TEMA 2: PRE-PROCESAMIENTO DE DATOS
ANÁLISIS DE CORRELACIONES
TRANSFORMACIÓN DEL TIPO DE VARIABLES
GENERACIÓN DE VARIABLES

Motivación

- La elección (generación) de las variables utilizadas en el aprendizaje del modelo es un paso fundamental para su éxito
  - Selección de variables: eligen algunas de las variables del data set inicial
    - Lo veremos más adelante, en esta lección vamos a ver el análisis de correlaciones
  - Transformación de variables:
    - Tipo de variable
    - Datos almacenados de la variable
  - Generación de variables: construyen nuevas variables a partir de las originales
    - PCA

Análisis de correlaciones

- Análisis de correlaciones para variables numéricas: su objetivo es cuantificar la fuerza con la que una variable se obtiene a partir de otra

$$r_{A,B} = \frac{\sum_{i=1}^{n}(A_i - \pi_A) * (B_i - \pi_B)}{n * \sigma_A * \sigma_B}$$

  - A_i, B_i: son los valores *i*-ésimos de las variables A y B
  - π_A, π_B: son la media de los valores de las variables A y B
  - σ_A, σ_B: son las desviaciones estándar de los valores de las variables A y B
- $r_{A,B} > 0$ → A y B están correlacionadas positivamente (ambas tienen comportamiento similar)
- $r_{A,B} = 0$ → A y B son independientes
- $r_{A,B} < 0$→ A y B están correlacionadas negativamente (si una variable crece, la otra decrece)

Matriz de correlaciones

	x	*x²*	*1/x*
x	1		
x²	0.98	1	
1/x	-0.9	-0.81	1

Análisis de correlaciones

- Análisis de correlaciones para variables categóricas: test de correlación χ^2

$$\chi^2{}_{A,B} = \sum_{\{i=1\}}^{C} \sum_{\{j=1\}}^{r} \frac{\left(o_{ij}-e_{ij}\right)^2}{e_{ij}}$$

  - $o_{ij}: cuenta\left(A = A_i, B = B_j\right)$
  - $e_{ij} = \frac{cuenta(A=A_i)*cuenta(B=B_j)}{n}$
  - C y r son el número de valores diferentes de A y B
- Consular nivel de significancia de la tabla χ^2 con $(r-1)*(C-1)$ grados de libertad
  - Si el valor de la tabla es menor que el calculado, las variables A y B están correlacionadas

Análisis de correlaciones

- Para mejorar la interpretación del test de correlación χ^2
- Coeficiente de contingencia de Cramer (V de Cramer)

$$V = \sqrt{\frac{\chi^2}{n * (q-1)}}$$

  - n es el número de ejemplos (observaciones)
  - $q = \min(r, C)$
- V está en el rango [0, 1]
  - 0: independencia
- Ejemplo: http://asignatura.us.es/dadpsico/apuntes/ChiCuadrado.pdf

Transformación del tipo de variable
Numérica a categórica

- Discretización: Transforma los valores de las variables numéricas en un número finito de intervalos
  - A cada intervalo se le asocia una etiqueta (categoría)
  - Los intervalos producen una partición sin solapamiento de los ejemplos
    - Se transforma un atributo numérico en uno categórico
    - Puede ser visto como un método de reducción de datos
      - De muchos valores a unas pocas categorías

- Algunos métodos de aprendizaje no funcionan con atributos numéricos por lo que es imprescindible realizar esta transformación

Transformación del tipo de variable
Numérica a categórica

- Definición del proceso de discretización para aprendizaje supervisado (clasificación)
  - Dado un dataset con N ejemplos y C clases
  - Un algoritmo de discretización transformará un atributo numérico A en m intervalos

$$D = \{[d_0, d_1], (d_1, d_2], \dots, (d_{m-1}, d_m]\}$$

    - donde d_0 y d_m son los valores mínimo y máximo, respectivamente
    - $d_i < d_{i+1}$ para $i = \{0, \dots, m-1\}$
  - Al resultado discreto D se le llama esquema de discretización del atributo A
  - Al conjunto de valores $P = \{d_1, \dots d_{m-1}\}$ se le llama conjunto de puntos de corte del atributo A

