Ciencia de datos

Para el sector público de salud

Módulo 4

Sesión 4: Modelos de clasificación

academia.opensaludlab.org

opensaludlab.org

Twitter / Instagram / LinkedIn

Elaboración de iniciativas de participantes

Presentaciones durante Junio 2021

Temática: libre, pero relacionada a la CD

Aplicar los conceptos aprendidos y las etapas de un proyecto de CD

Grupos de **2 a 5 personas (plazo 23.3.21)**

Twist: canal #CD Proyectos

Ok… vamos a lo que nos convoca

Machine Learning

El comienzo...

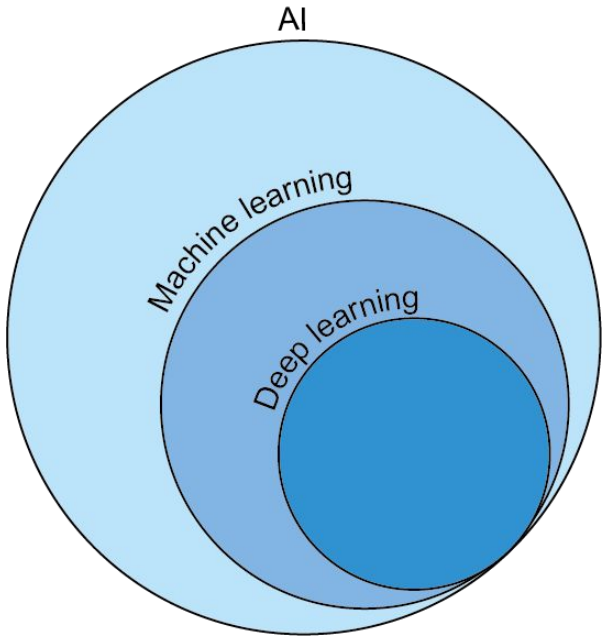
ML ¿Qué es y cuándo usarlo?

- Algoritmos de inteligencia artificial
- Perfectos para máxima predicción (modelos)
- En general son malos para explicar
- El mejor caso es cuando tengo al menos 1000 casos
- Si tengo millones de datos, mejor Deep Learning
- Tanto para clasificación como regresión

