

# Tema 2: Preparación de datos.

---



Universidad  
Francisco de Vitoria  
**UFV** Madrid

*Alberto Nogales*

*alberto.nogales@ufv.es*

*Curso 2021-2022*

# Índice

- Tipos de datos.
- Calidad de los datos.
- Técnicas de recopilación, almacén, limpieza y transformación de datos.
- Técnicas de exploración y selección de datos.
- Reducción de la dimensión.

# Preparación de datos

- Objetivo: organizarlos de manera que pueda ser procesados por los programas de construcción de modelos que hayan sido elegidos y, al mismo tiempo, asegurar que los datos se hallan de tal forma que se pueda obtener el mejor modelo posible del conjunto de datos.

# Tipos de datos

- Estructurados: información, presentada en forma de columnas con cabecera y filas que puede ser ordenada y procesada. Fácil acceso.
  - Tablas de datos.
  - Transacciones.

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

# Tipos de datos

- No estructurados: aquel que no tiene una estructura interna identifiable (el 80%). Se pueden organizar para hacer búsquedas sencillas. Se almacenan sin que el sistema entienda el formato.

- Texto.
- Imágenes.
- Videos.
- Sonido.

The university has 5600 students.  
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.  
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

# Tipos de datos

- Semiestructurados: aquellos que usan etiquetas para identificar distintos datos, permitiendo agruparlos y establecer jerarquías.
  - JSON.
  - XML.
  - NoSQL.

