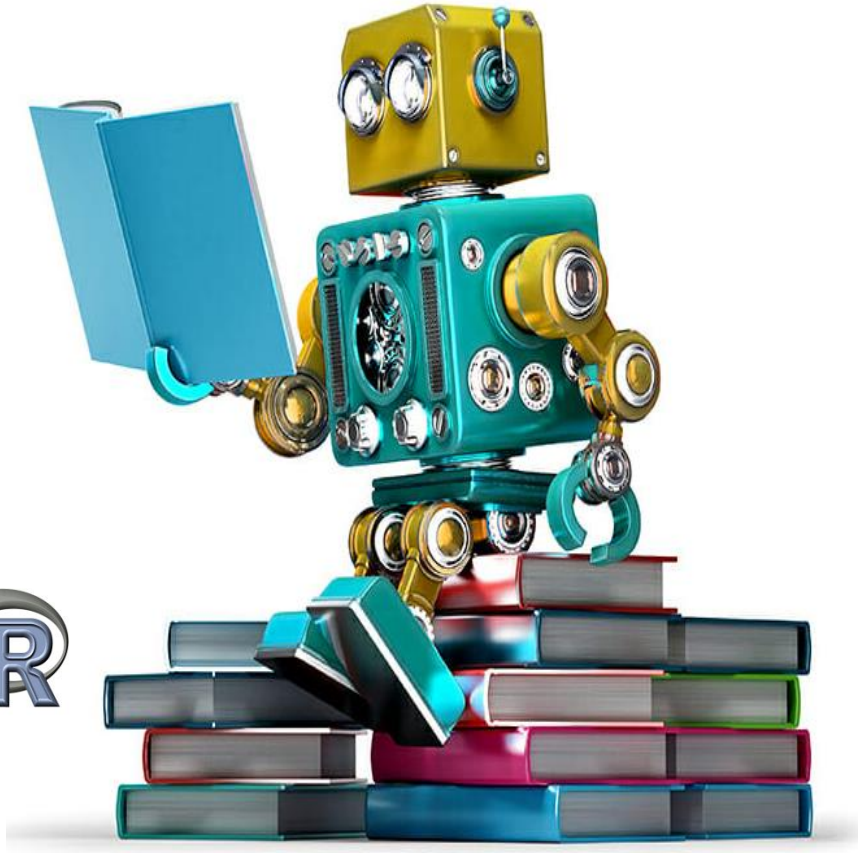
ALGORITMOS DE CLASIFICACIÓN : RANDOM FOREST



«Lo llaman suerte, pero es constancia. Lo llaman casualidad, pero es disciplina. Lo llaman genética pero es sacrificio. Ellos hablan , tú estudia.»»

San Marcos Data Science Community

Mentores

José Antonio Cárdenas Garro
ESTADÍSTICA
UNMSM
MSc in Data Science Candidate
Promotion "Erwin Kraenau Espinal"
Universidad Ricard Palma

Predictive Modelling Specialist Scotiabank

Especialidades : Predictive Modeling | Machine Learning | Advanced Statistics | Risk Modeling | Data Science

André Omar Chávez Panduro
ESTADÍSTICA
UNMSM
MSc in Data Science Candidate
Promotion "Erwin Kraenau Espinal"
Universidad Ricard Palma

Data Scientist Interbank

Especialidades : Predictive Modeling | Machine Learning | Advanced Statistics | Business Analytics| Data Science

Aldo Ray Chávez Panduro
INGENIERÍA DE SISTEMAS
UNMSM
Student of MSc in Data Science
Universidad Ricard Palma

Risk Management Specialist BBVA Continental

Especialidades : Big Data| Machine Learning | Programming | Risk Specialist – IFRS9 | Data Science

SMDSC

Agenda

- Definición de Clasificación.
- Ejemplos de clasificación.
- Bagging.
- Random Forest.

DEFINICIONES BÀSICAS

- **Conjunto de Datos (Data Set):** El total del conjunto de datos sobre los que queremos desarrollar un algoritmo de Machine Learning con el fin de obtener un modelo que lo represente lo mejor posible. Contendrá variables independientes y dependientes.

- **Variables Independientes (Features), (VI):** Aquellas columnas del Data Set que serán usadas por el algoritmo para generar un modelo que prediga lo mejor posible las variables dependientes.

- **Variables dependientes (Labels,Target), (VD):** Columna del data set que responde a una correlación de VI y que debe ser predicha por el futuro modelo

- **Conjunto de Datos de Entrenamiento (Training Set):** Subconjunto del Data Set que será utilizado para entrenar el modelo que se pretende generar.

- **Conjunto de Datos de Test (Test Set):** Subconjunto del data set que se le pasará al modelo una vez haya sido entrenado para comprobar, mediante el uso de diferentes métricas, sus indicadores más importantes de calidad.

CLASIFICACIÓN: DEFINICIÓN

- Dada una colección de registros (Conjunto de Entrenamiento) cada registro contiene un conjunto de variables (atributos) denominado x, con una variable (atributo) adicional que es la clase denominada y.

- El objetivo de la ***clasificación*** es encontrar un modelo (una función) para predecir la clase a la que pertenecería cada registro, esta asignación una clase se debe hacer con la mayor precisión posible.

