

TA2 Ejercicio 1 Estandarización / normalización



- descargar el dataset “wine” de UCI y abrir con excel
- revisar el problema planteado
- analizar la variable objetivo
- analizar los atributos y sus tipos de datos
- analizar técnicas para estandarizar / normalizar -
  - REMITIR UN DOC a la tarea UT02-TA2 en la webas con las decisiones de normalización / estandarización tomadas

2

2

---

---

---

---

---

---

---

TA2 Ejercicio 2



Utilizando el dataset “wine” , analizar los bloques apropiados para implementar las técnicas identificadas en el Ej. 1.

- Identificar los bloques de RM y escribir en un doc la descripción / parámetros de cada un (muy breve!)
- Armar al menos 2 líneas de proceso para comparar con y sin estandarización / normalización

3

3

---

---

---

---

---

---

---

Training vs. Test Error



- Queremos saber qué tan bien se comportará nuestro modelo!
- Lo hacemos midiendo la exactitud, para seleccionar los mejores algoritmos y luego ajustar los parámetros para lograr mejores resultados.
- Validación del modelo:
  - Error de entrenamiento (training error)
  - Error de test (test error)

4

4

---

---

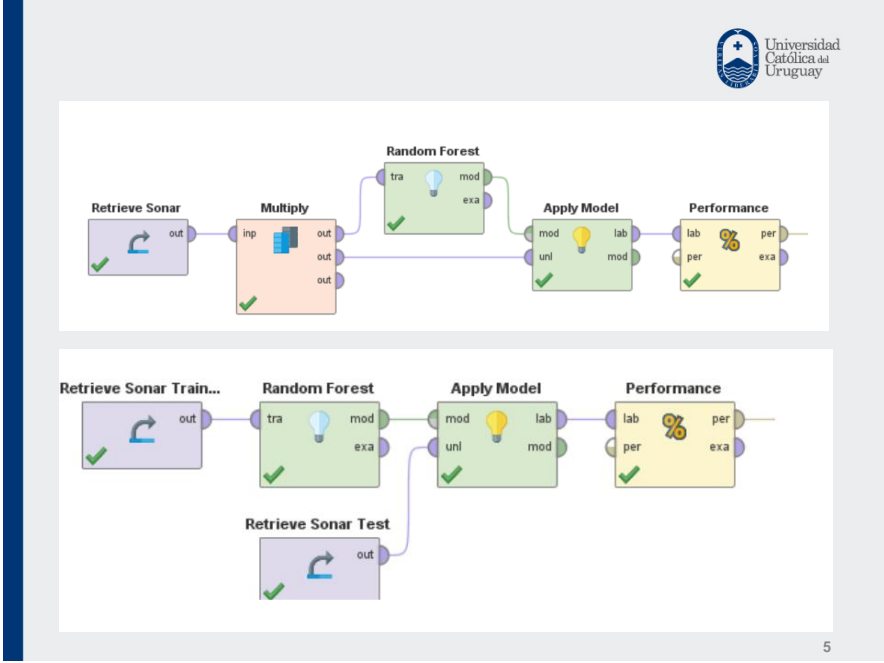
---

---

---

---

---



---

---

---

---


---

---

---

---

Por qué se debe ignorar / descartar el error de entrenamiento



- A pesar de que encontraremos multiples referencias al error de entrnamiento en la literature de ML, es una mala práctica
- Puede conducir a errores peligrosos
- Ejemplo con valores bidimensionales (rangos -1 a 1), igual cantidad de ejemplos “positivos” y “negativos”

