Ciencia de datos

Para el sector público de salud

Módulo 4

Sesión 5: Modelos de clasificación

academia.opensaludlab.org

opensaludlab.org

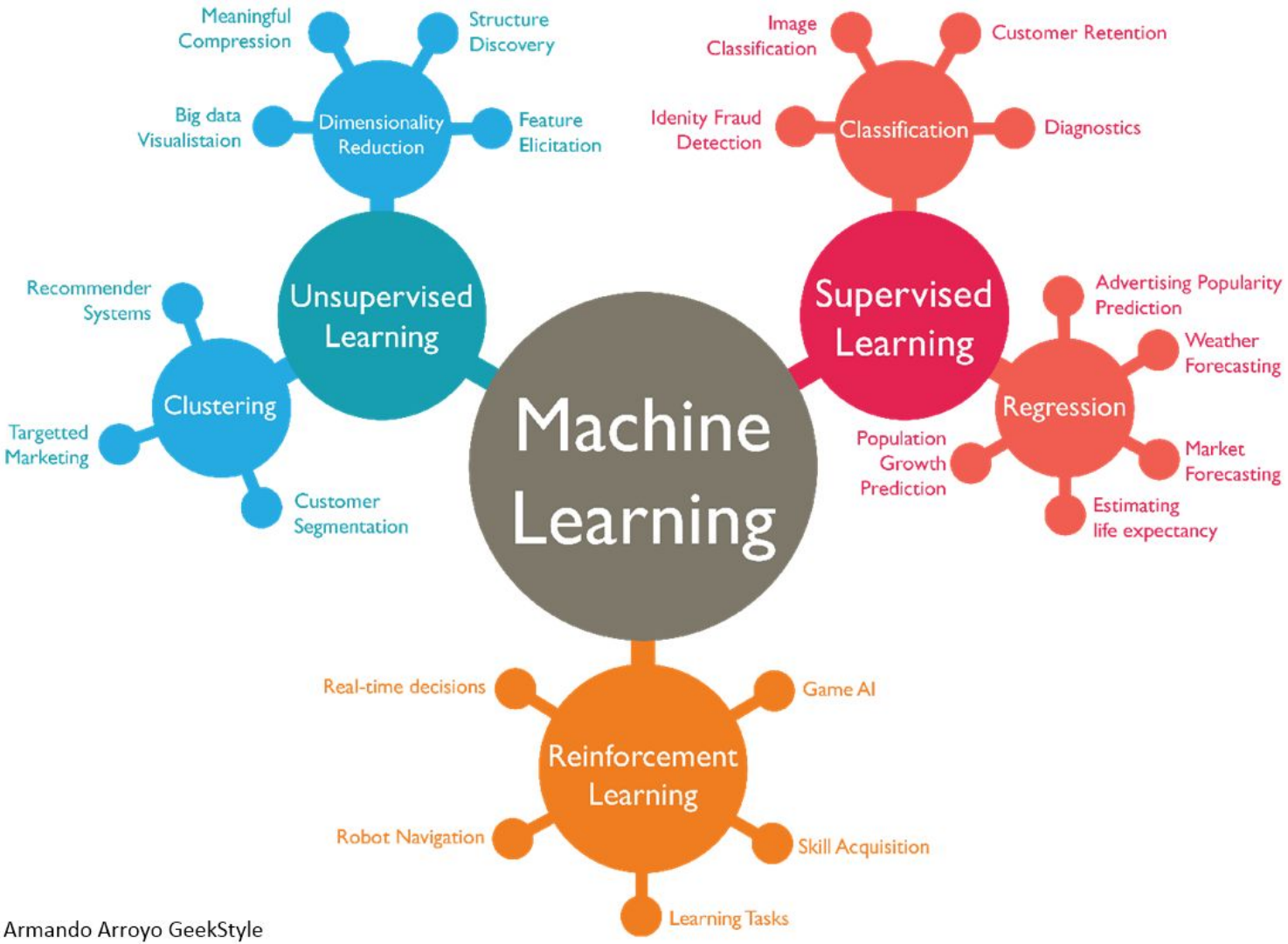
Twitter / Instagram / LinkedIn

Machine Learning

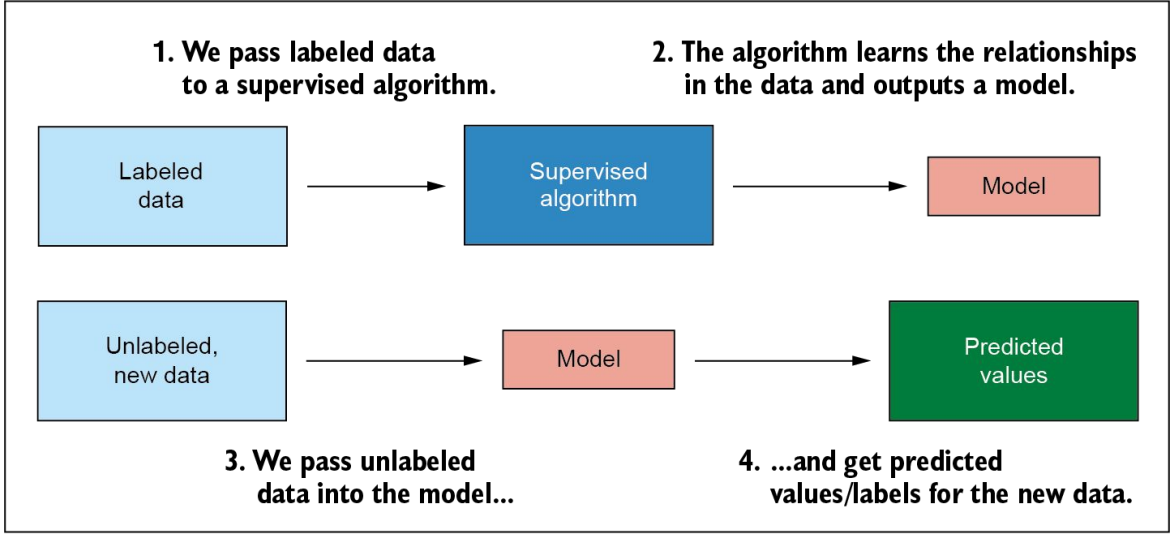
El comienzo...