- Un conjunto de prueba (tabla de testing) se utiliza para determinar la precisión del modelo. Por lo general, el conjunto de datos dado se divide en dos conjuntos al azar de el de entrenamiento y el de prueba.

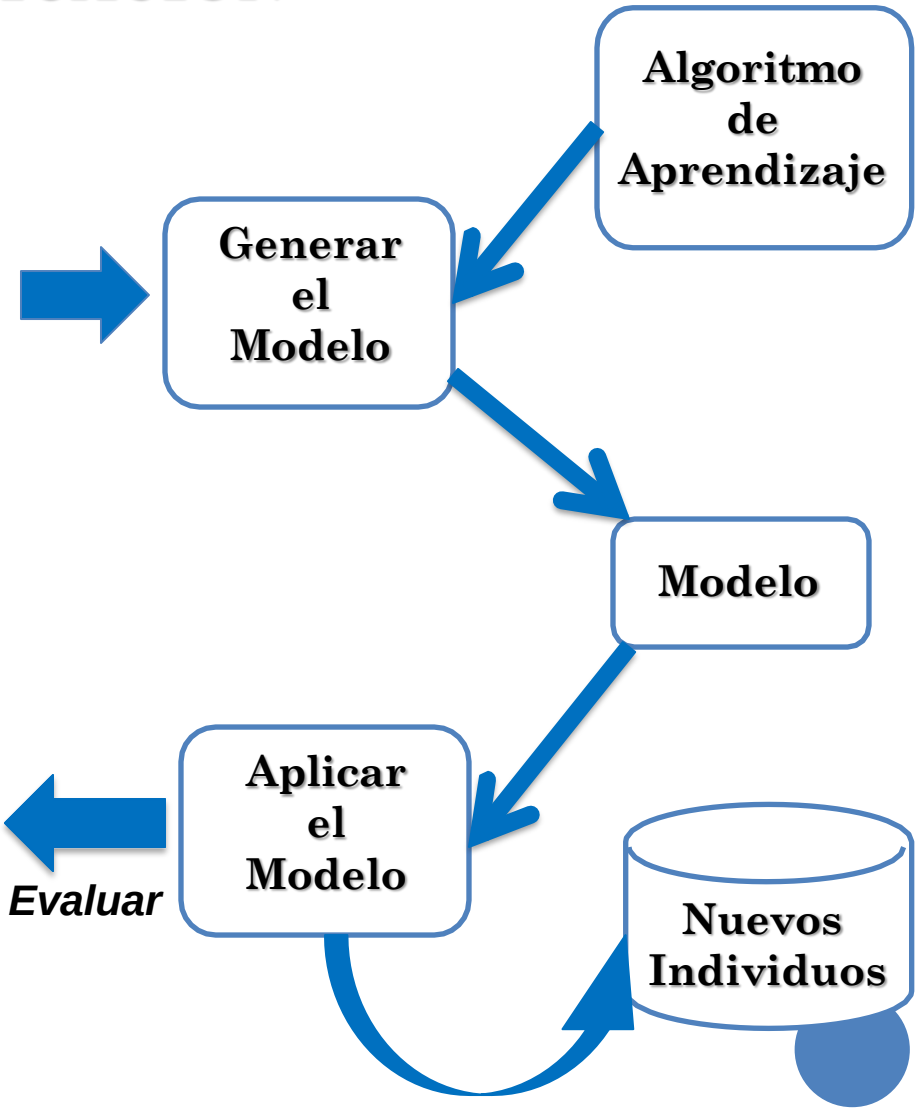
MODELO GENERAL DE LOS MÉTODOS DE CLASIFICACIÓN

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES	FRAUDE
1	SI	SOLTERO	S/ 1,000	NO
2	SI	CASADO	S/ 5,000	NO
3	NO	CASADO	S/ 3,500	SI
4	SI	VIUDO	S/ 4,500	NO
5	NO	SOLTERO	S/ 2,000	NO
6	NO	SOLTERO	S/ 1,500	SI

Tabla de Aprendizaje

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES	FRAUDE
7	SI	SOLTERO	S/ 4,000	NO
8	SI	CASADO	S/ 5,500	NO
9	NO	CASADO	S/ 6,500	SI