6

---

---

---

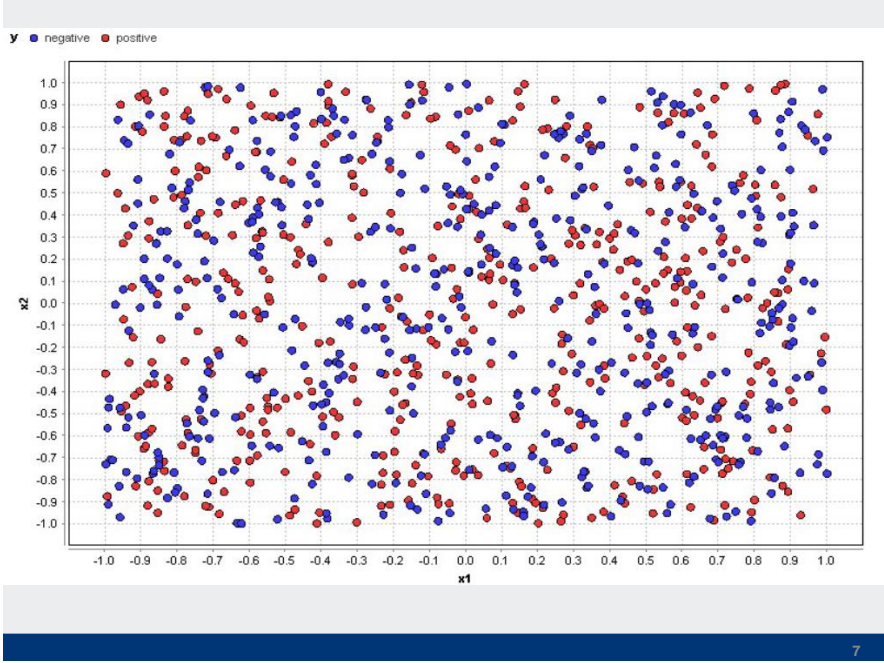
---

---

---

---

---



---

---

---

---

---

---

---

---



- 500 ejemplos positivos y 500 negativos, distribuidos aleatoriamente
- ¿puede construirse un modelo predictivo?
- ... lo mejor que se podría obtener sería 50%...
- Si aplicamos k-nn (con  $n = 1$ ), obtendremos una exactitud de 100%

8

8

---

---

---

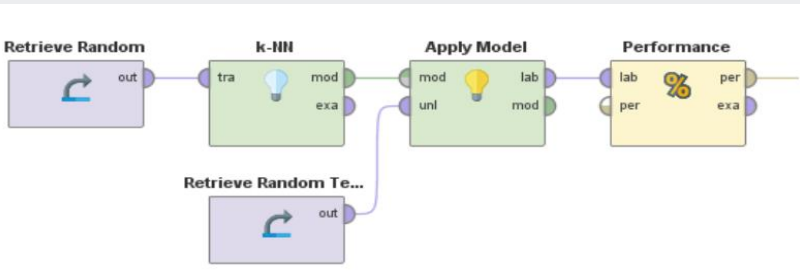
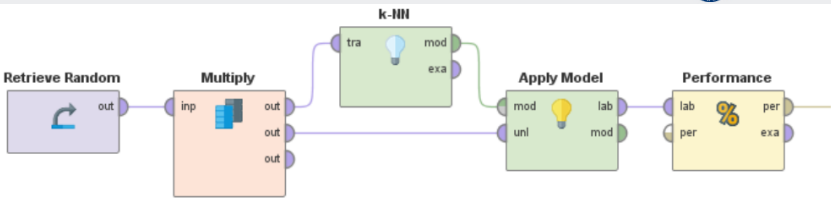
---

---

---

---

---



9

9

---

---

---

---

---

---

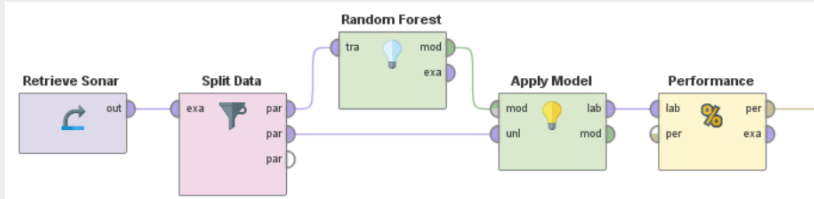
---

---

## Validación utilizando conjuntos hold-out



- Ya vimos que considerar el error de entrenamiento es una mala idea...
- Lo primero a tener en cuenta es que obtener o generar más datos puede ser difícil y costoso!
- Es práctica común entonces utilizar una parte de los datos disponibles para entrenamiento, y otra para test ("hold-out set" – "data split")



10

---

---

---

---

---

---

---

---

TA2 Ejercicio 3



- Generar un proceso con “split validation” (con parámetros estándar 70/30), para medir el rendimiento en los dos “canales” (con y sin normalización / estandarización)
- utilizar un algoritmo “k-nn” con parámetros estándar
- ejecutar los modelos y registrar los resultados de performance de los dos “canales” en una planilla simple comparativa

11

---

---

---

---

---

---

---