



Universidad de
los Andes



***“Construcción de un modelo de minería
de datos para la segmentación de clientes
en una empresa de telecomunicaciones”***

Dr. Sebastián Maldonado
Facultad de Ingeniería y Ciencias Aplicadas,
Universidad de los Andes

Gastón Rojas
Departamento de Ingeniería Industrial,
Universidad de Talca.

IX Congreso del Instituto Chileno de Investigación Operativa
Pucón, 26-29 Octubre 2011



Introducción y contexto

Problemática

Objetivos

Metodología

Selección de atributos

Métodos de segmentación

Aplicación y resultados

Conclusiones

INTRODUCCIÓN – Selección de Atributos

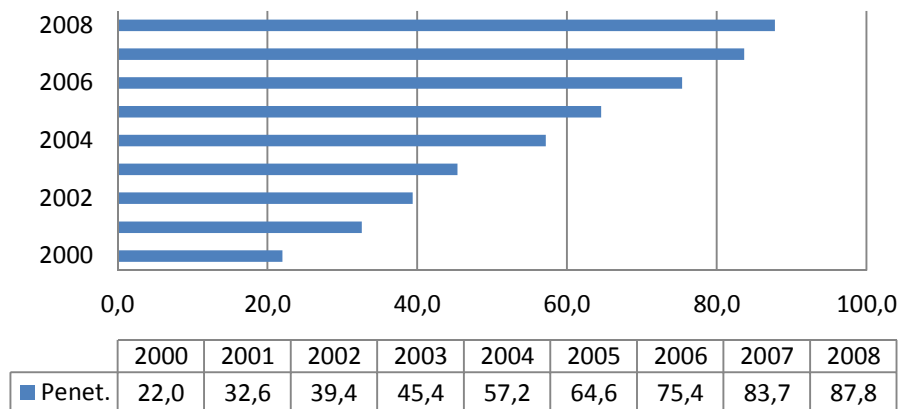
- Objetivo: elegir subconjunto de atributos relevantes, eliminando atributos que generan ruido y confunden al método.
- Ventajas: mejor representación y comprensión del modelo, entrenamiento más rápido, mejor predicción.
- Desafío: Problema NP-Hard



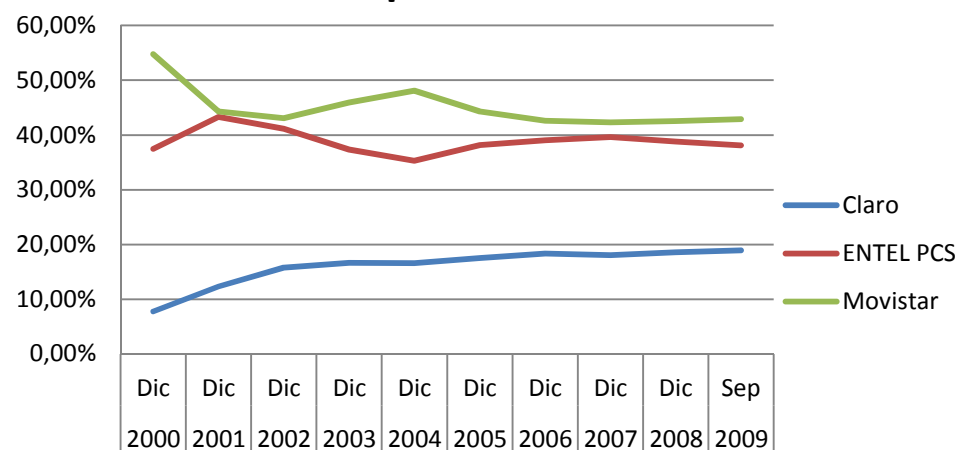
DESCRIPCIÓN CONTEXTO

- Inversión representa 4,8 % de inversión economía en el 2008.
- Aumento del PIB del 3,2% el 2008.
- Nivel penetración 87,8 cada 100 habitantes en el 2008.
- Tendencias en acceso internet y telefonía inalámbrica.