Tabla de Testing



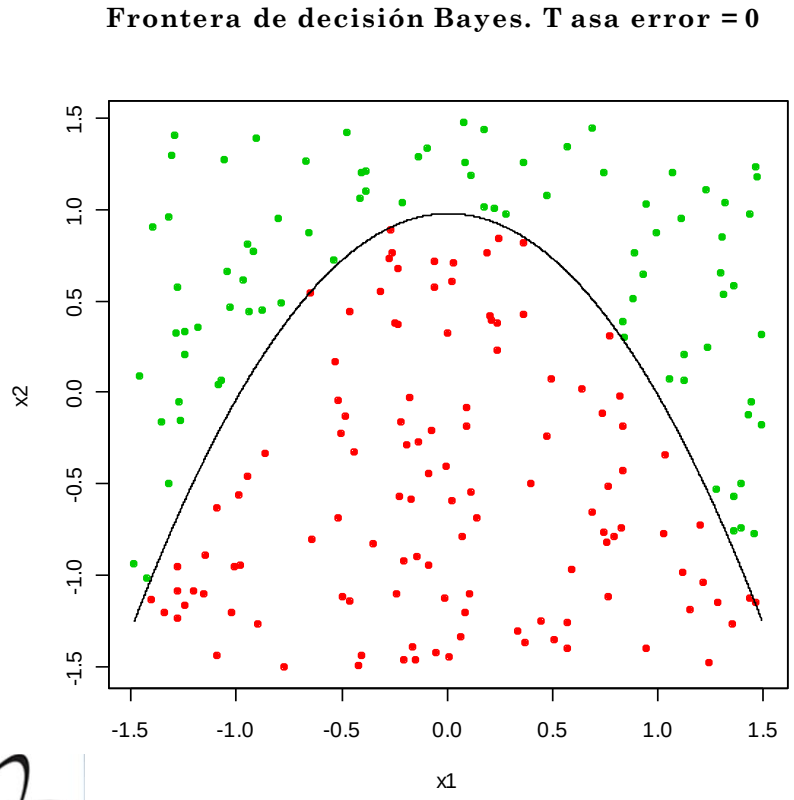
DEFINICIÓN DE CLASIFICACIÓN

- Dada una base de datos $D = \{t_1, t_2, \dots, t_n\}$ de tuplas o registros (individuos) y un conjunto de clases $C = \{C_1, C_2, \dots, C_m\}$, el **problema de la clasificación** es encontrar una función $f: D \rightarrow C$ tal que cada t_i es asignada una clase C_j.

- $f: D \rightarrow C$ podría ser una Red Neuronal, un Árbol de Decisión, un modelo basado en Análisis Discriminante, o una Red Beyesiana.

Ejemplo 1: Algoritmo Naive Bayes

- Problema de clasificación binaria. Frontera de decisión cuadrática

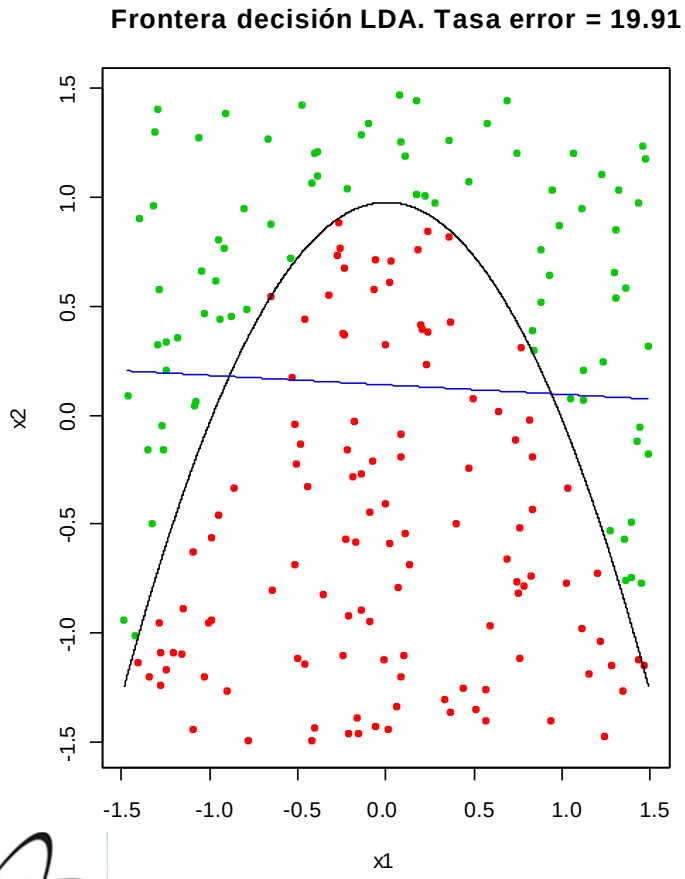


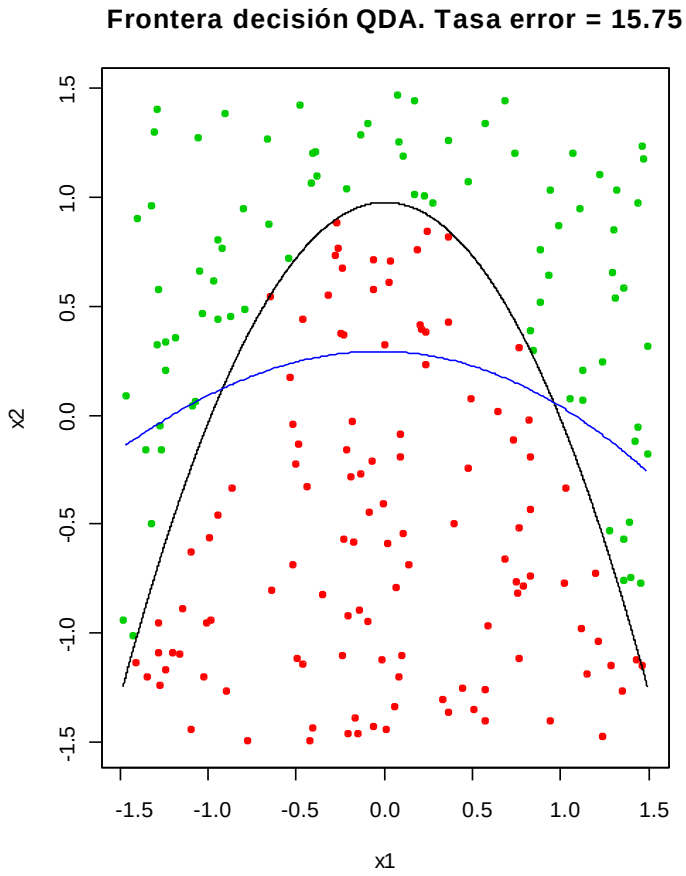
Se simularon 200 observaciones del vector bidimensional (x1,x2); ambas componentes del vector son variables independientes con una distribución uniforme en (-1.5,1.5)