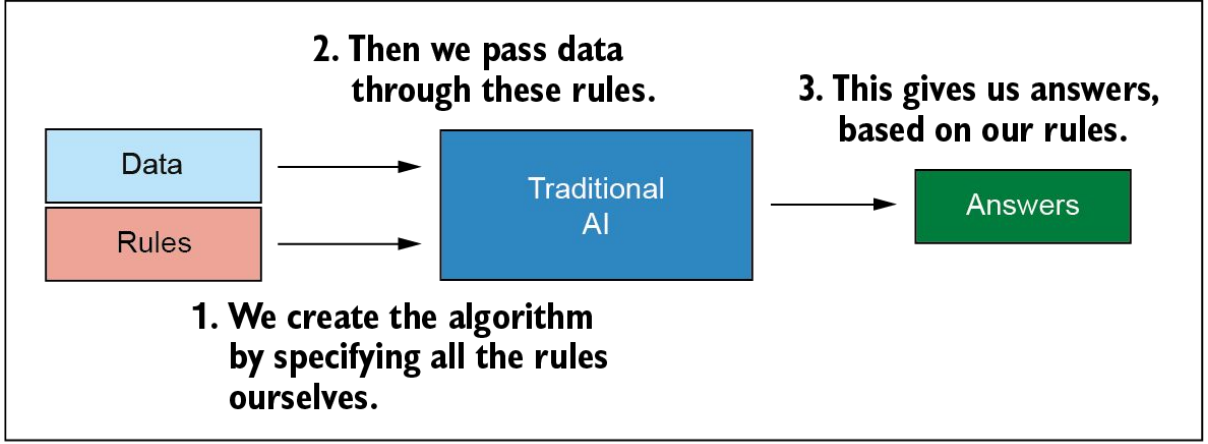
ML ó Aprendizaje automático

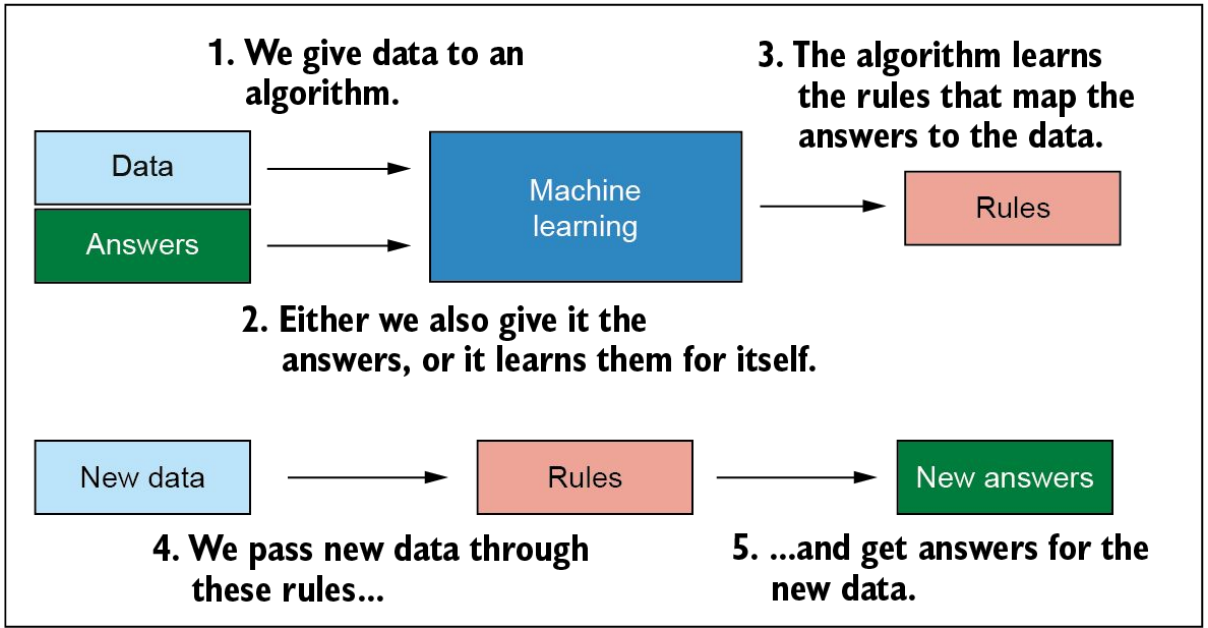
Es un subcampo de la inteligencia artificial donde los algoritmos “aprenden” patrones de los datos para realizar una tarea específica.