Transformación del tipo de variable
Numérica a categórica

Técnicas de binning

- Discretización de anchura igual
  - Se elige el número de intervalos: m
  - Se divide el rango del atributo en m intervalos de anchura fija
    - $anchura = (valor_{máximo} - valor_{mínimo})/m$
    - $d_{i+1} - d_i = anchura$ con $i = \{0, \ldots, m-1\}$
  - También se puede especificar la anchura
    - Se obtiene m a partir de ella
    - Un intervalo puede ser de anchura diferente
  - Ejemplo: variable con rango [0, 10] a discretizar en 4 categorías

0.2	1	1.2	1.5	2.2	4	4.2	5.1	7	7.3	9.8

  - $anchura = \frac{10-0}{4} = 2.5$

0.2	1	1.2	1.5	2.2	4	4.2	5.1	7	7.3	9.8

Transformación del tipo de variable Numérica a categórica

Técnicas de binning

- Discretización de frecuencia igual
  - Se elige el número de intervalos: m
  - Se distribuyen los n ejemplos en los intervalos de tal forma que cada uno tenga aproximadamente el mismo número de ejemplos
    - $nEj = n/m$
    - nEj es el número de ejemplos que contendrá cada intervalo (categoría)
  - Ejemplo: variable con rango [0, 10] a discretizar en 4 categorías
    - $nEj = \frac{11}{4} = 2.75$

0.2	1	1.2	1.5	2.2	4	4.2	5.1	7	7.3	9.8

0.2	1	1.2	1.5	2.2	4	4.2	5.1	7	7.3	9.8

Transformación del tipo de variable
Numérica a categórica

Técnicas de binning

- Discretización de frecuencia fija (FFD)
  1. Se elige la frecuencia de cada intervalo: nEj
  2. Se asignan los ejemplos (ordenados previamente) a la categoría hasta alcanzar nEj
  3. Se crea una nueva categoría y se vuelven a asignar ejemplos
     - Se repiten 2 y 3 hasta que no queden ejemplos
  - El último intervalo puede tener un número diferente de ejemplos

  - Ejemplo: variable con rango [0, 10]

0.2	1	1.2	1.5	2.2	4	4.2	5.1	7	7.3	9.8

- Deseamos 4 ejemplos en cada categoría

0.2	1	1.2	1.5	2.2	4	4.2	5.1	7	7.3	9.8

Transformación del tipo de variable Categórica a numérica

- ***Codificación ordinal***
  - Algunas técnicas de aprendizaje automático no soportan variables categóricas
  - Transformar las variables categóricas a variables numéricas
  - La codificación ordinal consiste en transformar cada valor categórico en una valor entero
    - Esta solución presenta dos problemas importantes
      - Se asume un orden de los valores categóricos
      - Los nuevos valores enteros pueden ser utilizados para operaciones posteriores
        - La variable inicial no lo permite

Transformación del tipo de variable
Categórica a numérica

- ***One hot encoding***
  - Para evitar los problemas anteriores una transformación muy habitual es generar un conjunto de variables binarias por cada variable categórica
    - Sea N el número de valores de la variable categórica
    - Se generan N variables binarias
      - Una por cada valor
      - Cada una contiene 1 como valor de los ejemplos con el valor correspondiente a la nueva variable binaria
        - 0 en el resto de posiciones
  - Esta transformación también se conoce como transformación 1-N
  - Problema: la variable original tiene una cardinalidad alta (muchos valores)
    - Implica generar muchas variables que conllevan un conjunto de datos muy disperso
      - Problemas de rendimiento y numéricos

Transformación del tipo de variable Categórica a numérica

- Ejemplo de one hot encoding

Color	Rojo	Verde	Azul
Rojo	1	0	0
Verde	0	1	0
Azul	0	0	1
Rojo	1	0	0
Azul	0	0	1
Azul	0	0	1

Transformación del tipo de variable
Categórica a numérica

- ***Codificación binaria (Binary encoding)***
  - Los valores categóricos se codifican utilizando codificación ordinal
  - Los números obtenidos se codifican en binario
  - Se generan tantas variables binarias como dígitos de la codificación binaria
    - El número binario se copia dígito a dígito a las nuevas variables

Color	Ordinal	Cod. Binaria	0	1
Rojo	0	00	0	0
Verde	1	01	0	1
Azul	2	10	1	0
Rojo	0	00	0	0
Azul	2	10	1	0
Azul	2	10	1	0

Transformación del tipo de variable
Categórica a numérica

- Codificación de conteo
  - Sustituye cada valor por el número de veces que aparece dicho valor en el dataset

