

# Essential Machine Learning with Python

[https://www.facebook.com/qslearningperu/?ref=page\\_internal](https://www.facebook.com/qslearningperu/?ref=page_internal)



QS Learning

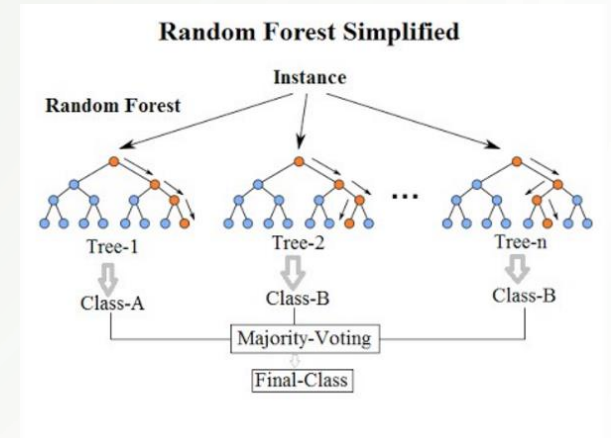
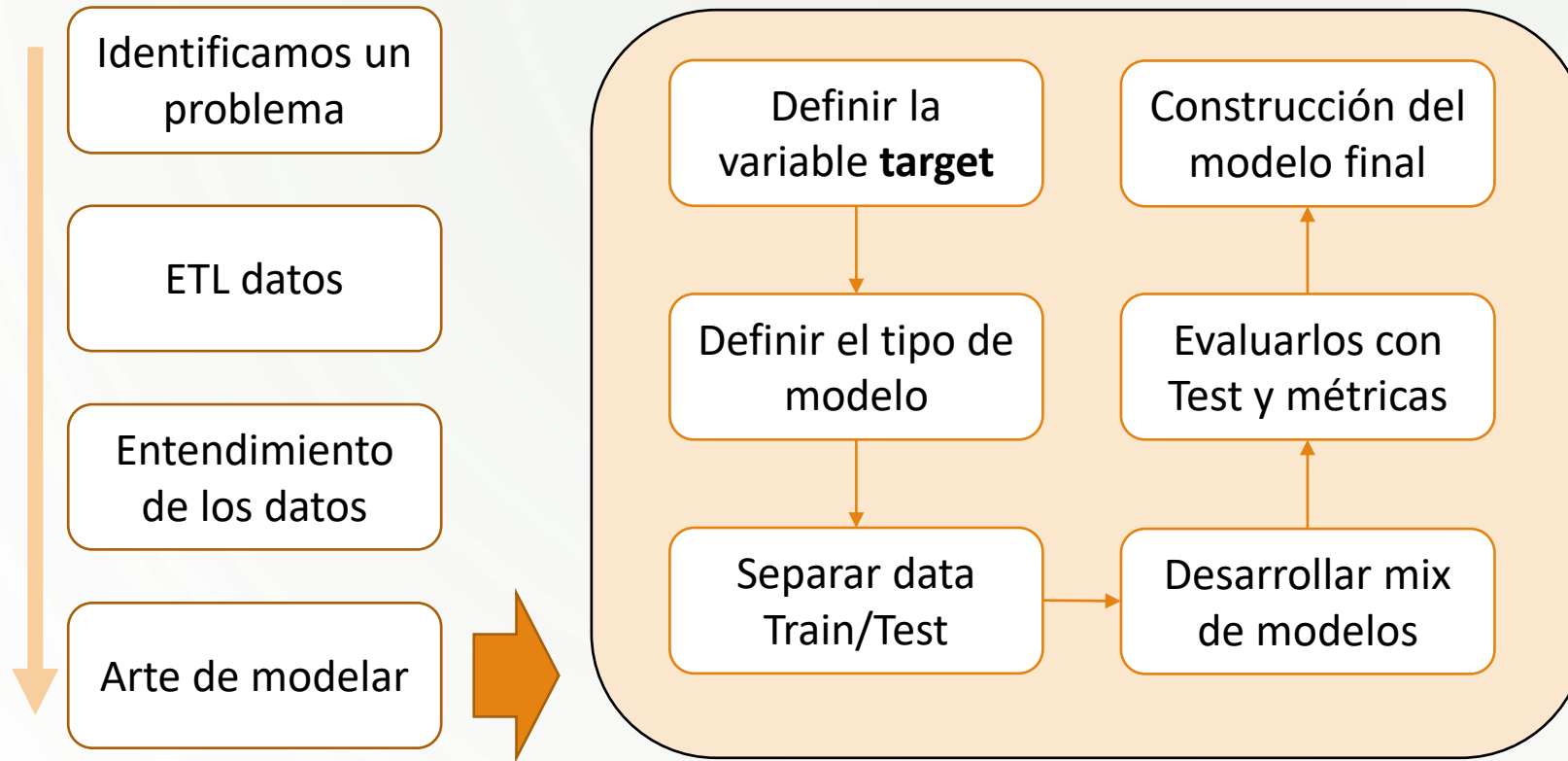