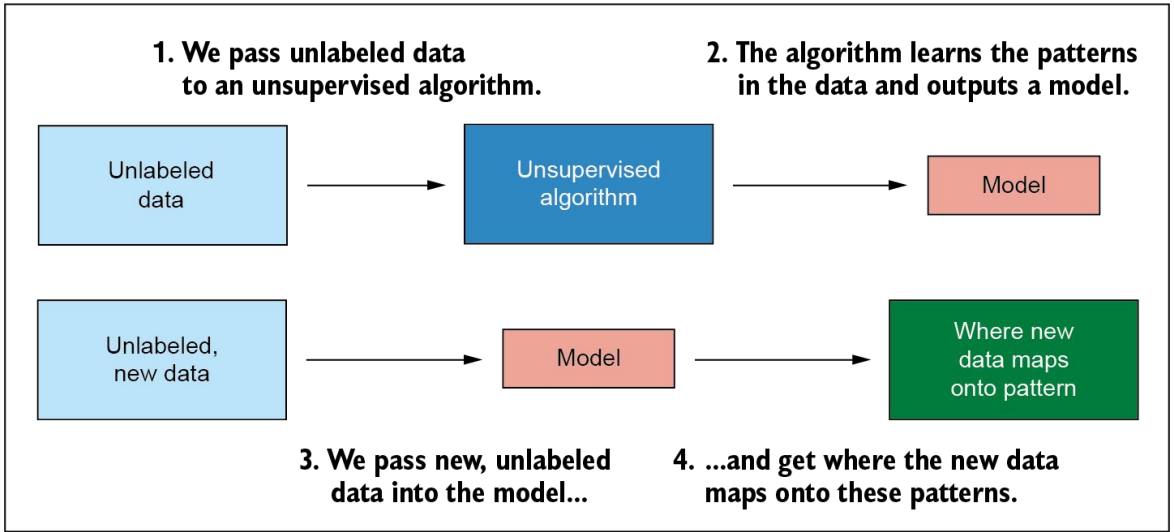

MACHINE LEARNING
R
ESTADÍSTICAS
PROGRAMACIÓN
YO

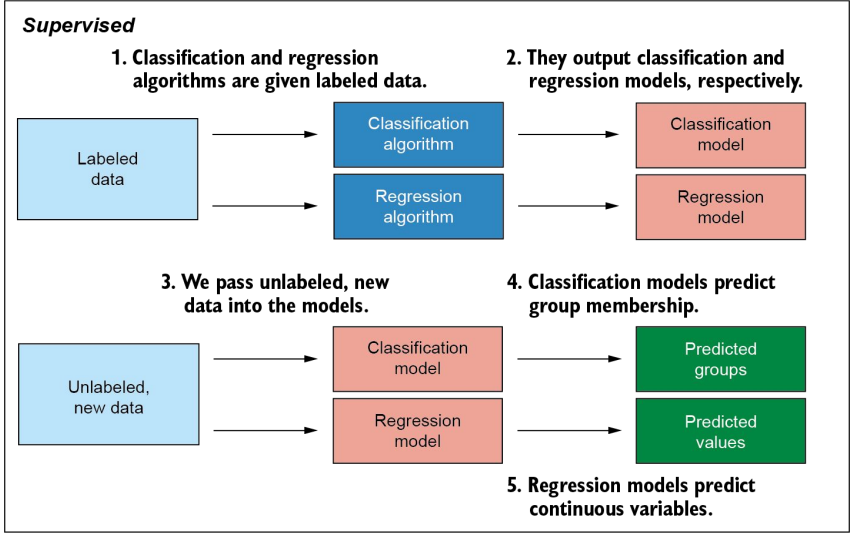

EVER GIVEN
EVERGREEN
عزت عادل
EZZAT ADEL

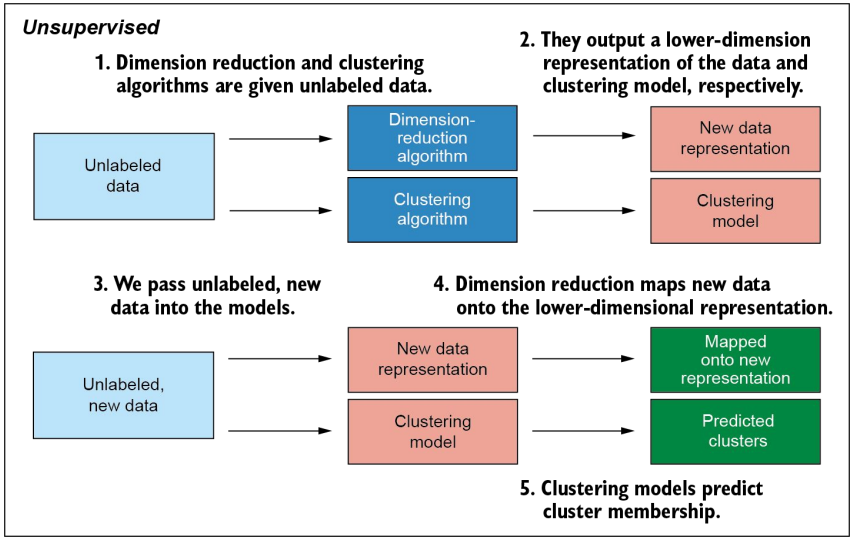


Armando Arroyo GeekStyle


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.