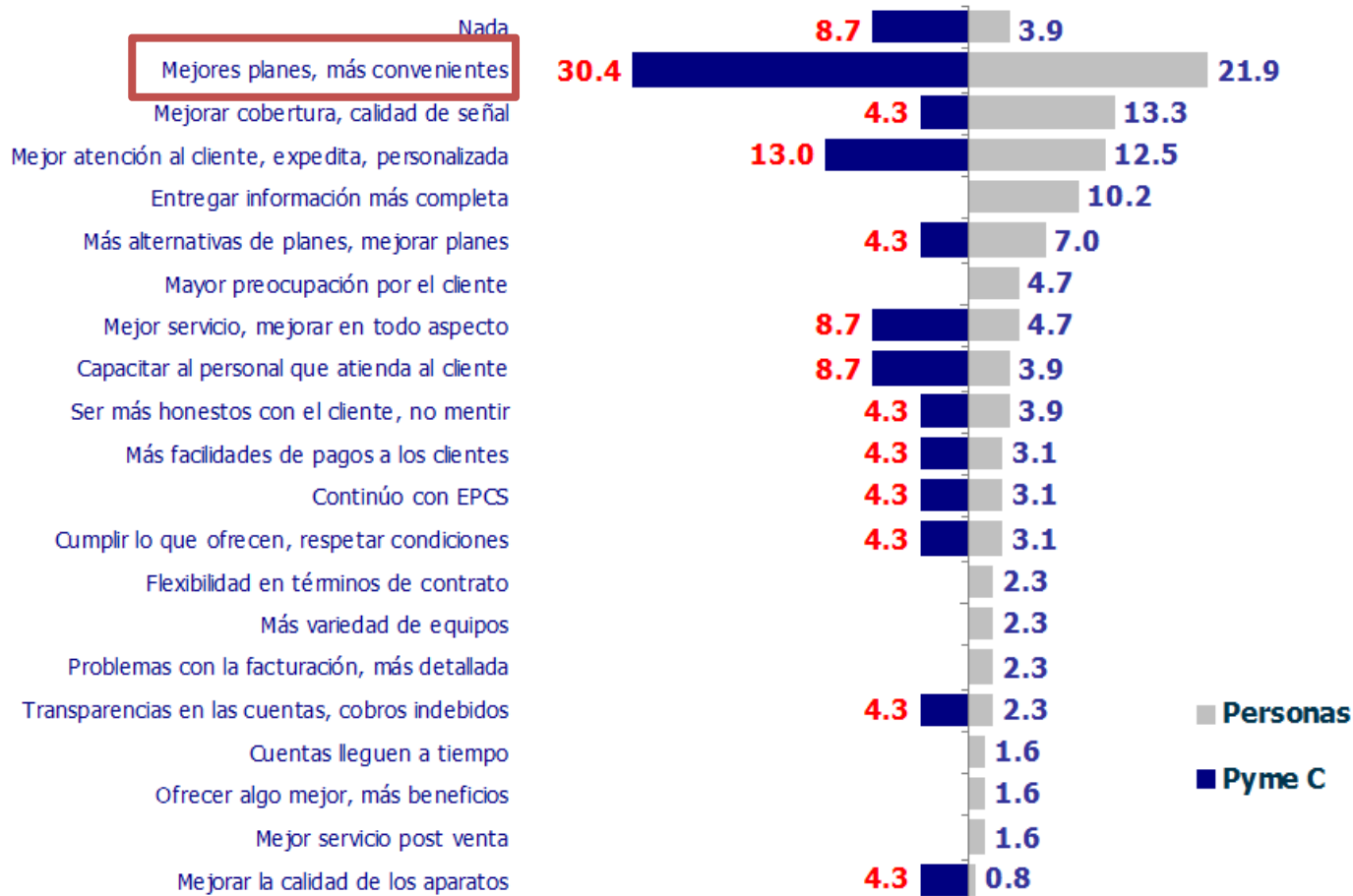
Penetración (n°abonados/100hab.)