Color	Color’
Rojo	2
Verde	3
Verde	3
Rojo	2
Verde	3
Azul	1

Transformación del tipo de variable Categórica a numérica

- **Transformación basada en la salida**
  - Variable categórica (2 clases)
    - Se calcula la probabilidad de una clase para cada valor de la variable categórica
      - Se asigna dicha probabilidad como valor numérico

Trend	Target	Trend_Encoded
Up	1	0.66
Up	1	0.66
Down	0	0.33
Flat	0	0.5
Down	1	0.33
Up	0	0.66
Down	0	0.33
Flat	0	0.5
Flat	1	0.5
Flat	1	0.5

	Target		
Trend	0	1	Probability (1)
Up	1	2	0.66
Down	2	1	0.33
Flat	2	2	0.5

Transformación del tipo de variable Categórica a numérica

- **Transformación basada en la salida**
  - Variable numérica
    - Se calcula la agregación de los valores de salida para cada valor de la variable categórica
      - Se asigna dicho valor agregado como valor numérico

Trend	Target	Trend_Encoded
Up	21	23.7
Up	24	23.7
Down	8	10.3
Flat	15	14.5
Down	11	10.3
Up	26	23.7
Down	12	10.3
Flat	16	14.5
Flat	14	14.5
Flat	13	14.5

Trend	Target - Average
Up	23.7
Down	10.3
Flat	14.5

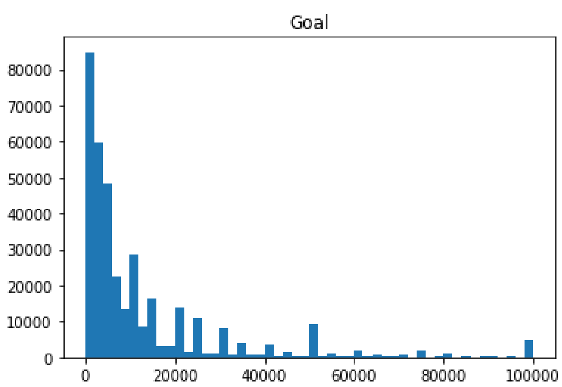
Transformación de los datos de una variable

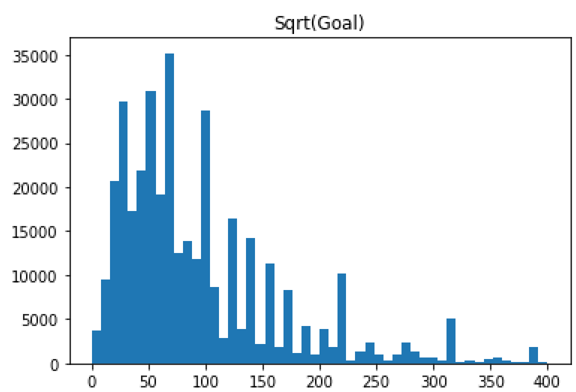
- Proceso por el que se cambia el contenido de una variable para que permita mejorar la calidad de los datos
  - Transformaciones habituales: raíz cuadrada, logaritmo
- Ejemplo: precio (goal) de los proyectos realizados

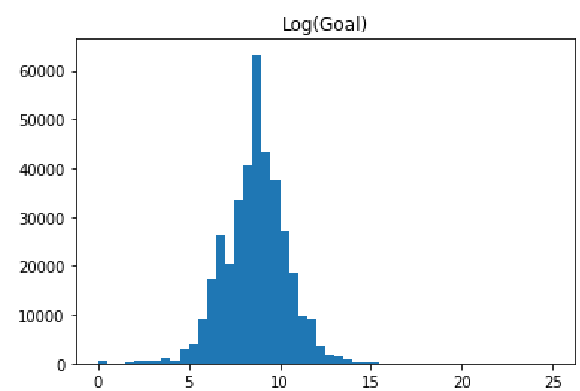
```
plt.hist(ks.goal, range=(0, 100000), bins=50);
plt.title('Goal');
```

```
plt.hist(np.sqrt(ks.goal), range=(0, 400), bins=50
plt.title('Sqrt(Goal)');
```

```
plt.hist(np.log(ks.goal), range=(0, 25), bins=50);
plt.title('Log(Goal)');
```