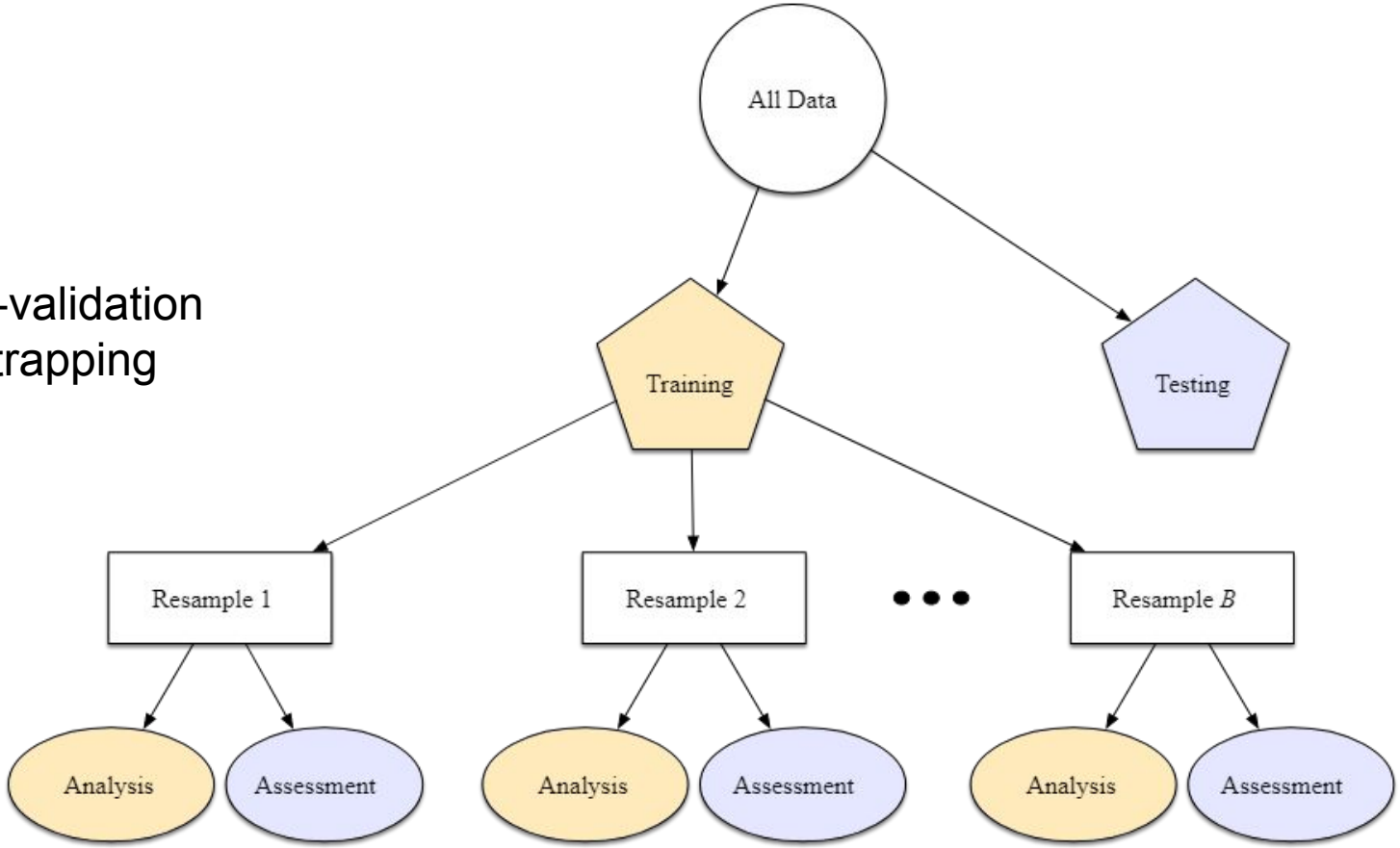

1. We pass unlabeled data to an unsupervised algorithm.
2. The algorithm learns the patterns in the data and outputs a model.
Unlabeled data
Unsupervised algorithm
Model
Unlabeled, new data
Model
Where new data maps onto pattern
3. We pass new, unlabeled data into the model...
4. ...and get where the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