La clasificación viene dada por:

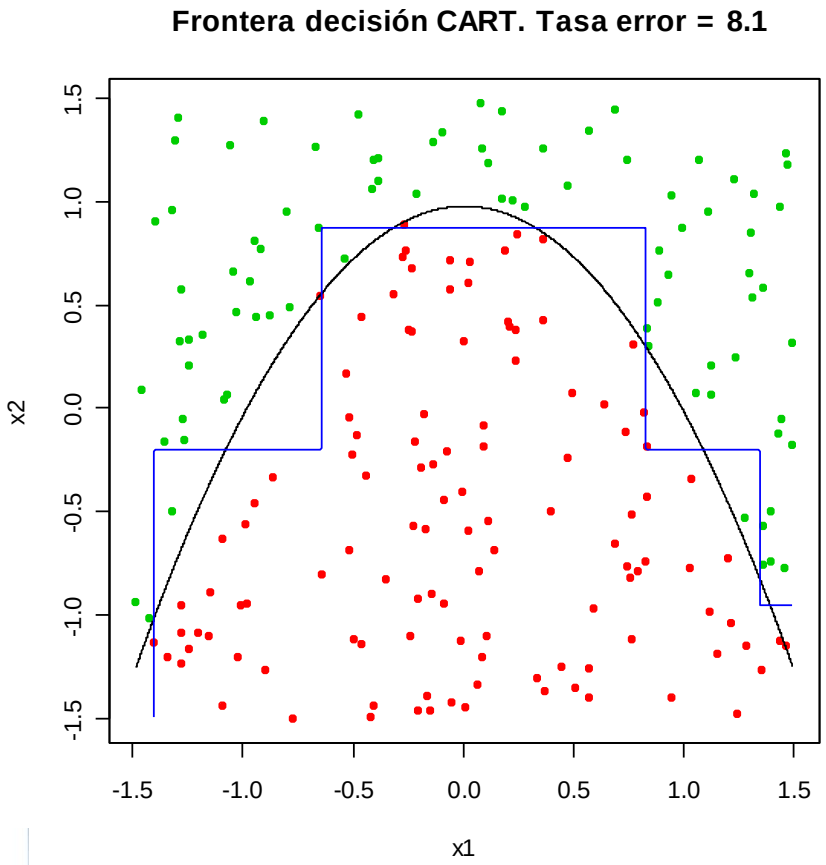
Clase 2: $x2 > 1-x1^2$
Clase 1: $x2 < 1-x1^2$

Ejemplo 2: Soluciones LDA y QDA

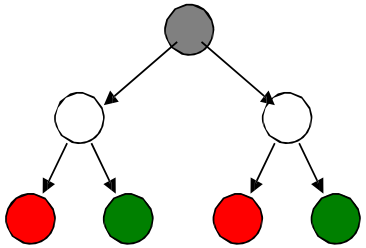




EJEMPLO 3: SOLUCIÓN ÁRBOL DE DECISIÓN

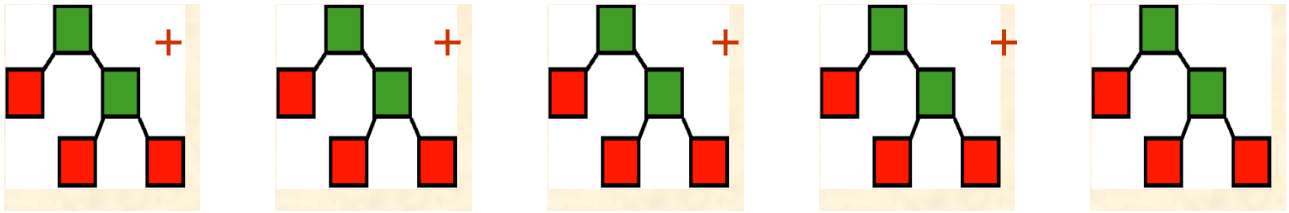


Solución obtenida mediante **CART**. Árbol sin podar



¿Qué es bagging?