+51 937 012 707 / +51 915 111 457



quants.admission@gmail.com

# Modelizacion



# MÉTODOS ENSAMBLADOS



“Solos podemos hacer muy poco y  
juntos podemos hacer mucho”  
**Helen Keller**

Lo mismo se aplica a la mayoría de los casos de la innovación con grandes impactos y con tecnologías avanzadas en el mundo.

El dominio del **Machine Learning** también está en la misma carrera para hacer predicciones y clasificaciones de una manera más precisa utilizando el llamado **ensemble method** y se ha demostrado que el **ensemble modeling** ofrece una de las formas más convincentes de construir modelos predictivos de alta precisión.

Los métodos de conjunto son modelos de aprendizaje que alcanzan el rendimiento al combinar las opiniones de múltiples de predictores base

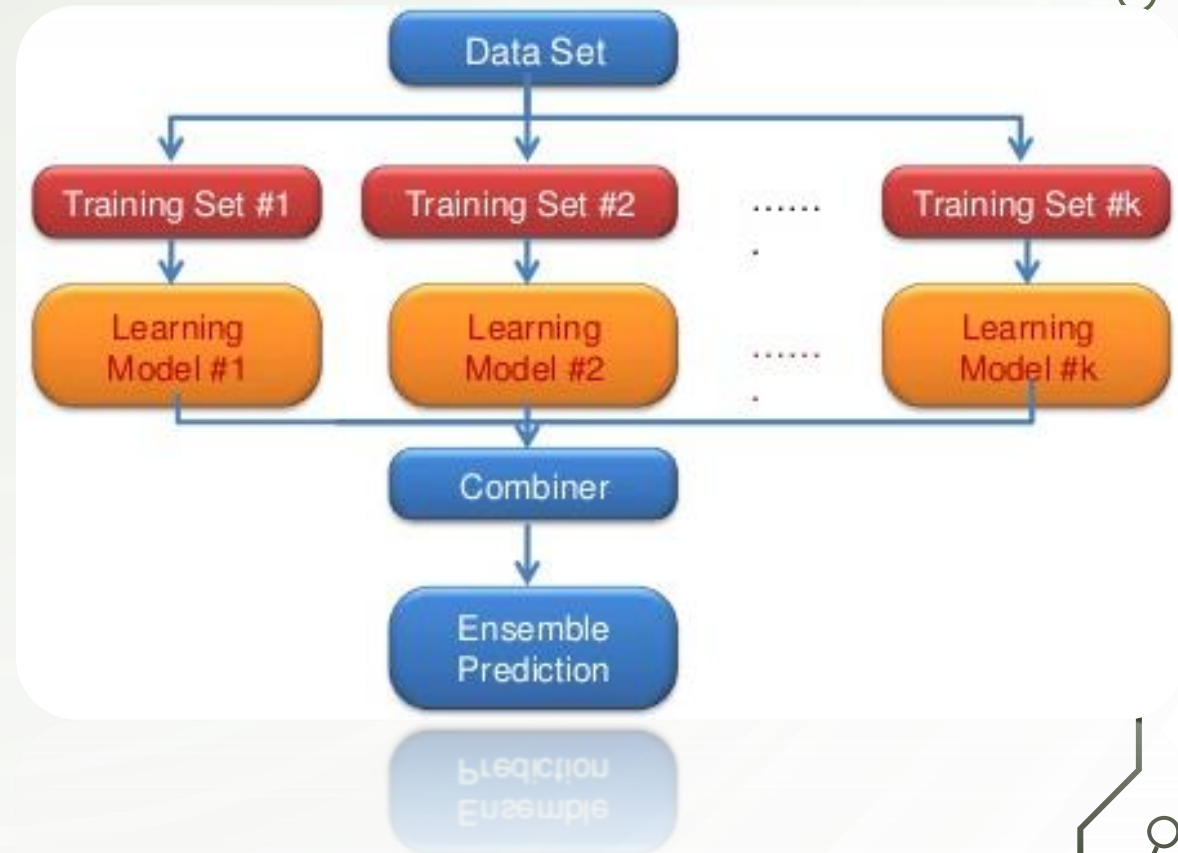
# ¿QUÉ SON LOS MODELOS ENSAMBLADOS?

