Clase 23 Técnicas avanzadas de modelamiento predictivo: Árboles de decisión y Random Forest

Diplomado en Análisis de Datos y Modelamiento Predictivo con Aprendizaje Automático para la Acuicultura

Dra. María Angélica Rueda Calderón

Pontificia Universidad Católica de Valparaíso

12 July 2023

PLAN DE LA CLASE

1.- Introducción

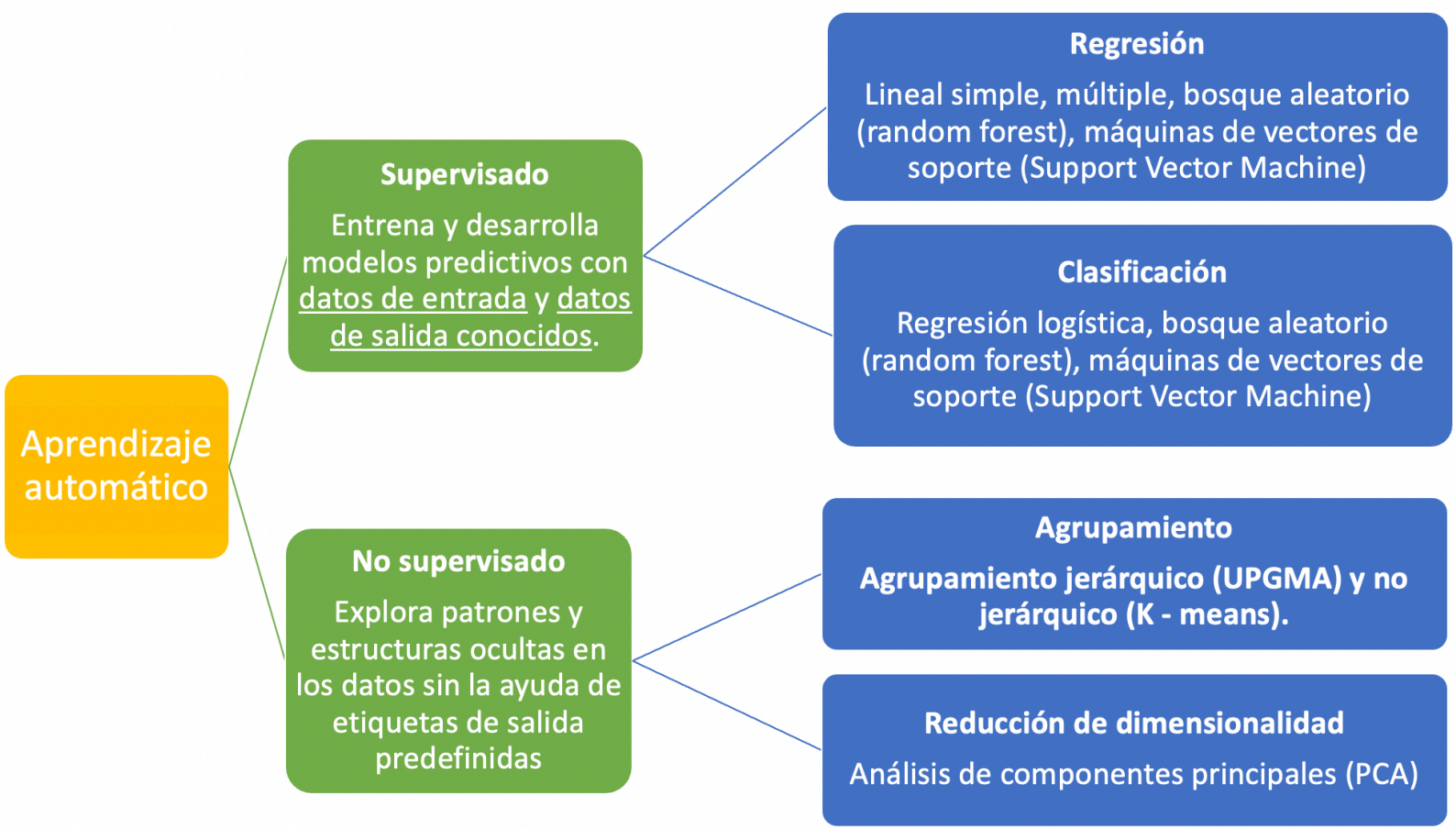
- Tipos de modelos predictivos.
- ¿Qué son los arboles de decisión?
- ¿Qué es el algoritmo de Random Forest?
- Métricas para medir el desempeño de un modelo de clasificación.
- Práctica con ejemplo de maduración temprana.

2.- Práctica con R y Rstudio cloud.

- Realizar modelamiento predictivo con Árboles de Decisión y Random Forest.
- Realizar gráficas avanzadas con ggplot.

MACHINE LEARNING Y MODELOS PREDICTIVOS

Existen diferentes modelos predictivos, los cuales se clasifican bajo el marco del aprendizaje automático, machine learning en inglés, como algoritmos o metodos supervisados y no supervisados.



ESTUDIO DE CASO: PREDICCIÓN DE HIPOXIA EN LAGUNA

Table 1
List of physicochemical and meteorological variables used as potential predictors of hypoxia.

