Minería de Datos
IIC2433

Ensembles

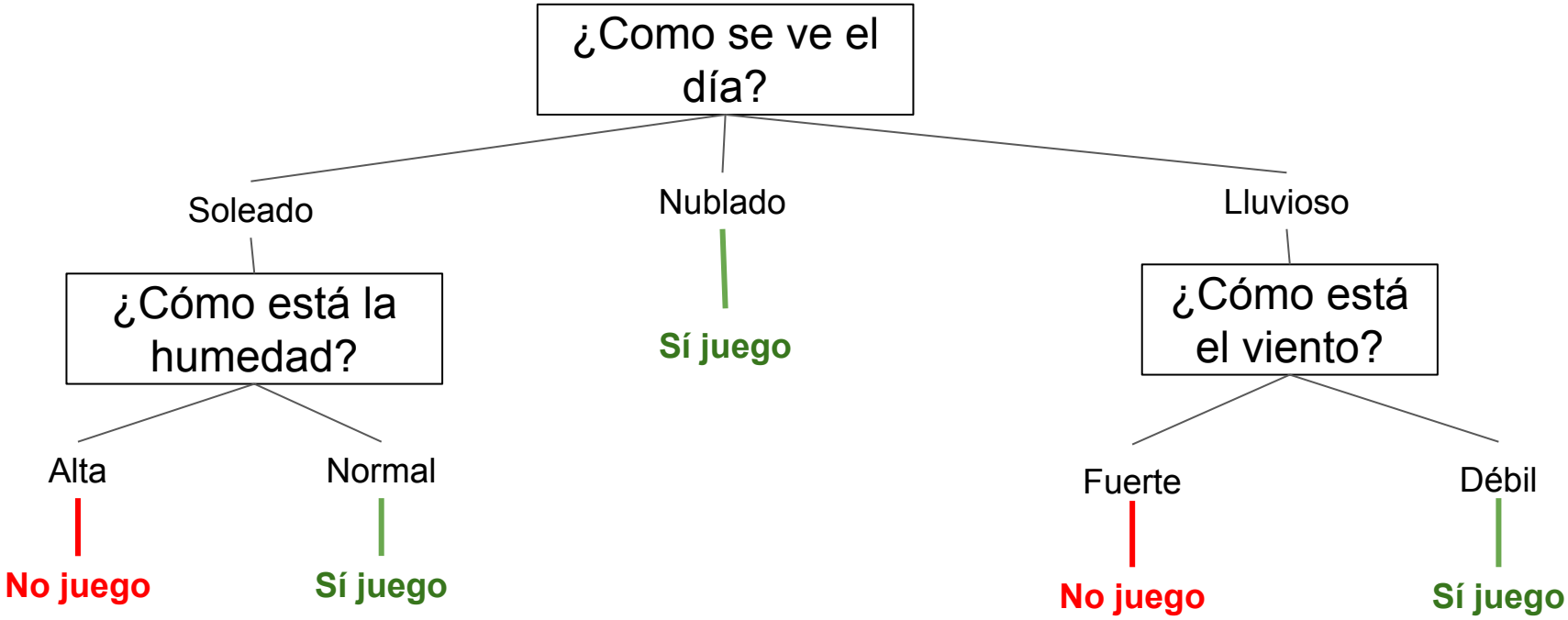
Vicente Domínguez

¿Qué veremos esta clase?

- El modelo Random Forest
- Qué es un Ensemble

Árbol de decisión

Recordando

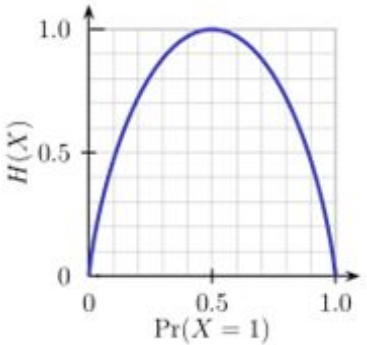


Árbol de decisión

•Los nodos del árbol representan variables, las ramas representan valores de las variables que permiten clasificar

•Las hojas del árbol corresponden a la clasificación

•En la construcción del árbol, se testea una variable a la vez, usando el concepto de Entropía (número de bits necesarios para transmitir un mensaje - o - nivel de incerteza respecto a un evento)



$$H = -\sum_{i=1}^{M} P_i \log_2 P_i$$

¿Problemas con este modelo?