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
  </Student>
  ...
</University>
```

# Tipos de datos

Propiedad	Datos estructurados	Datos semiestructurados	Datos no estructurados
Métodos de organización	Basado en BBDD relacionales.	Basado en datos etiquetados como XML.	Basado en caracteres, imágenes, etc.
Manejo de las transacciones	Muy desarrolladas. Aplica técnicas de concurrencia.	Adaptadas de las BBDD pero mejorables.	No hay transacción.
Actualización	Sobre tuplas, filas o tablas.	Sobre tuplas o grafos.	Se actualiza el archivo entero.
Flexibilidad	Dependiente del esquema, poco flexible.	Más flexible que estructurados pero menos que los no.	Muy flexible.
Escalabilidad	Difícil de escalar.	Más sencillo de escalar.	Muy fácil de escalar.
Robustez	Muy robusto.	Poco probado.	Nada robusto.
Rendimiento de consultas	Permite consultas muy complejas.	Consultas más sencillas.	Búsqueda en textos y poco mas.

# Tipos de datos

**Tablas de datos:** Consisten en una colección de registros con un conjunto de atributos cada uno.



# Tipos de datos

- Valor: representación simbólica de un atributo o característica de una entidad.
- Atributo: representación de una propiedad o característica.
- Registro: Una colección de atributos.
- Colección de registros: Tabla.
- Conjunto de tablas = Dataset.

# Tipos de datos

Clasificación de los atributos:

- Numéricos: Adoptan valores reales, enteros o naturales (infinitos valores de un conjunto numérico)
- Booleanos: “True” o “False que equivalen a 0 o 1.
- Categóricos: Aquellos que toman valores de un conjunto finito. [“Bajo”, “Medio”, “Alto”]

# Tipos de datos

- **Transacciones:** Cada registro o transacción (ticket) se compone de un conjunto de ítems.

ID	Items
1	pintura, brocha, cemento
2	cemento, ladrillo, cal, sierra
3	pintura, brocha, aguarrás, guantes

ID	Pintura	Brocha	Cemento	Ladrillo	Cal	Sierra	Aguarrás	Guantes
1	1	1	1	0	0	0	0	0
2	0	0	1	1	1	1	0	0
3	1	1	0	0	0	0	1	1

# Tipos de datos

- **Texto:** Un documento consiste en un vector de términos, cada término es una componente o atributo del vector. Ej: El valor representa la frecuencia del término en el documento.

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

# Tipos de datos

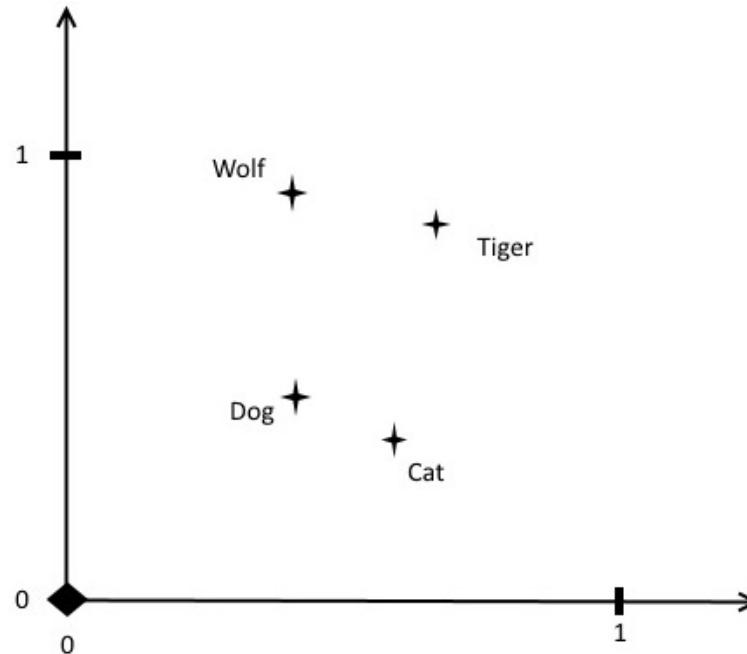
Otras maneras de representar texto.

## Diccionario

cat	the	quick	brown	fox	jumped	over	dog	bird	flew
0	1	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0	0
0	1	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0

...	kangaroo	house
...	0	0
...	0	0
...	0	0
...	0	0
...	0	0
...	0	0
...	0	0
...	0	0
...	0	0

## Embeddings



# Tipos de datos

- **Imágenes:** En blanco y negro, matriz de NxM (ancho y alto en píxeles) con valores entre 0 y 255. En color, igual pero con profundidad 3 (canales RGB).

170	238	85	255	221	0
68	136	17	170	119	68
221	0	238	136	0	255
119	255	85	170	136	238
238	17	221	68	119	255
85	170	119	221	17	136

# Tipos de datos

- **Videos:** En blanco y negro, X matrices (tantas como instantes de tiempo) de NxM (ancho y alto en píxeles) con valores entre 0 y 255. En color, igual pero con profundidad 3 (canales RGB).



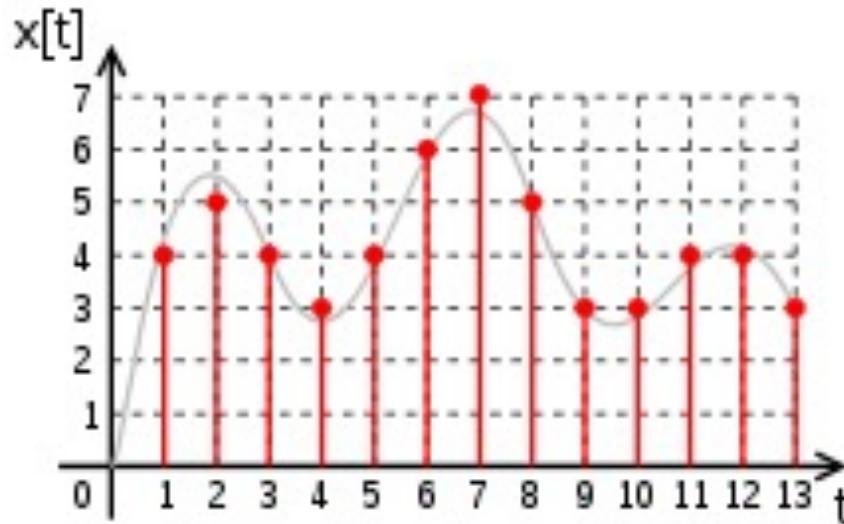
# Tipos de datos

- **Redes sociales:** La información que se genera en una red social, corresponde a un grafo. En él, los nodos serían usuarios y las aristas su manera de relacionarse. Los grafos se representan como matrices.

