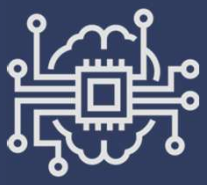


MACHINE LEARNING

MSC. RENZO CLAURE

1

MODELOS DE ENSAMBLE

MSC RENZO CLAURE

2

MACHINE LEARNING

MODELOS DE ENSAMBLE

MODELOS QUE COMBINAN VARIOS MODELOS

- Se combinan diferentes modelos, o varios modelos con diferentes parámetros
- Fortalecen las debilidades de otros modelos, reduciendo el sobreajuste
- Mejorar la eficacia de las predicciones, reduciendo o controlando el sobreajuste
- Reducen la comprensibilidad
- Exigen más procesamiento

MSC RENZO CLAURE

3

MACHINE LEARNING

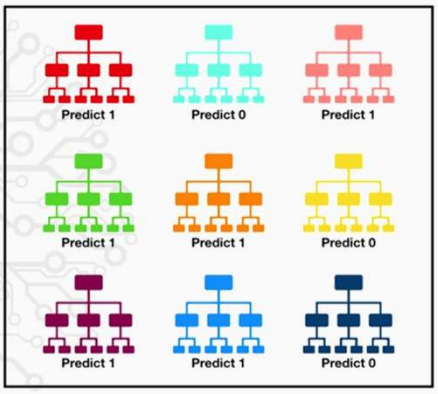
RANDOM FOREST

MODELOS DE ENSAMBLE CON ÁRBOLES DE DECISIÓN

- Un ensamble de varios árboles
- Son muy utilizados actualmente
- La debilidad de los árboles de decisión es el sobre ajuste
- Varios árboles reducen el sobreajuste
- Los árboles se entrenan en varias muestras sobre el mismo universo
- La variación o variabilidad de los datos está mejor representada

MSC RENZO CLAURE

4



MACHINE LEARNING

CONSTRUCCIÓN DE RANDOM FOREST

MUESTREO CON REEMPLAZAMIENTO

Time/h	Degradation of sample number/g					
	115#	201#	216#	235#	237#	243#
1 258	0	0	0	0	0	0
1 327	0.003 20	0.001 00	0.001 00	0.011 00	0.001 00	0.001 10
1 396	0.003 70	0.002 60	0.001 10	0.014 00	0.002 00	0.001 60
1 541	0.006 30	0.002 70	0.000 50	0.026 00	0.003 70	0.003 20
1 593	0.005 80	0.002 10	0	0.031 00	0.004 20	0.003 70
1 688	0.008 90	0.003 70	0.000 10	0.044 00	0.003 70	0.004 70
1 758	0.012 00	0.002 60	0.001 00	0.056 00	0.003 70	0.005 30
1 858	0.013 00	0.004 20	0.000 50	0.061 00	0.003 80	0.005 80
1 998	0.015 00	0.004 30	0.001 60	0.076 00	0.005 70	0.006 80
2 058	0.016 00	0.005 80	0.000 50	0.081 00	0.005 80	0.007 80
2 212	0.018 00	0.007 30	0.001 60	0.090 00	0.007 40	0.008 90
2 288	0.022 00	0.007 40	0.002 70	0.102 00	0.008 90	0.010 00
2 364	0.023 00	0.00750	0.002 10	0.113 00	0.009 00	0.011 00
2 503	0.025 00	0.007 20	0.002 70	0.126 00	0.010 00	0.011 60
2 598	0.027 00	0.006 90	0.002 00	0.134 00	0.010 50	0.012 60
2 672	0.029 00	0.007 90	0.002 10	0.140 00	0.011 00	0.013 00