Participación de mercado

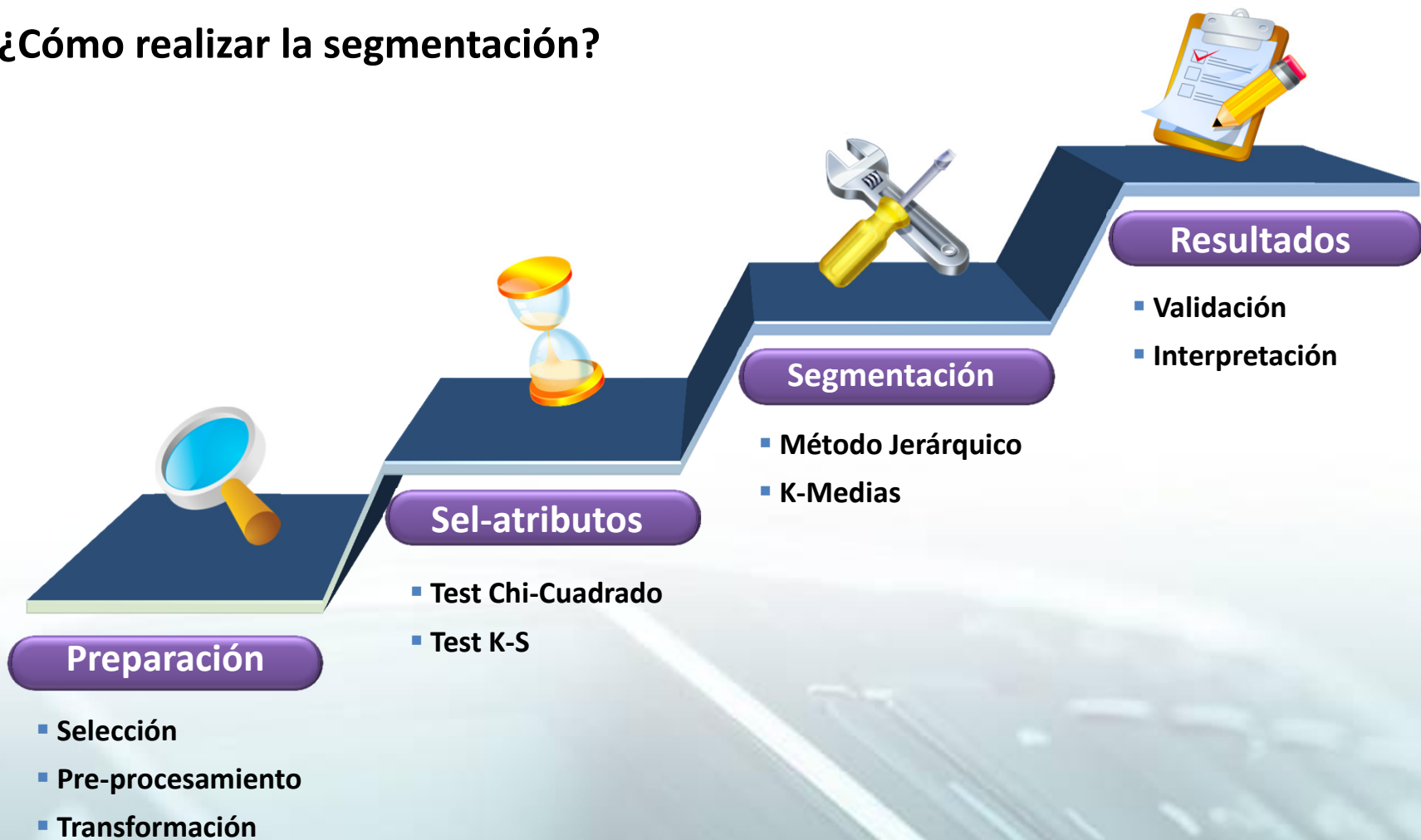


Problemática



METODOLOGÍA

¿Cómo realizar la segmentación?



Preparación

Selección

Pre-
procesamiento

Transformación

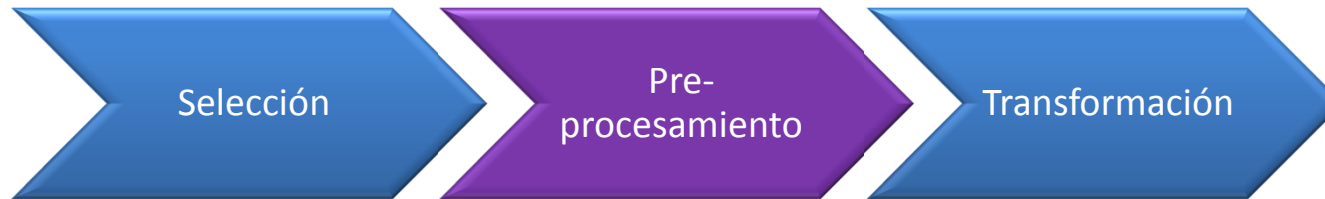
- Selección de servicio
- Extracción Base de Datos
- Grupo de estudio
- Selección Periodo
- Variables Comportamiento Servicio
- Variables Comportamiento Cliente

