# Introducción al cálculo numérico y ciencia de datos

## Clase 4

Mg. Víctor Viana

Dr. Diego Passarella

Tacuarembó – abril de 2023

EDUCACION PERMANENTE
Universidad de la República

SEDE TACUAREMBÓ
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

# Minería de datos (Data Mining)

- La minería de datos es el proceso de descubrir patrones, relaciones y tendencias útiles a partir de grandes conjuntos de datos.
- Es una disciplina interdisciplinaria que involucra la estadística, el aprendizaje automático, la inteligencia artificial y la informática.
- El objetivo de la minería de datos es convertir los datos brutos en información útil, a menudo utilizando técnicas de análisis estadístico y aprendizaje automático para identificar patrones o estructuras en los datos.
- Estos patrones pueden utilizarse para realizar predicciones, tomar decisiones informadas y optimizar procesos empresariales.

# Reducción de Dimensionalidad

- Trabajar con datasets muy anchos (mucha dimensiones) genera una gran cantidad de problemas, más allá del costo computacional.
- Esto se conoce como la “maldición de la dimensionalidad”.
- Al aumentar la dimensionalidad del espacio, la densidad de los datos baja exponencialmente, por lo que la estabilidad y robustez de las técnicas de análisis se compromete, y la significatividad estadística de los resultados se debilita.
- Se requiere recolectar una enorme cantidad de datos de entrenamiento para garantizar un cubrimiento razonable de los posibles casos.

# Reducción de Dimensionalidad(II)

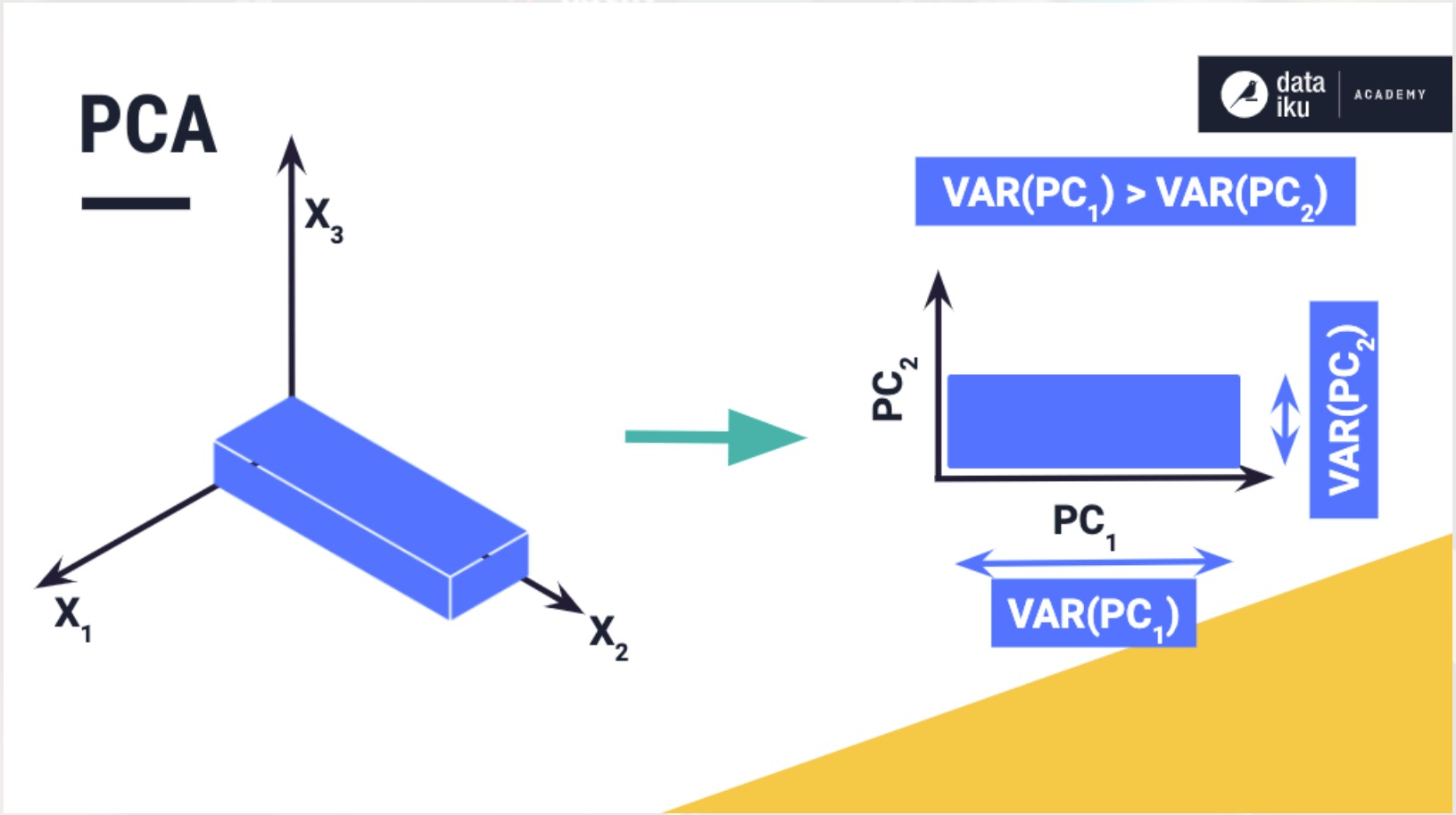
- Cuando la cantidad de dimensiones es muy alta las métricas de distancia pierden las propiedades intuitivas, y los métodos basados en distancias (k-NN por ejemplo) se sesgan muy rápidamente.
- Además, durante las tareas de “curado” del dataset y análisis exploratorio, muchas veces es necesario visualizar los datos de alguna manera razonablemente “accionable”.
- Finalmente, datos con muchos atributos tienen mucha probabilidad de tener atributos irrelevantes, valores faltantes o contaminados, etc., por lo que se generan modelos con mucha variancia y el sobreajuste es difícil de controlar.