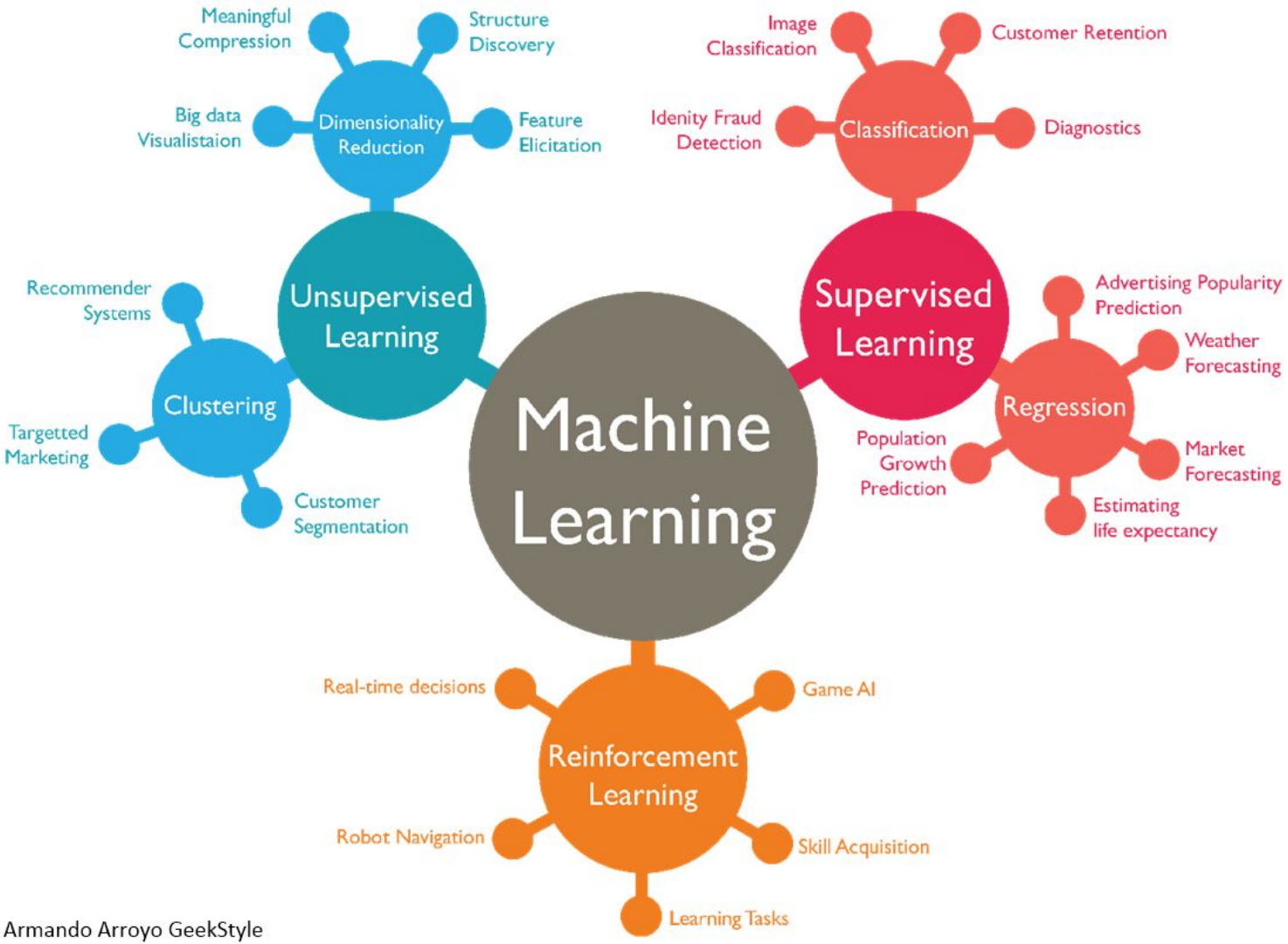
algoritmo != modelo



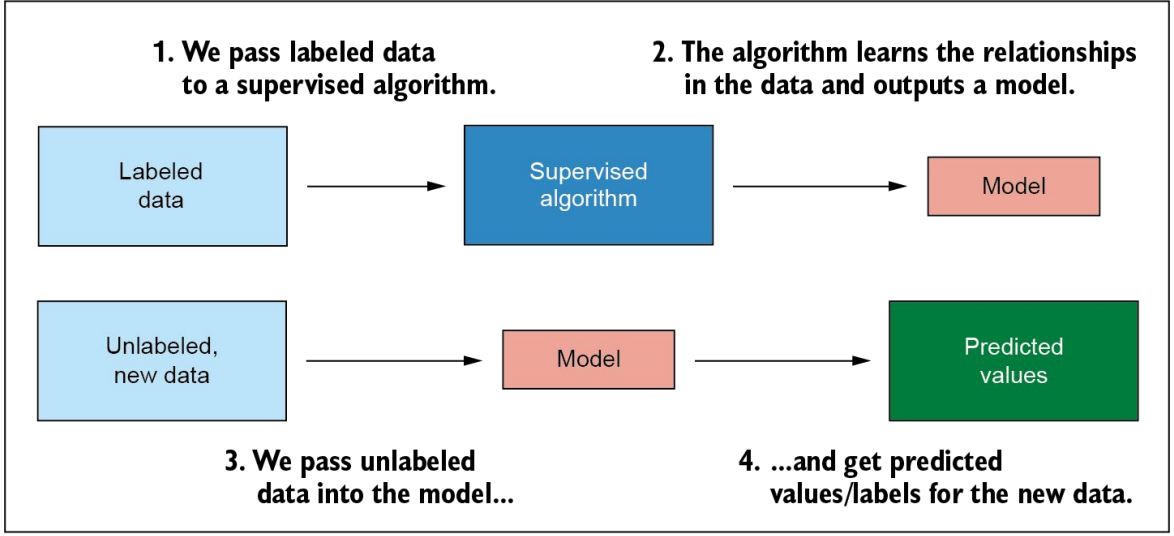
https://www.manning.com/books/machine-learning-with-r-the-tidyverse-and-mlr

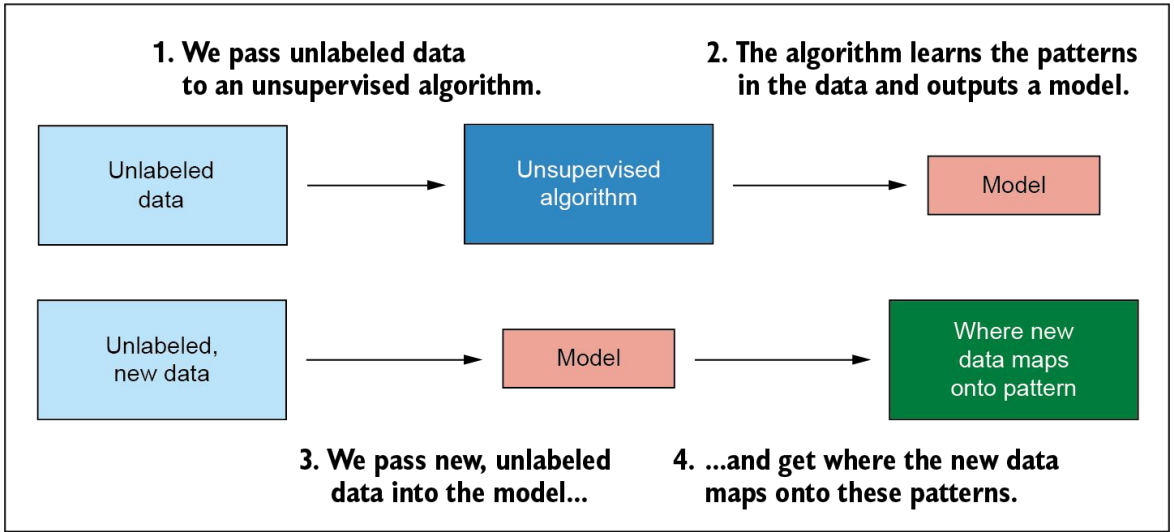

2. Then we pass data through these rules.
3. This gives us answers, based on our rules.
Data
Rules
Traditional AI
Answers
1. We create the algorithm by specifying all the rules ourselves.