	0	1	2	3
0	0	1	0	1
1	0	0	1	0
2	0	0	0	1
3	0	1	0	0

# Tipos de datos

- **Señales:** Dependiendo de si la señal es analógica/digital, de un canal o multicanal, etc, la complejidad varía. Simplificándolo, es un vector de N instantes de tiempo con los valores de la frecuencia.



# Calidad de los datos

- **Calidad de datos:** es una forma compleja de medir las propiedades de los datos desde diferentes perspectivas. Es un examen exhaustivo de la eficiencia, la fiabilidad y la conveniencia de los datos.



# Calidad de los datos

Se medirá de acuerdo a **6 dimensiones**. Entendiendo por **dimensión**, describir una **característica** de los datos que puede **medirse o evaluarse** en función de **normas definidas para determinar la calidad de los datos**.

# Calidad de los datos

**Las 6 dimensiones serían:**

- **Compleitud:** En algunos casos, los datos que no están son irrelevantes, pero cuando se vuelven necesarios para un proceso del negocio, éstos se vuelven críticos.
- **Conformidad:** Los datos que están en los campos de la tabla, deben estar en un formato estándar y legible.

# Calidad de los datos

- **Consistencia:** Al hacer el cruce de información con los registros, se debe evitar la información contradictoria.
- **Precisión / Exactitud:** Si los datos no son precisos, estos no pueden ser utilizados. Para detectar si lo son, se comparan con una fuente de referencia.

# Calidad de los datos

- **Duplicación:** Es importante saber si se tiene la misma información en formatos iguales o similares dentro de la tabla.
- **Integridad:** Otra dimensión de calidad importante radica en el hecho de saber si toda la información relevante de un registro está presente de forma que se pueda utilizar.

# Técnicas de selección

- **Registros:** sólo se eligen instancias completas representativas del total de los datos disponibles. Aleatoria o por filtros.

**¿Qué individuos del set de datos voy a usar?**

Seleccionamos los considerados jóvenes (<26)

Edad	Puesto de trabajo	Sueldo
→ 18	Camarero	14.000
→ 25	Abogado	20.000
45	Abogado	45.000

# Técnicas de selección

- **Atributos:** Se realiza porque haya atributos irrelevantes, redundantes o por exceso de dimensionalidad.

**¿Qué características de los datos necesito?**

Eliminar claves

DNI	Ciudad	País
41692347G	Madrid	España
83458290F	Viena	Austria

Eliminar atributos dependientes (Ciudad vs CP)

# Técnicas de limpieza

- **Datos incompletos:** No se ha recogido la información o no se ha proporcionado. Ver si es significante.

**Ejemplo:** la edad de un cliente es *null*.

**¿Cómo lo encuentro?**

-Buscar valores que sean *null*.

**¿Qué se hace?**

-No se tiene en cuenta el registro.

-No se tiene en cuenta el atributo, muchos *nulls*.

-Se completa calculando el valor medio del atributo o el más frecuente.

# Técnicas de limpieza

- **Datos redundantes:** Pueden ser consecuencias de la mezcla de dos o más conjuntos de datos.

**Ejemplo:** Cliente con el mismo DNI y distinto apellido.

**¿Cómo lo encuentro?**

- Obtener frecuencias de atributos únicos.

**¿Qué se hace?**

- Dejar uno, el más reciente.

# Técnicas de limpieza

- **Datos incorrectos o inconsistentes:** Se dan porque no hay procesos de control de errores.

**Ejemplo:** Cliente de 45 años con Carnet Joven.

**¿Cómo se encuentra?**

-Crear una regla para cada caso.