Generación de variables

- La normalización puede no ser suficiente para mejorar el modelo aprendido
- Puede ser beneficioso agregar la información de varias variables
  - Transformaciones lineales
    - Sea $B = \{B_1, \dots, B_m\}$ un subconjunto de m variables del conjunto total de variables $A = \{A_1, \dots, A_n\}$ con $m \leq n$

$$Z = r_1 * B_1 + r_2 * B_2 + \ \dots + r_m * B_m$$

      - donde r_i es el peso de la variable i-ésima de B
      - caso simple (media aritmética) $r_i = \frac{1}{m} \ \forall \ i \in \{1, \dots, m\}$
  - Transformaciones polinómicas
  - Transformaciones no polinómicas
  - Interacciones entre variables discretas
    - Unión de los términos de las diferentes variables

Generación de variables: PCA

- El algoritmo Principal Components Analysis (PCA) es uno de los métodos más antiguos y utilizados para transformar los datos y reducir su dimensionalidad

- Técnica multi-variante que transforma las variables originales $(X_1, X_2, \dots, X_n)$
  - En otro conjunto de variables $(CP_1, CP_2, \dots, CP_n)$
  - Las nuevas variables CP_i se denominan componentes principales
    - Son perpendiculares entre ellas
    - Forman una nueva base con un nuevo origen de coordenadas

- Para realizar la proyección de un ejemplo X en cada componentes principal hay que realizar una combinación lineal de las variables iniciales
  - $CP_1 = (a_{11}, a_{12}, \dots, a_{1n})$
    - $w_{CP_1} = a_{11} * X_1 + a_{12} * X_2 + \cdots + a_{1n} * X_n$
  - …
  - $CP_n = (a_{n1}, a_{n2}, \dots, a_{nn})$
    - $w_{CP_n} = a_{n1} * X_1 + a_{n2} * X_2 + \cdots + a_{nn} * X_n$

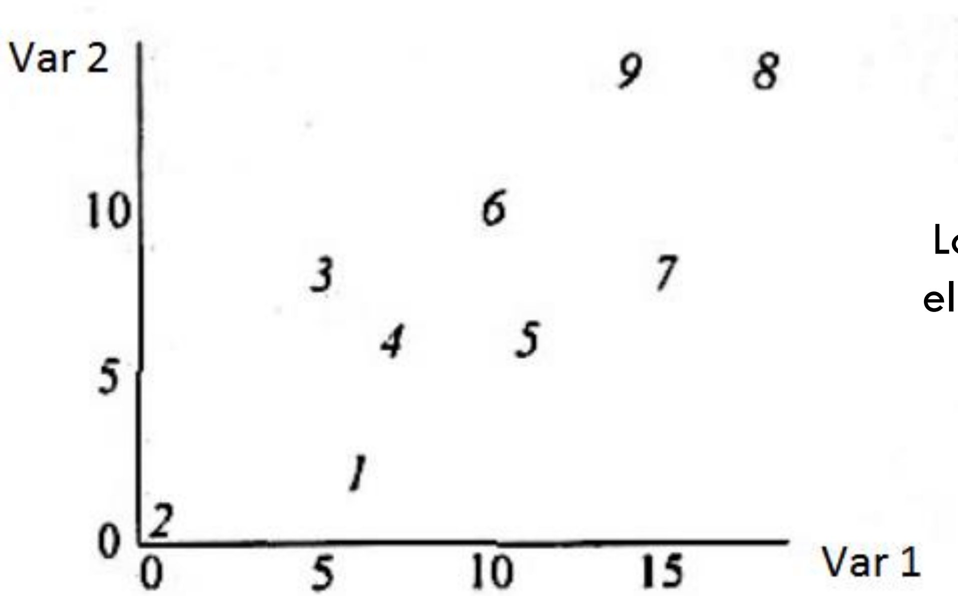
    - $a_{i,j}$ es el coeficiente correspondiente a la variable j en el componente i

- Sintetizan la mayor parte de la información contenida en los datos originales
  - La mayor parte de la varianza

