UNIVERSIDAD DE SANTIAGO DE CHILE

DIINF

Ayudantía

Análisis de datos – PEP 1

Ayudante Gustavo Hurtado

UNIVERSIDAD DE SANTIAGO DE CHILE

Contenido incluido en la PEP 1

- Capítulo I – Introducción.
- Capítulo II – Análisis de Componentes Principales (ACP).
- Capítulo III – Análisis de Agrupamientos.
- Capítulo IV - Análisis Discriminante.

UNIVERSIDAD DE SANTIAGO DE CHILE

Capítulo I – Introducción

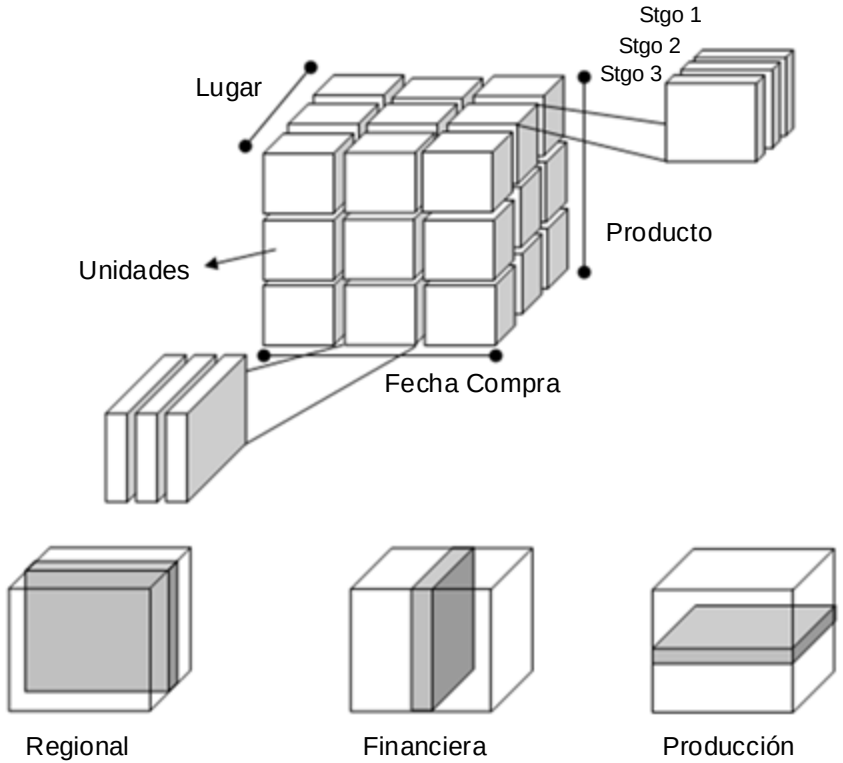
- Análisis de Datos y Minería de Datos.
- Obtención de conocimiento en Bases de Datos (proceso KDD).
- Modelos lineales de regresión y aprendizaje no lineal.
- Hipótesis de los modelos basados en aprendizaje.
- Bases de datos operacionales y analíticas.
- Datawarehouse.

Capítulo I – Introducción

Datawarehouse

Ej: Análisis de una librería con tiendas a nivel nacional.

Producto	Lugar	Fecha Compra	Unidades
CD	Santiago1	Mes 1	1500
Libro	Linares	Mes 1	150
Revista	Temuco1	Mes 1	506
CD	Santiago2	Mes 2	1020
CD	Santiago3	Mes 3	1567



Pregunta 3. Datawarehouse – Visitadores médicos

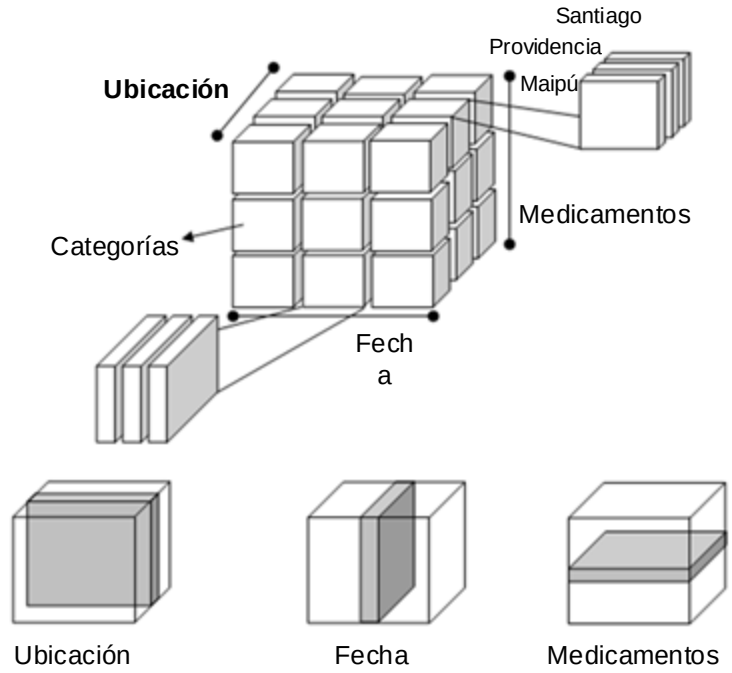
Se requiere construir un Datawarehouse para un laboratorio de medicamentos que lleve la información de sus visitadores médicos.