**¿Qué se hace?**

-Eliminar el registro. Excepto si tenemos una manera de corregir el dato.

# Técnicas de limpieza

- **Errores de transcripción:** Muy típicos con el uso de mayúsculas y minúsculas. Formatos de fechas.  
**Ejemplo:** BARCELONA y Barcelona se podrían considerar distintas ciudades.

**¿Cómo se encuentra?**

- Ordenar los valores de un atributo alfabéticamente

**¿Qué se hace?**

- Unificar el formato.

# Técnicas de limpieza

- **Variaciones en las referencias a los mismos conceptos:** En datos categóricos usar distintas etiquetas para referirse a lo mismo.

**Ejemplo:** "Profesión liberal" vs "Autónomo".

**¿Qué se hace?**

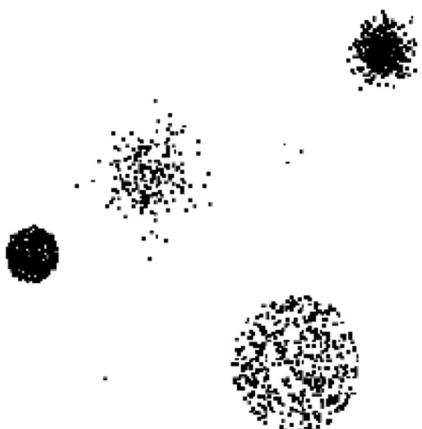
-Obtener una lista única de los valores de un atributo categórico.

**¿Qué se hace?**

- Elegir una etiqueta.

# Técnicas de limpieza

- **Valores atípicos (Outliers):** Aparece un valor que no se relaciona en nada con el resto, totalmente diferente y aislado, se puede considerar que es una anomalía en la medición. ¿Si son muchos valores ya no son atípicos?.



No siempre es conveniente eliminarlos.

# Transformación de datos

- **Convertir valores continuos a categóricos:** Asignar una categoría a cada rango de valores que necesitemos, fijando bien una correspondencia entre los valores numéricos y la categoría.

**¿Cómo puedo usar un modelo supervisado con una variable objetivo que es numérica?**

**Ejemplo:** Edad<18=“Joven” y Edad>18=“Mayor”

# Transformación de datos

- **Datos categóricos a numéricos:** Proceso inverso al anterior. Con intervalos o valores únicos

**¿Cómo puedo usar un modelo predictivo si la variable a predecir es categórica?**

**Ejemplo:** “Rojo”=1, “Verde”=2, etc.

# Transformación de datos

- **Simplificación de valores:** dividir todos los valores por una constante.  
**Ejemplo:** Dividir los sueldos entre mil.
- **Agrupación de valores continuos:** asignar una “etiqueta” numérica a un evento.  
**Ejemplo:** Agrupar compras de franjas horarias.
- **Añadir una etiqueta nueva:** debido a la evolución del ámbito.  
**Ejemplo:** Manera preferida de contacto con cliente.

# Transformación de datos

- **Expansión de un atributo:** cuando el valor de un atributo puede adoptar los valores en un conjunto limitado de categorías.

Póliza	Riesgo	Coste
1525	Alto	650
7235	Baja	250
5324	Medio	400

Póliza	Alto	Medio	Bajo	Coste
1525	1	0	0	650
7235	0	0	1	250
5324	0	1	0	400

# Transformación de datos

- **Derivación de datos:** utilizar los atributos de los datos existentes para derivar nuevos atributos.

Id	Altura	Peso
24567	1,90	88
32456	1,85	92
33345	1,78	65

Id	Altura	Peso	IMC
24567	1,90	88	24,4
32456	1,85	92	26,9
33345	1,78	65	23,0

# Transformación de datos

- **Fusión de datos:** combinar dos set de datos en uno.

Id	Altura	Peso
24567	1,90	88
32456	1,85	92
33345	1,78	65

Id	Ciudad	Profesión
24567	Málaga	Funcionario
32456	Madrid	Profesor
33345	Salamanca	Carpintero

Id	Altura	Peso	Ciudad	Profesión
24567	1,90	88	Málaga	Funcionario
32456	1,85	92	Madrid	Profesor
33345	1,78	65	Salamanca	Carpintero

# Transformación de datos: Normalización

- **Normalización:** consiste en situar los datos sobre una escala de valores equivalente que permita la comparación de atributos que toman valores en dominios o rangos diferentes.