Variables	Range	Mean	Skewness
Water Temperature (Tw)	7.9–32.0 °C	21.4	−0.17
pH	7.4–9.0 dimensionless	8.1	0.19
Salinity (Sal)	19.8–51.7 psu	37.2	−0.37
Chlorophyll (Chl-a)	0.43–5.5 µg/L	3.3	0.35
Relative Humidity (RH)	36.7–91.3%	71.8	−0.78
Atmospheric Pressure (ATM)	996.0–1033.4 hPa	1014.4	0.5
Solar Radiation (SolRad)	30.7–604.8 W/m^3	393.7	−0.77
East-west wind (EW-wind)[a]	−5.5–11.7 m/s	0.48	1.10
North-south wind (NS-wind)[a]	−8.16–4.6 m/s	−0.06	−1.33

[a] The positive values of variables EW-wind and NS-wind indicate wind velocities from east to west and from north to south, respectively, whereas negative values indicate wind velocities in the opposite directions.

Fuente: Politikos et al. 2021

MÉTODOS DE PREDICCIÓN

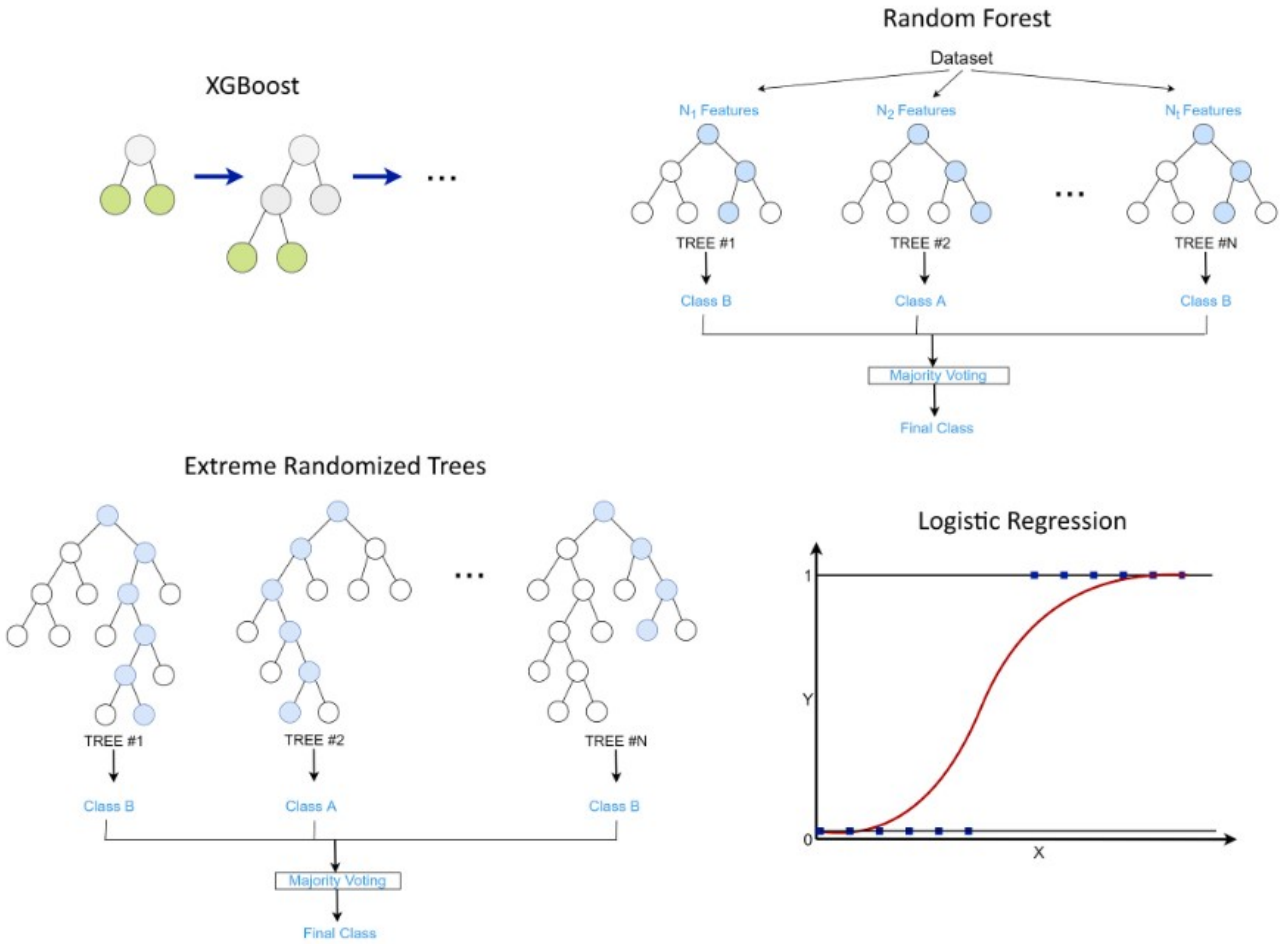


Fig. 3. Schematic view of XGBoost, Random Forest, Extreme Randomized Trees and Logistic Regression.