Son métodos que intentan sobrepasar las limitaciones de modelos individuales mediante la agregación de un conjunto de estimadores base.

Estos métodos ensamblados suelen ser bastante más precisos que el estimador base.

Para que el ensamblado de un modelo funcione, el estimador base tiene que cumplir varios requisitos:

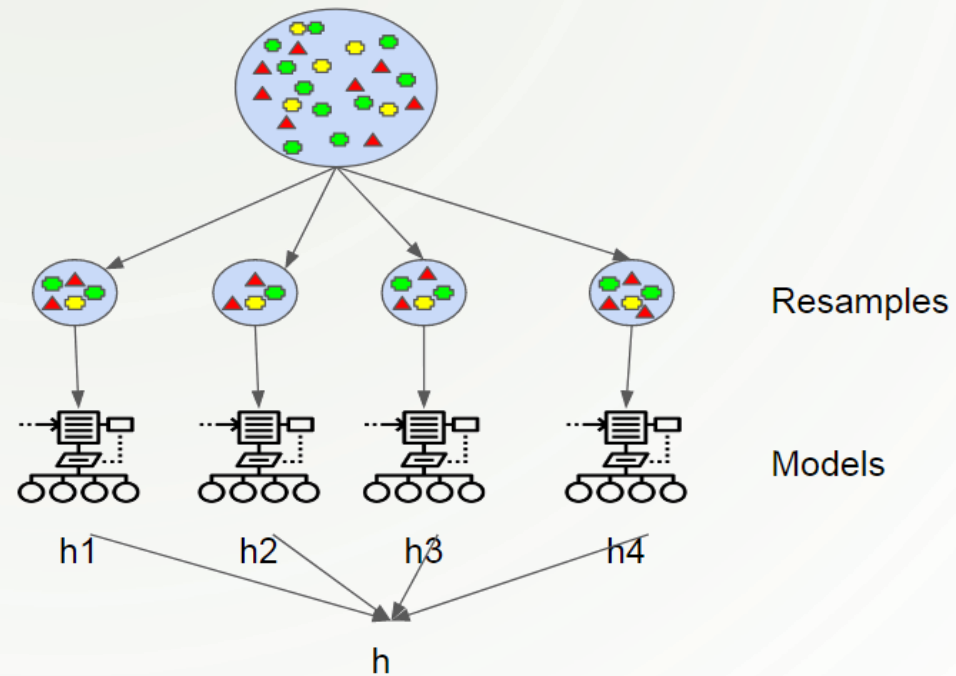
- El estimador base tiene que ser preciso, al menos más preciso que el tirar una moneda al aire.
- Los estimadores base tienen que ser diversos, sus errores deben ocurrir en distintas observaciones.



# ENSEMBLE MODELING

Existen diversas técnicas de ensamblado de modelos, de los cuales revisaremos lo siguientes:

- Bagging
- Boosting
- Stacking
- Bosques aleatorios
- XGBoost





# BAGGING

La agrupación o la agregación Bootstrap de datos de muestra se puede utilizar para reducir el problema de sobreajuste y también de reducir la varianza.

Es un caso especial de promediar modelos que son entrenados con una muestra del dataset completo, siempre con el mismo tamaño.

Al tomar la muestra uniformemente y con reemplazo sin generar datos arbitrarios. Se espera que el tipo de muestreo bootstrap tenga la fracción  $(1 - 1/e)$  que es casi equivalente a 63.2% de registros únicos y otros duplicados. Este concepto se utiliza principalmente en los métodos de árbol de decisión.

Cada uno de los modelos se entrenan independientemente.