**¿Están mis valores numéricos en diferentes rangos?**

- Es necesaria antes de aplicar algunos métodos.
- Sin normalización los métodos quedan sesgados por la influencia de los atributos con valores más altos.

# Transformación de datos: Normalización

- **Por máximos:** encontrar el valor máximo del atributo a normalizar X y dividir los valores por este.

$$Z_i = X_i / X_{\max}$$

- **Por la diferencia:** compensar el efecto de la distancia del valor que tratamos con respecto al máximo de los valores observados.

$$Z_i = |X_i - X_{\min}| / |X_{\max} - X_{\min}|$$

# Transformación de datos: Discretización

- **Discretización:** consiste en establecer un criterio por el medio del cual se pueden dividir los valores de un atributos en dos o más conjuntos disjuntos.
  - Se realiza por motivos de coste computacional.
  - El tiempo necesario es menor con datos discretizados.
  - En general necesitan menos espacio de almacenamiento.
  - Los modelos que usan datos continuos pesan más.
  - La comprensión de los modelos es más sencilla usando menos valores para describir un atributo.

# Transformación de datos: Discretización

- **Partición en intervalos de la misma amplitud:**

Dado un atributo  $X$  numérico con valor mínimo  $X_{\min}$  y máximo  $X_{\max}$ :

- 1)Fijar un número  $K$  de intervalos a alcanzar.
- 2)Dividir el rango de valor de valores  $(X_{\min}, X_{\max})$  en  $k$  intervalos  $\{(X_{\min 1}, X_{\max 1}) \dots (X_{\min k}, X_{\max k})\}$   
De manera que la distancia entre valor mínimo y máximo de un intervalo sea  $(X_{\max} - X_{\min})/k$

# Transformación de datos: Discretización

- **Ejemplo:** Los años de experiencia trabajando. Los que menos tiempo llevan es de menos de un año (0) y los que mas años, 16.

$$X_{\min}=0, X_{\max}=16$$

1)  $k=4$  intervalos

2) Distancia $=(16-0)/4=4$

$$X_{\min_1}=X_{\min}=0$$

$$X_{\max_k}=X_{\max}=16$$

$$(0,3) (4,7) (8,11) (12,16)$$

# Transformación de datos: Discretización

- **Partición en intervalos de igual frecuencia:** En el caso anterior podría darse el caso de que algunos intervalos tuviesen un único elemento. ¿Es un caso atípico? Podemos obtener los intervalos según la frecuencia (Número de valores  $n$  dividido por el Número de intervalos  $k$ ).

# Transformación de datos: Discretización

- **Ejemplo:** La edad de los trabajadores de una empresa.

edad={55,22,27,40,28,22,28,31,27,31,31,55}

1) Los ordeno de menor a mayor.

{22,22,27,27,28,28,31,31,31,40,55,55}

2) Obtenemos sus frecuencias.

{22:2,27:2,28:2,31:3,40:1,55:2}

3) Queremos 3 intervalos. La frecuencia **deseada** sería  
 $12/3=4$

(22,22,27,27) (28,28,31,31,31) (40,55,55)

# Reducción de la dimensionalidad

Asegurar la calidad del modelo resultante:

- Trabajando con menos atributos.

Problemas de la alta dimensionalidad:

- Coste de procesamiento y almacenamiento.
- Atributos relevantes e irrelevantes.
- Calculo de distancias complejo.
- Data sparsity.
- La maldición de la dimensionalidad.



# Reducción de la dimensionalidad

Nos podemos dar cuenta debido a que:

1. El programa de construcción del modelo elegido no puede tratar la cantidad de datos de la que disponemos.
2. El programa puede tratarlos, pero el tiempo requerido para construir el modelo es inaceptablemente alto.