# Reducción de Dimensionalidad(III)

- La **reducción de la dimensionalidad** consiste en transformar el dataset original en otro de menor cantidad de atributos (tabla “más angosta”), pero que retenga las propiedades significativas del original con respecto al propósito de análisis.
- Muy crudamente, estos métodos pueden catalogarse en lineales y no lineales, y varias de las técnicas que ya conocemos pueden utilizarse directa o indirectamente para reducir dimensionalidad.

# Análisis de Componentes Principales

- El **Análisis de Componentes Principales** (**PCA** en ingles) es una técnica de **reducción de dimensionalidad** utilizada en estadística y aprendizaje automático para extraer información importante de conjuntos de datos complejos.
- El objetivo del PCA es reducir el número de variables en un conjunto de datos, manteniendo al mismo tiempo la mayor cantidad posible de información contenida en los datos.
- El PCA logra esto transformando las variables originales en un conjunto más pequeño de variables llamadas componentes principales, que son combinaciones lineales de las variables originales.




Extraído de: https://knowledge.dataiku.com/latest/ml-analytics/statistics/concept-principal-component-analysis-pca.html

# Análisis de Componentes Principales(II)

- El PCA es útil para representar y visualizar datos en un espacio dimensional reducido de variables no correlacionadas que maximizan las variaciones existentes en los datos.
- Estos atributos del PCA también lo hacen útil para el preprocesamiento de datos (o procesamiento de características) previo a la construcción de modelos, ya que reducir las características de un conjunto de datos puede mejorar el rendimiento del entrenamiento.

# Varianza Explicada

- La varianza explicada es una medida que se utiliza en estadística para evaluar la cantidad de variabilidad en un conjunto de datos que es explicada por un modelo o por una variable.
- En el contexto del PCA, la varianza explicada se refiere a la cantidad de variabilidad en los datos originales que es explicada por cada componente principal en el modelo.
- Cada componente principal en un modelo de PCA explica una fracción de la varianza total en los datos.
- La varianza explicada se calcula como la proporción de la varianza total de los datos que se explica por un componente principal específico.

# Varianza Explicada(II)

- La varianza explicada es importante porque nos ayuda a determinar cuánta información se puede extraer de un conjunto de datos utilizando el modelo del PCA.
- También puede ayudarnos a determinar cuántos componentes principales debemos incluir en el modelo para explicar una cantidad razonable de variabilidad en los datos.