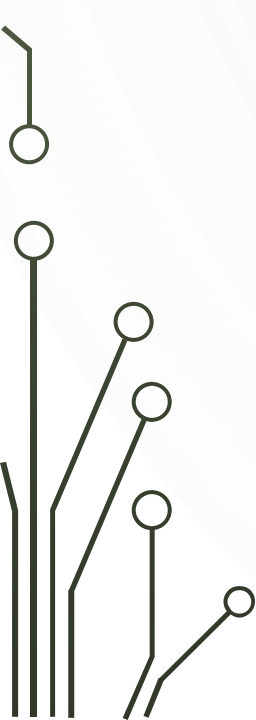


# BOOSTING

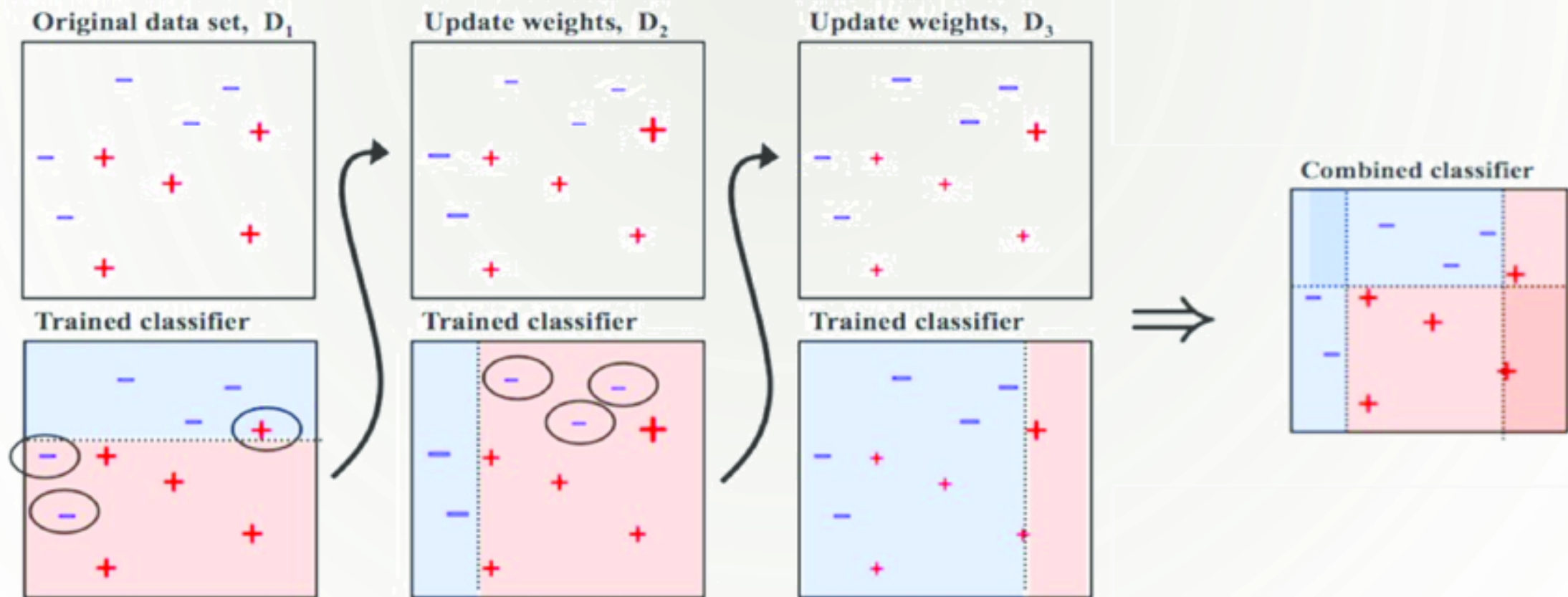
Los **Boosting** son un método ensamblado iterativo que cambia la distribución de muestro de los datos de entrenamiento iterativamente. En cada iteración se busca aprender de los errores en la iteración anterior.

Una de las primeras implementaciones del **Boosting** es el algoritmo llamado **AdaBoost**.

Algoritmo **AdaBoost**:

1. Tomar una muestra (con reemplazo) del dataset de entrenamiento, donde todas las observaciones tienen la misma probabilidad de ser elegidas.
  2. Entrenar un estimador base (cb1)
  3. Calcular el error de entrenamiento  $e(cb1)$  y marcar las predicciones erradas.
  4. Tomar una nueva muestra del dataset, donde aquellas observaciones clasificadas incorrectamente por cb1 tengan un peso igual al error (forzamos a que sean seleccionados)
  5. Repetir 3 y 4 k veces (para usar k estimadores base)
  6. Predecir en base a una votación ponderada, donde los pesos de cada estimador base dependen inversamente de su error de entrenamiento.
- 