Plan
Región
Fecha inicio
Fecha compañía
Tráfico
Profesión
Nivel Educación
Estado Civil
Edad
Hijos
~~Cargo~~

< 10 %



Preparación



➤ Matriz

➤ Reducción y creación

V. Categóricas

Pymes = { A, B, C, C Alto Valor }

Plan Voz = { suscripción – cuenta controlada }

~~Región~~ = 13 regiones = { Metropolitana - Región }

Fecha ingreso compañía = { Nuevo o antiguo }



V. Continuas

~~Tráficos~~ – { Promedio, máximo, mínimo, contadores }

Preparación

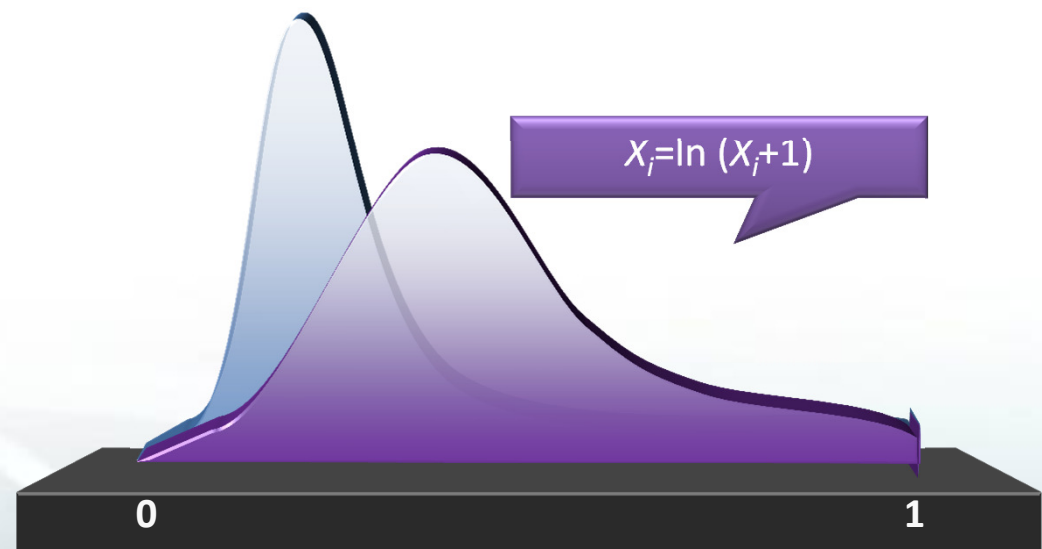


Variables Categóricas

Variables
Binarias

1 ; SI pertenece
0 ; NO pertenece

Variables Continuas



$$x'_i = \left(\frac{x_i - \min(x)}{\max(x) - \min(x)} \right)$$

SELECCIÓN DE ATRIBUTOS

¿Cómo saber que atributos son relevantes en el estudio?

Relevantes con respecto ... ¿a qué?