- ¿Qué ocurre cuando tenemos muchos atributos?
- ¿Cuánto cuesta construir un árbol?

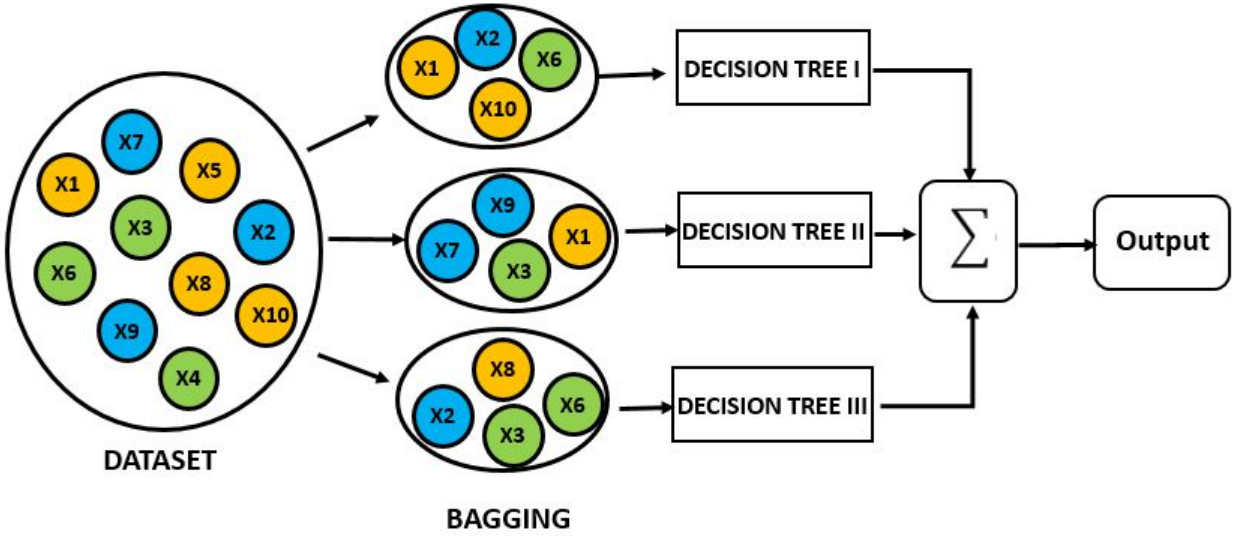
¿Cómo mejorar este modelo?

- Una parte importante del área de minería de datos es la de buscar mejoras a los modelos actuales.
- Ahora que tenemos uno de nuestros primeros modelos complejos de clasificación, ¿cómo podríamos mejorarlo?
- ¿Qué ocurre cuando tenemos muchos atributos?

¿Cómo mejorar este modelo?

- Leo Breiman propuso una mejora al árbol decisión, el cual es un modelo con muy buenos resultados y rendimiento.
- Actualmente es muy utilizado en la industria y academia.

