Ciencia de Datos

- **Módulo 4**

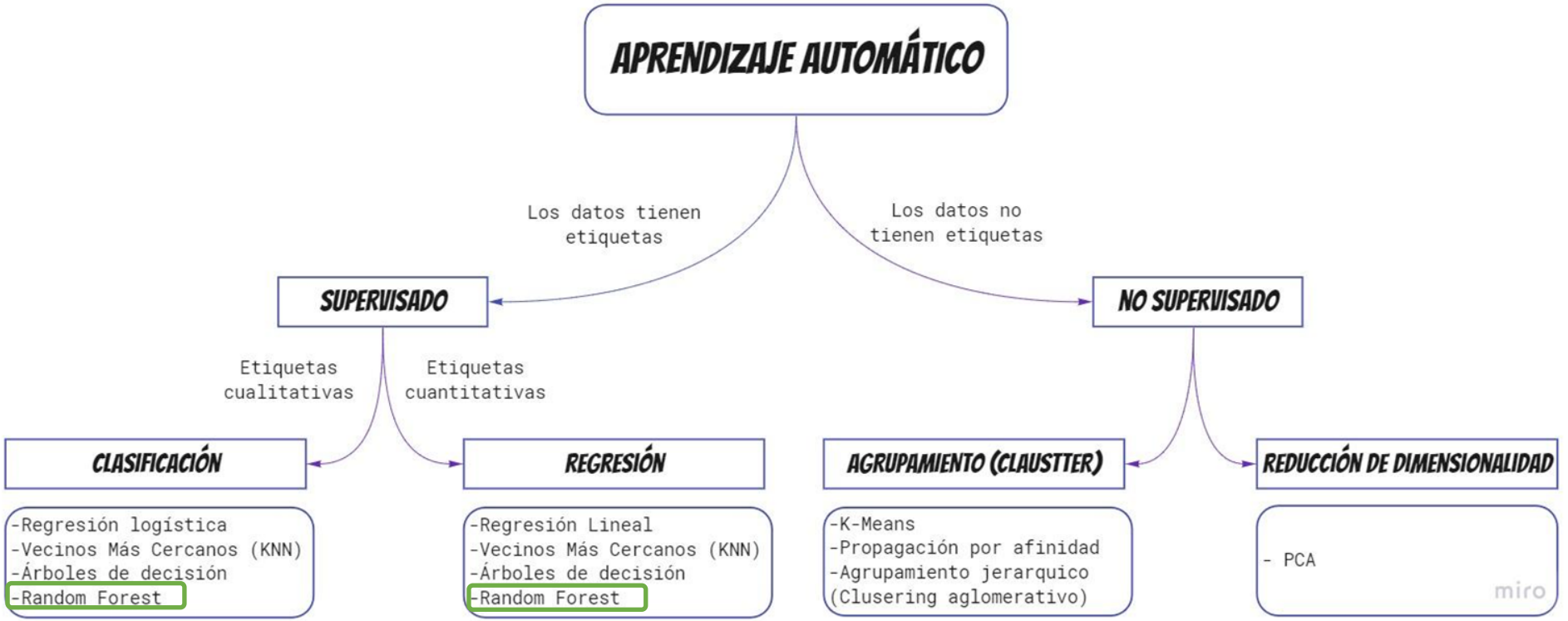
Random Forest

UTNFra

GyEE Gestión y Evaluación Educativa

<aprendé programando Virtual>

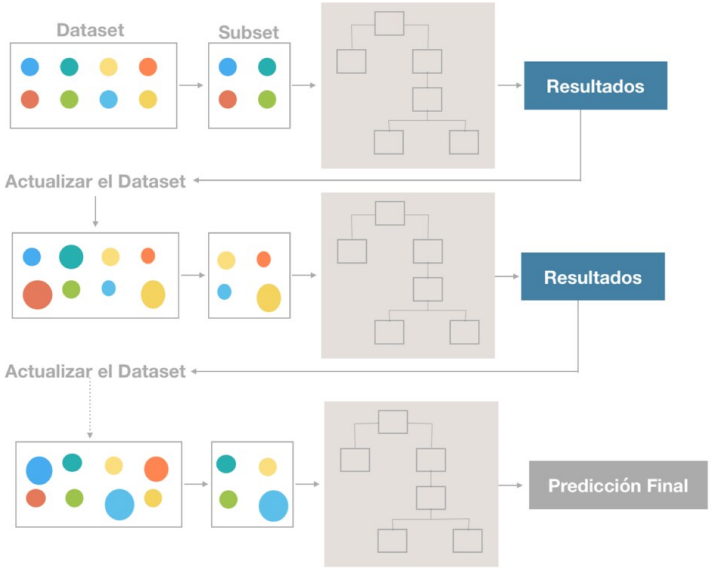
SECUNDARIA del futuro


APRENDIZAJE AUTOMÁTICO
Los datos tienen etiquetas
Los datos no tienen etiquetas
SUPERVISADO
NO SUPERVISADO
Etiquetas cualitativas
Etiquetas cuantitativas
CLASIFICACIÓN
REGRESIÓN
AGRUPAMIENTO (CLAUSTTER)
REDUCCIÓN DE DIMENSIONALIDAD
-Regresión logística
-Vecinos Más Cercanos (KNN)
-Árboles de decisión
-Random Forest
-Regresión Lineal
-Vecinos Más Cercanos (KNN)
-Árboles de decisión
-Random Forest
-K-Means
-Propagación por afinidad
-Agrupamiento jerarquico
(Clusering aglomerativo)
- PCA
miro

Ensambles

Muchas veces entrenamos muchos modelos para resolver un mismo problema descubriendo que cada modelo es bueno en alguna sección del mismo, pero ninguno en la totalidad. La principal hipótesis es que cuando los modelos se combinan correctamente podemos obtener modelos más precisos y / o robustos. Esto es lo que se conoce como ensambles de modelos.