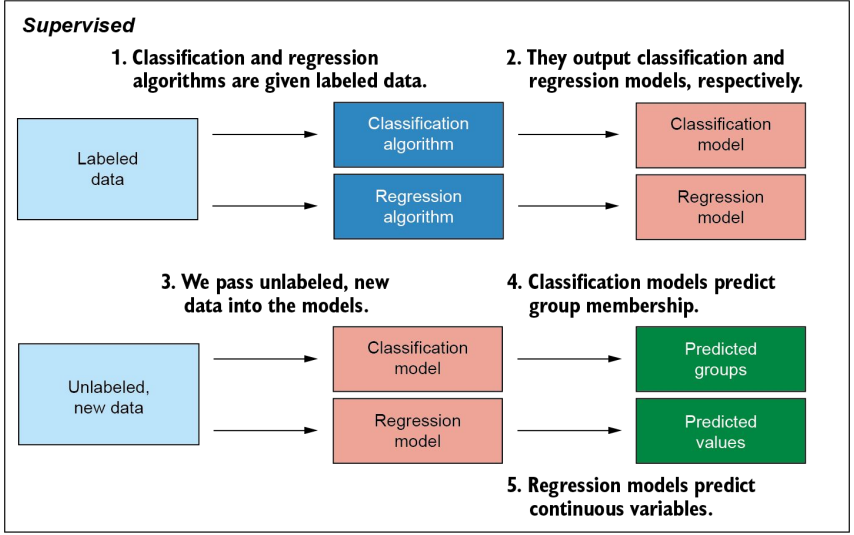

1. We give data to an algorithm.
3. The algorithm learns the rules that map the answers to the data.
Data
Answers
Machine learning
Rules
2. Either we also give it the answers, or it learns them for itself.
New data
Rules
New answers
4. We pass new data through these rules...
5. ...and get answers for the new data.