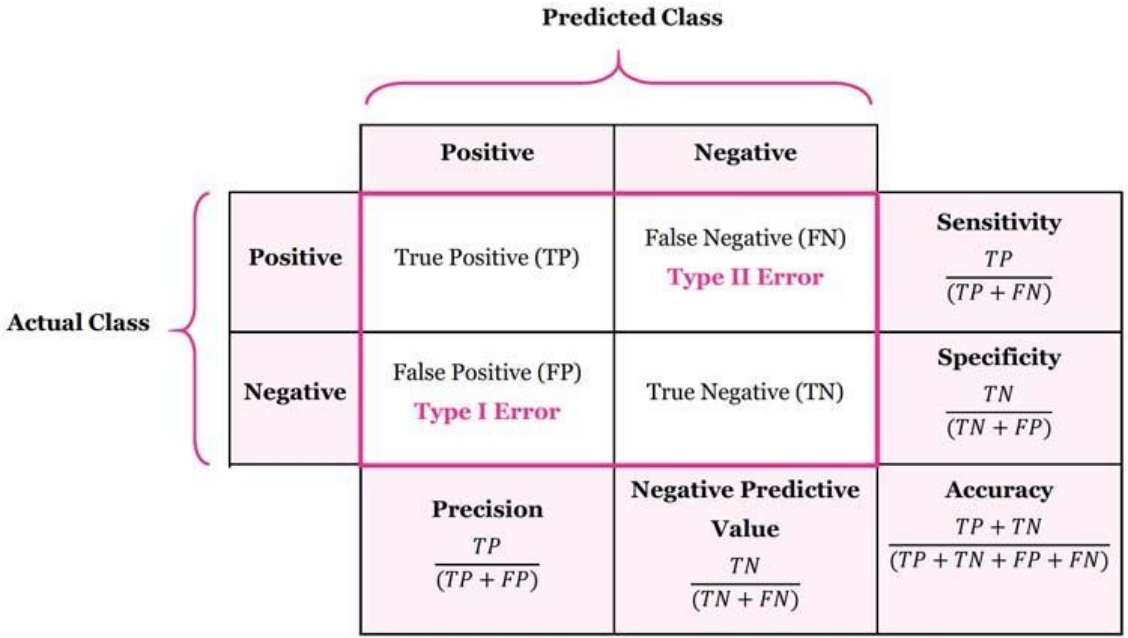
Cross-validation
Bootstrapping



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

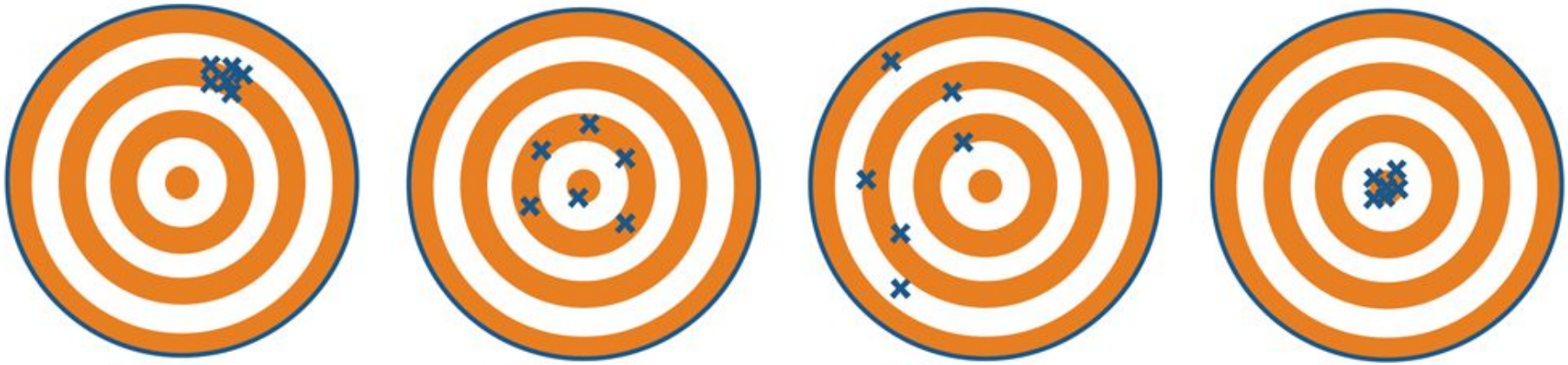
Métricas

Matriz de confusión



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

Precision vs Accuracy



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.