Random Forest

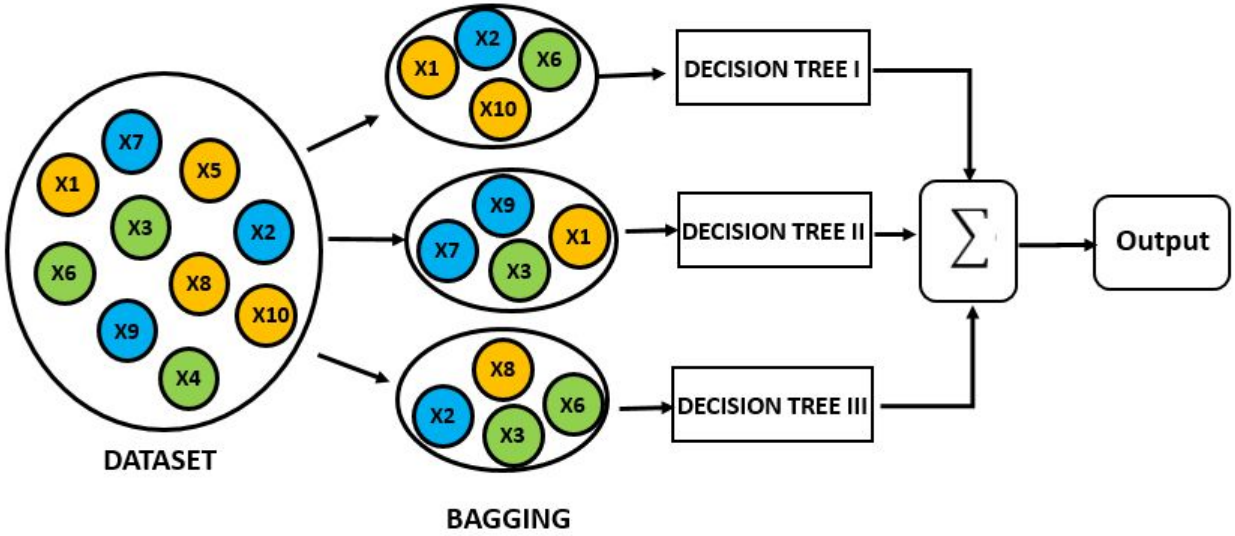


https://medium.com/greyatom/a-trip-to-random-forest-5c30d8250d6a

Random Forest

- Modelo basado en los árboles de decisión.
- Como su nombre lo dice, genera un bosque o selva de ellos para tomar una decisión.
- Aparte de eso, cada árbol está formado por un subconjunto de los atributos totales.
- Finalmente, para clasificar se genera una votación entre todos los árboles.

Random Forest



https://medium.com/greyatom/a-trip-to-random-forest-5c30d8250d6a

Random Forest
Paso 1

- Se definen los parámetros del algoritmo, estos son:
  - **n_estimators**: la cantidad de árboles a utilizar
  - **max_features**: cantidad máxima de atributos a utilizar por cada árbol
  - **max_depth**: profundidad máxima de cada árbol
  - **criteria**: criterio para elegir atributos

Random Forest
Paso 2

- Se hace el proceso de *bagging* el cual se realiza el *bootstraping*:
  - ***bootstraping***: se hace un oversample o sobre muestra de los datos para cada árbol.
- Luego de cada uno de estos pasos se genera un *bag* de datos para cada árbol.

Random Forest
Paso 3

- Se entrena cada árbol con su bag, generando un conjunto de árboles entrenados para clasificar o generar alguna regresión sobre un conjunto de datos.
- Cada vez que se vaya a elegir un nodo, este toma un sub set de features de tamaño **max_features, y elige un atributo entre ellos.**