# Reducción de la dimensionalidad

- **Reducción del número de atributos:** la reducción del número de atributos consiste en encontrar un subconjunto de los atributos originales que permita obtener modelos de la misma calidad que los que se obtendrían utilizando todos los atributos.

# Reducción de la dimensionalidad

**1. Selección de atributos:** saber que subconjunto de atributos puede generar un modelo con la misma calidad. Eliminar atributos irrelevantes.

Usar la fuerza bruta: probar con todas las combinaciones de conjuntos con atributos  $n-1$ . Después con  $n-2$  y así hasta que no haya mejoras.

**2. Extracción de atributos:** Encontrar una transformación que lleva el espacio de medidas  $p$  en un espacio de características de dimensión menor.

# Reducción de la dimensionalidad

Una explicación visual:

<https://youtu.be/wvsE8jm1GzE>

<https://github.com/zalandoresearch/fashion-mnist/blob/master/doc/img/embedding.gif>



# Análisis de Componentes Principales (PCA)

**Transformar** un conjunto de **variables**, a las que se denomina originales, en un **nuevo conjunto** de variables denominadas **componentes principales**. Estas últimas se caracterizan por estar **incorreladas** entre sí y, además, **pueden ordenarse** de acuerdo con la información que llevan incorporada.

$V_1$	$V_2$	...	$V_P$
$X_{11}$	$X_{12}$	...	$X_{1P}$
...	...	...	...
$X_{N1}$	$X_{N2}$	...	$X_{NP}$

**Siempre son valores continuos.**

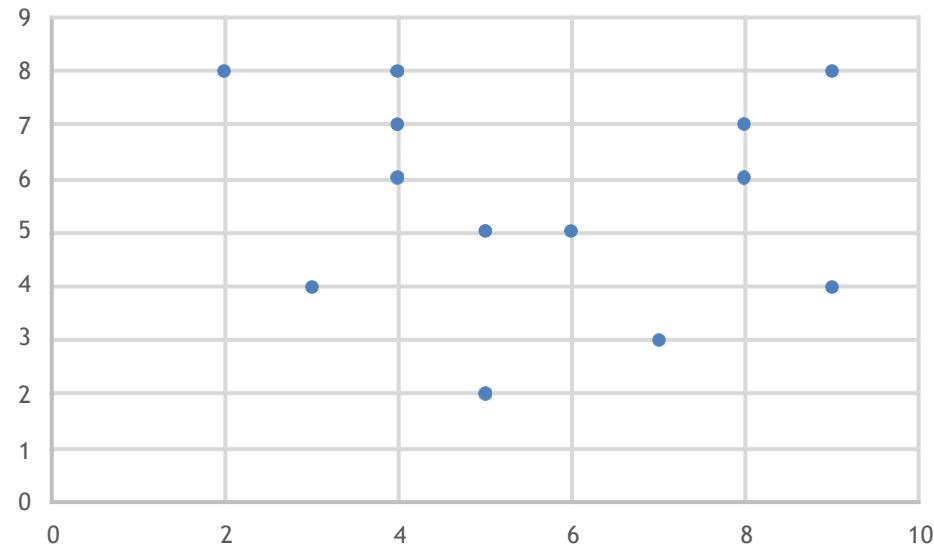
# Análisis de Componentes Principales (PCA)

Partiendo de un set de datos de alumnos y sus notas.

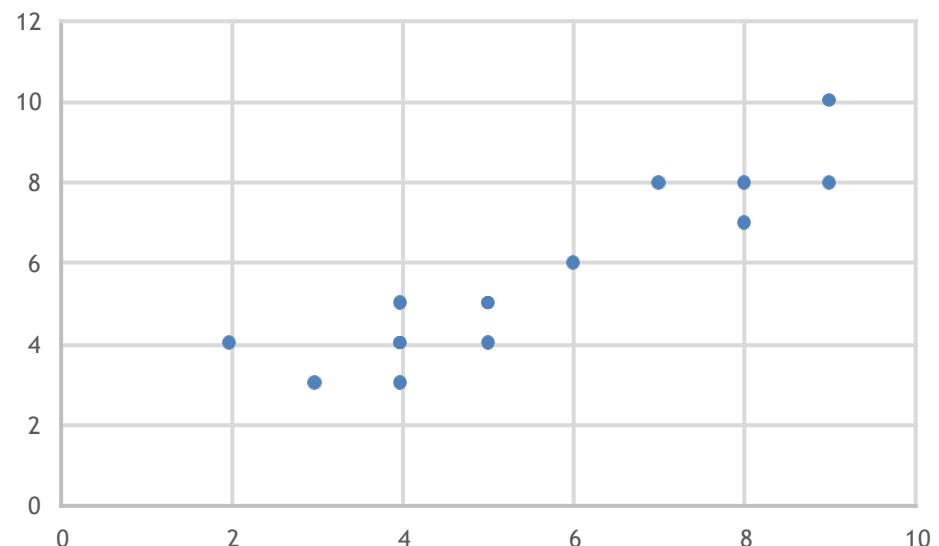
ALUMNO	LENGUA	MATEMÁTICAS	FÍSICA	INGLÉS	FILOSOFÍA	HISTORIA	QUÍMICA	E. FÍSICA
1	5	5	5	5	5	5	5	5
2	7	4	3	8	4	7	3	8
3	5	8	7	6	5	6	7	5
4	7	2	4	8	7	7	3	6
5	8	9	10	8	8	7	9	4
6	4	9	8	4	3	4	7	5
7	6	4	4	6	5	5	3	7
8	4	7	8	3	3	2	8	3
9	5	5	4	5	6	5	5	1
10	7	4	5	7	8	8	4	6
11	7	8	8	7	7	6	7	9
12	4	3	3	4	3	2	1	4
13	7	4	4	7	8	7	4	5
14	3	5	5	2	3	3	5	7
15	5	6	6	5	5	5	6	6

# Análisis de Componentes Principales (PCA)

Podemos hacer representaciones de 2D y 3D e interpretarlas.



**Matemáticas (X) vs Inglés (Y)**



**Matemáticas (X) vs Física (Y)**

# Análisis de Componentes Principales (PCA)

- 1.- Estandarizar los valores. Para ello se calcula la matriz de correlaciones. Se divide cada valor por la media muestral y se divide por la desviación estándar.
- 2.- Al obtener los componentes principales, habrá tantos como variables originales. Se calculará la matriz los autovalores y los autovectores. Con los primeros se calculan el porcentaje de varianza y el porcentaje acumulado. A partir de los segundos los **componentes**.

# Análisis de Componentes Principales (PCA)

El **porcentaje de varianza** se calcula teniendo en cuenta los autovalores con una regla de tres respecto al sumatorio de los autovalores.

$$\left. \begin{array}{l} 8,00 - 100\% \\ 3,71403 - X \end{array} \right\} X = (3,71403 * 100) / 8 = 46,380$$

El **porcentaje acumulado**, sumando el de varianza de la componente anterior.

Para el componente 1:  $46,380 + 0 = 46,380$

Para el componente 2:  $46,380 + 35,760 = 82,140$

Componente	Autovalores	% de varianza	% acumulado
1	3,71043	46,380	46,380
2	2,86078	35,760	82,140

# Análisis de Componentes Principales (PCA)

Componente	Autovalores	% de varianza	% acumulado
1	3,71043		
2	2,86078		
3	0,953481		
4	0,215574		
5	0,151316		
6	0,0628091		
7	0,0317443		
8	0,0138659		