z1	x2	x3	x4	z1	x2	x3	x4	z1	x2	x3	x4
v11	v21	v31	v41	v11	v21	v31	v41	v11	v21	v31	v41
v12	v22	v32	v42	v12	v22	v32	v42	v12	v22	v32	v42
v13	v23	v33	v43	v13	v23	v33	v43	v13	v23	v33	v43
v14	v24	v34	v44	v14	v24	v34	v44	v14	v24	v34	v44
v15	v25	v35	v45	v15	v25	v35	v45	v15	v25	v35	v45
v16	v26	v36	v46	v16	v26	v36	v46	v16	v26	v36	v46
v17	v27	v37	v47	v17	v27	v37	v47	v17	v27	v37	v47
v18	v28	v38	v48	v18	v28	v38	v48	v18	v28	v38	v48
v19	v29	v39	v49	v19	v29	v39	v49	v19	v29	v39	v49
v20	v30	v40	v50	v20	v30	v40	v50	v20	v30	v40	v50
v21	v31	v41	v51	v21	v31	v41	v51	v21	v31	v41	v51
v22	v32	v42	v52	v22	v32	v42	v52	v22	v32	v42	v52
v23	v33	v43	v53	v23	v33	v43	v53	v23	v33	v43	v53
v24	v34	v44	v54	v24	v34	v44	v54	v24	v34	v44	v54
v25	v35	v45	v55	v25	v35	v45	v55	v25	v35	v45	v55
v26	v36	v46	v56	v26	v36	v46	v56	v26	v36	v46	v56
v27	v37	v47	v57	v27	v37	v47	v57	v27	v37	v47	v57
v28	v38	v48	v58	v28	v38	v48	v58	v28	v38	v48	v58
v29	v39	v49	v59	v29	v39	v49	v59	v29	v39	v49	v59
v30	v40	v50	v60	v30	v40	v50	v60	v30	v40	v50	v60
v31	v41	v51	v61	v31	v41	v51	v61	v31	v41	v51	v61
v32	v42	v52	v62	v32	v42	v52	v62	v32	v42	v52	v62
v33	v43	v53	v63	v33	v43	v53	v63	v33	v43	v53	v63
v34	v44	v54	v64	v34	v44	v54	v64	v34	v44	v54	v64
v35	v45	v55	v65	v35	v45	v55	v65	v35	v45	v55	v65
v36	v46	v56	v66	v36	v46	v56	v66	v36	v46	v56	v66
v37	v47	v57	v67	v37	v47	v57	v67	v37	v47	v57	v67
v38	v48	v58	v68	v38	v48	v58	v68	v38	v48	v58	v68
v39	v49	v59	v69	v39	v49	v59	v69	v39	v49	v59	v69
v40	v50	v60	v70	v40	v50	v60	v70	v40	v50	v60	v70

z1	x2	x3	x4	z1	x2	x3	x4	z1	x2	x3	x4
v11	v21	v31	v41	v11	v21	v31	v41	v11	v21	v31	v41
v12	v22	v32	v42	v12	v22	v32	v42	v12	v22	v32	v42
v13	v23	v33	v43	v13	v23	v33	v43	v13	v23	v33	v43
v14	v24	v34	v44	v14	v24	v34	v44	v14	v24	v34	v44
v15	v25	v35	v45	v15	v25	v35	v45	v15	v25	v35	v45
v16	v26	v36	v46	v16	v26	v36	v46	v16	v26	v36	v46
v17	v27	v37	v47	v17	v27	v37	v47	v17	v27	v37	v47
v18	v28	v38	v48	v18	v28	v38	v48	v18	v28	v38	v48
v19	v29	v39	v49	v19	v29	v39	v49	v19	v29	v39	v49
v20	v30	v40	v50	v20	v30	v40	v50	v20	v30	v40	v50
v21	v31	v41	v51	v21	v31	v41	v51	v21	v31	v41	v51
v22	v32	v42	v52	v22	v32	v42	v52	v22	v32	v42	v52
v23	v33	v43	v53	v23	v33	v43	v53	v23	v33	v43	v53
v24	v34	v44	v54	v24	v34	v44	v54	v24	v34	v44	v54
v25	v35	v45	v55	v25	v35	v45	v55	v25	v35	v45	v55
v26	v36	v46	v56	v26	v36	v46	v56	v26	v36	v46	v56
v27	v37	v47	v57	v27	v37	v47	v57	v27	v37	v47	v57
v28	v38	v48	v58	v28	v38	v48	v58	v28	v38	v48	v58
v29	v39	v49	v59	v29	v39	v49	v59	v29	v39	v49	v59
v30	v40	v50	v60	v30	v40	v50	v60	v30	v40	v50	v60
v31	v41	v51	v61	v31	v41	v51	v61	v31	v41	v51	v61
v32	v42	v52	v62	v32	v42	v52	v62	v32	v42	v52	v62
v33	v43	v53	v63	v33	v43	v53	v63	v33	v43	v53	v63
v34	v44	v54	v64	v34	v44	v54	v64	v34	v44	v54	v64
v35	v45	v55	v65	v35	v45	v55	v65	v35	v45	v55	v65
v36	v46	v56	v66	v36	v46	v56	v66	v36	v46	v56	v66
v37	v47	v57	v67	v37	v47	v57	v67	v37	v47	v57	v67
v38	v48	v58	v68	v38	v48	v58	v68	v38	v48	v58	v68
v39	v49	v59	v69	v39	v49	v59	v69	v39	v49	v59	v69
v40	v50	v60	v70	v40	v50	v60	v70	v40	v50	v60	v70