Generación de variables: PCA

- Idea intuitiva

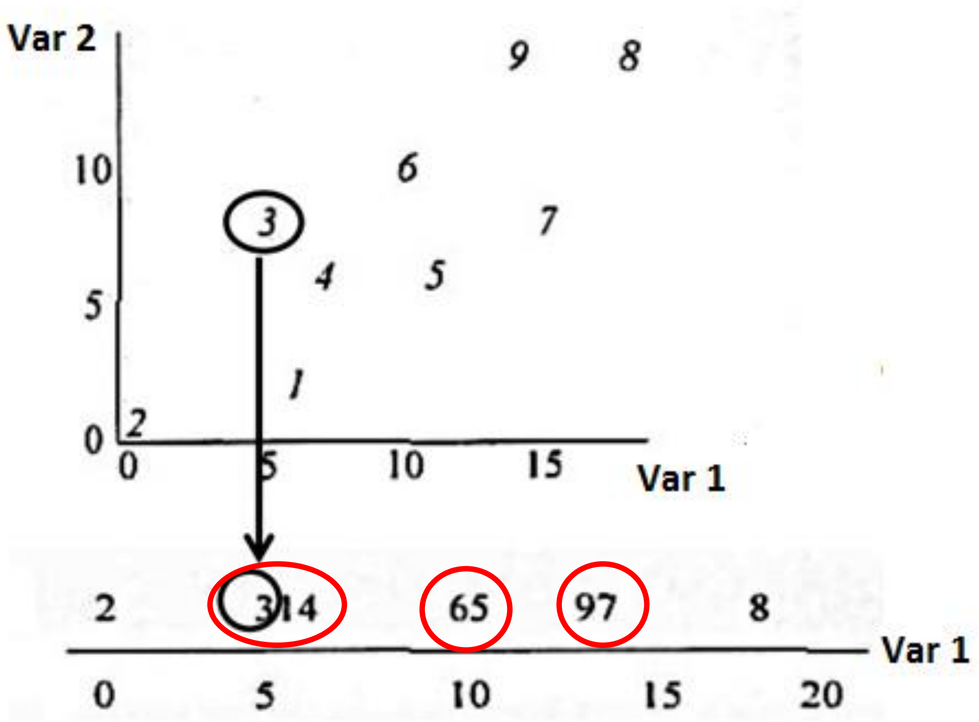
Ejemplo	1	2	3	4	5	6	7	8	9
Var 1	6	0	5	7	11	10	15	18	14
Var 2	2	0	8	6	6	10	8	14	14



Los números representan el índice de cada ejemplo

Generación de variables: PCA

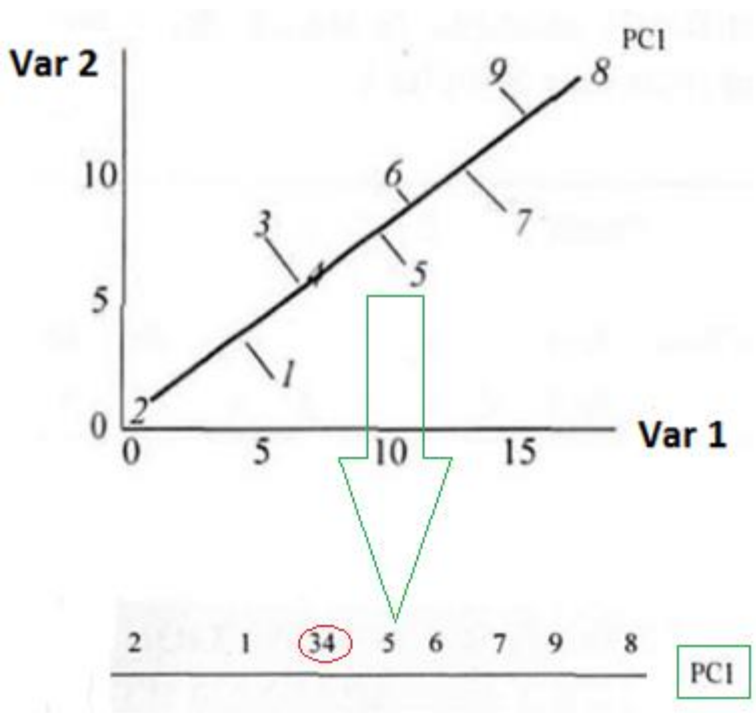
- Para comprobar la separabilidad de los ejemplos
  - Proyección de los datos a una dimension
    - Por ejemplo utilizando la variable 1
    - Con la variable 2 tenemos una situación similar



Datos poco separables

Generación de variables: PCA

- Sería bueno transformar los datos de tal forma que se mejorase la separabilidad de los ejemplos
  - Idea: generar una variable que maximice la varianza de los ejemplos si son proyectados perpendicualrmente a ella