# Varianza Explicada - Ejemplo

- Supongamos que tenemos un conjunto de datos con 4 variables y deseamos realizar un PCA para reducir la dimensionalidad de los datos.
- Al aplicar el PCA, obtenemos los siguientes resultados:
  - El primer componente principal explica el 50% de la varianza total en los datos.
  - El segundo componente principal explica el 30% de la varianza total en los datos.
  - El tercer componente principal explica el 15% de la varianza total en los datos.
  - El cuarto componente principal explica el 5% de la varianza total en los datos.

# Varianza Explicada – Ejemplo (II)

- La varianza explicada en este caso se puede interpretar como la proporción de la varianza total en los datos que se explica por cada componente principal.
- Por ejemplo, el primer componente principal explica el 50% de la varianza total en los datos, lo que significa que el 50% de la variabilidad en los datos se puede explicar mediante una combinación lineal de las variables originales que forman el primer componente principal.
- En términos de la interpretación práctica de la varianza explicada, podríamos decidir incluir solo los primeros dos componentes principales en nuestro modelo de PCA, ya que juntos explican el 80% de la varianza total en los datos.
- Al hacer esto, reduciríamos la dimensionalidad de los datos de 4 variables originales a solo 2 componentes principales, lo que podría facilitar el análisis y la interpretación de los datos.

# Vectores propios

Un vector propio es un vector no nulo que se mantiene en la misma dirección después de ser transformado por una matriz.

Matemáticamente, un vector propio v de una matriz A se define como:

$$A * v = \lambda * v$$

donde λ es un escalar conocido como valor propio (o autovalor) de A asociado con el vector propio v.

# Vectores propios(II)

- Los vectores propios son importantes en muchas aplicaciones de matemáticas, ciencias e ingeniería, ya que a menudo representan las direcciones principales de variabilidad o transformaciones importantes en un conjunto de datos.
- En el análisis de componentes principales (PCA), los vectores propios de la matriz de covarianza de los datos representan las direcciones principales de variabilidad en los datos, que se utilizan para construir los componentes principales del modelo.

# El algoritmo de PCA y ejemplos

Partimos de conocer un conjunto de ***n*** datos, los cuales son descriptos a través de valores en ***p***-dimensiones. La base de ese espacio ***p***-dimensional se denota por:

$$\{u_j\}, con\ 1 \leq j \leq p$$

un dato cualquiera se puede expresar como:

$$d^{(i)} = x_1^{(i)} u_1 + x_2^{(i)} u_2 + \cdots + x_p^{(i)} u_p, con\ 1 \leq i \leq n$$

con $\left\{x_j^{(i)}\right\}$ los coeficientes que almacenan la información de $d^{(i)}$ en cada una de las ***p***-dimensiones.

# Algoritmo PCA

Toda la información de los ***n*** datos representados en la base de ***p*** dimensiones, se puede almacenar en una matriz de datos observados ***X***, de dimensión *[n x p]*

$$X = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & \cdots & x_p^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & \cdots & x_p^{(2)} \\ x_1^{(3)} & x_2^{(3)} & x_3^{(3)} & & x_p^{(3)} \\ & \vdots & & \ddots & \vdots \\ x_1^{(n)} & x_2^{(n)} & x_3^{(n)} & \cdots & x_p^{(n)} \end{pmatrix}$$

Donde para cada columna se puede calcular el valor medio de los datos en esa dimensión

$$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_j^{(i)}$$

# Algoritmo PCA

Con la matriz de datos observados y los valores medios de cada dimensión se calcula la matriz de covarianza ***S***, matriz cuadrada de tamaño *[p x p]*, cuyos elementos son:

$$S_{rt} = \frac{1}{n}\sum_{i=1}^{n}\left(x_r^{(i)} - \bar{x}_r\right)\left(x_t^{(i)} - \bar{x}_t\right), con\ 1 \leq r, t \leq p$$

Cuando *r* = *t* tenemos la varianza de esa dimensión, mientras que cuando r ≠ t se tiene la covarianza entre esas dos dimensiones.

