Módulo 3: Aprendizaje Supervisado

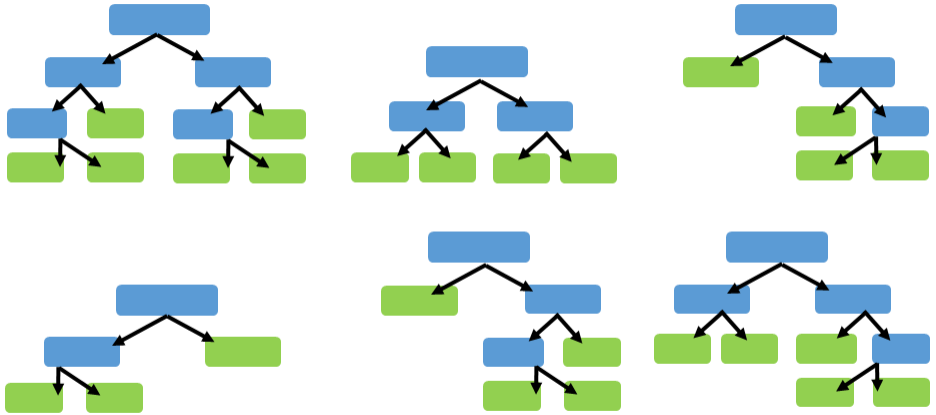
3.5. Random Forest y Gradient Boosting

Rafael Zambrano
rafazamb@gmail.com

Random Forest

- Los árboles de decisión son fáciles de construir, usar e interpretar, pero son imprecisos y tienden a causar overfitting

- Random Forest combina la simplicidad de los árboles de decisión y mejora la precisión



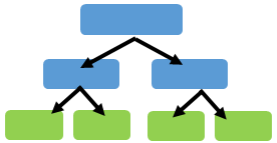
Random Forest

Funcionamiento

1) A partir de los datos originales, crea nuevos datos, escogiendo filas aleatorias con repetición
2) Crea un árbol de decisión para cada nuevo conjunto de datos, escogiendo columnas aleatorias para cada nodo del árbol

Cliente	Edad	Trabaja	Hipoteca	Ingresos
A	32	SÍ	SÍ	Altos
B	25	SÍ	SÍ	Altos
C	48	NO	NO	Bajos
D	67	NO	SÍ	Bajos
E	18	SÍ	NO	Bajos

Cliente	Edad	Trabaja	Hipoteca	Ingresos
A	32	SÍ	SÍ	Altos
B	25	SÍ	SÍ	Altos
A	32	SÍ	SÍ	Altos
B	25	SÍ	SÍ	Altos
D	67	NO	SÍ	Bajos



Random Forest

Uso

- Para predecir con nuevos datos, cada árbol va a dar un resultado diferente
- Se escoge la opción más votada

Cliente	Edad	Trabaja	Hipoteca	Ingresos
Z	46	SÍ	NO	

Árbol #	Predicción
1	Altos
2	Altos
3	Bajos
…	…

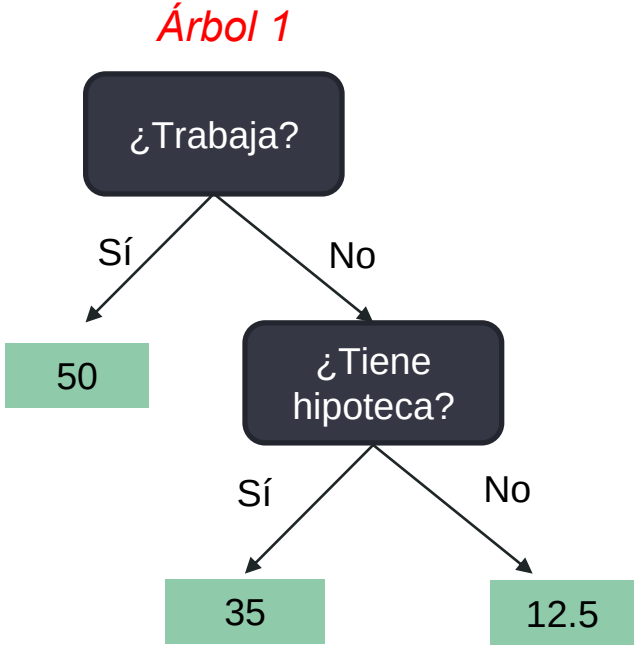
Ingresos Altos	Ingresos Bajos
90 árboles	10 árboles