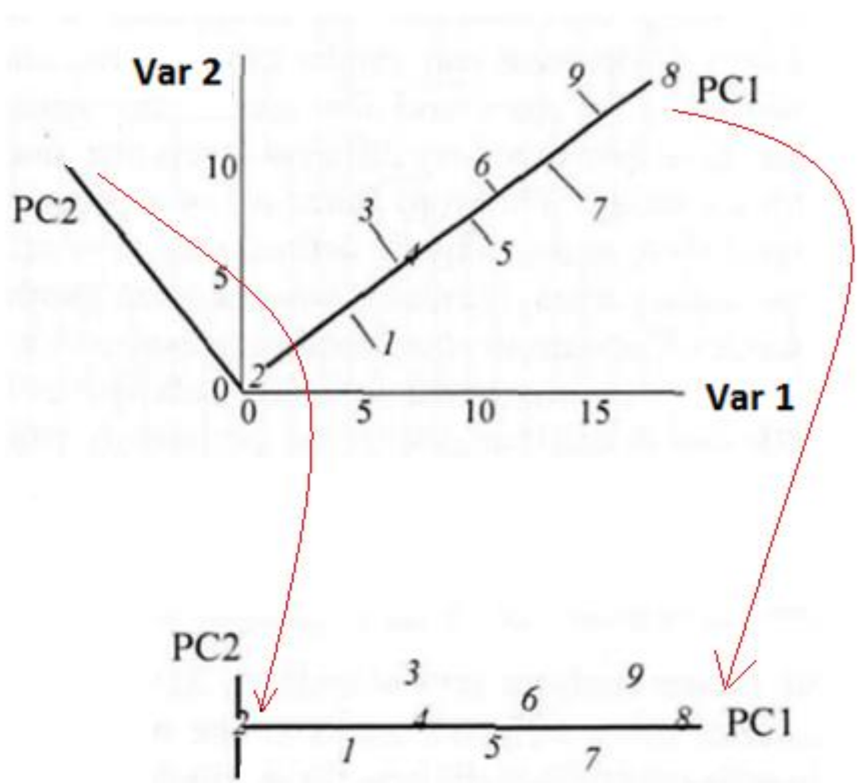


PC1: Componente principal 1

Datos más separables

Generación de variables: PCA

- Podemos generar otra variable (componente principal) ya que originalmente teníamos dos variables
  - Restricción: debe ser perpendicular a la primera para que formen una base
  - También maximiza la varianza de los ejemplos si se proyectan hacia ella



Generación de variables: PCA

- Procedimiento para generar las variables (componentes principales)
  - Normalizar cada variable para que tenga media 0 y desviación estándar 1
    - De esta forma las variables con rangos más grandes no se verán favorecidas (variarían más)
    - Se obtiene el dataset normalizado DN
  - Calcular la matriz de covarianzas, $C = DN^T * DN$
    - Dimensión: número de variables por número de variables
  - Obtener los vectores y valores propios de C
    - Se obtienen tantos como variables
    - Cada vector propio es una componente principal
    - Cada valor propio está asociado a un vector propio
      - Representa la importancia (varianza) del vector propio

Generación de variables: PCA

- Para reducir el número de variables
  - Se ordenan los vectores propios de acuerdo a sus valores propios
  - Se escogen aquellos que representen un determinado porcentaje de la varianza
    - Se normalizan los valores propios
      - Cada valor propio se divide por la suma de todos los valores propios
    - Se escogen tantos como sea necesarios para alcanzar la varianza deseada
      - Se van acumulado los valores propios normalizados asociados a los vectores propios utilizados
    - Finalmente, los ejemplos originales se proyectan sobre las nuevas variables para obtener los nuevos valores
      - Por cada ejemplo se recorren todas las componentes principales
        - Producto matricial entre el ejemplo y el componente principal: un valor (coordenada)
        - Al final, para cada ejemplo, tenemos tantos valores como componentes

- Habitualmente se escogen las componentes principales necesarias para mantener el 95% o más de la varianza del dataset original
- El PCA es útil cuando existen muchas variables independientes con una correlación alta

Generación de variables: PCA

- Análisis de la influencia de las variables originales en un componente principal
  - Realizar la correlación entre
    - Los ejemplos proyectados sobre el componente principal
    - Los ejemplos en cada variable original
  - Los variables originales con mayor correlación (positiva o negativa) serán las más influyentes en el componente principal