PERFORMANCE DE LOS MODELOS

F1-Score : Es una métrica que sirve para medir el desempeño del modelo de clasificación. Un valor alto indica (Buen desempeño del modelo predictivo).

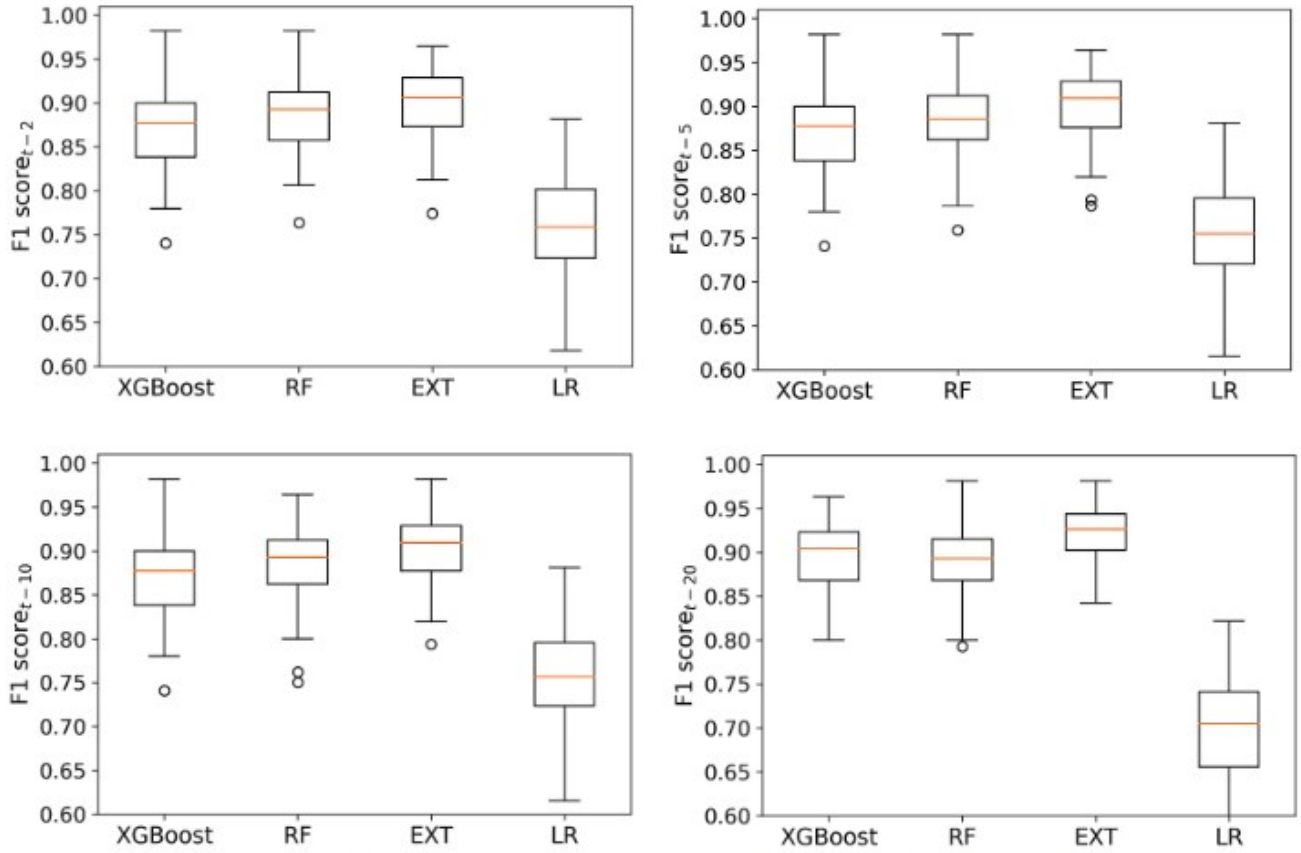


Fig. 4. Boxplots of F1 score metrics to statistically compare the performance of XGBoost, Random Forest (RF), Extremely Randomized Trees (EXT) and Logistic Regression (LR) for time lags of 2,5,10 and 20 days.