Métricas

Clasificación

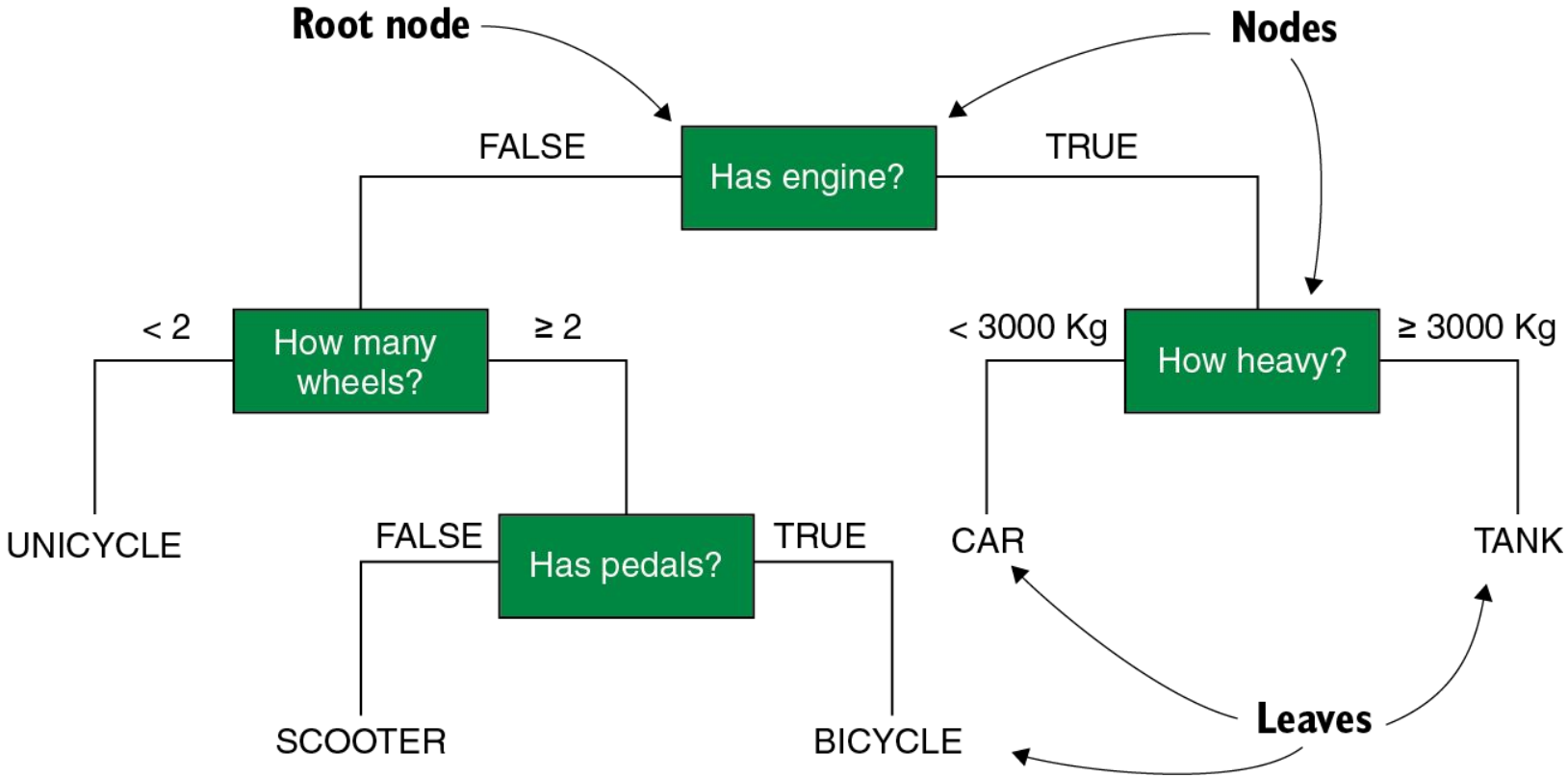
- Accuracy
- Kappa

Regresión

- RMSE
- MAE
- R2

Machine Learning
Aprendizaje supervisado

Árboles de decisión



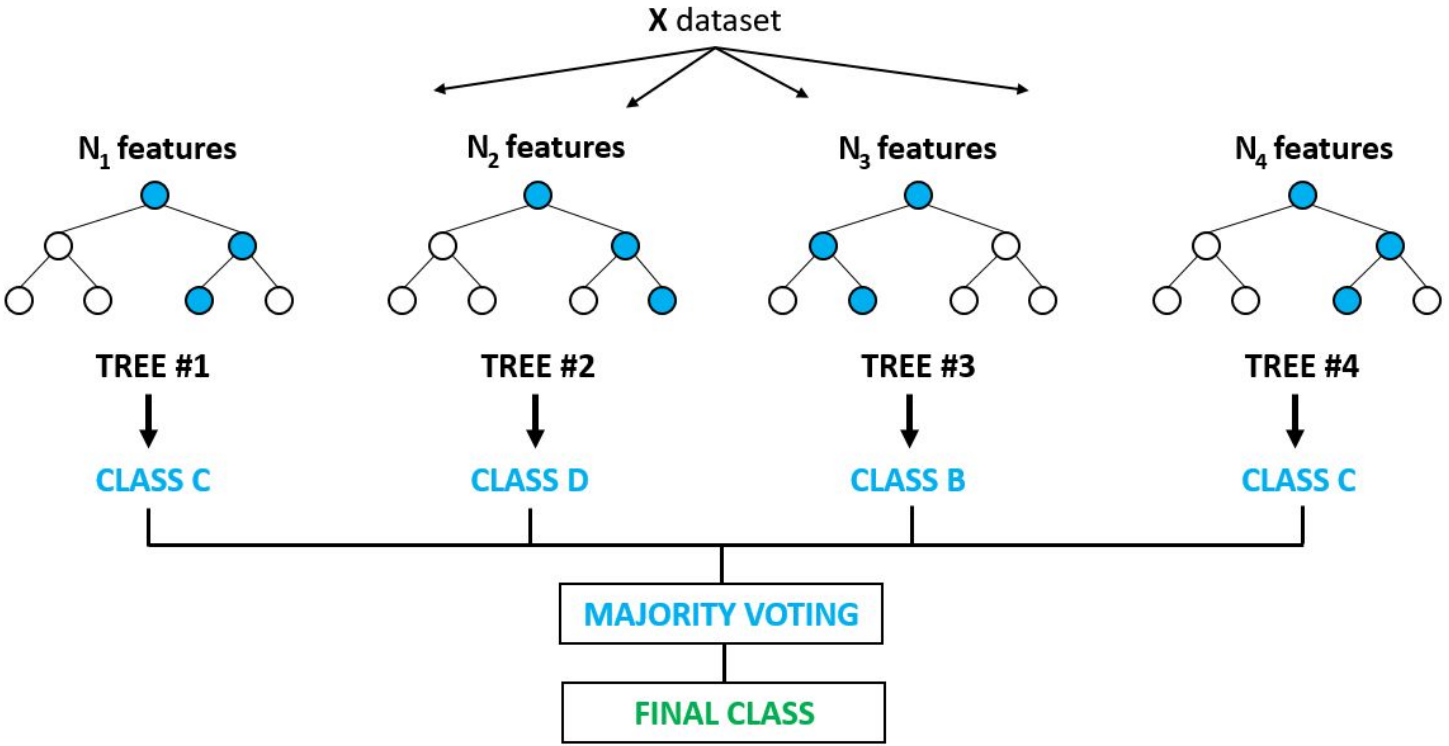
Árboles de decisión

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.

http://www.milbo.org/rpart-plot/prp.pdf

Random Forest



Random Forest

Hiperparámetros

- **mtry:** n° de predictores a muestrearse en cada split de árbol
- **min_n:** n° de observaciones necesarias para seguir dividiendo nodos