Considerando que el Datawarehouse es de interés para conocer la penetración territorial y de productos, tome como elemento de evaluación las visitas realizadas a los médicos y las muestras entregadas a cada facultativo.

Identifique dimensiones y las medidas, realice un esquema gráfico y presente un ejemplo de dato atómico.

Pregunta 3. Datawarehouse – Visitadores médicos

Visitadores	Médicos	Especialidades	Ubicación	Fecha visita	Muestras
Visitador 2	Médico 1	General	Santiago	dd/MM/aaaa	Medicamento 2
Visitador 3	Médico 2	Cirugía	Maipú	dd/MM/aaaa	Medicamento 3
Visitador 1	Médico 1	Oftalmología	Providencia	dd/MM/aaaa	Medicamento 1
…	…	…	…	…	…
Visitador 1	Médico 3	Cirugía	Maipú	dd/MM/aaaa	Medicamento 3



Principales dimensiones: Ubicación Fecha Medicamentos

UNIVERSIDAD DE SANTIAGO DE CHILE

Capítulo II – Análisis de Componentes Principales (ACP)

- Fundamentos matemáticos del modelo Análisis de Componentes Principales.
- Aplicar el modelo PCA en un conjunto de datos perteneciente a un problema específico.

UNIVERSIDAD DE SANTIAGO DE CHILE

Capítulo II – Análisis de Componentes Principales (ACP)

Sea el siguiente conjunto de datos de notas de estudiantes:

Estudiante	Matemat.	Lenguaje	Arte
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

$$\mathbf{A} = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$

1. Calcular la media para cada dimensión del conjunto de datos.

$$\bar{\mathbf{A}} = [\, 66 \quad 60 \quad 60 \,]$$

UNIVERSIDAD DE SANTIAGO DE CHILE

Capítulo II – Análisis de Componentes Principales (ACP)

2. Centrar los datos (con respecto a la media).
3. Calcular matriz de covarianza. $cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Xi - \bar{x})(Yi - \bar{y})$

	Matemáticas	Lenguaje	Arte
Matemáticas	630.0	450.0	225.0
Lenguaje	450.0	450.0	0.0
Arte	225.0	0.0	900.0

2. Calcular matriz de correlación.

1.000000	0.845154	0.298807
0.845154	1.000000	0.000000
0.298807	0.000000	1.000000

Capítulo II – Análisis de Componentes Principales (ACP)

5. Obtener *valores propios* y *vectores propios*.

λ_1=1.397, λ_2=1.00, λ_3=0.214

$$\sum_{i=1}^{p} \lambda_i = \boldsymbol{P}$$

	Vector 1	Vector 2	Vector 3
L	-0.7071068	0.0000000	-0.7071068
M	-0.6415992	0.4203581	0.6415992
A	-0.2972381	-0.9073583	0.2972381