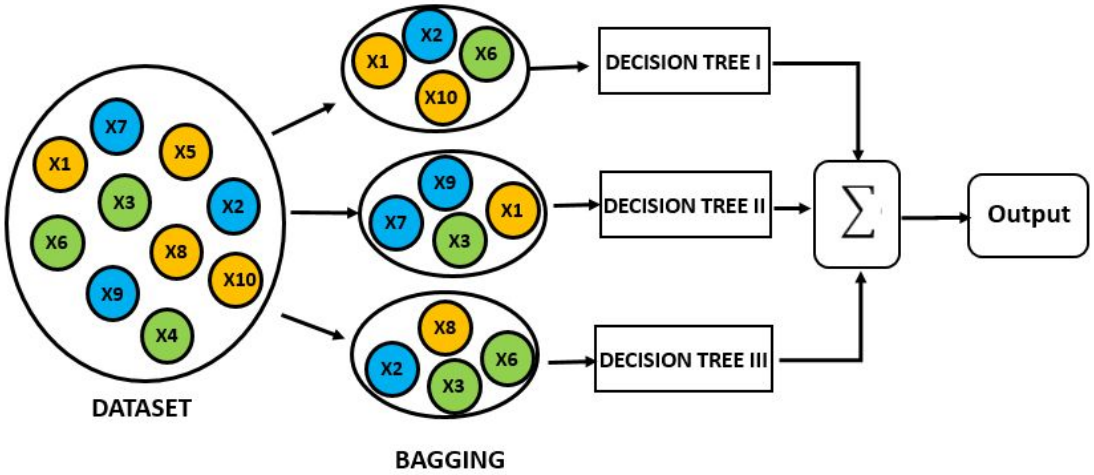
Random Forest
Paso 3

- Para medir el rendimiento del algoritmo se puede obtener el *Out of Bag (OOB) score.*
- Para obtener el OOB score, lo que se hace es ver cada dato, y clasificarlo por cada árbol que no lo usó para entrenar. Se genera un votación entre ellos y se mide el error de clasificación.
- Para más detalles del OOB score, les recomiendo revisar este post

Random Forest

Paso 4

- Luego de obtener un buen OOB score, el algoritmo cuando reciba un nuevo dato lo va a clasificar en base una votación entre todos los árboles entrenados del bosque.



Random Forest

Ventajas

- Evita el sobre ajuste
- Funciona bien con grandes cantidades de datos.
- Funciona bien con una gran cantidad de atributos.
- Puede ejecutarse de forma paralela cada árbol, entrenando de forma eficiente.

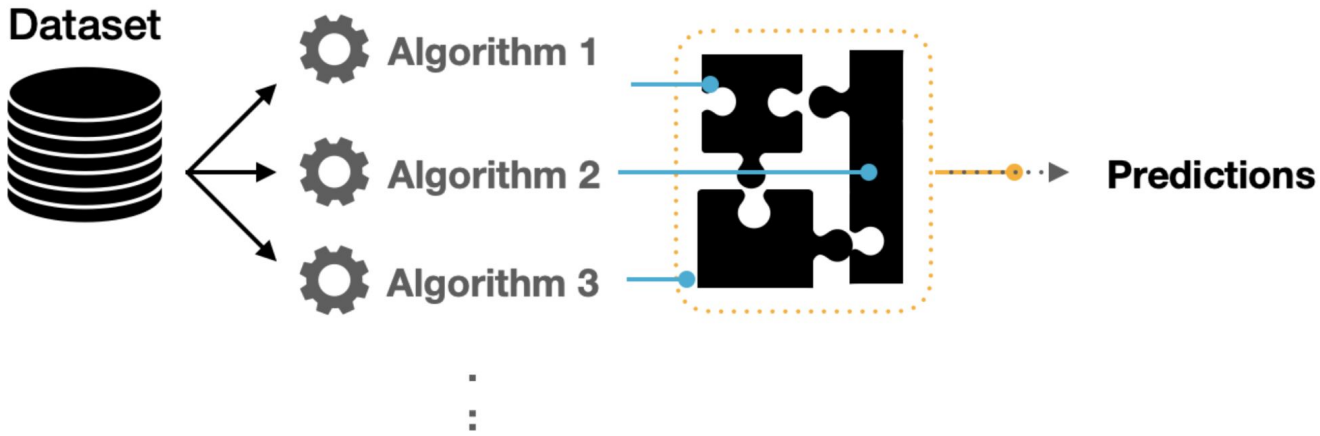
Random Forest

Desventajas

- Es difícil de interpretar, a diferencia del árbol de decisión
- Tarda más en generarse, es computacionalmente más costoso
- Si es que los datos son ruidosos, puede sobre ajustarse al ruido

Random Forest
Ensemble

- El modelo Random Forest es un Ensemble
- ¿Qué es un Ensemble?



Fuente: https://towardsdatascience.com/ensemble-models-5a62d4f4cb0c

Ensemble Learning

- Plantea que una mayor diversidad de modelos mejoran el performance general.
- Se pueden generar varias combinaciones, ya sea votaciones, ponderaciones o incluso, un aprendizaje continuo.