# Análisis de Componentes Principales (PCA)

Componente	Autovalores	% de varianza	% acumulado
1	3,71043	46,380	46,380
2	2,86078	37,760	82,140
3	0,953481	11,919	94,059
4	0,215574	2,695	96,753
5	0,151316	1,891	98,645
6	0,0628091	0,785	99,430
7	0,0317443	0,397	99,827
8	0,0138659	0,173	100,00

# Análisis de Componentes Principales (PCA)

Asignatura	Componente 1	Componente 2
Lengua	0,500113	0,0853043
Matemáticas	-0,112909	0,555049
Física	-0,0517681	0,574789
Inglés	0,498752	0,036556
Filosofía	0,450292	0,121881
Historia	0,49264	0,0635768
Química	-0,0726488	0,573763
E. Física	0,187002	-0,0694516

Con los **autovectores**, mirar el **valor absoluto** de los coeficientes. Los que tienen un **valor alto** influirán en la componente. Mirar los **signos** de los coeficientes de valor alto. En este caso todos son **positivos**, influyen de forma **directamente proporcional** (positivamente).

La formula para llevar los datos a la primera componente sería:

$$C_1 = 0,500113 * X_1 - 0,112909 * X_2 - 0,0517681 * X_3 + \dots + 0,187002 * X_n$$

# Análisis de Componentes Principales (PCA)

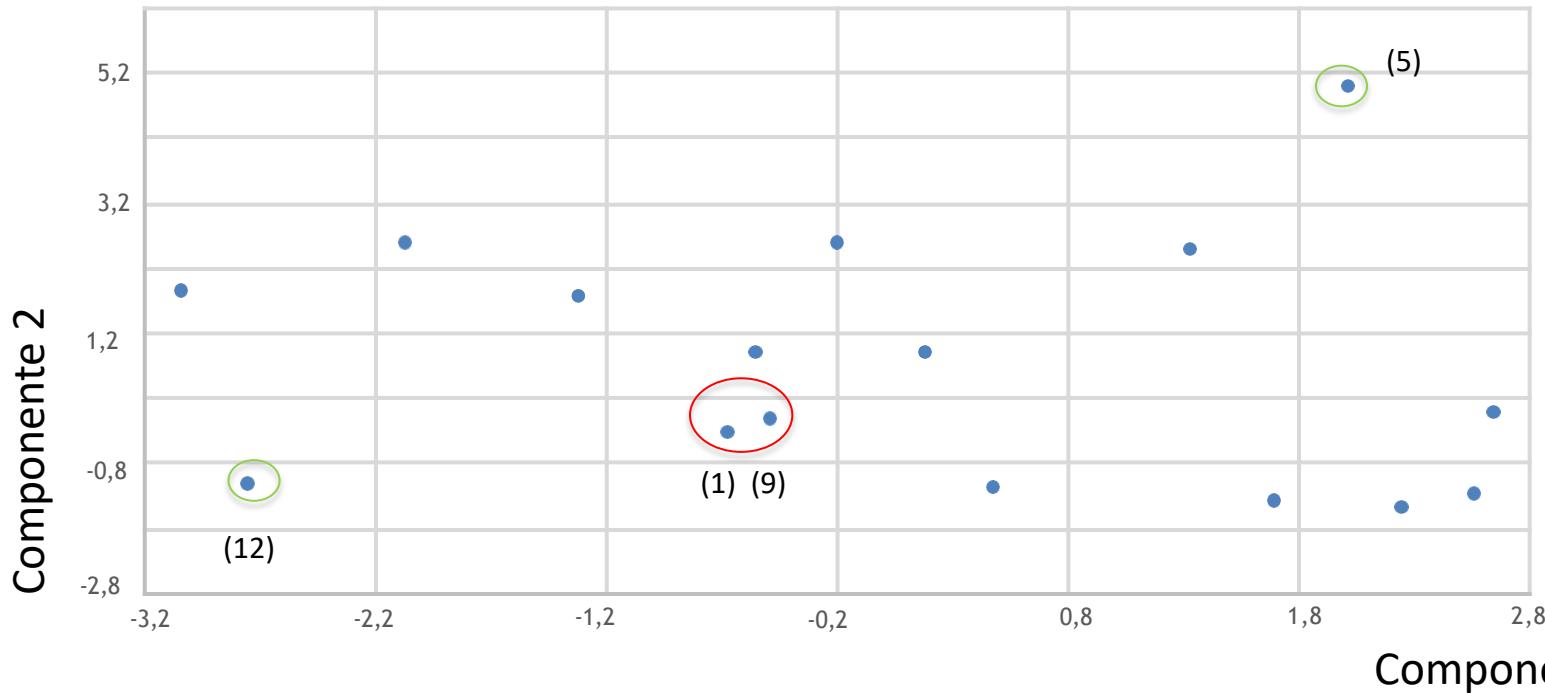
Asignatura	Componente 1	Componente 2
Lengua	0,500113	0,0853043
Matemáticas	-0,112909	0,555049
Física	-0,0517681	0,574789
Inglés	0,498752	0,036556
Filosofía	0,450292	0,121881
Historia	0,49264	0,0635768
Química	-0,0726488	0,573763
E. Física	0,187002	-0,0694516

## Interpretaciones:

- En la primera componente, influye mas: Lengua, Inglés, Filosofía, Historia. Y lo hacen positivamente. Los estudiantes son buenos en esas asignaturas. ¿Por lo tanto?
- En la segunda componente, influye mas: Matemáticas, Física, Química. ¿Cómo lo hacen? ¿Qué se puede decir de la componente?
- ¿Qué ocurre con la variable de Educación Física?

# Análisis de Componentes Principales (PCA)

Para representar al primer alumno, se aplican las formulas de ambas componentes a sus valores originales.



Alumno 1 y 9 están próximos. Alumno 5 y 12 están en extremos.

# Análisis de Componentes Principales (PCA)

Comprobamos las hipótesis con los datos originales.

ALUMNO	LENGUA	MATEMÁTICAS	FÍSICA	INGLÉS	FILOSOFÍA	HISTORIA	QUÍMICA	E. FÍSICA
1	5	5	5	5	5	5	5	5
2	7	4	3	8	4	7	3	8
3	5	8	7	6	5	6	7	5
4	7	2	4	8	7	7	3	6
5	8	9	10	8	8	7	9	4
6	4	9	8	4	3	4	7	5
7	6	4	4	6	5	5	3	7
8	4	7	8	3	3	2	8	3
9	5	5	4	5	6	5	5	1
10	7	4	5	7	8	8	4	6
11	7	8	8	7	7	6	7	9
12	4	3	3	4	3	2	1	4
13	7	4	4	7	8	7	4	5
14	3	5	5	2	3	3	5	7
15	5	6	6	5	5	5	6	6

# Análisis de Componentes Principales (PCA)

## Caso de las variables climatológicas de ciudades.

Temperatura Media	Precipitaciones	Velocidad Viento	Humedad	Temperatura Mínima Media	Temperatura Máxima Media	Altitud
14,7	450,2	2,5	67	2,5	20,17	290

Componente	Autovalores	% de varianza	% acumulado
1	3,61041		
2	1,8083		
3	0,834191		
4	0,461938		
5	0,207834		
6	0,0714179		
7	0,00590614		

# Análisis de Componentes Principales (PCA)

## Caso de las variables climatológicas de ciudades.

Temperatura Media	Precipitaciones	Velocidad Viento	Humedad	Temperatura Mínima Media	Temperatura Máxima Media	Altitud
14,7	450,2	2,5	67	2,5	20,17	290

Componente	Autovalores	% de varianza	% acumulado
1	3,61041	51,577	51,577
2	1,8083	25,833	77,410
3	0,834191	11,917	89,327
4	0,461938	6,599	95,926
5	0,207834	2,969	98,895
6	0,0714179	1,020	99,916
7	0,00590614	0,084	100,00

# Análisis de Componentes Principales (PCA)

Variable	Componente 1	Componente 2