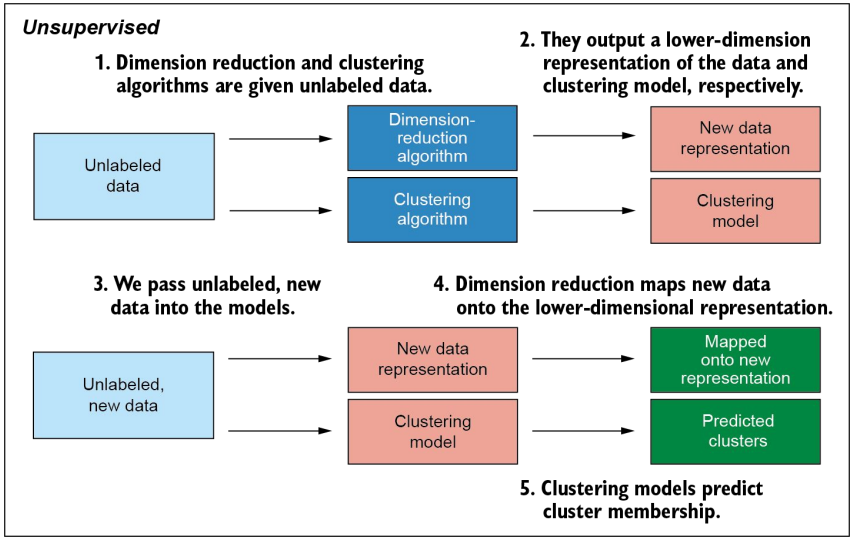


Armando Arroyo GeekStyle

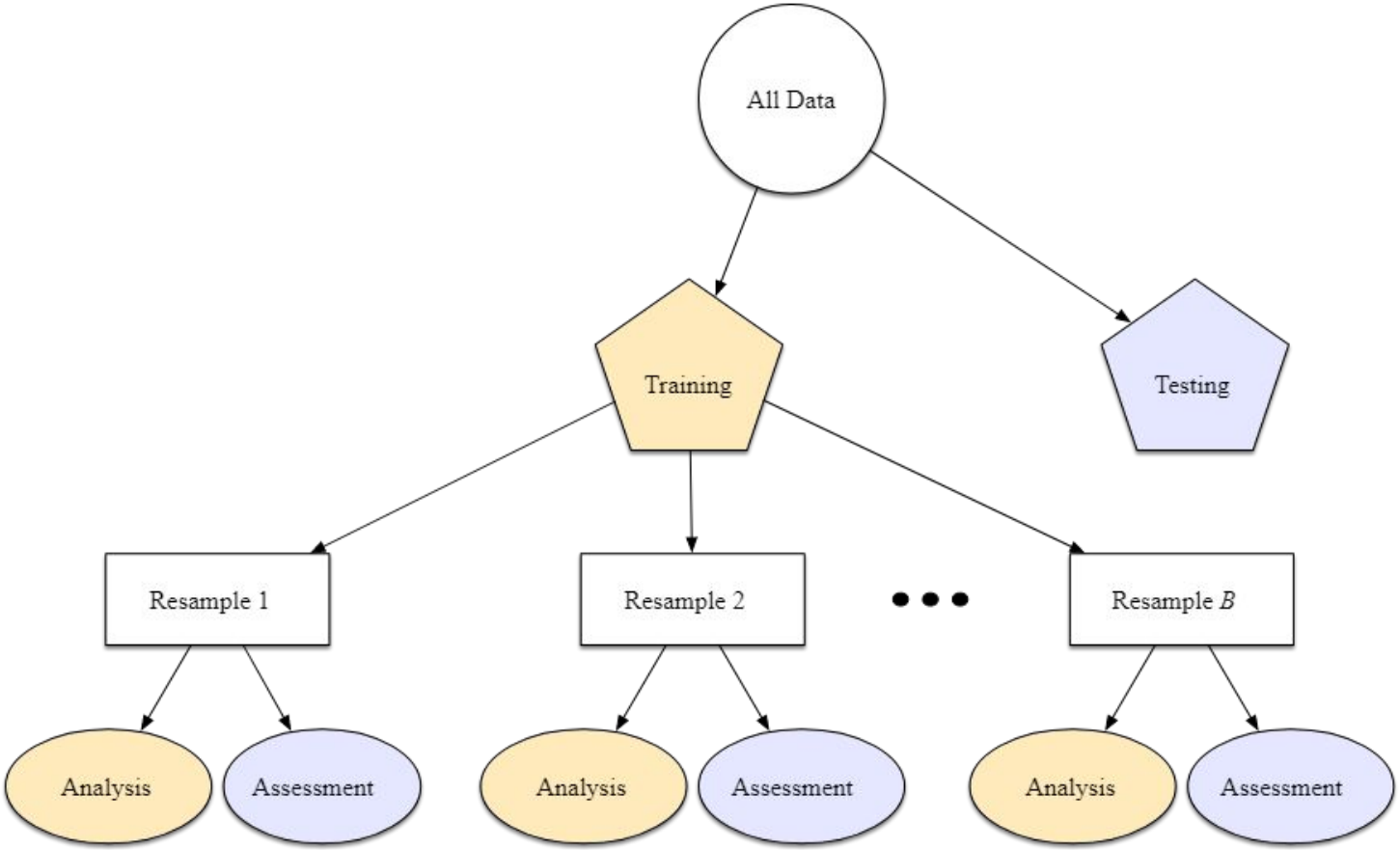

1. We pass labeled data to a supervised algorithm.
2. The algorithm learns the relationships in the data and outputs a model.
Labeled data
Supervised algorithm
Model
Unlabeled, new data
Model
Predicted values
3. We pass unlabeled data into the model...
4. ...and get predicted values/labels for the new data.


1. We pass unlabeled data to an unsupervised algorithm.
2. The algorithm learns the patterns in the data and outputs a model.
Unlabeled data
Unsupervised algorithm
Model
Unlabeled, new data
Model
Where new data maps onto pattern
3. We pass new, unlabeled data into the model...
4. ...and get where the new data maps onto these patterns.


Supervised
1. Classification and regression algorithms are given labeled data.
2. They output classification and regression models, respectively.
Labeled data
Classification algorithm
Regression algorithm
Classification model
Regression model
3. We pass unlabeled, new data into the models.
4. Classification models predict group membership.
Unlabeled, new data
Classification model
Regression model
Predicted groups
Predicted values
5. Regression models predict continuous variables.


Unsupervised
1. Dimension reduction and clustering algorithms are given unlabeled data.
2. They output a lower-dimension representation of the data and clustering model, respectively.
Unlabeled data
Dimension-reduction algorithm
Clustering algorithm
New data representation
Clustering model
3. We pass unlabeled, new data into the models.
4. Dimension reduction maps new data onto the lower-dimensional representation.
Unlabeled, new data
New data representation
Clustering model
Mapped onto new representation
Predicted clusters
5. Clustering models predict cluster membership.