MSC RENZO CLAURE

MACHINE LEARNING

CONSTRUCCIÓN DE RANDOM FOREST

CONSTRUCCIÓN DE ÁRBOLES, MUESTREO DE VARIABLES

z1	x2	x3	x4	z1	x2	x3	x4	z1	x2	x3	x4
v11	v21	v31	v41	v11	v21	v31	v41	v11	v21	v31	v41
v12	v22	v32	v42	v12	v22	v32	v42	v12	v22	v32	v42
v13	v23	v33	v43	v13	v23	v33	v43	v13	v23	v33	v43
v14	v24	v34	v44	v14	v24	v34	v44	v14	v24	v34	v44
v15	v25	v35	v45	v15	v25	v35	v45	v15	v25	v35	v45
v16	v26	v36	v46	v16	v26	v36	v46	v16	v26	v36	v46
v17	v27	v37	v47	v17	v27	v37	v47	v17	v27	v37	v47
v18	v28	v38	v48	v18	v28	v38	v48	v18	v28	v38	v48
v19	v29	v39	v49	v19	v29	v39	v49	v19	v29	v39	v49
v20	v30	v40	v50	v20	v30	v40	v50	v20	v30	v40	v50
v21	v31	v41	v51	v21	v31	v41	v51	v21	v31	v41	v51
v22	v32	v42	v52	v22	v32	v42	v52	v22	v32	v42	v52
v23	v33	v43	v53	v23	v33	v43	v53	v23	v33	v43	v53
v24	v34	v44	v54	v24	v34	v44	v54	v24	v34	v44	v54
v25	v35	v45	v55	v25	v35	v45	v55	v25	v35	v45	v55
v26	v36	v46	v56	v26	v36	v46	v56	v26	v36	v46	v56
v27	v37	v47	v57	v27	v37	v47	v57	v27	v37	v47	v57
v28	v38	v48	v58	v28	v38	v48	v58	v28	v38	v48	v58
v29	v39	v49	v59	v29	v39	v49	v59	v29	v39	v49	v59
v30	v40	v50	v60	v30	v40	v50	v60	v30	v40	v50	v60
v31	v41	v51	v61	v31	v41	v51	v61	v31	v41	v51	v61
v32	v42	v52	v62	v32	v42	v52	v62	v32	v42	v52	v62
v33	v43	v53	v63	v33	v43	v53	v63	v33	v43	v53	v63
v34	v44	v54	v64	v34	v44	v54	v64	v34	v44	v54	v64
v35	v45	v55	v65	v35	v45	v55	v65	v35	v45	v55	v65
v36	v46	v56	v66	v36	v46	v56	v66	v36	v46	v56	v66
v37	v47	v57	v67	v37	v47	v57	v67	v37	v47	v57	v67
v38	v48	v58	v68	v38	v48	v58	v68	v38	v48	v58	v68
v39	v49	v59	v69	v39	v49	v59	v69	v39	v49	v59	v69
v40	v50	v60	v70	v40	v50	v60	v70	v40	v50	v60	v70