Temperatura media	0,521913	-0,0282158
Precipitaciones	-0,0551481	0,590282
Humedad	0,0470988	0,621072
Velocidad viento	0,1257	-0,486115
Temp. Media max.	0,461348	0,0936507
Temp. Media min.	0,49067	-0,0585732
Altitud	-0,502939	-0,128579

¿Que podemos decir de las componentes?

# Análisis de Componentes Principales (PCA)

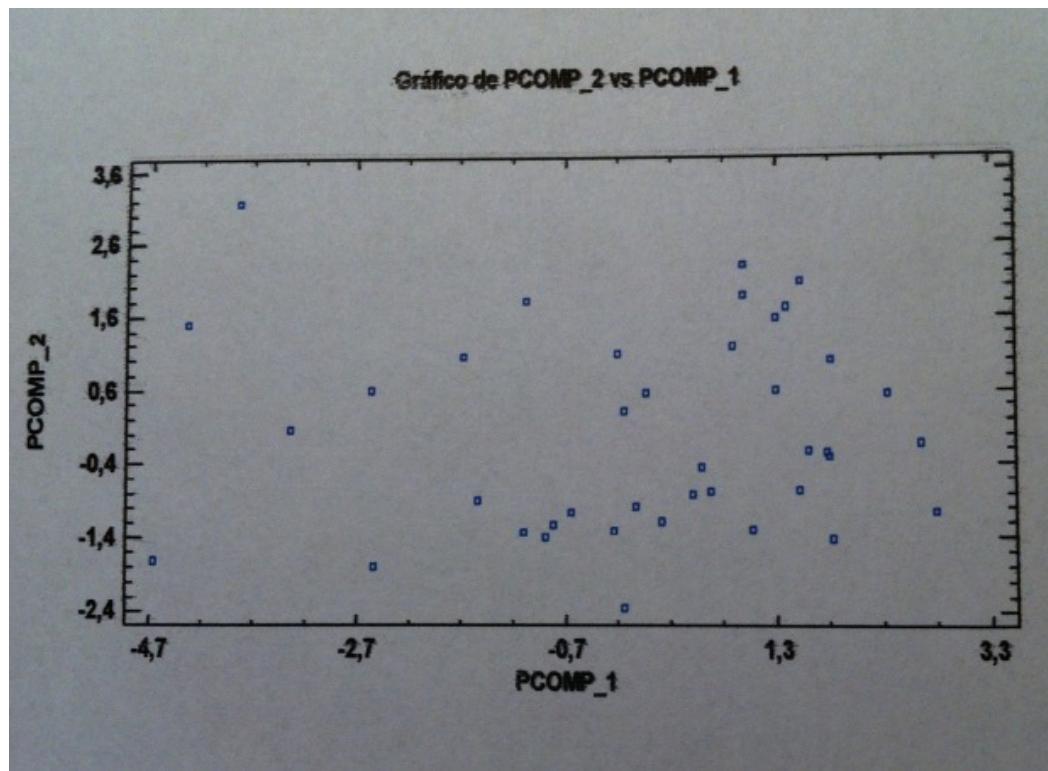
Variable	Componente 1	Componente 2
Temperatura media	0,521913	-0,0282158
Precipitaciones	-0,0551481	0,590282
Humedad	0,0470988	0,621072
Velocidad viento	0,1257	-0,486115
Temp. Media max.	0,461348	0,0936507
Temp. Media min.	0,49067	-0,0585732
Altitud	-0,502939	-0,128579

En la primera componente influyen: temperatura media, máxima, mínima y altitud (**Es negativa!!!!**).

En la segunda componente: precipitaciones, humedad y velocidad del viento.

# Análisis de Componentes Principales (PCA)

Al llevarlo a nuestro nuevo espacio con 2 componentes.  
¿Qué podemos interpretar?



# Análisis de Componentes Principales (PCA)

## Caso de las variables de equipos de fútbol.

Equipo	Puntos	Ganados	Empatados	Perdidos	Goles a favor	Goles en contra
Barcelona	87	27	6	5	105	35
R.Madrid	78	25	3	10	83	52
Sevilla	70	21	7	10	54	39
Atlético	67	20	7	11	80	57
Villarreal	65	18	11	9	61	54
Valencia	62	18	8	12	68	54
Deportivo	58	16	10	12	48	47
Málaga	55	15	10	13	55	59
Mallorca	51	14	9	15	53	60
Espanyol	47	12	11	15	46	49

# Análisis de Componentes Principales (PCA)

## Caso de las variables de equipos de fútbol.

Almería	46	13	7	18	45	61
Racing	46	12	10	16	49	48
Athletic	44	12	8	18	47	62
Sporting	43	14	1	23	47	79
Osasuna	43	10	13	15	41	47
Valladolid	43	12	7	19	46	58
Getafe	42	10	12	16	50	56
Betis	42	10	12	16	51	58
Numancia	35	10	5	23	38	69
Recre	33	8	9	21	34	57

# Análisis de Componentes Principales (PCA)

Caso de las variables de equipos de fútbol.

Componente	Autovalores	% de varianza	% acumulado
1	4,107		
2	1,511		
3	0,269		
4	0,113		
5	0,0000...		
6	0,0000...		

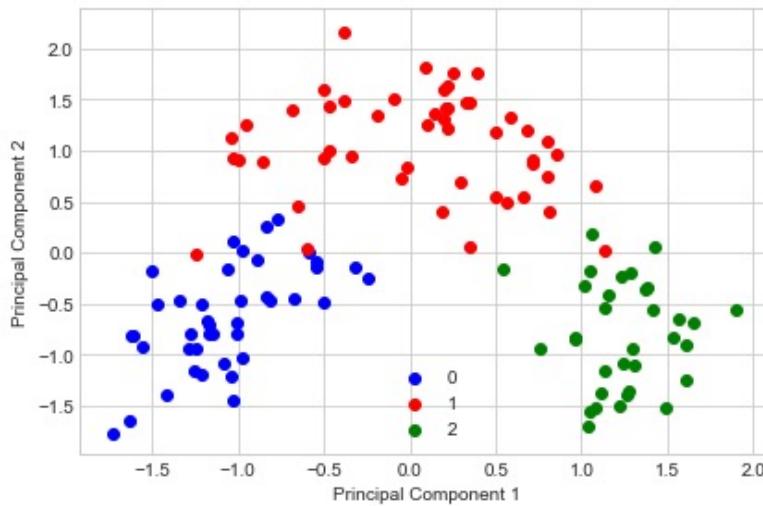
# Análisis de Componentes Principales (PCA)

Caso de las variables de equipos de fútbol.

Variable	Componente 1	Componente 2
Puntos	0,989	0,096
Ganados	0,955	0,276
Empatados	-0,181	-0,953
Perdidos	-0,953	0,313
GolesFavor	0,919	0,198
GolesContra	-0,682	0,617

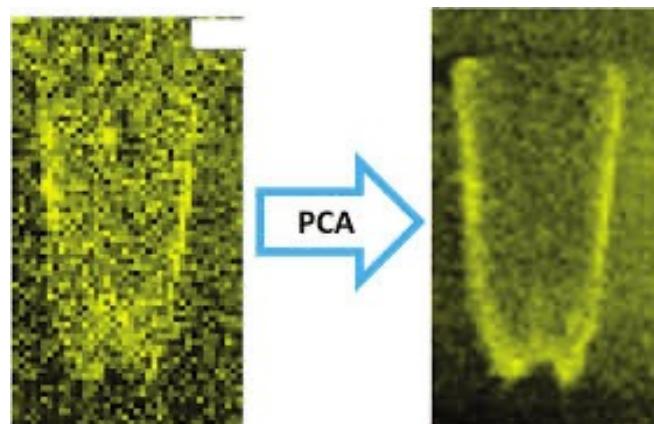
# Análisis de Componentes Principales (PCA)

- **Uso 1:** Reducción de la dimensionalidad. Obtenemos un conjunto de componentes que nos permite mapear los datos a otro espacio. Los componentes están ordenados de acuerdo a su importancia. Se proyectan los datos en unos pocos componentes.



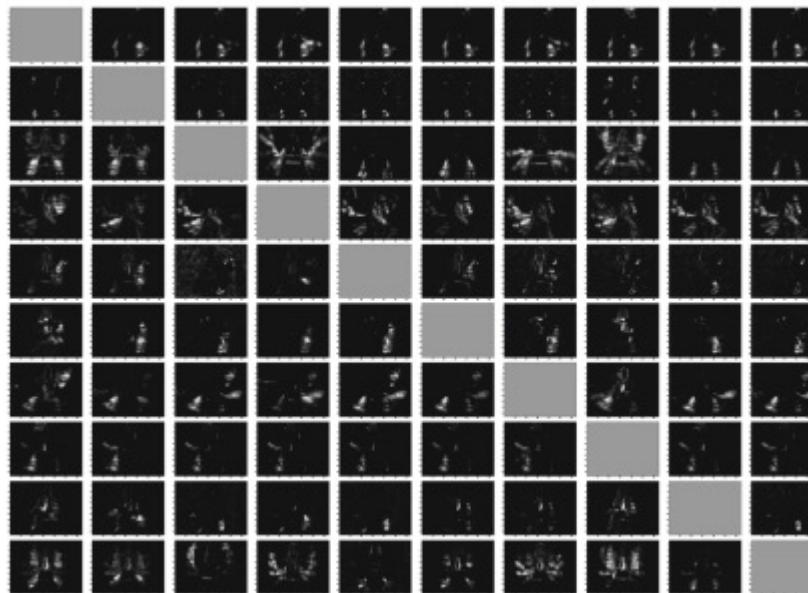
# Análisis de Componentes Principales (PCA)

- **Uso 2:** Eliminar ruido, información no relevante de los datos. Obtener las componentes principales y reconstruir los datos originales.



# Análisis de Componentes Principales (PCA)

- **Uso 3:** Clasificar. Aplicando PCA a cada clase. Al clasificar un nuevo individuo, el que tenga el error mínimo (diferencia con respecto al resto), nos dirá que clase le corresponde.



# Reducción del dataset

- **Reducción del número de casos:** consiste en encontrar una muestra, un subconjunto original de casos, que muestre un comportamiento parecido.

Obtención incremental de la muestra. Empezar por un conjunto de casos aleatorios (10%). Evaluar los resultados e incrementar el número de casos.