MATRIZ DE CONFUSIÓN

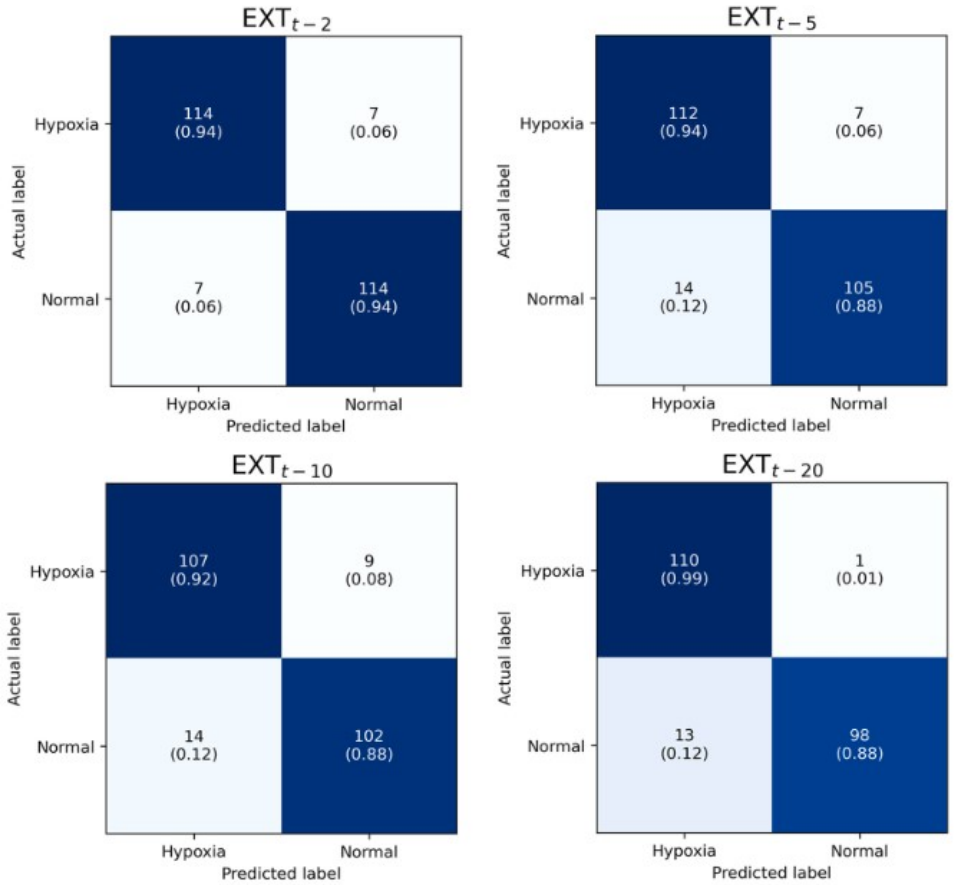


Fig. A1. Confusion matrix of "Normal" and "Hypoxia" classes on the testing set with EXT_{t-2}, EXT_{t-5}, EXT_{t-10}, and EXT_{t-20} algorithms. Each row and column represent the two classes. In the main diagonal boxes, we observe the number of instances that were correctly labeled for each class, whereas and in non-diagonal boxes, we observe the number of instances that were misclassified. The number in parenthesis show the proportion of correct and incorrect classifications per row, which is the recall value.

Fuente: Politikos et al. 2021

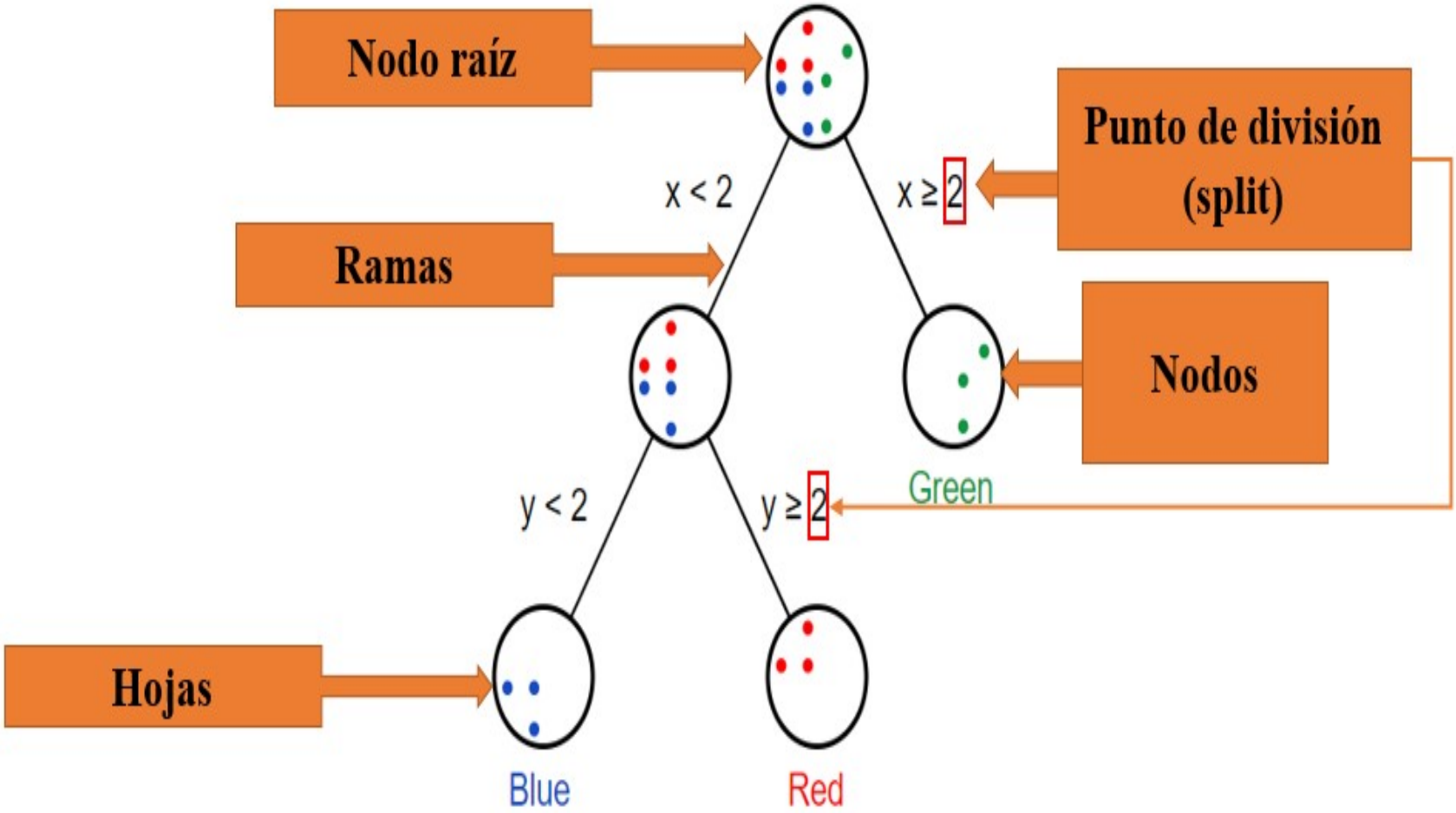
¿QUÉ SON LOS ÁRBOLES DE DECISIÓN?