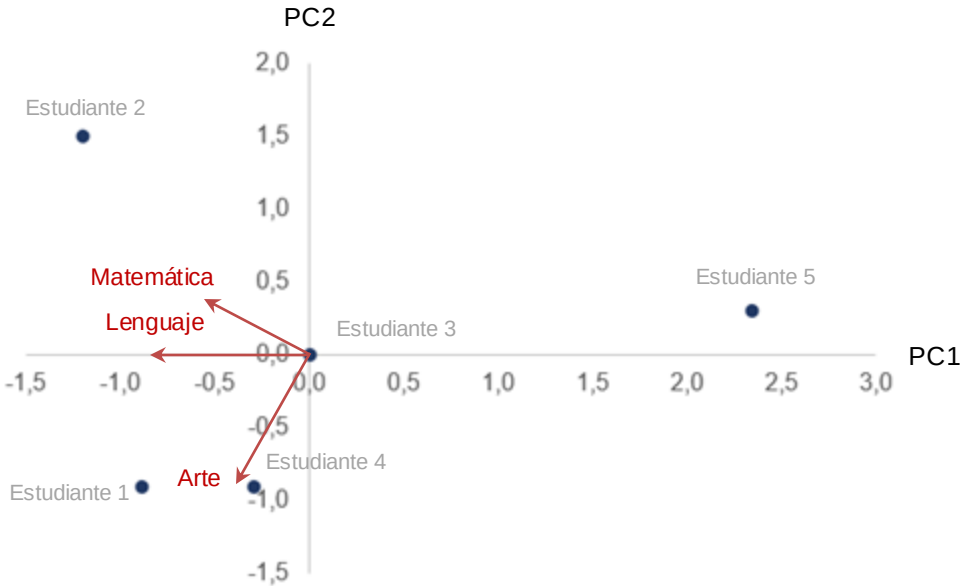
5. Ordenar los vectores propios a partir de valores propios ($\lambda_1 > \lambda_2 > \lambda_3$).
6. Normalizar muestras originales.

Capítulo II – Análisis de Componentes Principales (ACP)

Matriz de transformación

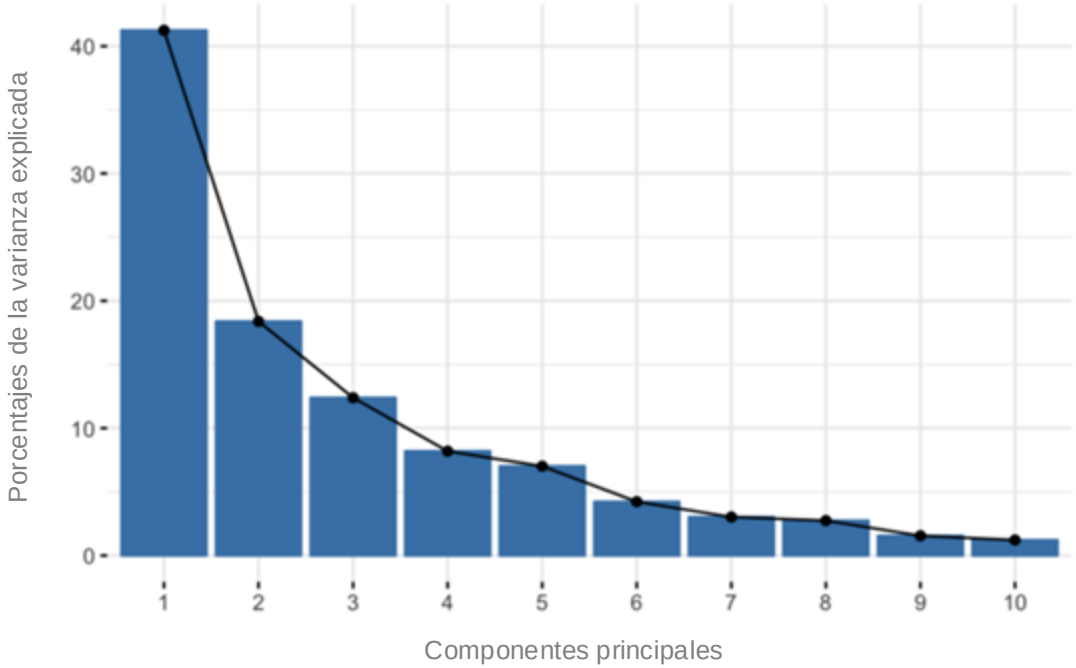
	PC1	PC2
L	-0.7071068	0.0000000
M	-0.6415992	0.4203581
A	-0.2972381	-0.9073583

	PC1	PC2
Estudiante 1	-0.88248220	-0.9073583
Estudiante 2	-1.19536437	1.5018344
Estudiante 3	0.01194376	0.0000000
Estudiante 4	-0.28529430	-0.9073583
Estudiante 5	2.35119711	0.3128822



UNIVERSIDAD DE SANTIAGO DE CHILE

Capítulo II – Análisis de Componentes Principales (ACP)



Capítulo II – Análisis de Componentes Principales (ACP)

Pregunta 2. Caracterización billetes falsos

Para caracterizar billetes falsos, los bancos suizos realizaron un análisis que consistía en tomar medidas de los billetes. Para el análisis se disponía de tres grupos diferentes de billetes. *Originales de papel, originales de plástico y billetes falsos*. Cada billete fue caracterizado por las siguientes variables:

LON	:	Longitud del billete.
LD	:	Largo de la Diagonal del billete.
AI	:	Ancho Izquierdo del billete.
AD	:	Ancho Derecho del billete.
AMI	:	Ancho Margen Inferior del billete
AMS	:	Ancho Margen Superior del billete.