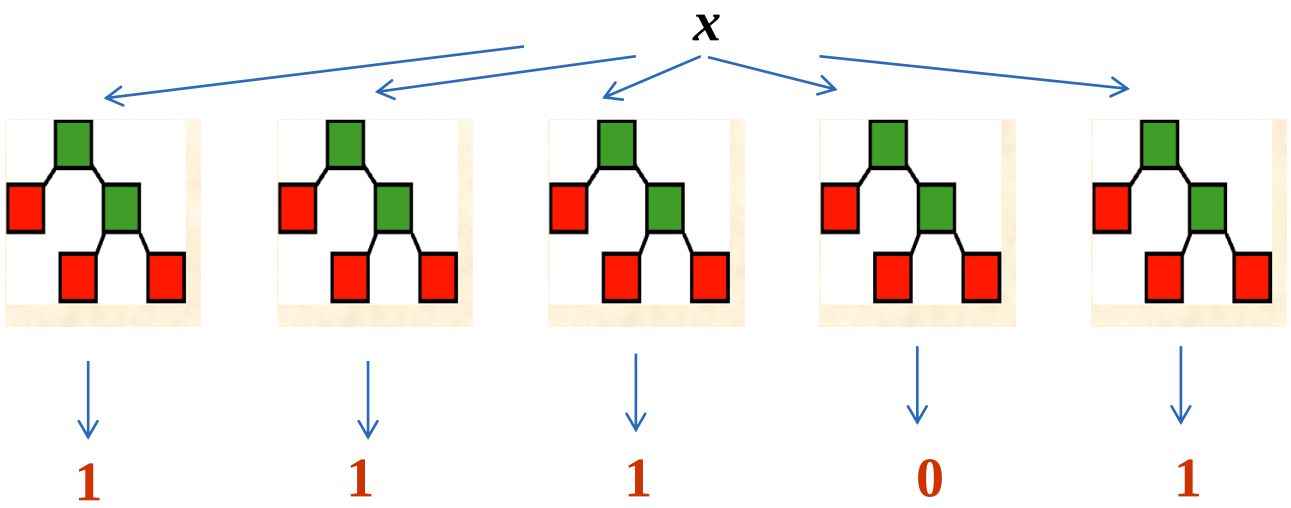
- Bagging quiere decir bootstrap aggregation. Introducido por Leo Breiman (Berkeley) en 1996
- La idea es simple. Si tienes las opiniones de un comité de expertos, considéralas todas para tomar una decisión
- Se extraen muestras bootstrap del conjunto de datos. Para cada muestra, se obtiene un modelo de predicción. El nuevo predictor "bagging" se construye mediante agregación



- El objetivo es reducir la inestabilidad

Bagging Para Clasificación

- Si el problema es de **clasificación** bagging clasificará cada nueva observación por mayoría.
  Por ejemplo:

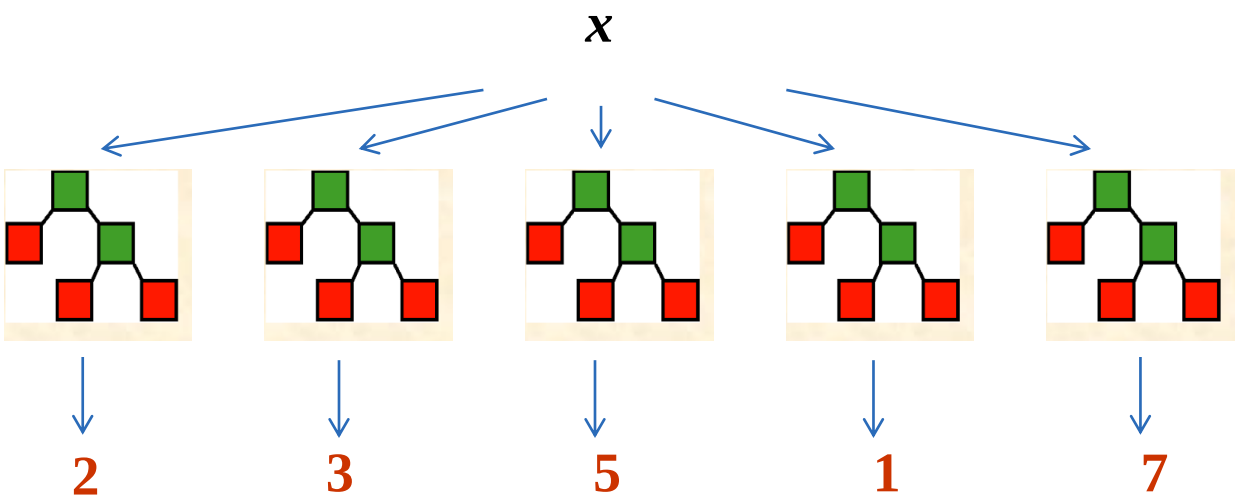


- La clase 1 recibió cuatro votos. La clase 0 un voto.
- El predictor bagging clasificará ***x*** en la clase 1.