- Los árboles de decisión son un tipo de modelo de aprendizaje automático supervisado.
- Son estructuras de tipo árbol que representan un conjunto de reglas de decisión y predicción.
- Pueden manejar problemas de clasificación y regresión.
- Los arboles de decisión son la base de muchos modelos predictivos incluyendo Random forest.

PARTES DEL ÁRBOL DE DECISIÓN

- **Nodo raíz:** Es el punto de partida del árbol y no tiene padres. Representa la característica o variable predictora que mejor separa los datos en función de la variable objetivo (**Var. Respuesta**).
- **Ramas:** Son las conexiones que salen del nodo raíz o de otros nodos y representan las diferentes opciones o caminos que se pueden tomar en función de los valores de una variable predictora.
- **Hojas:** Son los ***nodos terminales*** del árbol y no tienen hijos. Representan las decisiones finales o las clasificaciones de las instancias de datos.
- **Punto de división:** Es el criterio utilizado para dividir los datos en dos ramas durante la construcción del árbol.

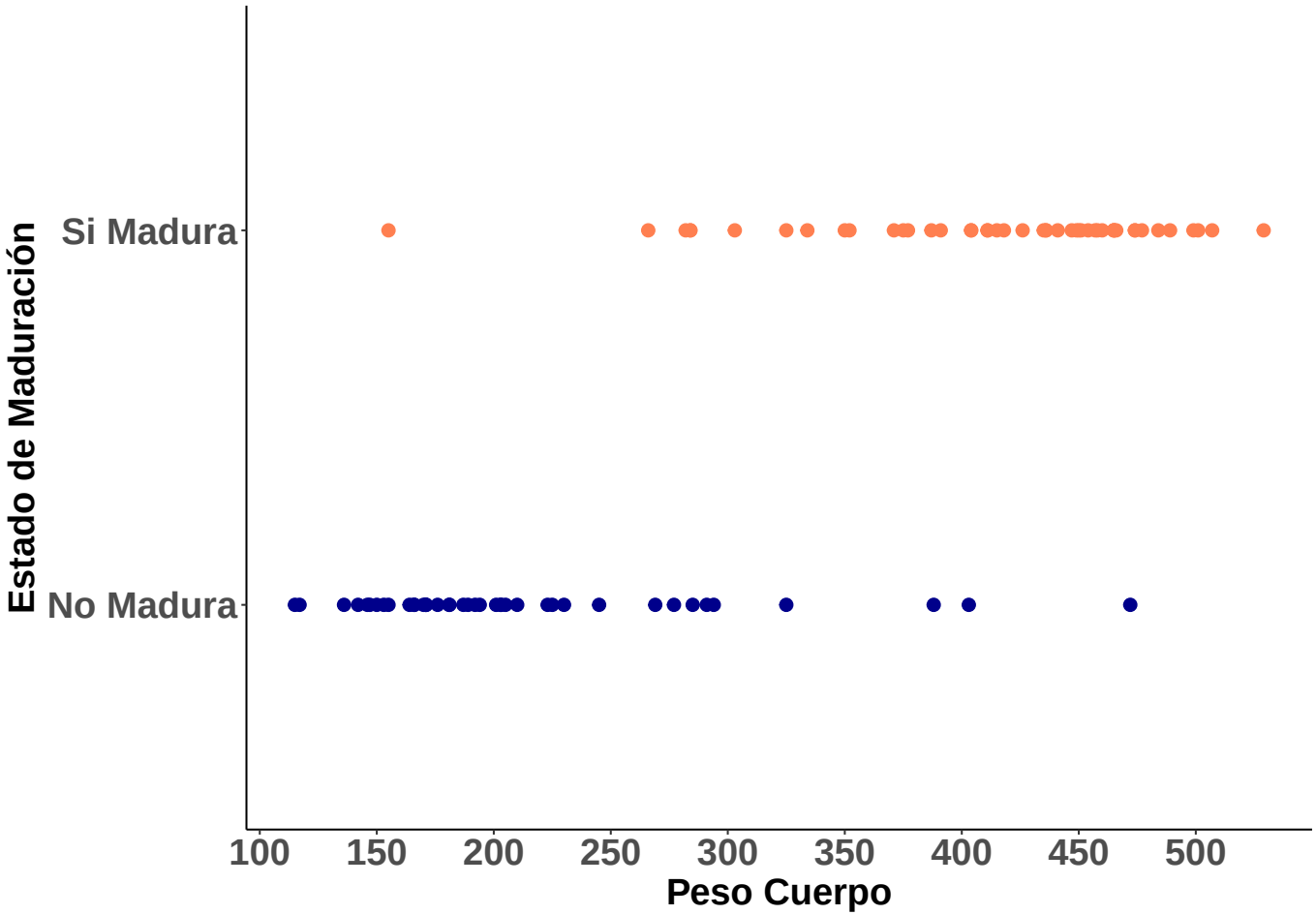
PARTES DEL ÁRBOL DE DECISIÓN



ÍNDICE DE GINI

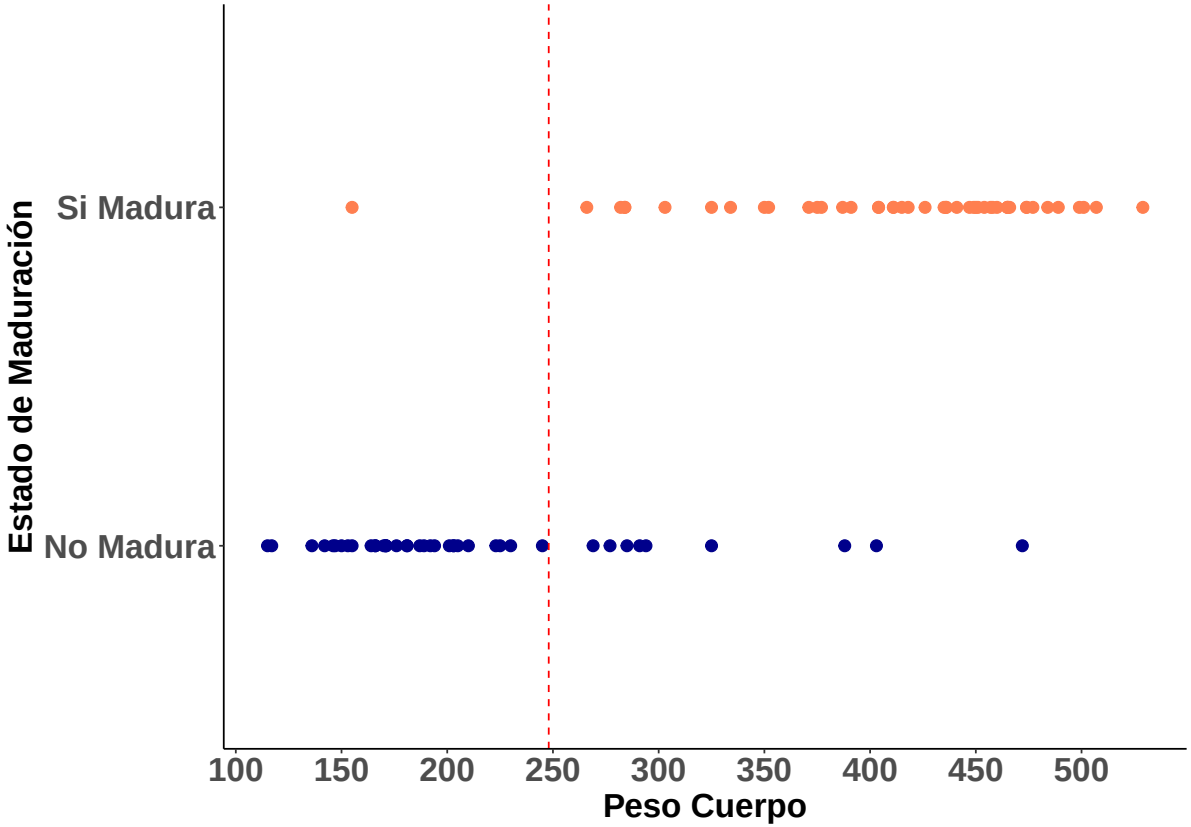
- Cuando se construye un árbol de decisión, el algoritmo de Gini busca dividir los datos en cada nodo de la manera que minimice la impureza total de las ramas resultantes.
- El índice de Gini se utiliza para medir qué tan mezcladas están las clases en un nodo específico y se calcula considerando la probabilidad de selección aleatoria de una muestra y clasificarla incorrectamente.
- Cuanto más bajo sea el valor del índice de Gini, más puro será el nodo y mejor será la separación de las clases en ese nodo.