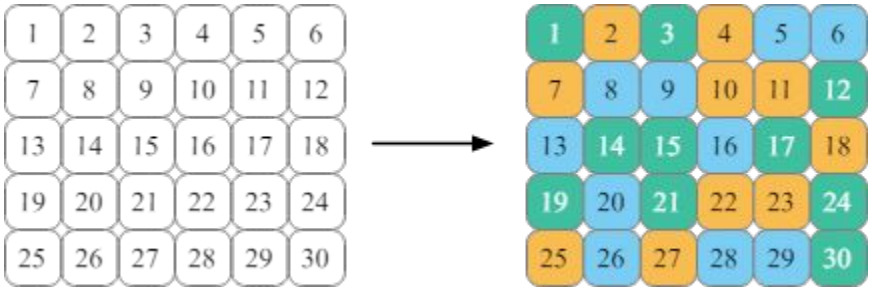
Resampleo



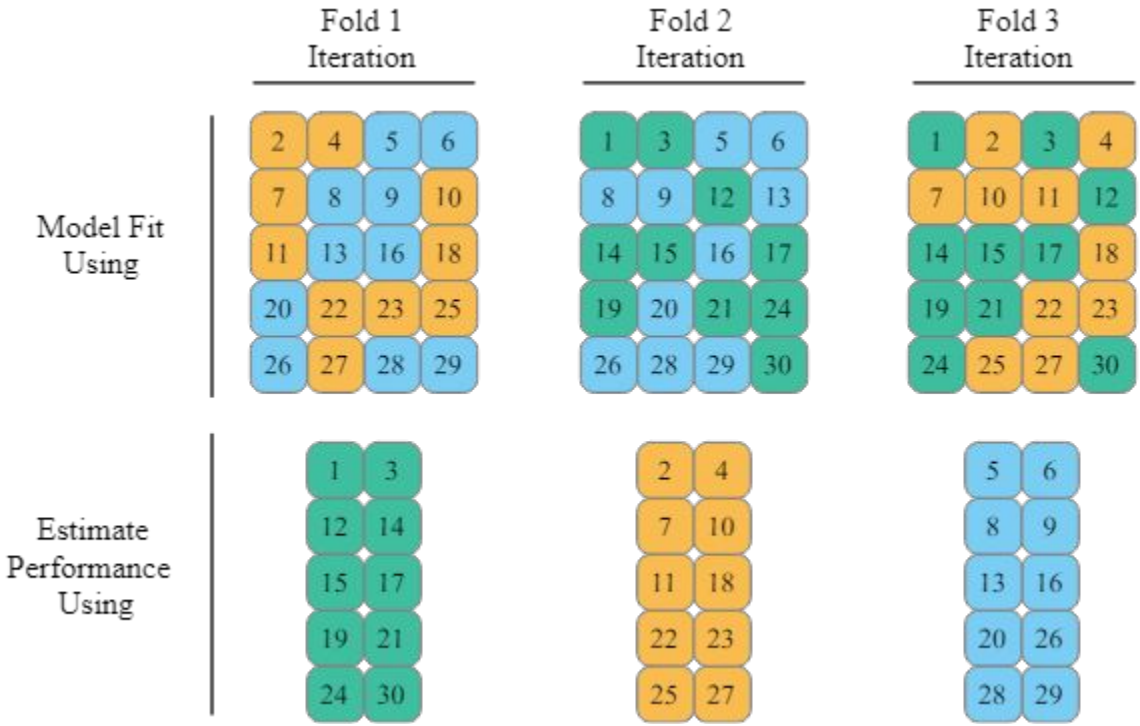
https://www.tmwr.org/resampling.html

Cross-validation

La validación cruzada es un método de remuestreo bien establecido. Si bien hay una serie de variaciones, el método de validación cruzada más común es la validación cruzada en V. Los datos se dividen aleatoriamente en V conjuntos de aproximadamente el mismo tamaño (llamados Folds o "pliegues"). A modo de ilustración, a continuación se muestra V = 3 para un conjunto de datos de treinta puntos de ajuste de entrenamiento con asignaciones de pliegues (folds) aleatorias. El número dentro de los símbolos es el número de muestra:

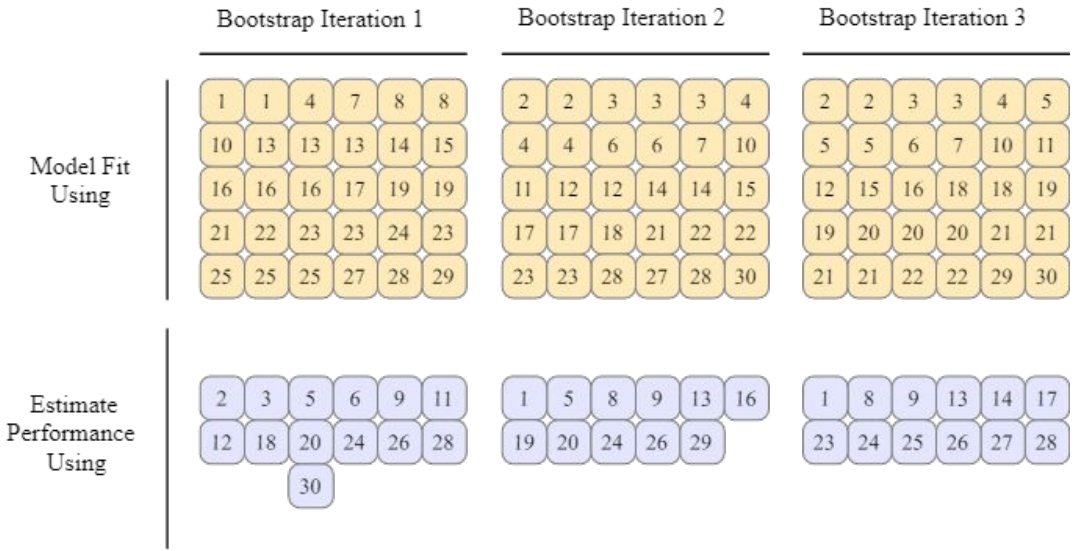


https://www.tmwr.org/resampling.html


Fold 1
Iteration
Fold 2
Iteration
Fold 3
Iteration
Model Fit
Using
2 4 5 6
7 8 9 10
11 13 16 18
20 22 23 25
26 27 28 29
1 3 5 6
8 9 12 13
14 15 16 17
19 20 21 24
26 28 29 30
1 2 3 4
7 10 11 12
14 15 17 18
19 21 22 23
24 25 27 30
Estimate
Performance
Using
1 3
12 14
15 17
19 21
24 30
2 4
7 10
11 18
22 23
25 27
5 6
8 9
13 16
20 26
28 29