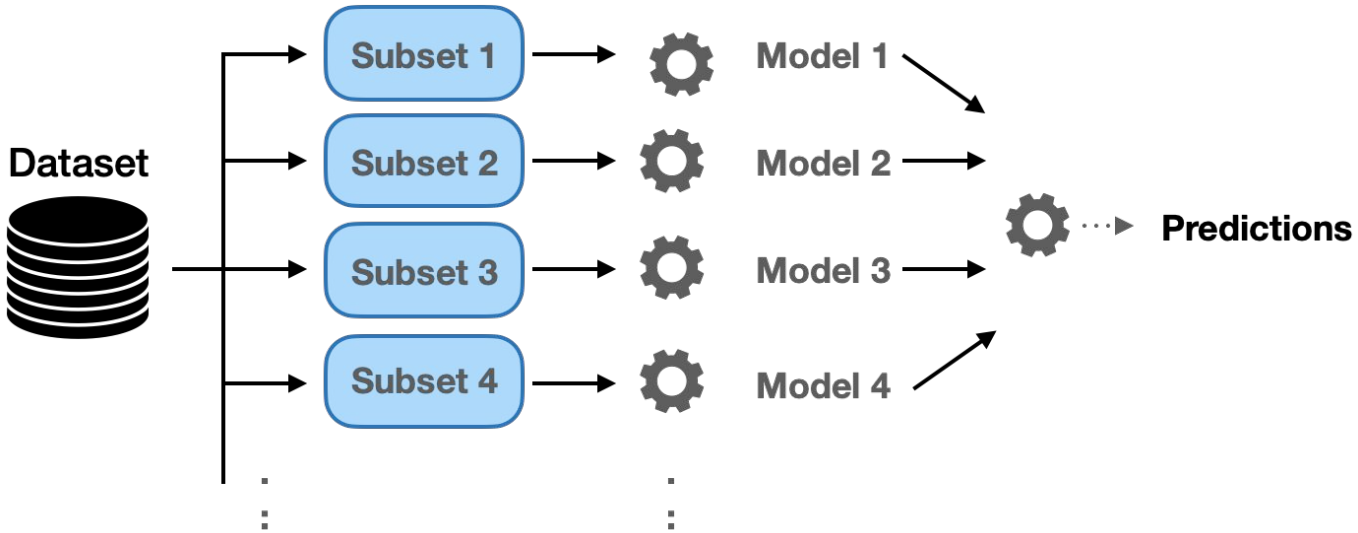
Ensemble Learning

Tipos

- En la actualidad cada vez se han generado formas más creativas de combinar modelos.
- Entre las clásicas, existen las siguientes:
  - Bagging
  - Boosting
  - Stacking

Ensemble Learning

Bagging



Fuente: https://towardsdatascience.com/ensemble-models-5a62d4f4cb0c

Ensemble Learning

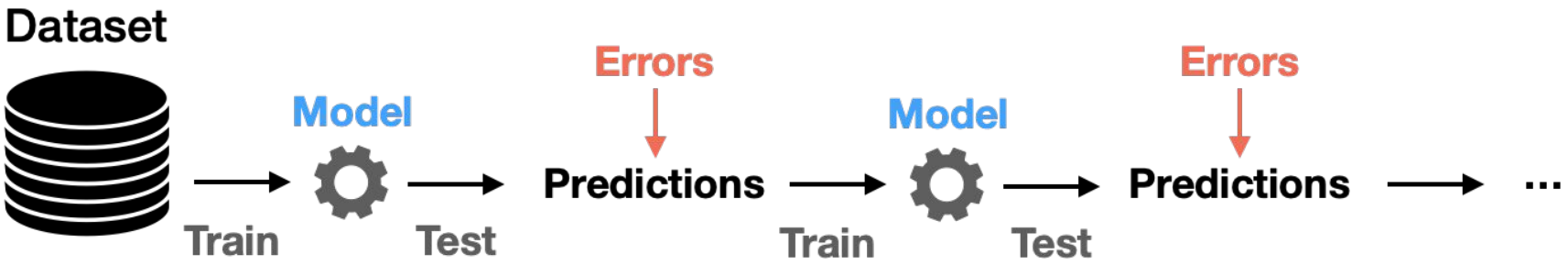
Bagging

- Consiste en generar K sub sets del dataset y entrenar K modelos con esos sub sets
- En general se utiliza bootstraping lo cual en el caso ideal samplea un ~63% de los datos.