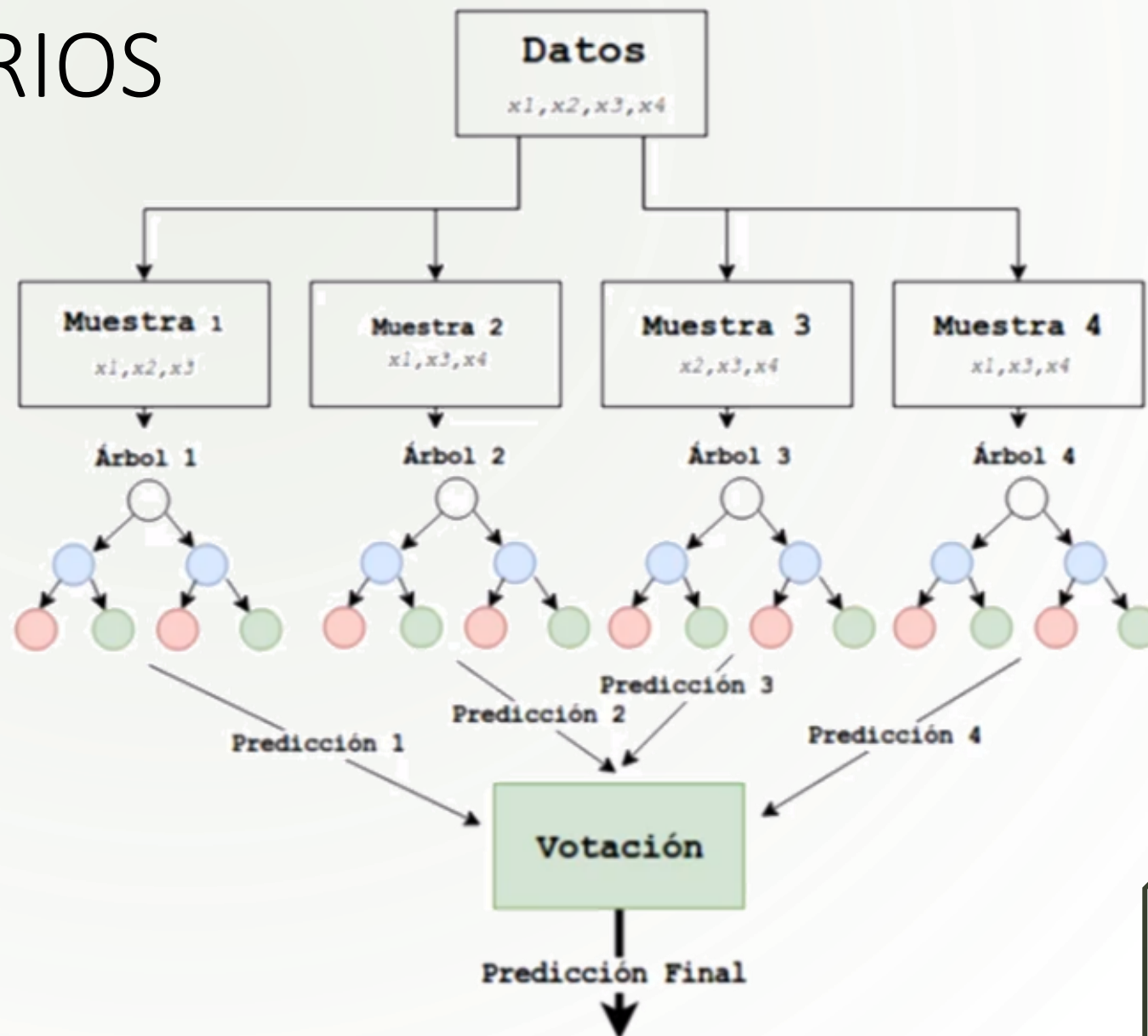
# BOOSTING





# BOSQUES ALEATORIOS

Los **Bosques aleatoriso (Random Forest)** al igual que los métodos ensamblados anteriores trabaja con muestras del dataset total, la única diferencia es que va muestreando también la cantidad de variables que usará en cada estimador base (Árboles de decisión)