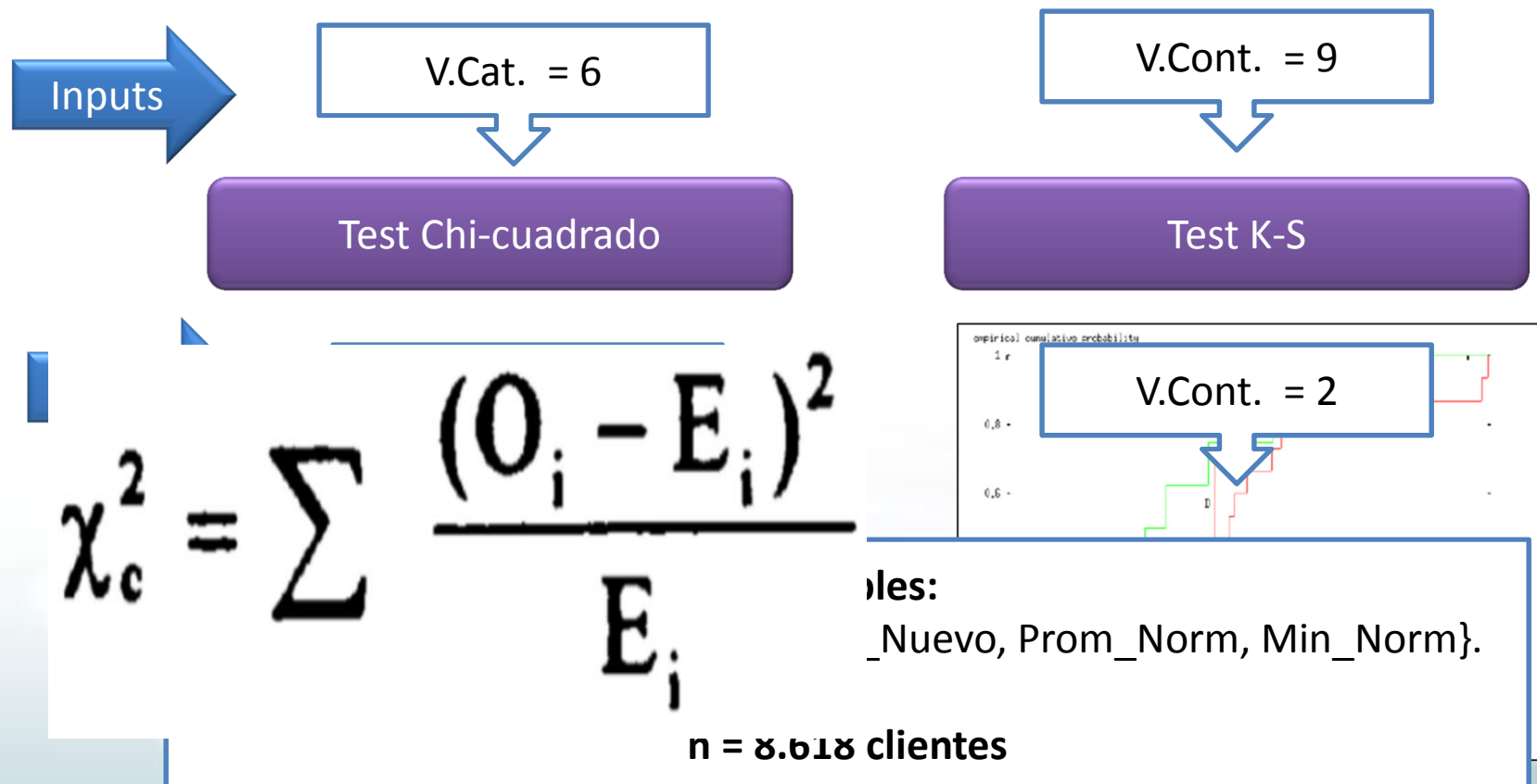
Creación de variable que identifique el comportamiento de los clientes, el cual indique las características deseadas por la compañía ...

“Continuidad en el servicio”

MES1	MES2	MES3	MES4	MES5	MES6	MES7	Etiqueta	Var_Obj
a	a	a	a	a	a	a	Bueno	1
a	a	s	a	s	a	a	Malo	0
a	a	a	a	s	a	a	Malo	0
a	a	a	s	s	d	d	Malo	0

SELECCIÓN DE ATRIBUTOS

Métodos de Filtro: eliminación de atributos irrelevantes de forma previa al algoritmo, estudiando propiedades estadísticas (correlación con variable objetivo).



SEGMENTACIÓN



SEGMENTACIÓN

Características

Método Jerárquico

- No hay decisión del número de clusters.
- Problemas cuando los datos contienen alto nivel de error.
- Puede ser muy lento

K-Medias

- Más rápido, más fiable
- Es necesario especificar el número de clusters (arbitrario).
- Es necesario establecer la semilla inicial (arbitrario).
- Los algoritmos son sensibles a la semilla inicial.

Metodo sugerido o utilizado

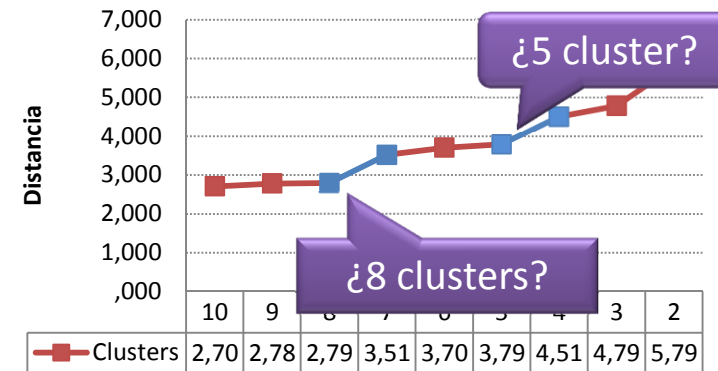
1. Primero ejecutar un método jerárquico para definir el número de clusters
2. Luego utilizar el procedimiento *k*-medias para formar los clusters