$$\lim_{N\to\infty}\left(1-\frac{1}{N}\right)^{N}=e^{-1}=0.368$$

Ensemble Learning

Boosting



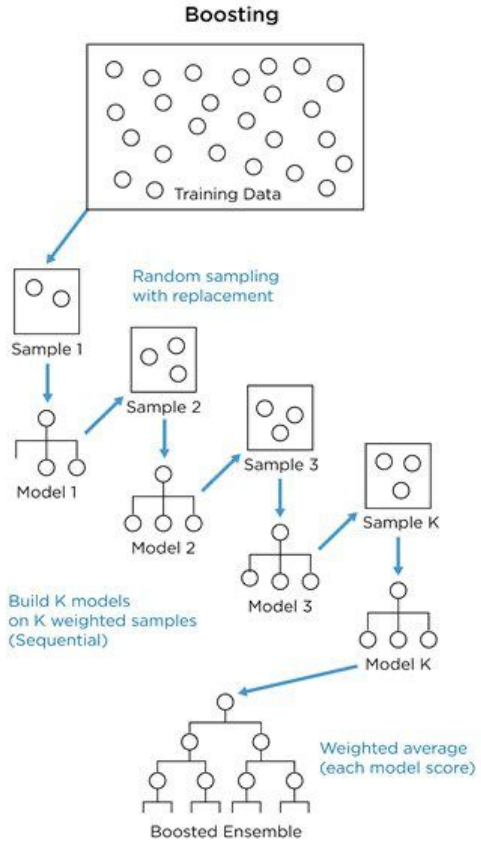
Fuente: https://towardsdatascience.com/ensemble-models-5a62d4f4cb0c

Ensemble Learning
Boosting

- Consiste en generar un conjunto de modelos, que aprenden de forma **secuencial.**
- Las formas más conocidas de aprender son dos:
  - Adaptative Boosting (AdaBoost)
  - Gradient Boosting (XGBoost)