EJERCICIO: ÁRBOL DE DECISIÓN MADURACIÓN TEMPRANA



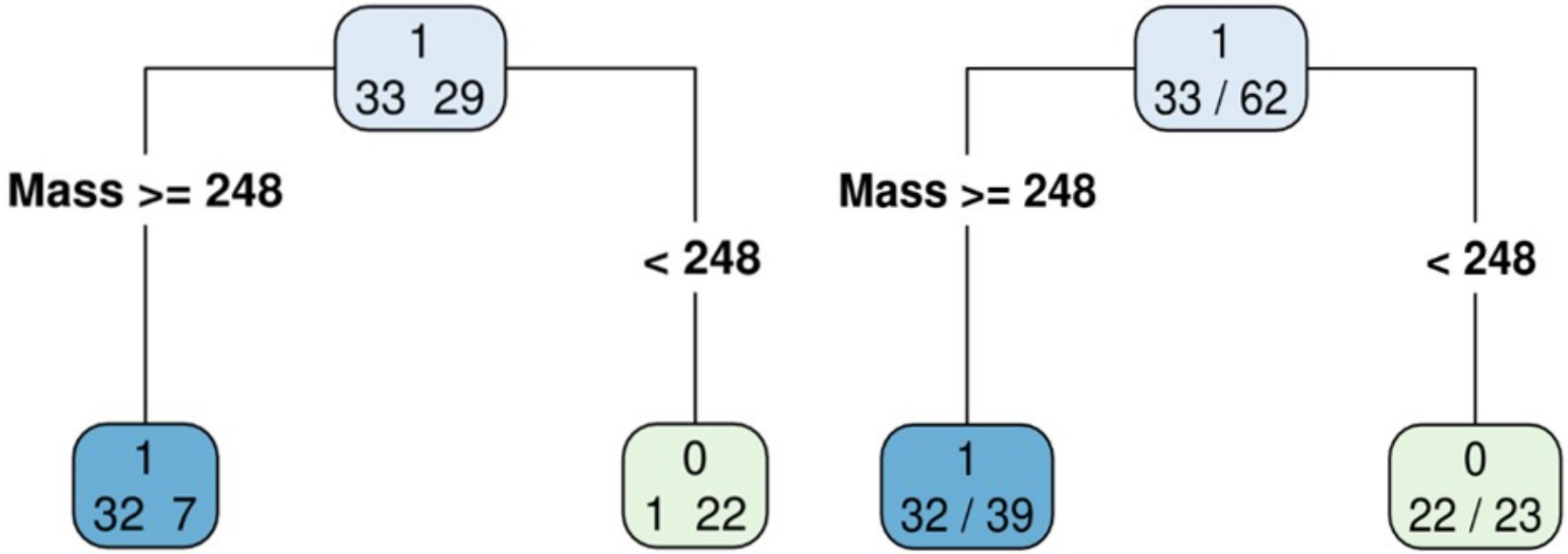
IDENTIFICAR PARTICIÓN EN ÁRBOL DE DECISIÓN

	count	ncat	improve	index	adj
Mass	62	1	17.47	248	0



ÁRBOL DE DECISIÓN (CLASIFICACIÓN)

train_data: 62 (**70%**) Maduros:33 Inmaduros:29



MATRIZ DE CONFUSIÓN

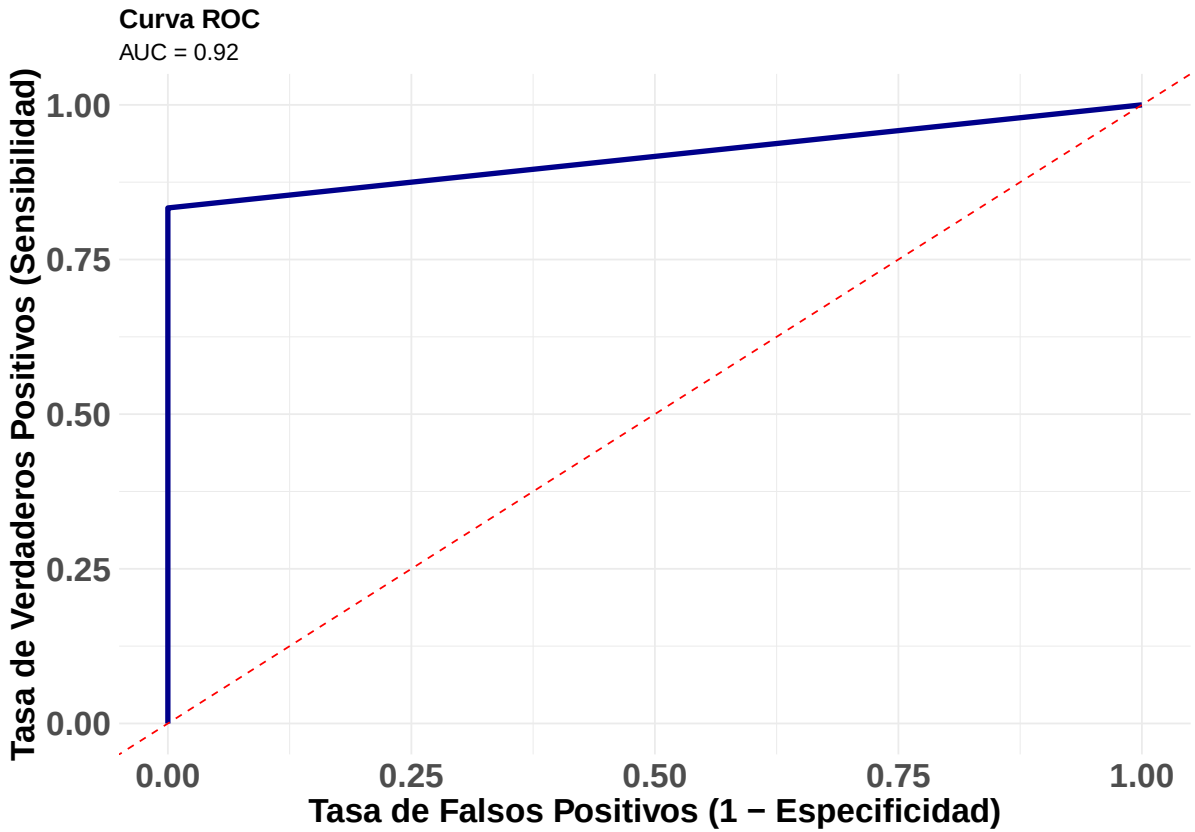
Función para hacer matriz de confusión `confusionMatrix()`