El proceso de crear nuevos datos (Bootsrapping) y agregarlos para tomar una decisión se conoce como **Bagging**

Gradient Boosting

- Estos algoritmos son similares a Random Forest, pero en lugar de crear árboles aleatorios, cada nuevo árbol mejora al anterior

Cliente	Edad	Trabaja	Hipoteca	Ingresos	Error
A	32	SÍ	SÍ	90	40
B	25	SÍ	SÍ	50	0
C	48	NO	NO	25	12.5
D	67	NO	SÍ	35	0
E	18	SÍ	NO	10	-40

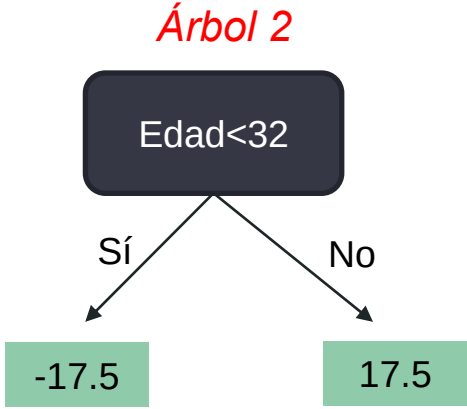


Gradient Boosting

- Estos algoritmos son similares a Random Forest, pero en lugar de crear árboles aleatorios, cada nuevo árbol mejora al anterior

- Por ejemplo, si cada nuevo árbol intenta predecir el error cometido por el árbol anterior

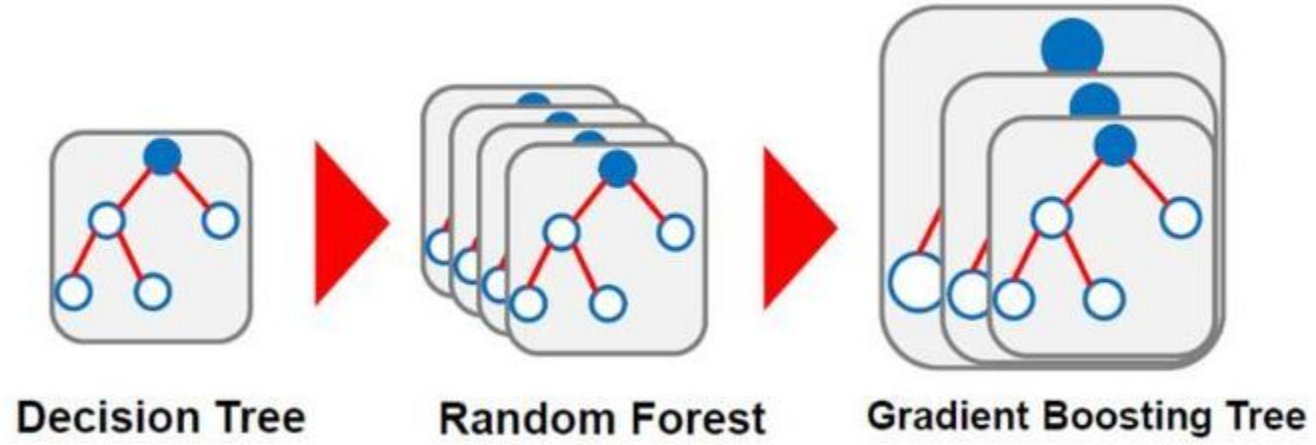
Cliente	Edad	Trabaja	Hipoteca	Ingresos	Error 1	Error 2
A	32	SÍ	SÍ	90	40	22.5
B	25	SÍ	SÍ	50	0	17.5
C	48	NO	NO	25	12.5	-5
D	67	NO	SÍ	35	0	-17.5
E	18	SÍ	NO	10	-40	-22.5
				Error medio	18,5	17



Gradient Boosting

- El algoritmo más conocido de Gradient Boosting es el **XGBoost** (eXtreme Gradient Boosting)

- Hoy en día, son las mejores técnicas de Machine Learning en modelos predictivos (sin considerar Deep Learning)



Random Forest y XGBoost en R

- En R, podemos crear modelos de random forest con la librería `randomForest(formula, data, ntree)`

- Para utilizar gradient boosting, podemos emplear la librería `caret`

```
train(formula, data, method = "gbm")
train(formula, data, method = "xgbTree")
```

¡Gracias!

Contacto: Rafael Zambrano

rafazamb@gmail.com