La varianza siempre es positiva y cuanto mayor, nos indica mayor variabilidad en esa dimensión

La covarianza entre dos dimensiones puede ser positiva, nula o negativa

# Algoritmo PCA

Con la matriz de datos observados y los valores medios de cada dimensión se calcula la matriz de covarianza ***S***, matriz cuadrada de tamaño *[p x p]*, cuyos elementos son:

$$S_{rt} = \frac{1}{n}\sum_{i=1}^{n}\left(x_r^{(i)} - \bar{x}_r\right)\left(x_t^{(i)} - \bar{x}_t\right), con\ 1 \leq r, t \leq p$$

Cuando *r* = *t* tenemos la varianza de esa dimensión, mientras que cuando r ≠ t se tiene la covarianza entre esas dos dimensiones.

La varianza siempre es positiva y cuanto mayor, nos indica mayor variabilidad en esa dimensión

La covarianza entre dos dimensiones puede ser positiva, nula o negativa

# Algoritmo PCA

La matriz de covarianza es simétrica y a su vez, se opta por normalizarla, de forma de pasar a ser la matriz de correlación ***R*** entre dimensiones

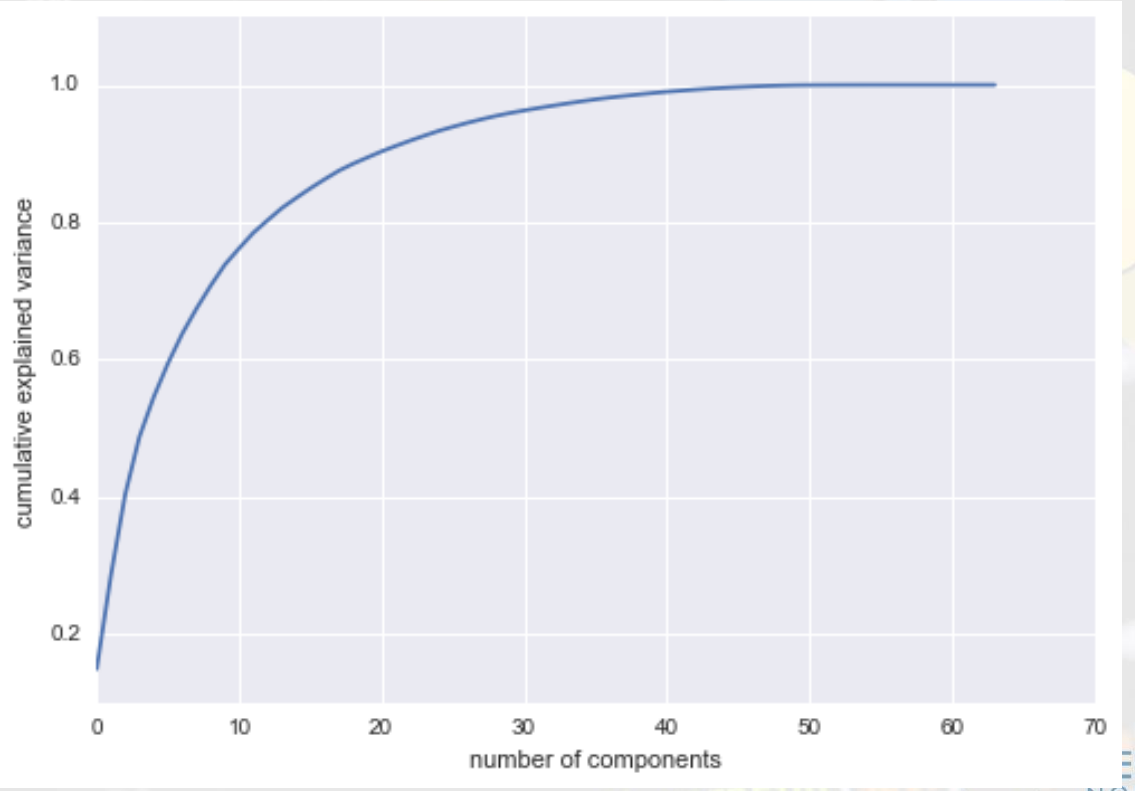
$$-1 \leq R_{tr} \leq 1, para\ t \neq r\ y\ R_{tt} = 1$$