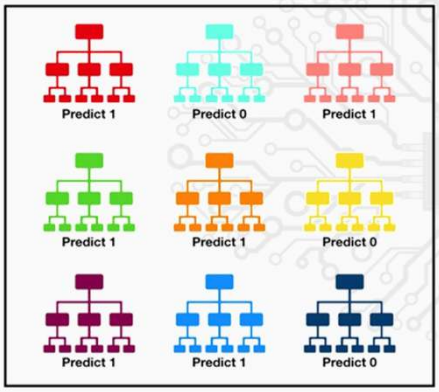
z1	x2	x3	x4	z1	x2	x3	x4	z1	x2	x3	x4
v11	v21	v31	v41	v11	v21	v31	v41	v11	v21	v31	v41
v12	v22	v32	v42	v12	v22	v32	v42	v12	v22	v32	v42
v13	v23	v33	v43	v13	v23	v33	v43	v13	v23	v33	v43
v14	v24	v34	v44	v14	v24	v34	v44	v14	v24	v34	v44
v15	v25	v35	v45	v15	v25	v35	v45	v15	v25	v35	v45
v16	v26	v36	v46	v16	v26	v36	v46	v16	v26	v36	v46
v17	v27	v37	v47	v17	v27	v37	v47	v17	v27	v37	v47
v18	v28	v38	v48	v18	v28	v38	v48	v18	v28	v38	v48
v19	v29	v39	v49	v19	v29	v39	v49	v19	v29	v39	v49
v20	v30	v40	v50	v20	v30	v40	v50	v20	v30	v40	v50
v21	v31	v41	v51	v21	v31	v41	v51	v21	v31	v41	v51
v22	v32	v42	v52	v22	v32	v42	v52	v22	v32	v42	v52
v23	v33	v43	v53	v23	v33	v43	v53	v23	v33	v43	v53
v24	v34	v44	v54	v24	v34	v44	v54	v24	v34	v44	v54
v25	v35	v45	v55	v25	v35	v45	v55	v25	v35	v45	v55
v26	v36	v46	v56	v26	v36	v46	v56	v26	v36	v46	v56
v27	v37	v47	v57	v27	v37	v47	v57	v27	v37	v47	v57
v28	v38	v48	v58	v28	v38	v48	v58	v28	v38	v48	v58
v29	v39	v49	v59	v29	v39	v49	v59	v29	v39	v49	v59
v30	v40	v50	v60	v30	v40	v50	v60	v30	v40	v50	v60
v31	v41	v51	v61	v31	v41	v51	v61	v31	v41	v51	v61
v32	v42	v52	v62	v32	v42	v52	v62	v32	v42	v52	v62
v33	v43	v53	v63	v33	v43	v53	v63	v33	v43	v53	v63
v34	v44	v54	v64	v34	v44	v54	v64	v34	v44	v54	v64
v35	v45	v55	v65	v35	v45	v55	v65	v35	v45	v55	v65
v36	v46	v56	v66	v36	v46	v56	v66	v36	v46	v56	v66
v37	v47	v57	v67	v37	v47	v57	v67	v37	v47	v57	v67
v38	v48	v58	v68	v38	v48	v58	v68	v38	v48	v58	v68
v39	v49	v59	v69	v39	v49	v59	v69	v39	v49	v59	v69
v40	v50	v60	v70	v40	v50	v60	v70	v40	v50	v60	v70

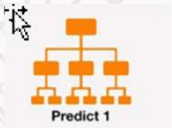
MSC RENZO CLAURE

MACHINE LEARNING

CONSTRUCCIÓN DE RANDOM FOREST

CONSTRUCCIÓN DE ÁRBOLES, MUESTREO DE VARIABLES





MSC RENZO CLAURE

7

MACHINE LEARNING

RANDOM FOREST

PARÁMETROS

- Max_features: se puede configurar la cantidad de variables a considerar para el ajuste. Más variables, toma más tiempo en aprender, pero mejora la precisión, afecta en la generalización
- N_estimators: es el número de árboles creados en el ensamble, a mayor cantidad de datos mayor debe ser este número para mejorar la precisión, pero consume más recursos
- Max_depth: configura la profundidad del árbol, por defecto llega a nodos puros o de frecuencia 2
- N_Jobs: cuantos cores usar en paralelo
- Random_state: ???

