Bienvenidos a Data Mining y Machine Learning!!!

Dra. Marcela Riccillo

ITBA

Datos

¿Qué es Data Mining y Machine Learning?

Modelos

Etapas

Primero que todo, entender el problema

Variables

Variable a Predecir

V1	V2	V3	V4	Sí/NO
x	x	x	x	x
x	x	x	x	x
x	x	x	x	x
x	x	x	x	x
x	x	x	x	x

Variables Predictoras

Casos a modelar

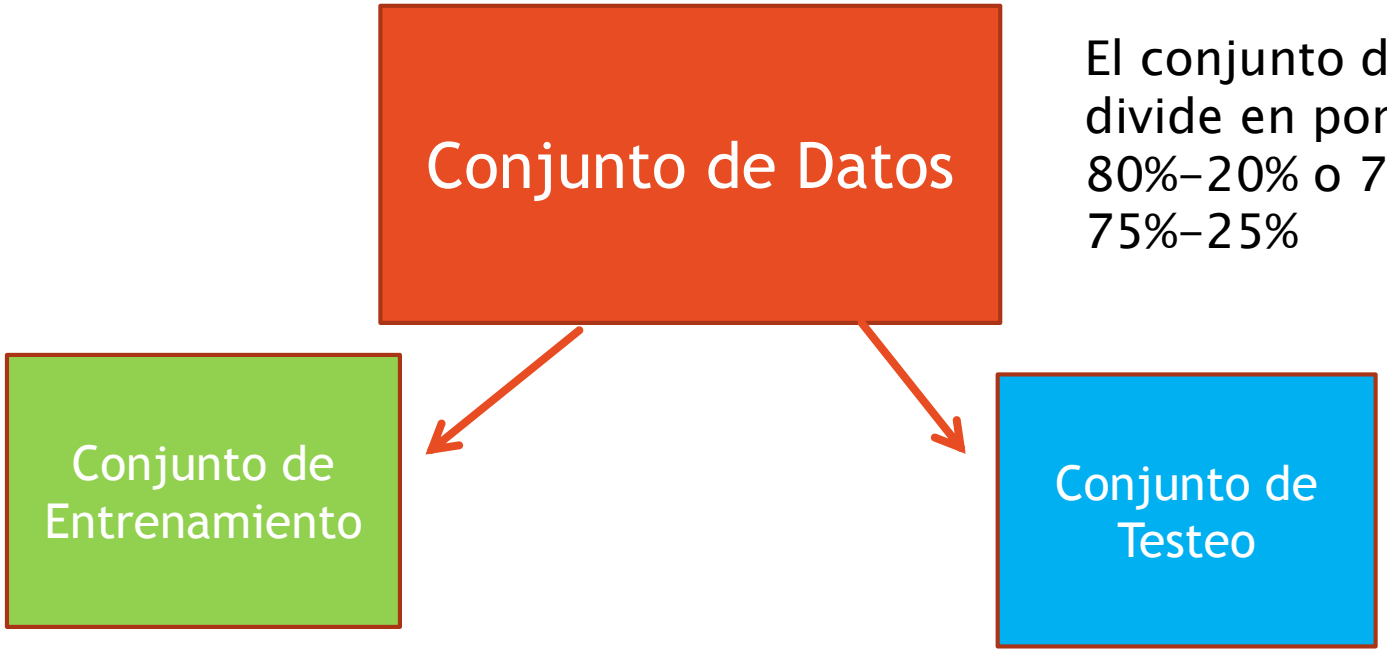
Regresión

Variable a predecir cuantitativa

Clasificación

Variable a predecir cualitativa

Conjuntos

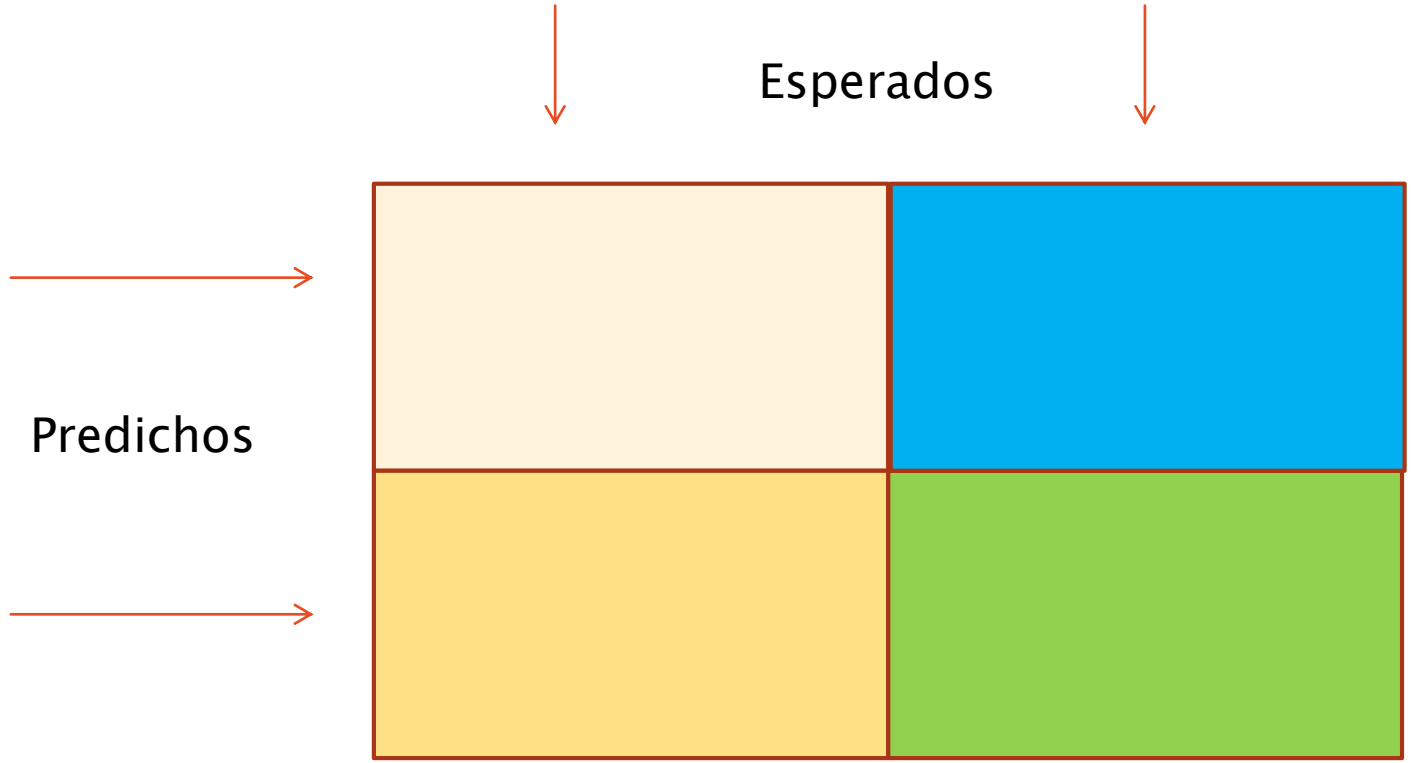


El conjunto de datos se divide en por ejemplo 80%–20% o 70%–30% o 75%–25%

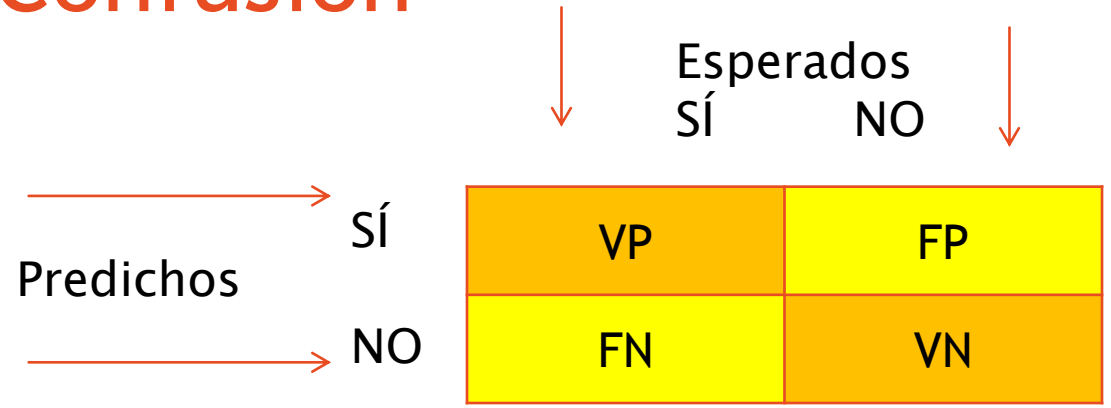
El Conjunto de testeo debe ser lo suficientemente
+ adecuado en tamaño
+ random
+ adecuado en variedad
+ independiente del conjunto de Entrenamiento

Detección de outliers

Matriz de Confusión



Matriz de Confusión



Accuracy = VP + VN / Todos
(Exactitud)
Accuracy = VP + VN / (VP + VN + FP + FN)

Sensibilidad = VP / (VP + FN)
(Recall)
Especificidad = VN / (VN + FP)

Práctica (1/3)

Dada una base:

Mostrar un resumen de las características de los datos

¿Qué tipos de variables se pueden identificar?

¿Cuántas variables tiene la base? ¿Cuántos registros? ¿Cuál es la variable a predecir?

¿Con qué variables podría hacer un histograma y con cuáles un gráfico de barras?

Realizar gráficos de dispersión coloreando por clases

Práctica (2/3)

Realizar gráficos de dispersión en 3D coloreando por clases

Realizar una matriz de gráficos de dispersión coloreando por clases

¿Qué es un outlier? ¿Qué métodos se pueden utilizar para detectar outliers?

Práctica (3/3)

Particionar los datos en un conjunto de entrenamiento y uno de testeo

¿Alcanza con analizar solamente el accuracy obtenido?