SEGMENTACIÓN: Método Jerárquico

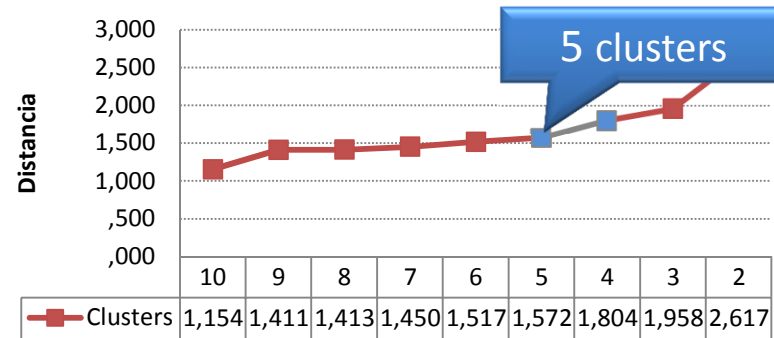
Single linkage



Complete linkage



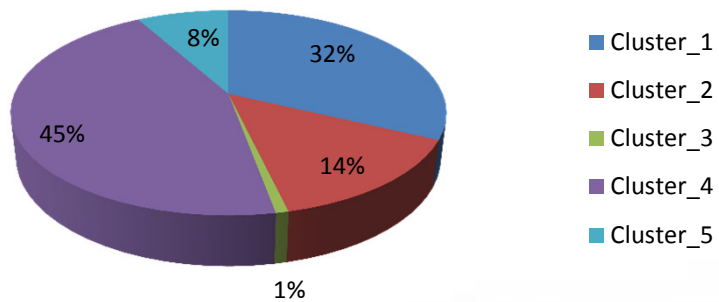
Average linkage



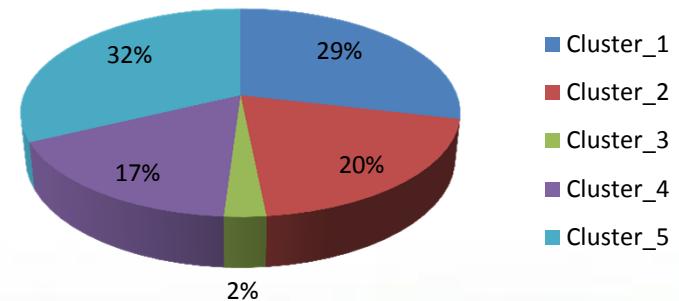
SEGMENTACIÓN: Método Jerárquico

- Tipo Aglomerativo
- Inicio de n agrupaciones y finaliza en 1 agrupación.
- Media de distancia euclidiana {single, complete y average linkage}