SMDSC

BAGGING PARA REGRESIÓN

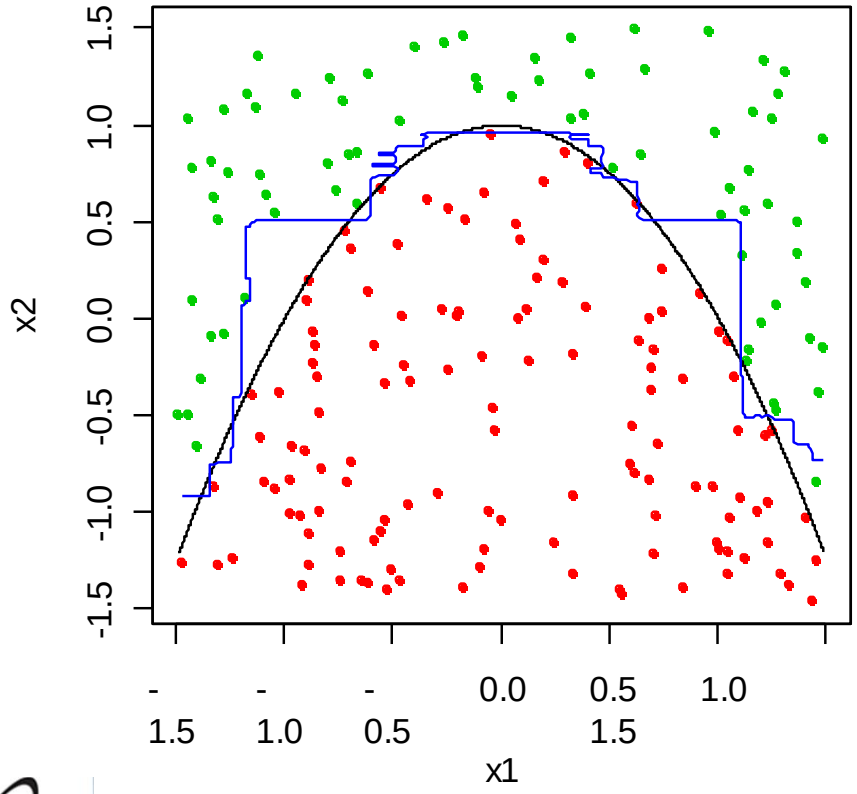
- Si el problema es de **regresión** la predicción bagging se obtiene promediando las predicciones de todos los modelos. Por ejemplo:



- La predicción bagging será: *(2+3+5+1+7)/5 = 3.6*
- Cuando la variable respuesta es binaria 0/1, el bagging para regresión se reduce al criterio de clasificar por mayoría

