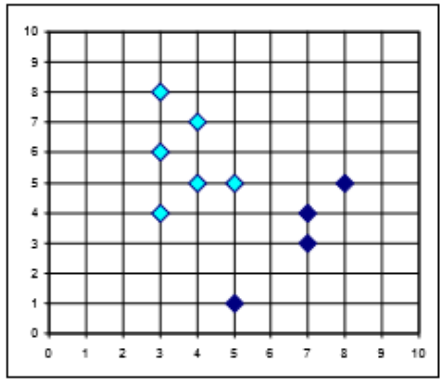
Proceso K-medias

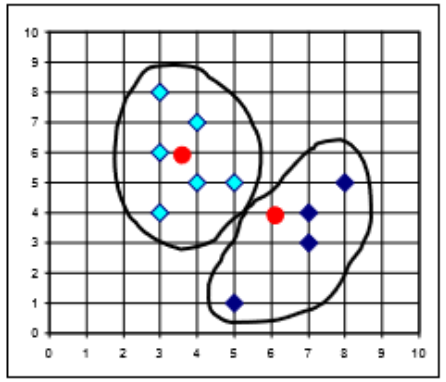
Paso 1: Para K=2
Arma al azar 2 grupos.





Paso 2:
- Luego calcula el centro de gravedad de cada cluster.
- Calcula la distancia de todos los puntos contra esos centros de gravedad y si un punto le queda más cerca de otro centro de gravedad lo cambia.





Paso 3:
Luego recalcula los centros de gravedad y se reasigna hasta que el método se Estabilice.

Ejemplo de cálculo de centro de gravedad

ID	Nombre	Matematicas	Ciencias	Espanol	Historia	EdFisica	Cluster
1	Lucia	7	6,5	9,2	8,6	8	C3
2	Pedro	7,5	9,4	7,3	7	7	C1
3	Ines	7,6	9,2	8	8	7,5	C1
4	Luis	5	6,5	6,5	7	9	C2
5	Andres	6	6	7,8	8,9	7,3	C3
6	Ana	7,8	9,6	7,7	8	6,5	C1
7	Carlos	6,3	6,4	8,2	9	7,2	C3
8	Jose	7,9	9,7	7,5	8	6	C1
9	Sonia	6	6	6,5	5,5	8,7	C2
10	Maria	6,8	7,2	8,7	9	7	C3

Promedio por variable

Centro de Gravedad Total de la Nube de Puntos

Matematicas	Ciencias	Espanol	Historia	EdFisica
6,8	7,7	7,7	7,9	7,4

Promedio por variable según cada cluster

Centro de Gravedad C1

Matematicas	Ciencias	Espanol	Historia	EdFisica
7,7	9,5	7,6	7,8	6,8

Centro de Gravedad C2

Matematicas	Ciencias	Espanol	Historia	EdFisica
5,5	6,3	6,5	6,3	8,9

Centro de Gravedad C3

Matematicas	Ciencias	Espanol	Historia	EdFisica
6,5	6,5	8,5	8,9	7,4

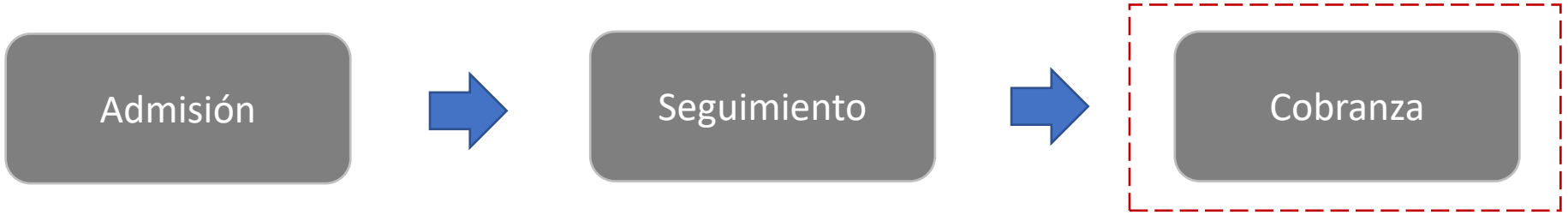
Aplicaciones

Caso: Clientes de Tarjeta de Crédito

- Se tiene un conjunto de clientes con tarjeta de crédito de un banco top en el Perú.
- Se busca segmentar a los tarjeta-habientes de acuerdo a diversas variables como: antigüedad, ticket promedio y edad.
- Una vez segmentado el conjunto de clientes, describir a los grupos encontrados.

Caso: Cobranzas

Etapas Gestión de Riesgos



Componentes e indicadores claves

- Score de admisión (buró)
- Sobre-endeudamiento
- Calificación SBS
- Variables de admisión (políticas de crédito)

- Indicadores mora, mora real
- Cosechas
- Provisiones
- Indicadores globales: RCG, Costo de riesgo
- Segmentación PD

- Indicadores efectividades por tramos de atraso
- Castigos
- Segmentación

Caso: Cobranzas

Variables

- V1
- V2
- V3
- V4
- V5

Metodología

Mejor perfil de recuperación

Segmento	Nro. Créditos	%
C1	25,000	36%
C2	27,000	40%
C3	10,000	15%
C4	6,500	9%
	68,500	

Caso: Cobranzas

Matriz de Cobranzas

Grupo	Nro. Créditos por Rango Saldo Capital			Total
	< S/3,000	[S/3,000 - S/7,000>	[S/7,000 a más	
C1				
C2				
C3				
C4				
Total				

Bibliografía

Caso: Cobranzas