Capítulo II – Análisis de Componentes Principales (ACP)

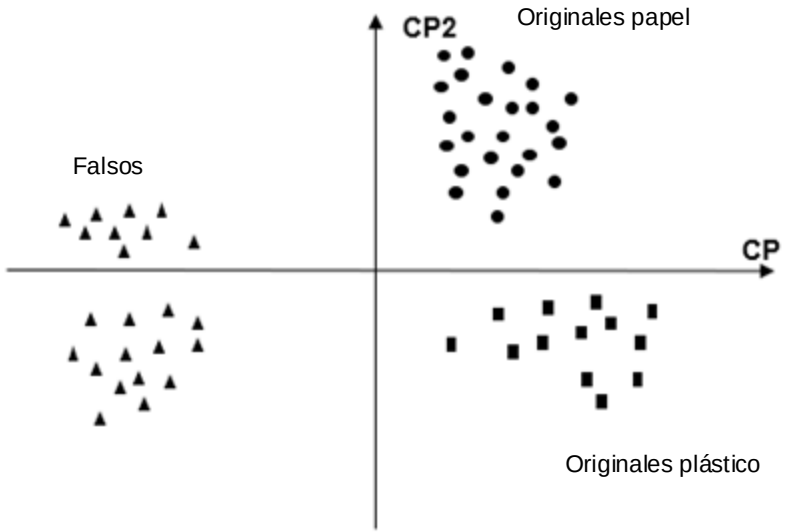
Pregunta 2. Caracterización billetes falsos

A continuación, se realizó un análisis de componentes principales con los siguientes resultados:

- Valores propios: 2,58; 1,34; 0,76; 0,56; 0,50; 0,26.

- Vectores propios para las dos primeras componentes:

	Componente 1	Componente 2
LON	0,395	0,799
LD	0,207	0,345
AI	0,445	-0,263
AD	0,411	-0,375
AMI	0,347	-0,072
AMS	0,560	-0,163



Capítulo II – Análisis de Componentes Principales (ACP)

Pregunta 2. Caracterización billetes falsos

Se realizó un trabajo similar con monedas midiendo 6 variables de tamaño de éstas y los valores propios del análisis fueron los siguientes: 1,96; 1,54; 1,09; 0,73; 0,40; 0,28.

a) Determine el porcentaje de validez del análisis.

b) Interprete cada una de las componentes.

c) Identifique las principales características de los billetes originales.

d) Determine si existen diferencias entre las falsificaciones.

e) Para el análisis de las monedas, ¿se logrará tener una precisión similar a la de los billetes?

UNIVERSIDAD DE SANTIAGO DE CHILE

Capítulo III – Análisis de Agrupamientos

- Agrupamientos jerárquicos y no jerárquicos.
- Similaridad en espacios n-dimensionales como concepto de distancia.
- Comprender la estructuración de un agrupamiento jerárquico.
- Medidas de similaridad y su aplicación a la agrupación.
- Algoritmos básicos de los agrupamientos.
- Medidas de calidad para evaluar agrupamientos.

UNIVERSIDAD DE SANTIAGO DE CHILE

Capítulo III – Análisis de Agrupamientos

Pregunta 1. Agrupamiento de genes

En una gran variedad de problemas de bioinformática, se requiere determinar grupos de genes que intervienen en una determinada enfermedad. Los genes están formados por un alfabeto básico que contiene 4 letras (bases) A, G, C, T. Muchas veces con un trozo de gen (aproximadamente 8 bases "letras") es posible caracterizar un gen.

Para la siguiente tabla 1 determine:

a) Un método para medir distancia entre genes.

b) Construya una matriz de distancia entre los genes.

c) Determine el dendograma que caracteriza este conjunto de genes.

d) Determine 2, 3 o 4 grupos según corresponda.

