# Reducción de datos

Héctor Fernando Gómez

Universidad del Caribe
Cancún, México

Febrero, 2021

# Contenido



# Contenido



# Introducción

En los proyectos en los que se dispone de una gran cantidad de datos es necesario reducir su volumen para eficientar su análisis.

En esta presentación exploraremos dos estrategias generales:

1. Muestreo
2. Reducción de la dimensionalidad

# Contenido



# Muestreo

Encontramos dos estrategias principales:

# Muestreo

Encontramos dos estrategias principales:

1. **Muestreo sesgado** con el que se busca compensar la presencia de una clase de observaciones, poco frecuente, sobre la que se tenga especial interés.

# Muestreo

Encontramos dos estrategias principales:

1. **Muestreo sesgado** con el que se busca compensar la presencia de una clase de observaciones, poco frecuente, sobre la que se tenga especial interés.
2. **Muestreo estratificado** con la que se busca preservar la proporción de observaciones de las diferentes clases que componen al conjunto original

# Contenido



# Reducción de la dimensionalidad

El objetivo es el de seleccionar un subconjunto de variables que brinden la suficiente información para conseguir los objetivos del modelado estadístico. Algunas opciones:

# Reducción de la dimensionalidad

El objetivo es el de seleccionar un subconjunto de variables que brinden la suficiente información para conseguir los objetivos del modelado estadístico. Algunas opciones:

1. Selección por significancia estadística

# Reducción de la dimensionalidad

El objetivo es el de seleccionar un subconjunto de variables que brinden la suficiente información para conseguir los objetivos del modelado estadístico. Algunas opciones:

1. Selección por significancia estadística
2. Factor de inflación de la varianza.

# Reducción de la dimensionalidad

El objetivo es el de seleccionar un subconjunto de variables que brinden la suficiente información para conseguir los objetivos del modelado estadístico. Algunas opciones:

1. Selección por significancia estadística
2. Factor de inflación de la varianza.
3. Análisis de componentes principales.

# Significancia estadística

Evaluada a través de pruebas de hipótesis a partir del siguiente proceso:

# Significancia estadística

Evaluada a través de pruebas de hipótesis a partir del siguiente proceso:

1. Suponer un modelo generativo.

# Significancia estadística

Evaluada a través de pruebas de hipótesis a partir del siguiente proceso:

1. Suponer un modelo generativo.
2. Identificar un estadístico que pueda asociarse con las variables.

# Significancia estadística

Evaluada a través de pruebas de hipótesis a partir del siguiente proceso:

1. Suponer un modelo generativo.
2. Identificar un estadístico que pueda asociarse con las variables.
3. Diseñar una prueba de hipótesis a partir de la distribución del estadístico (identificar hipótesis nula y alternativa).

# Significancia estadística

Evaluada a través de pruebas de hipótesis a partir del siguiente proceso:

1. Suponer un modelo generativo.
2. Identificar un estadístico que pueda asociarse con las variables.
3. Diseñar una prueba de hipótesis a partir de la distribución del estadístico (identificar hipótesis nula y alternativa).
4. Evaluar el estadístico sobre los datos y tomar una decisión acerca de la validez de la hipótesis.

# Factor de inflación de la varianza

Supón que se dispone de un conjunto de predictores, el **factor de inflación de la varianza** permite identificar a predictores redundantes. Para un predictor $X_i$ se define como:

$$FIV(X_i) = \frac{1}{1 - R_i^2}$$

siendo $R_i$ el coeficiente de determinación que se obtiene al modelar la variable $X - i$ en función del resto de los predictores.

# Análisis de componentes principales

Las ideas principales de la técnica de **análisis de componentes principales** son las siguientes:

# Análisis de componentes principales

Las ideas principales de la técnica de **análisis de componentes principales** son las siguientes:

- Buscamos generar **nuevas variables** aplicando una **transformación lineal** a las originales.

# Análisis de componentes principales

Las ideas principales de la técnica de **análisis de componentes principales** son las siguientes:

- Buscamos generar **nuevas variables** aplicando una **transformación lineal** a las originales.
- Geométricamente, las transformaciones definen **nuevos ejes coordenados.**

# Análisis de componentes principales

Las ideas principales de la técnica de **análisis de componentes principales** son las siguientes:

- Buscamos generar **nuevas variables** aplicando una **transformación lineal** a las originales.
- Geométricamente, las transformaciones definen **nuevos ejes coordenados.**
- Elegimos la transformación exigiendo que la **información proyectada** sobre los nuevos ejes coordenados sea **máxima**