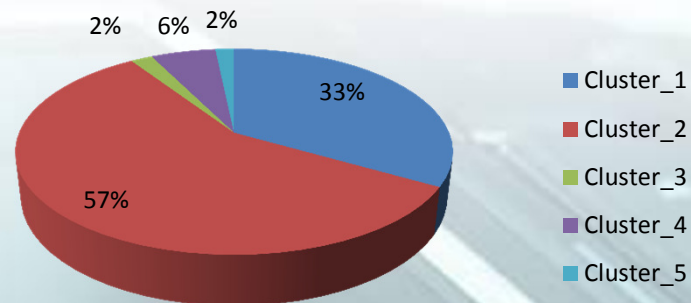
Single Linkage



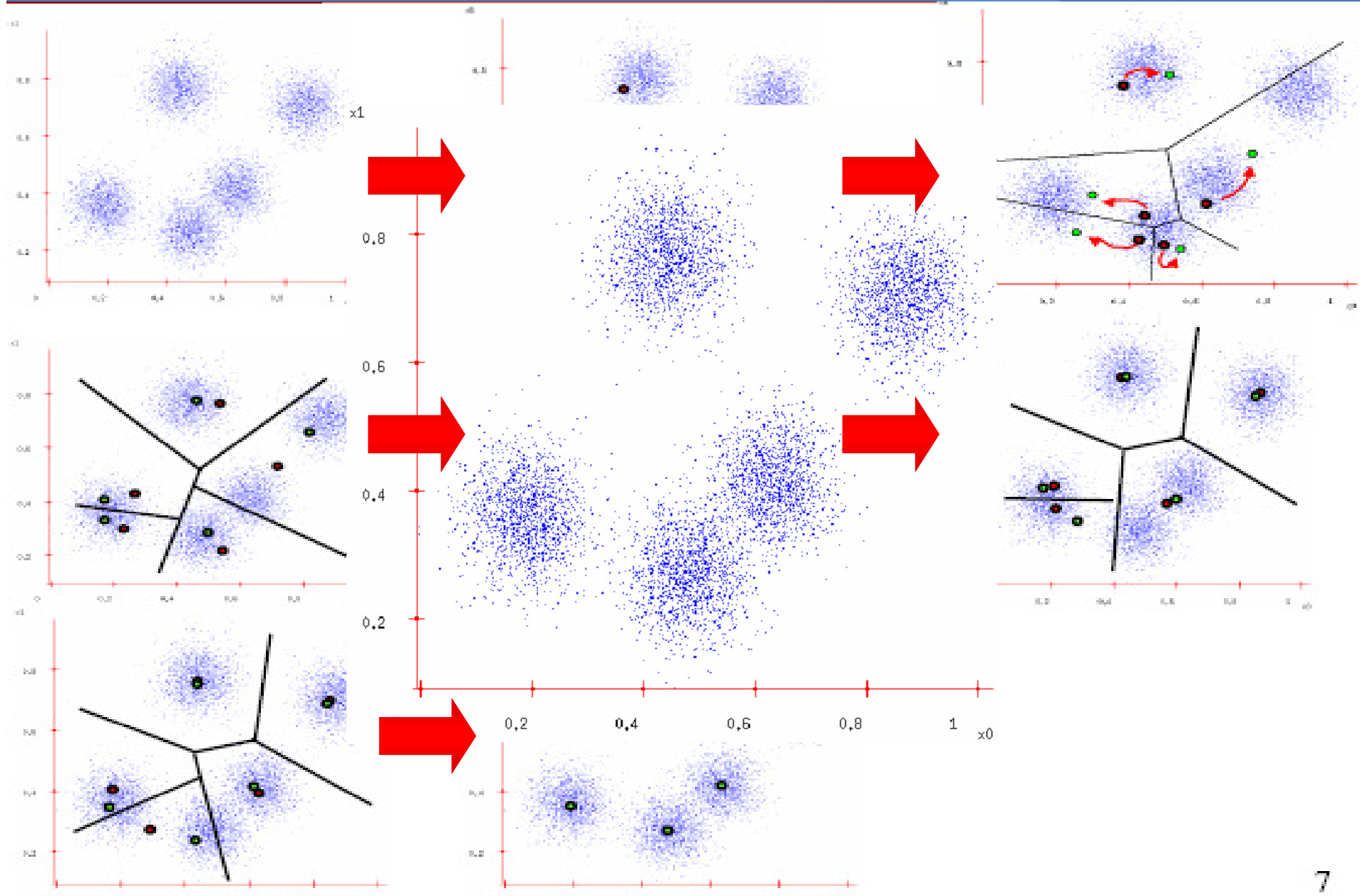
Complete Linkage



Average Linkage

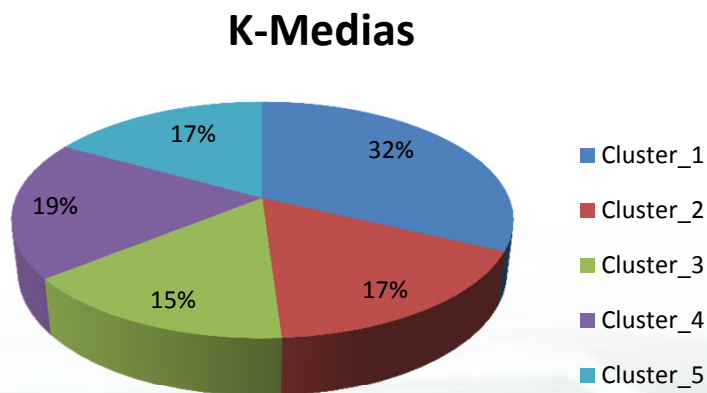


SEGMENTACIÓN: K-medias



SEGMENTACIÓN: K-medias

- Cantidad 5 Clusters
- Inicialización aleatorio de centroides
- Asignación de n observaciones al centroide más cercano
- Actualización de centroides



Análisis de sensibilidad

Cambio en la inicialización de centroides.