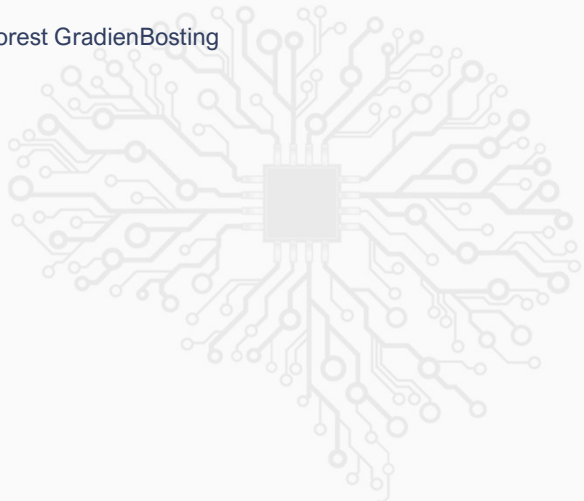
MSC RENZO CLAURE

8

MACHINE LEARNING

RANDOM FOREST
APLICACIÓN

- NB_14 RandomForest GradienBosting



MSC RENZO CLAURE

9

MACHINE LEARNING

RANDOM FOREST
ASPECTOS POSITIVOS Y NEGATIVOS

PROS
- No requiere estandarización
- Muy buenas predicciones y robusto a la generalización con una cantidad de datos aceptable
- Pueden configurarse parámetros de Pre-poda y Post-poda
- Puede configurarse su ejecución en procesadores en paralelo

CONTRA
- Consume más recursos
- No son modelos comprensibles
- En problemas con alta dimensionalidad, consume demasiados recursos y tiende al sobreajuste, deben considerarse en esos casos de preferencia modelos multiliniales

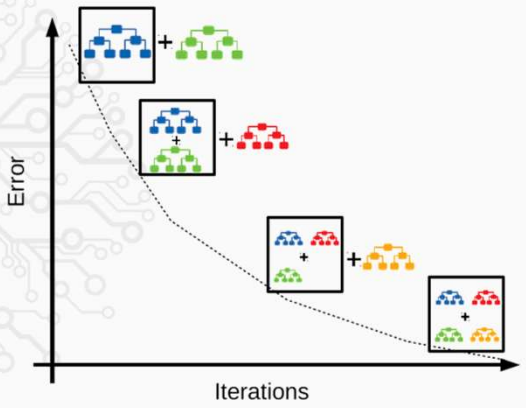
MSC RENZO CLAURE

10

MACHINE LEARNING

GRADIENT BOSTING
CONSTRUCCIÓN

- Se crean n árboles de entrenamiento, conocidos como aprendices débiles (weak learners)
- Cada construcción se basa en el anterior resultado y los mejora
- El gradiente o ratio de aprendizaje controla la razón de mejora de errores con respecto de los modelos previos
  - Learning rate alto, modelos más complejos y poco comprensibles
  - Learning bajo, modelos más simples y comprensibles



MSC RENZO CLAURE

11

MACHINE LEARNING

GRADIENT BOSTING
PARÁMETROS

- Learning rate: es el grado de mejora mínimo exigido con respecto al árbol anterior
- N_estimators: es el número de árboles creados en el ensamble (aprendices débiles
- Max_depth: configura la profundidad del árbol, por defecto llega a nodos puros o de frecuencia 2
- Es mejor ajustar primero N estimators, ya que es el que utiliza más recursos

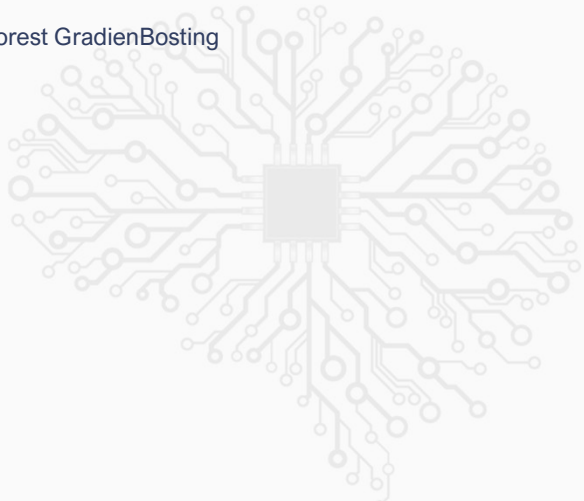
MSC RENZO CLAURE

12

MACHINE LEARNING

GRADIENT BOSTING
CONSTRUCCIÓN

- NB_14 RandomForest GradienBosting



MSC RENZO CLAURE

13

MACHINE LEARNING

GRADIENT BOSTING
VENYAJAS Y DESVENTAJAS

VENTAJAS
- Mejora la generalización y la exactitud
- Consume menos recursos que random forest
- Tampoco requiere normalización
- Tiene las mismas ventajas de un árbol de decisión

DESVENTAJAS
- El tunning requiere más recursos
- Es muy poco comprensible
- No es muy bueno para sets de alta dimensionalidad (también afecta a RanfomForest)
- Los cambios en el learning rate deben ser cuidadosos