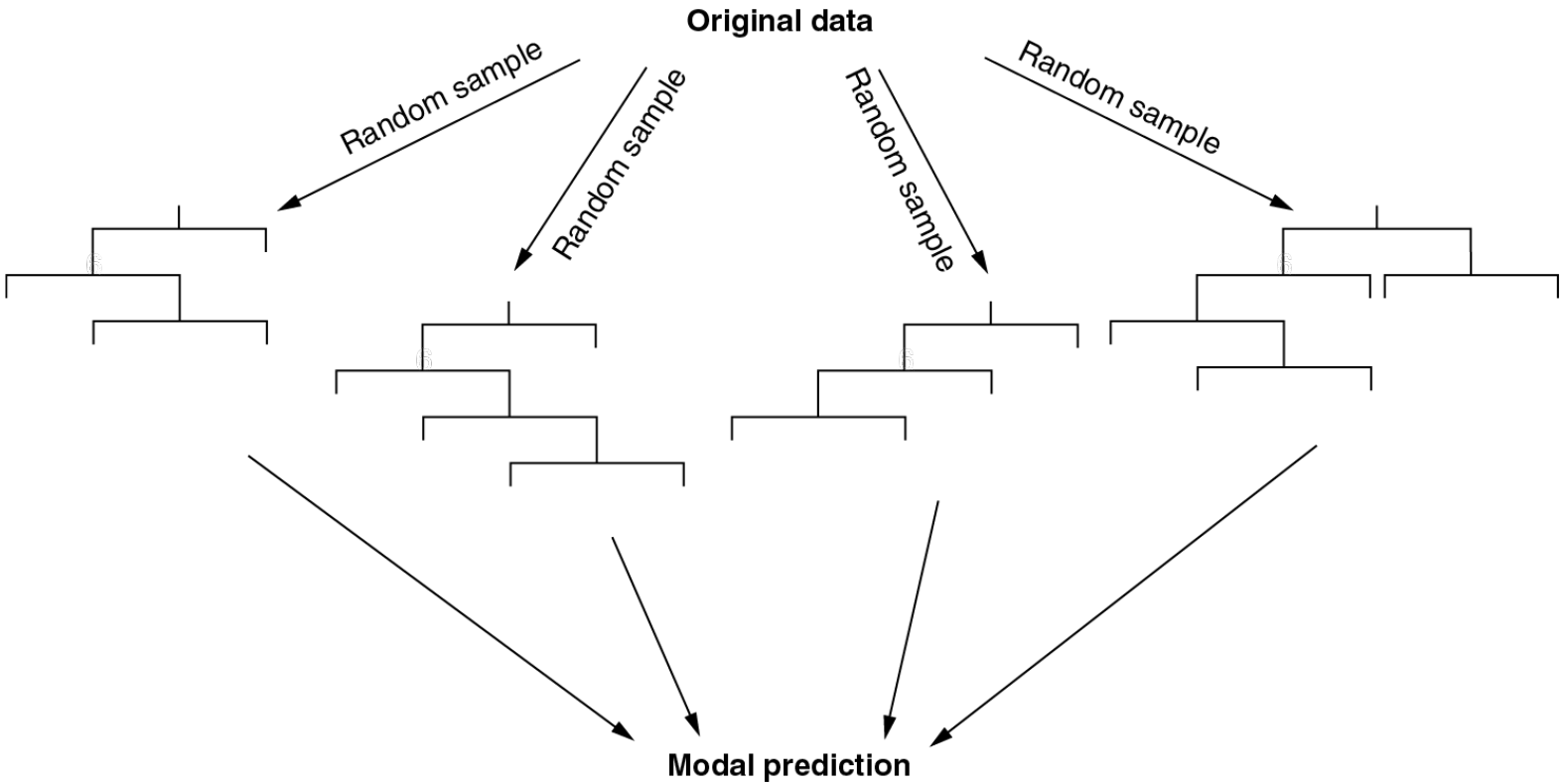
Bootstrapping

Un Bootstrap es una muestra que tiene el mismo tamaño que el conjunto de entrenamiento, pero se extrae con reemplazo. Esto significa que algunos puntos de datos del conjunto de entrenamiento se seleccionan varias veces para el conjunto de análisis.