Tabla 1

Gen 1	G	A	T	A	C	A	T	T
Gen 2	G	A	T	A	C	A	T	A
Gen 3	G	A	T	A	C	T	A	C
Gen 4	C	T	A	A	G	G	G	G
Gen 5	C	T	C	A	G	G	G	G
Gen 6	G	A	T	T	T	C	C	G
Gen 7	G	A	T	T	A	C	C	G

UNIVERSIDAD DE SANTIAGO DE CHILE

Capítulo III – Análisis de Agrupamientos

Gen 1	G	A	T	A	C	A	T	T
Gen 2	G	A	T	A	C	A	T	A

Gen 1	G	A	T	A	C	A	T	T
Gen 3	G	A	T	A	C	T	A	C

Gen 1	G	A	T	A	C	A	T	T
Gen 4	C	T	A	A	G	G	G	G

Gen 1	G	A	T	A	C	A	T	T
Gen 5	C	T	C	A	G	G	G	G

Gen 1	G	A	T	A	C	A	T	T
Gen 6	G	A	T	T	T	C	C	G

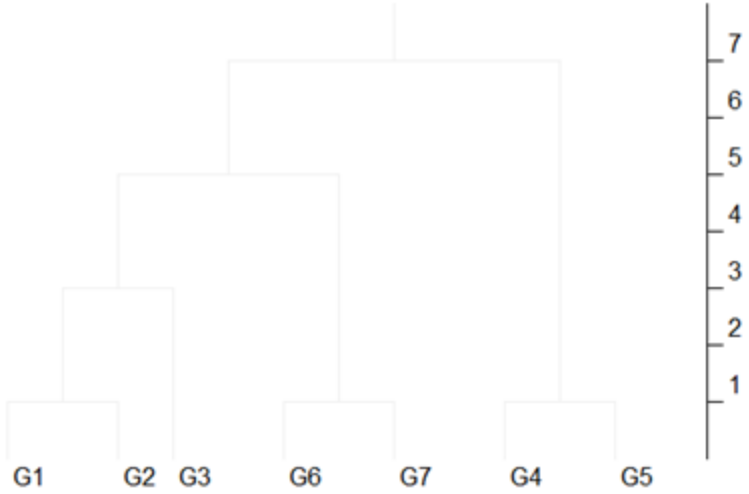
Gen 1	G	A	T	A	C	A	T	T
Gen 7	G	A	T	T	A	C	C	G

Matriz de distancia

	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7
Gen 1	0						
Gen 2		0					
Gen 3			0				
Gen 4				0			
Gen 5					0		
Gen 6						0	
Gen 7							0

UNIVERSIDAD DE SANTIAGO DE CHILE

Capítulo III – Análisis de Agrupamientos

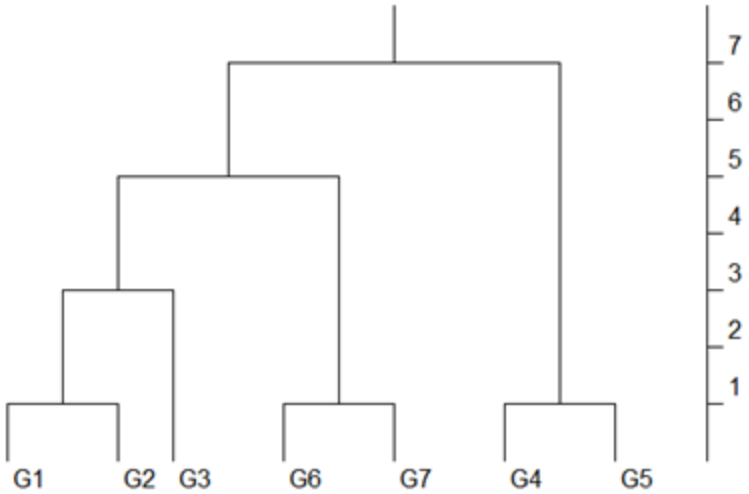


Matriz de distancia

	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7
Gen 1	0	1	3	7	7	5	5
Gen 2		0	3	7	7	5	5
Gen 3			0	7	7	5	5
Gen 4				0	1	7	7
Gen 5					0	7	7
Gen 6						0	1
Gen 7							0

UNIVERSIDAD DE SANTIAGO DE CHILE

Capítulo III – Análisis de Agrupamientos



Matriz de distancia

	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7
Gen 1	0	1	3	7	7	5	5
Gen 2		0	3	7	7	5	5
Gen 3			0	7	7	5	5
Gen 4				0	1	7	7
Gen 5					0	7	7
Gen 6						0	1
Gen 7							0

UNIVERSIDAD DE SANTIAGO DE CHILE

Capítulo IV - Análisis Discriminante

- Métodos de clasificación basados en razón de probabilidades (logística).
- Métodos de clasificación paramétricos y no paramétricos.
- Métodos de clasificación basados en discriminación lineal.
- Método no paramétrico de discriminación (clasificación y regresión).
- Metodología de evaluación de la clasificación binaria.