MSC RENZO CLAURE

14

MACHINE LEARNING

XGBoost

- Diseñada para ser más eficiente en cuanto a tiempo de ejecución y consumo de memoria, además de ofrecer regularización para prevenir el overfitting.
- Implementación: XGBoost es una biblioteca altamente optimizada y más avanzada que la implementación básica de Gradient Boosting. Introduce características como:
  - Regularización: Para evitar el sobreajuste, XGBoost introduce penalizaciones en la complejidad del modelo (L1 y L2 regularización).
  - Manejo de valores faltantes: XGBoost tiene métodos integrados para lidiar con valores faltantes de manera efectiva.
  - Optimización de memoria: está diseñado para ser eficiente en términos de uso de memoria, lo que permite entrenar modelos con grandes conjuntos de datos.
  - Paralelización: XGBoost puede entrenar modelos en paralelo, lo que acelera significativamente el proceso de entrenamiento.

MSC RENZO CLAURE

15

Filtración de datos

data leakage

MSC RENZO CLAURE

16

MACHINE LEARNING

Filtración de datos

data leakage

- Cuando la información de entrenamiento se filtra en el set de comprobación. Esta filtración sesgará el modelo entrenado, ocasionando un resultado engañoso
- Resultados demasiado buenos pueden ser síntoma de que en los modelos se filtraron dato que contienen las “respuestas” de la variable independiente
- Por ejemplo: el pago de una factura, El modelo trata de predecir que clientes no pagarán la factura, para evitar darles más crédito. Si sabemos que solo entregaremos el producto si la factura esta pagada, no podemos introducir la variable “entregado si/no” como variable independiente.

MSC RENZO CLAURE

17

MACHINE LEARNING

Filtrado de datos

ejemplos

- Predicción de Mora
- Diagnósticos para predecir enfermedades:
  - Si se incluye en el modelo que el paciente ya fue ingresado previamente por la misma enfermedad
  - Si el identificador del paciente está codificado según su afectación médica
- Apertura de una nueva cuenta en banca
  - Si se incluye en el modelo una variable que depende de un dato que solo se proporciona si el cliente se suscribe a la nueva cuenta

MSC RENZO CLAURE

18

MACHINE LEARNING

Filtración de datos

ejemplos

- Al entrenar los modelos
  - Cuando se normalizan los datos de fomra incorrecta, por ejemplo tomando la distribucón de todo el set de datos, no solo de la muestra de entrenamiento
  - Al codificar los datos, pueden ingresarse datos que ya saben la respuesta, tengan cuidado con las transformaciones
  - En las series de tiempo, no incluir ningún datos futro, un modelo no puede adivinar que va a pasar
- Al seleccionar variables
  - Cuando se quitar variables que revelan la respuesta, deben quitarse todas las relacionadas
  - Tener cuidado con los datos externos

MSC RENZO CLAURE

19

MACHINE LEARNING

Filtración de datos

detección

- Antes
  - Validar correlaciones sospechosamente altas
  - Explorar adecuadamente los datos
- Después
  - Son demasiado buenos los resultados, existen algunas variables con factores o pesos muy elevados?
  - Revisar si los resultados posteriores tienen comportamientos extraños
- Monitorear los modelos implementados con pruebas piloto
  - Realizar pruebas piloto, si los datos son demasiado malos revisar la construcción

MSC RENZO CLAURE

20

MACHINE LEARNING

Filtración de datos

deteccion

- Antes
  - Validar correlaciones sospechosamente altas
  - Explorar adecuadamente los datos
- Después
  - Son demasiado buenos los resultados, existen algunas variables con factores o pesos muy elevados?
  - Revisar si los resultados posteriores tienen comportamientos extraños
- Monitorear los modelos implementados con pruebas piloto
  - Realizar pruebas piloto, si los datos son demasiado malos revisar la construcción

MSC RENZO CLAURE

21

MACHINE LEARNING

Filtración de datos

atenuación

- La normalización de los datos debe hacerse de forma separada para cada muestra de entrenamiento y debe aplicarse la misma a su muestra de comprobación
- Con datos de tiempo, se deben tener las mismas ventanas de tiempo
- Es muy común reservar una parte de los datos para una Validación final
  - Solo si se tienen datos suficientes
  - Sirven como una prueba real
  - Permite ahcer una validación real de los datos

MSC RENZO CLAURE

22