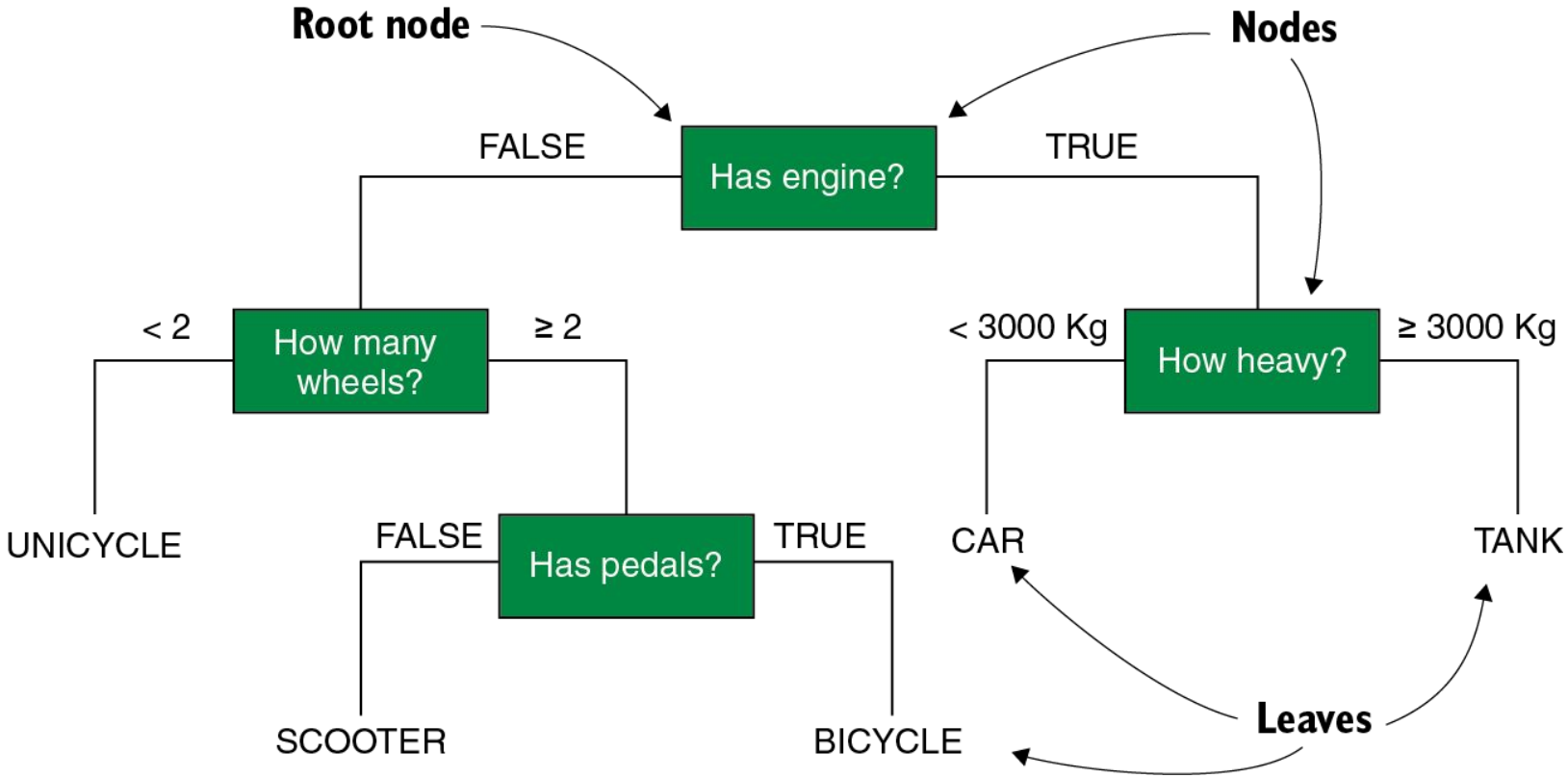
https://www.tmwr.org/resampling.html

Bootstrapping



Machine Learning
Aprendizaje supervisado

Árboles de decisión



Árboles de decisión

Hiperparámetros

- **min_n:** mínimo de muestras por nodos
- **tree_depth:** pone límite a la profundidad máxima de un árbol. Es un método para detener el algoritmo y evitar *overfiting*
- **cost_complexity:** costo o penalización a los errores de los árboles más complejos. Es un parámetro de parada. Si adopta el enfoque de construir árboles realmente profundos, el valor predeterminado de 0.01 podría ser demasiado restrictivo.

Random Forest

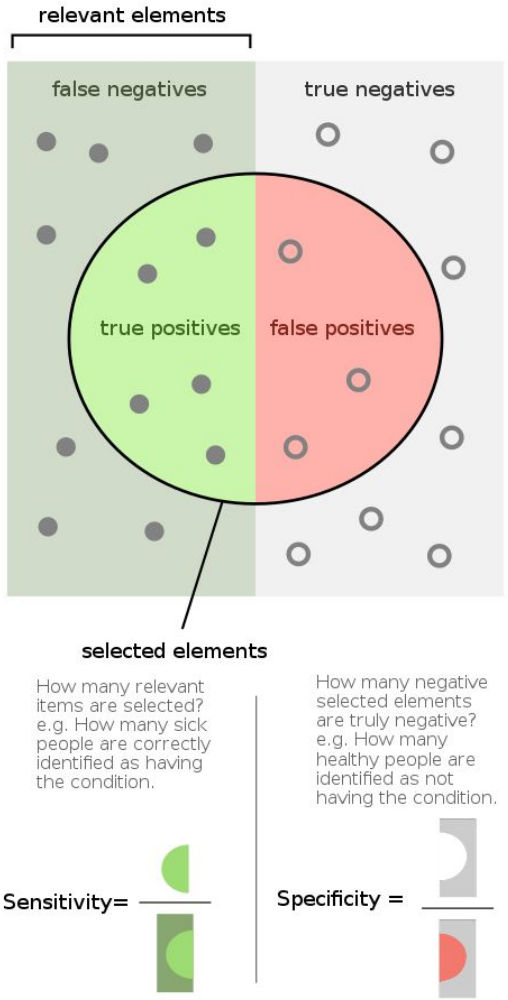
Hiperparámetros

- **mtry:** n° de predictores a muestrearse en cada split de árbol
- **min_n:** n° de observaciones necesarias para seguir dividiendo nodos

Métricas

Matriz de confusión

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$



https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Métricas

RMSE (raíz del error cuadrático medio)
https://es.wikipedia.org/wiki/Ra%C3%ADz_del_error_cuadr%C3%A1tico_medio

Mide la cantidad de error que hay entre dos conjuntos de datos. En otras palabras, compara un valor predicho y un valor observado o conocido.A diferencia del error absoluto medio (MAE), utilizamos RMSE en una variedad de aplicaciones cuando comparamos dos conjuntos de datos.

MAE (error absoluto medio)
https://es.wikipedia.org/wiki/Error_absoluto_medio

Permite evaluar la diferencia entre dos variables continuas. Sirve para cuantificar la precisión de una técnica de predicción.

Métricas

Curva ROC (Receiver Operating Characteristic)
https://es.wikipedia.org/wiki/Curva_ROC

Una curva ROC es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación.

Kappa https://es.wikipedia.org/wiki/Coeficiente_kappa_de_Cohen

Es una medida estadística que ajusta el efecto del azar en la proporción de la concordancia observada para elementos cualitativos (variables categóricas). En general se cree que es una medida más robusta que el simple cálculo del porcentaje de concordancia, ya que κ tiene en cuenta el acuerdo que ocurre por azar.

Veamos algo de código...