De la matriz ***R*** se obtienen los ***p*** autovalores γ, con sus correspondientes autovectores ***w*** (que son ortonormales entre sí, resultan de una combinación lineal de la base ***v*** y a su vez forman una base del espacio de ***p***-dimensiones original)

$$\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_p$$

Con estos valores podemos calcular la varianza explicada acumulada a medida que consideramos más proyecciones

# Algoritmo PCA – Varianza explicada acumulada




Extraído de:
https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.09-Principal-Component-Analysis.ipynb#scrollTo=br1VcQus6eQ1

# Algoritmo PCA

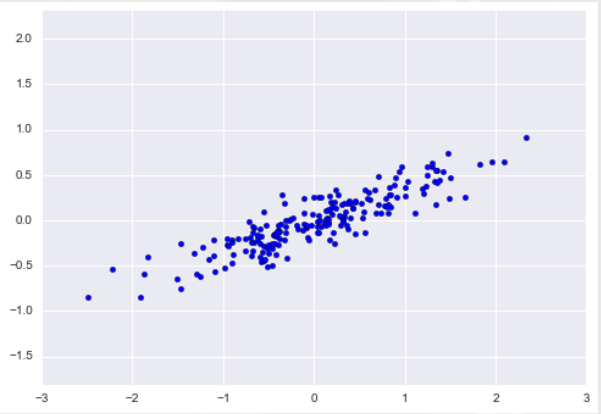
A partir de la curva de varianza explicada acumulada, podemos decidir cuántos ejes principales considerar en el análisis para generar la matriz de proyección ***W***.

Supongamos que los primeros ***q*** autovalores describen el 80% de la varianza de los datos. La matriz de proyección ***W*** estará formada por los autovectores correspondientes a los ***q*** autovalores más grandes, dispuestos en columnas, dando lugar a una matriz de dimensión *[p x q]*.

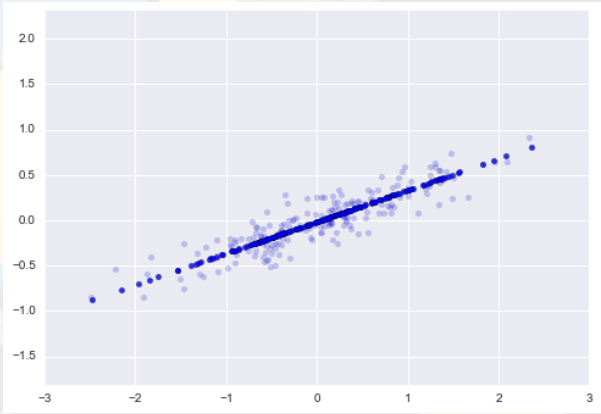
Finalmente, los datos proyectados en la base de componentes principales (reducida) resulta de:

$$Y = X\,W$$
$$[n \times q] = [n \times p][p \times q]$$

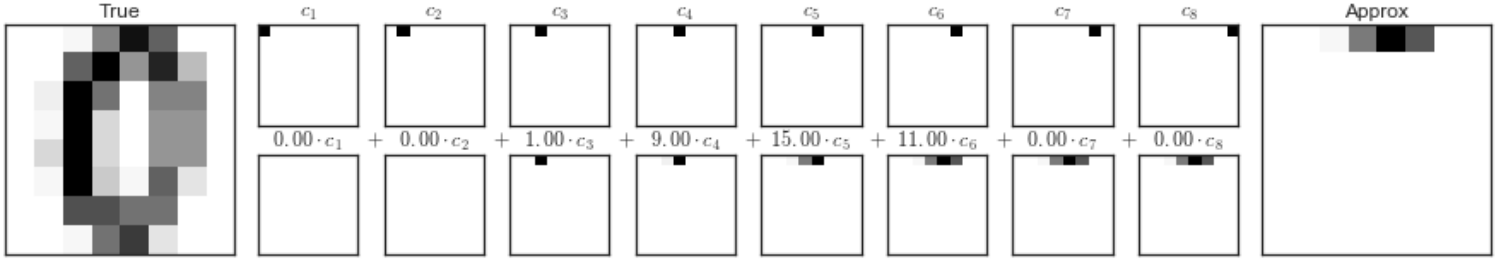
# Algoritmo PCA - Ejemplos

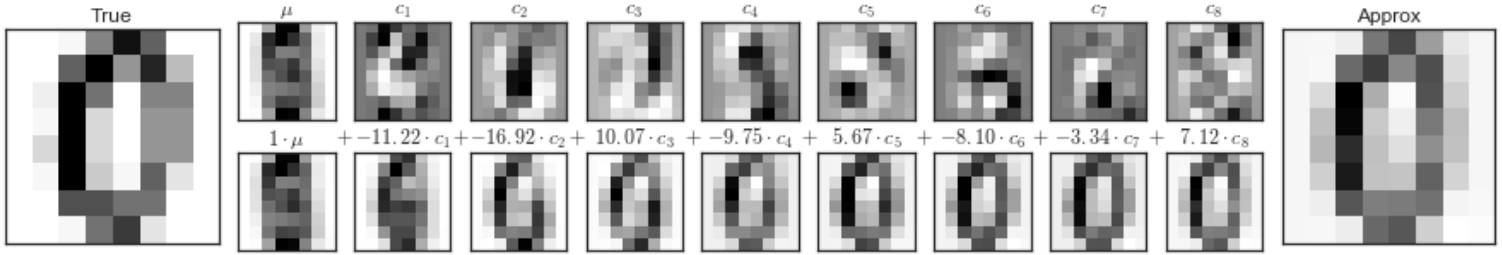




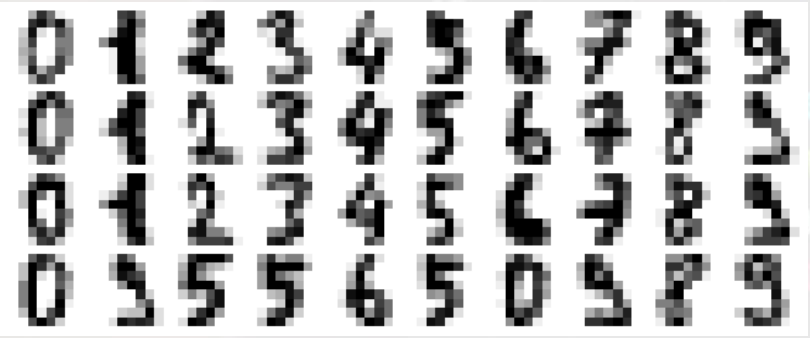


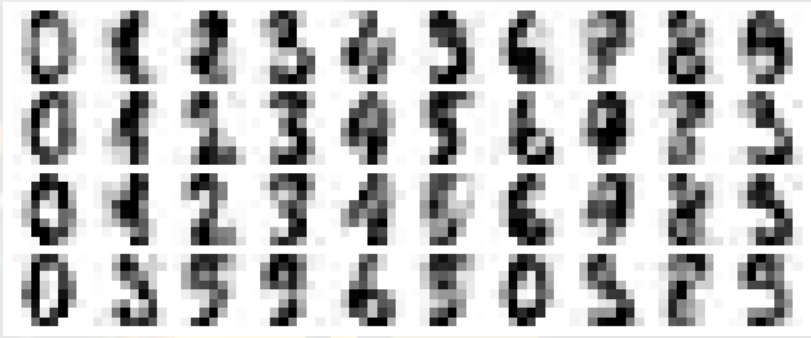
# Algoritmo PCA - Ejemplos



# Algoritmo PCA - Ejemplos





# Algoritmo PCA - Ejemplos

# Ejemplo – Iris flowers

Ver ejemplo en el Notebook: icncd_clase4_ejemplo_iris.ipynb

# Ejemplo Mínimo de PCA

Ver ejemplo en el Notebook: icncd_clase4_ejemplo1_pca.ipynb

# Ejemplo Mínimo de PCA

Ver ejemplo en el Notebook: icncd_clase4_ejemplo_diabetes.ipynb