- nodesize: n° mínimo de casos permitidos en una hoja (parecido a minbucket en rpart)
- **maxnodes:** n° máximo de hojas permitidas

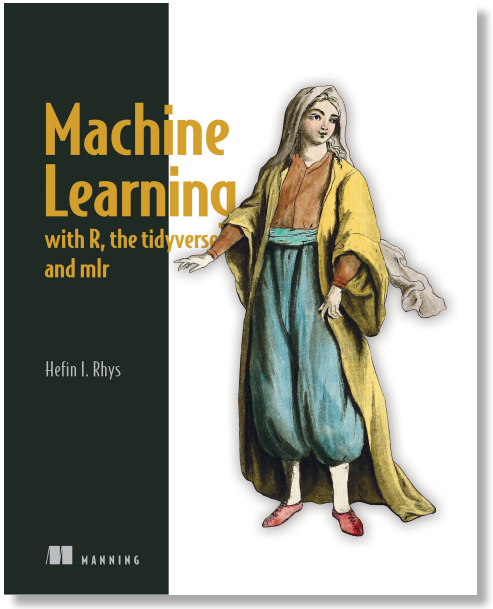
CARET (Classification And REgression Training)

Modelos disponibles

https://topepo.github.io/caret/available-models.html

Documentación oficial

https://topepo.github.io/caret/index.html



Machine Learning with R, the tidyverse, and mlr

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

Veamos algo de código...