90% permanencia en los clusters.

SEGMENTACIÓN: K-medias

Reporte

Atributos	Cluster _1	Cluster _2	Cluster _3	Cluster _4	Cluster _5
SI_Metropolitana	1.00	1.00	0.00	0.75	0.00
Si_SS	0.96	0.82	0.77	0.90	0.93
SI_Pyme	1.00	0.00	0.00	0.75	1.00
SI_Nuevo	0.00	0.00	0.23	1.00	0.00
Prom_Norm	0.27	0.30	0.26	0.25	0.23
Min_Norm	0.15	0.12	0.09	0.11	0.12

Medias de atributos

SI_Metropolitana	Si_SS	SI_Pyme	SI_Nuevo	Prom_red	Min_red
0.63	0.89	0.63	0.22	0.26	0.12

Resultados

Interpretación

Atributos	Cluster _1	Cluster _2	Cluster _3	Cluster _4	Cluster _5
SI_Metropolitana	X	X		-	
Si_SS	X	X		X	X
SI_Pyme	X			X	X
SI_Nuevo				X	
Prom_Norm	Media	Alta	Alto	Media	Baja
Min_Norm	Media	Media	Bajo	Media	Media

Cluster 1: Está conformados por Pymes que se encontraban en la compañía, pertenecientes a la Región Metropolitana, los que mayoritariamente posee un Plan de Suscripción. Además posee una utilización promedio en ocupar el servicio de plan de datos.

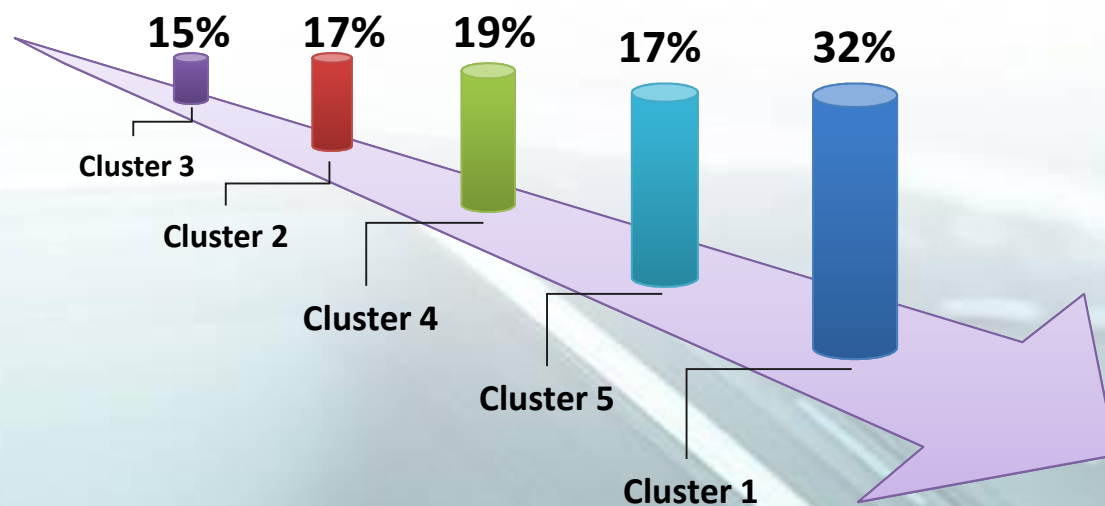
Resultados

➤ Utilización del conocimiento

➤ Variable objetivo.

➤ Etiquetación

Grupos	Malo	Bueno	Clientes	% mantenerse
Cluster_1	332	2,435	2,767	88%
Cluster_5	216	1,239	1,455	85%
Cluster_4	316	1,313	1,629	81%
Cluster_2	325	1,128	1,453	78%
Cluster_3	408	906	1,314	69%



CONCLUSIONES

- Modelo capaz de segmentar grupos de clientes de acuerdo a sus características.
- Resultados confirman supuestos de comportamiento.
- Grupos poseen características marcadas.
- Facilita aplicación de estrategias de marketing.

Trabajo Futuro:

- Actualizar información del cliente.
- Realizar un modelo de predicción de fuga (clasificación).
- Aumentar periodo de estudio (Ciclo de vida).