Ensemble Learning
Boosting

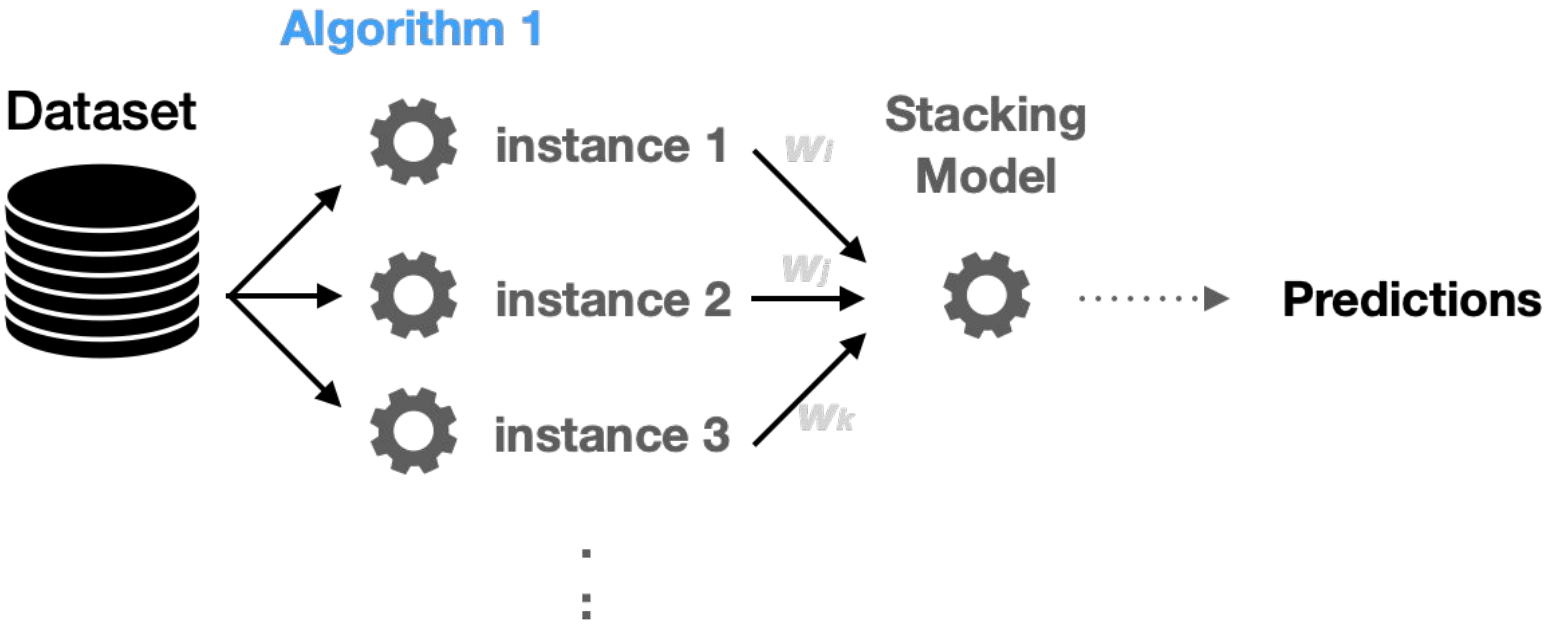
- Se entrenan un conjunto de modelos conocidos como weak learners
- Se puede utilizar una función de pérdida que busque minimizar el error entre la predicción del modelo anterior y el modelo actual.
- Este error es conocido como el residual y es el que se busca minimizar



Fuente: https://www.kdnuggets.com/2019/09/ensemble-learning.html

Ensemble Learning

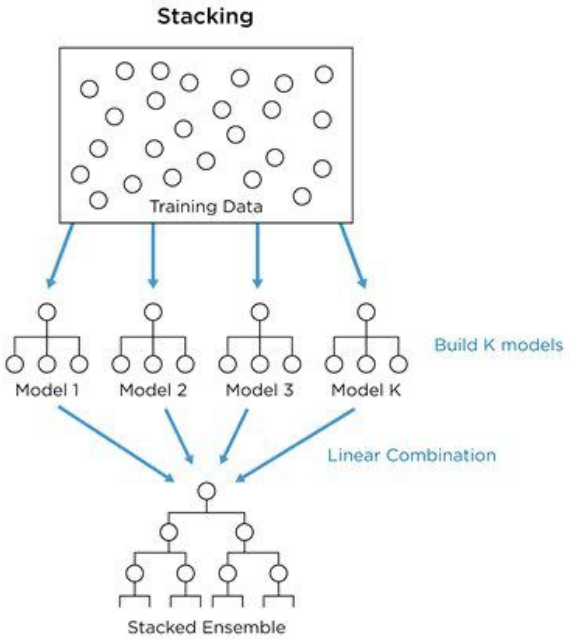
Stacking



Fuente: https://towardsdatascience.com/ensemble-models-5a62d4f4cb0c

Ensemble Learning
Stacking

- Consiste en entrenar un montón de modelos con el dataset completo.
- Pueden ser modelos completamente distintos
- La predicción realiza en base a una votación ponderada, en caso de clasificación
- En regresión puede ser una combinación lineal de las predicciones



Fuente: https://www.kdnuggets.com/2019/09/ensemble-learning.html

Ensemble Learning
Resumen

- Muchos de los algoritmos de ensembles se utilizan mucho en la actualidad.
- Hay librerías especializadas para utilizar este tipo de modelos, incluso optimizadas en cuanto a temas de performance o memoria utilizada.
- Un ejemplo es LightGBM