Es decir, combinando distintos modelos podemos tener un modelo mejor



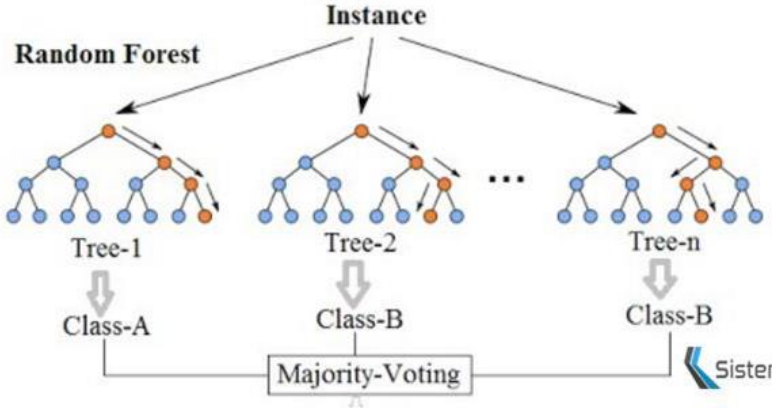
Fuente: https://aprendeia.com/metodos-de-ensamble-de-modelos-machine-learning-ensemble-methods-en-espanol/

Random Forest

Uno de los modelos más utilizados que usan la técnica de Ensamble(Bagging) es **Random Forest** que, como su nombre lo indica, es un conjunto aleatorio de Árboles de decisión.

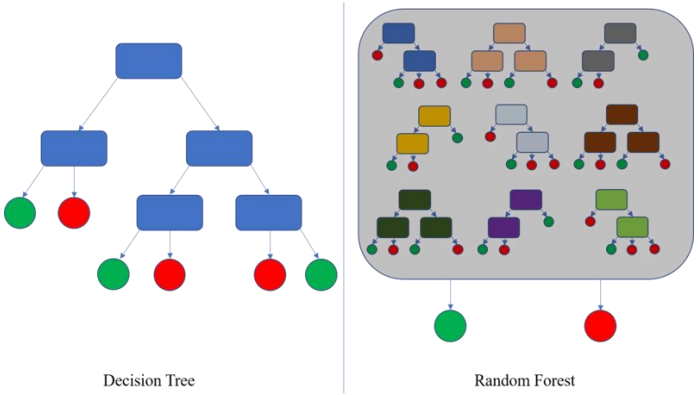
Los Árboles de decisión son modelos rápidos en su entrenamiento y fácilmente interpretables, pero muchas veces tienen el problema del Sobreajuste.

Una manera de resolver este problema es la combinación de muchos Árboles de manera aleatoria.



Fuente: https://sistemasvirtual.com/random-forests/

Random Forest



Fuente: https://commons.wikimedia.org/wiki/File:Decision_Tree_vs._Random_Forest.png

El entrenamiento de cada Árbol no se realiza con todo el conjunto de datos de entrenamiento, en su lugar, para cada árbol se seleccionan aleatoriamente **un grupo de observaciones y un grupo de variables**, de forma que no existan dos árboles que vean exactamente la misma información.

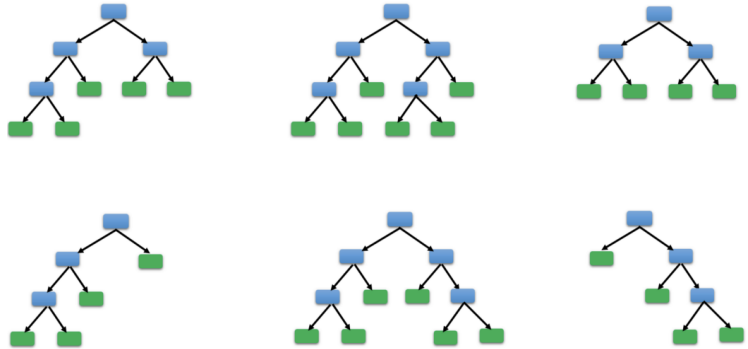
Para las predicciones, cada árbol genera su propia predicción. Si es un problema de regresión, se hace un promedio de todas las predicciones, en el caso de ser un problema de clasificación se hace una “votación”, ganando la clase que la mayor cantidad de árboles escogió.

Random Forest

Sobre un conjunto de datos – Dataset – se realiza una selección aleatoria de registros (con reposición) creando distintos Dataset.

Asimismo, sobre cada uno de esos Daset se seleccionan variables al azar para construir los Árboles de decisión.

De esta manera se construyen X cantidad de Árboles cada uno con una composición distinta de datos y de variables.



Fuente: https://fhernanb.github.io/libro_mod_pred/rand-forests.html

Al darle datos nuevos, cada Árbol realiza una predicción: en caso de Clasificación la predicción con más votos será la seleccionada, en el caso de una Regresión el promedio.

Random Forest

Vídeo

- https://www.youtube.com/watch?v=v6VJ2RO66Ag

Documentación Scikit-Learn

- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Adicional

- https://www.youtube.com/watch?v=J4Wdy0Wc_xQ
- https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205