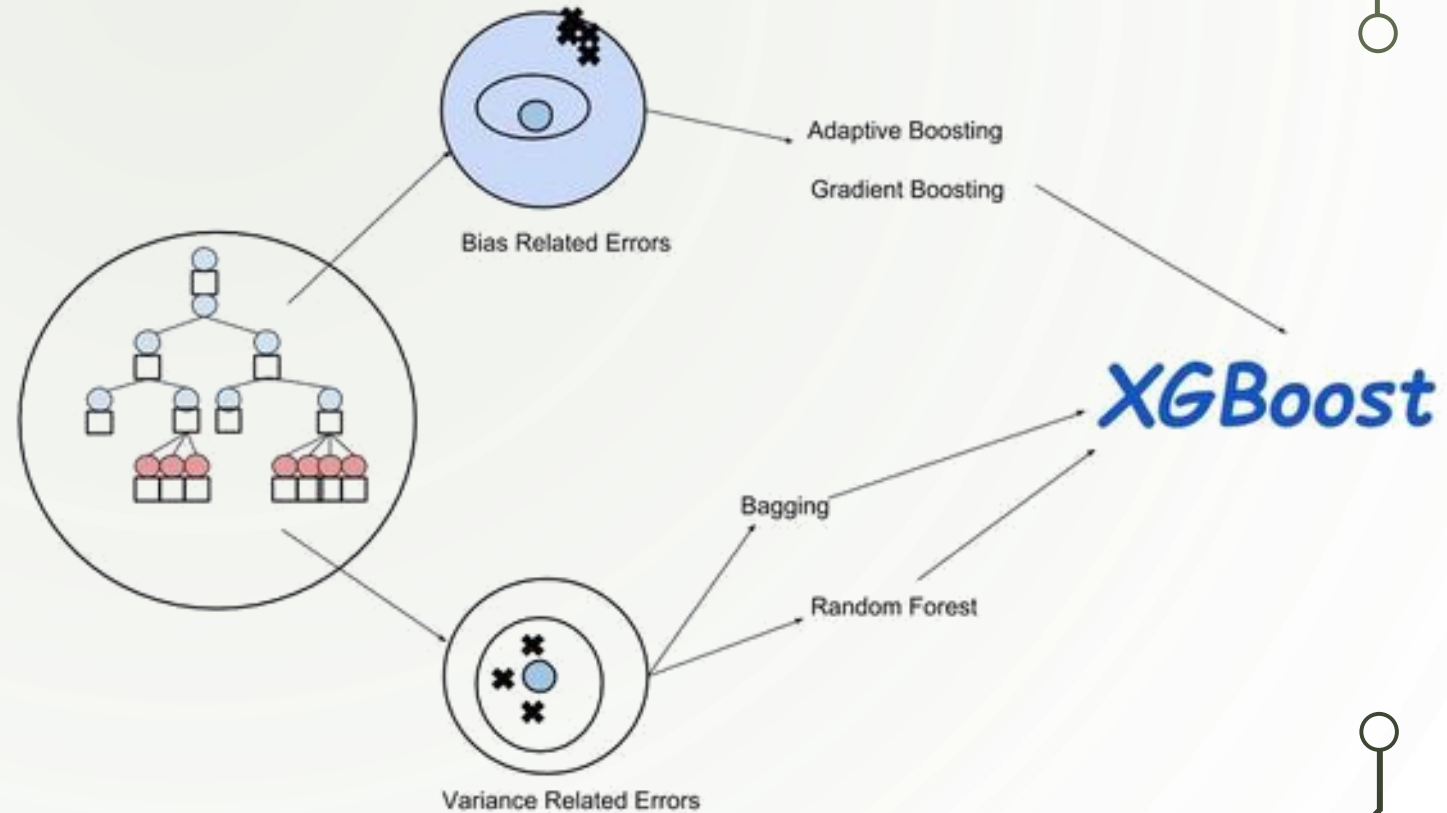


# XGBOOST

XGBoost es un algoritmo que ha estado dominando recientemente las competencias de aprendizaje automático aplicado y Kaggle para datos estructurados o tabulares.

XGBoost es una implementación de árboles de decisión mejorados con gradiente diseñados para la velocidad y el rendimiento.

Es una mezcla de boosting y bagging pero con la característica de estar optimizado para dataset extensos.



# STACKING

El **STACKING** es un enfoque diferente de combinar varios modelos con el concepto de meta-aprendizaje. Aunque el enfoque no tiene una fórmula empírica para la función de peso, la funcionalidad final es la misma que la de bagging y boosting.

Stacking considera los siguientes pasos:

1. Dividir el dataset en dos grupos (train, test).
2. Entrenar distintos estimadores base con train.
3. Poner a prueba los modelos con test.
4. Usando las predicciones del paso anterior como entradas y las respuestas correctas como salidas, entrene un estimador de nivel superior.