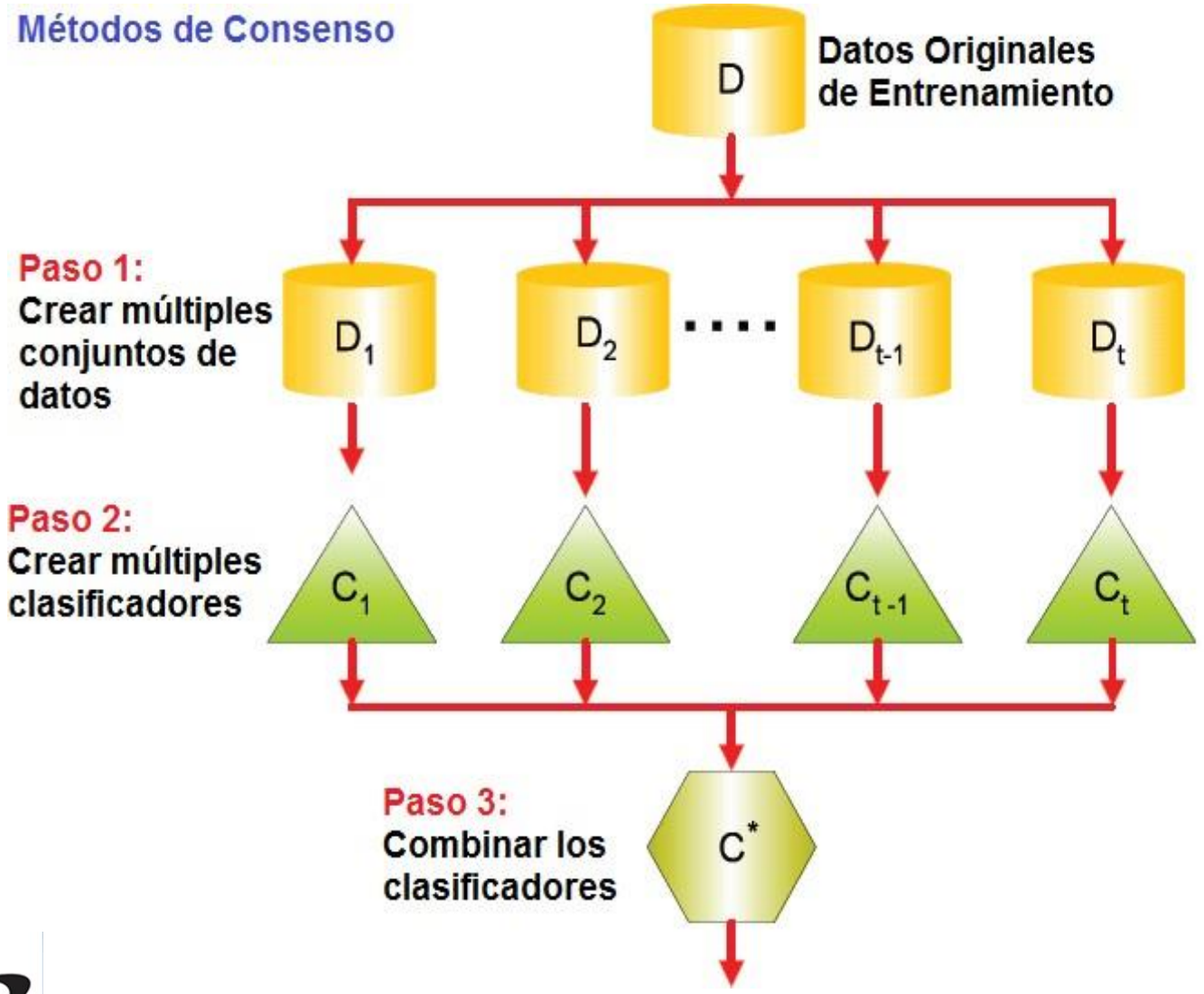
EJEMPLO 4: EL EFECTO BAGGING

Frontera decisión BAGGING. Tasa error = 5.8



- Se emplearon **50** muestras bootstrap.
- Se ha reducido la inestabilidad de CART.
- En este caso también se ha mejorado en capacidad predictiva.

Métodos de Consenso



Bosques Aleatorios (Random Forest)

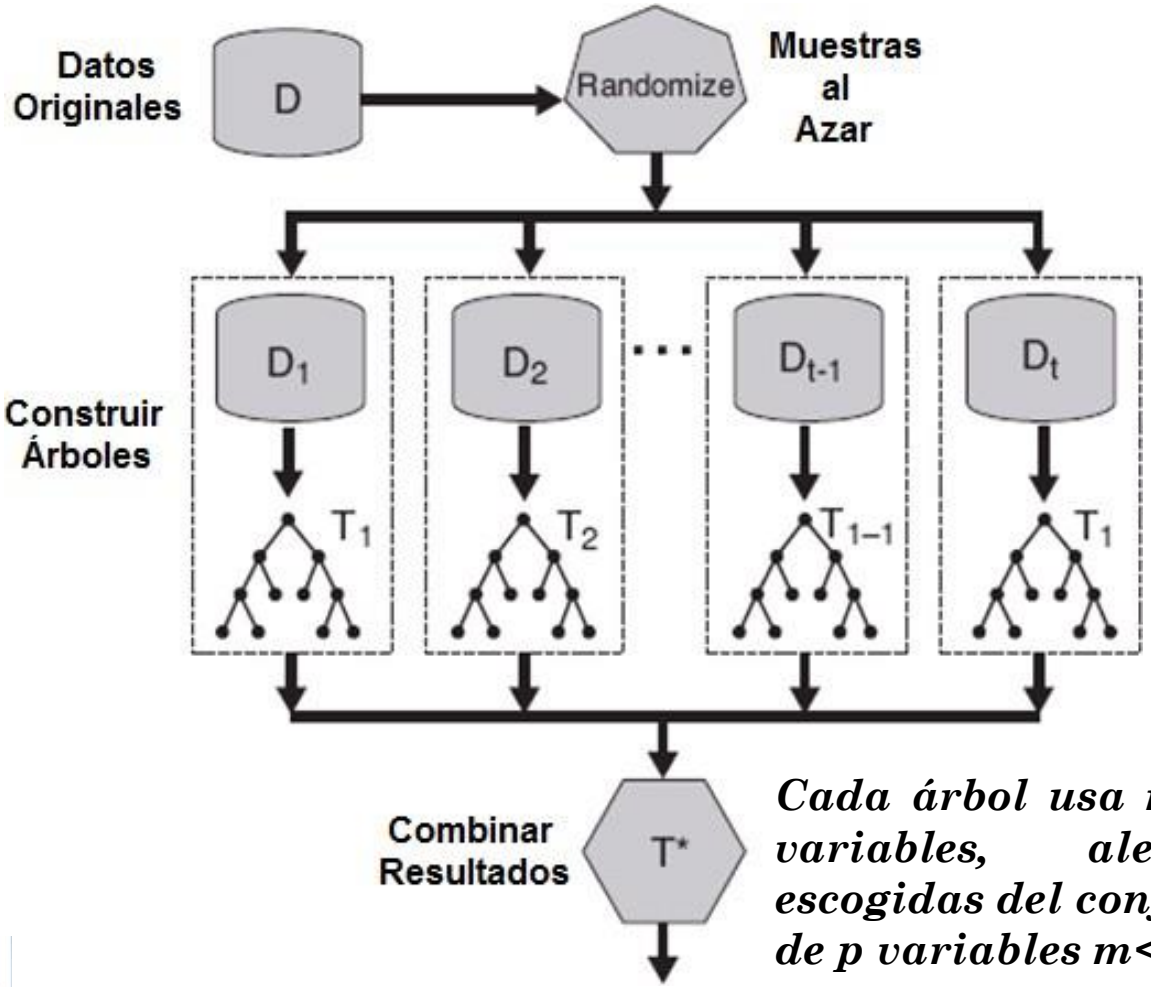
- El caso en el que todos los clasificadores del Método de Consenso son Árboles dicho método se denomina Bosques Aleatorios (Random Forest).

Random Forests o Bosques Aleatorios (RF)

- Desarrollado por Leo Breiman (Berkeley) en 2001
- Tiene su base en
  - ✓ La predicción con CART
  - ✓ La agregación de modelos de árbol
  - ✓ Bootstrap Aggregation (Bagging)

- Comercializado por Salford Systems en la herramienta **RandomForests™.**
- Implementado por Andy Liaw y Matthew Wiener en la librería **randomForest** del entorno R de programación.

BOSQUES ALEATORIOS (RANDOM FOREST)



Cada árbol usa m diferentes variables, aleatoriamente escogidas del conjunto de p variables m<p (m=mtry)

El mecanismo de Random Forest

- La aleatorización se introduce en el mecanismo de aprendizaje a través de dos vías: el **remuestreo** y la **aleatorización en la selección del corte** en cada nodo

- Se toman *B* muestras bootstrap del conjunto de datos para construir *B* árboles sin podar. Esto corresponde a la fase bagging y proporciona el bosque de árboles
- Para construir cada árbol del bosque, RF busca el corte en cada nodo entre un conjunto de *R* variables predictoras que han sido seleccionadas al azar
- Por defecto B = 500 y R = sqrt(nº predictores).

SMDSC

La estimación del error con RF

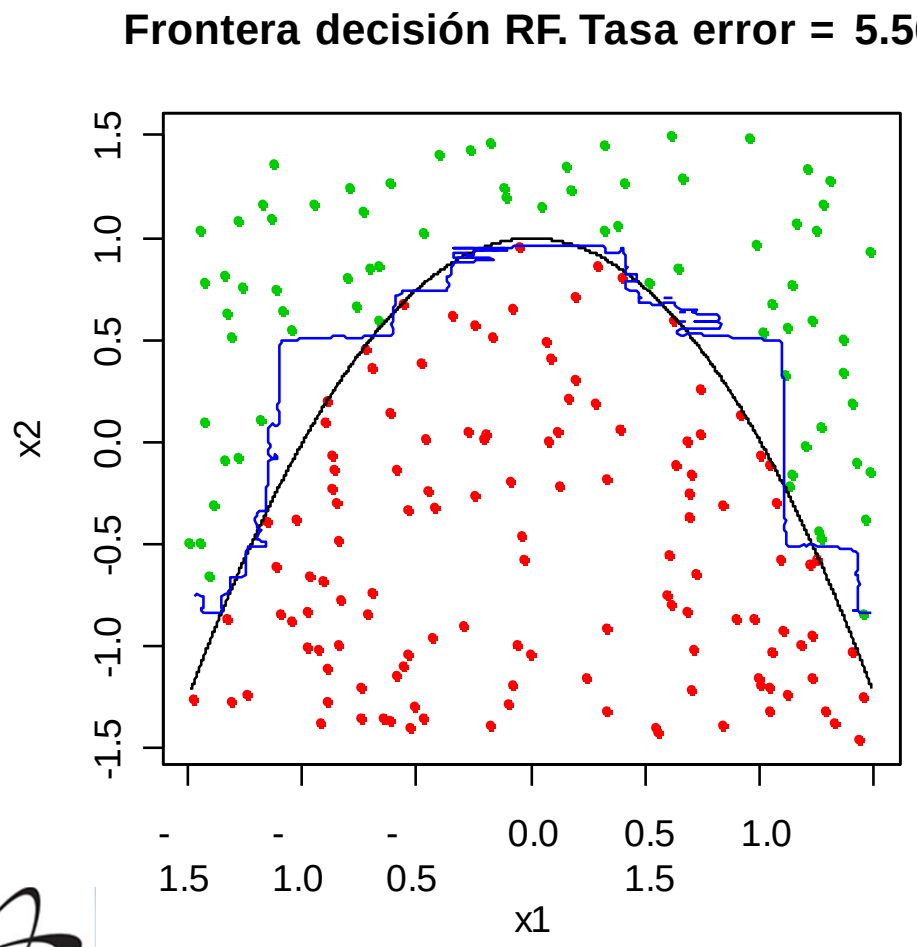
- Se define la tasa de error **out of bag** (OOB_i) de una observación x_i como el error obtenido al ser clasificada por los árboles del bosque construidos sin su intervención.

- La estimación **OOB** del error es el promedio de todos los OOB_i para todas las observaciones del conjunto de datos.

- Es mejor estimador que el error aparente. Parecida a la estimación por validación cruzada.

- La medida se puede extrapolar al problema de regresión describiéndola en términos del ECM.

Ejemplo 5: Solución con Random Forest



- Se empleó un bosque con **5000** árboles
- **¿Qué ha ocurrido con la tasa de error?** Comparar con lda, qda, CART y bagging

SMDSC

Documentación sobre Random Forest

- Página Web de **Leo Breiman:** http://www.stat.berkeley.edu/users/breiman/ Fallecido en julio de 2005
- Página Web de **Adele Cutler:** http://www.math.usu.edu/~adele/
- Página Web de **Salford Systems:** http://salford-systems.com/ Versión comercial. White papers y muchas aplicaciones de RF en consultoría

¡Gracias!

San Marcos Data Science Community

Auspicio : Escuela Académica Profesional de Estadística
San Marcos Data Science Community.

¿Preguntas?
Realicemos el taller