Table 2: Matriz de Confusión

	Maduro	Inmaduro
Maduro	14	2
Inmaduro	0	10

$$ACC = \frac{(14+10)}{(14+0+2+10)} = 0.92$$
$$SEN = \frac{(14)}{(14+0)} = 1$$
$$SPE = \frac{(10)}{(10+2)} = 0.83$$

AREA BAJO LA CURVA ROC (AUC)



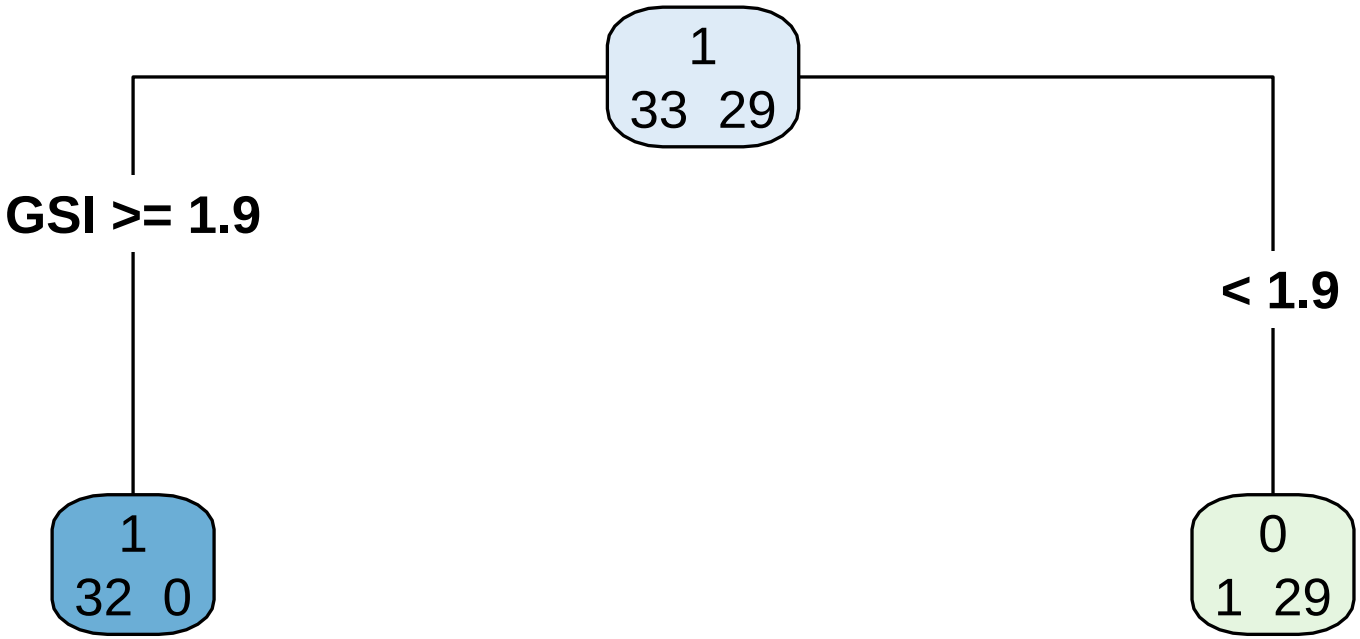
EJERCICIO ÁRBOL DE DECISIÓN CON MÚLTIPLES VARIABLES

n: 88, train_data: 62 (**70%**) y test_data: 26 (**30%**).

Table 3: Tabla de datos

Mass	SGR	Length	GSI	Maturation
391	1.64	30.8	2.271	1
203	1.23	26.3	0.049	0
474	1.77	31.8	2.266	1
436	1.72	31.8	2.255	1
282	1.93	27.8	2.241	1
277	1.08	29	0.065	0
171	1.11	25.2	0.041	0
171	1.06	25.5	0.041	0
284	1.76	28.2	1.57	1
466	1.72	32.6	2.693	1

ÁRBOL DE DECISIÓN



MATRIZ DE CONFUSIÓN CONSIDERANDO TODAS LA VARIABLES PREDICTORAS

Table 4: Matriz de Confusión AD

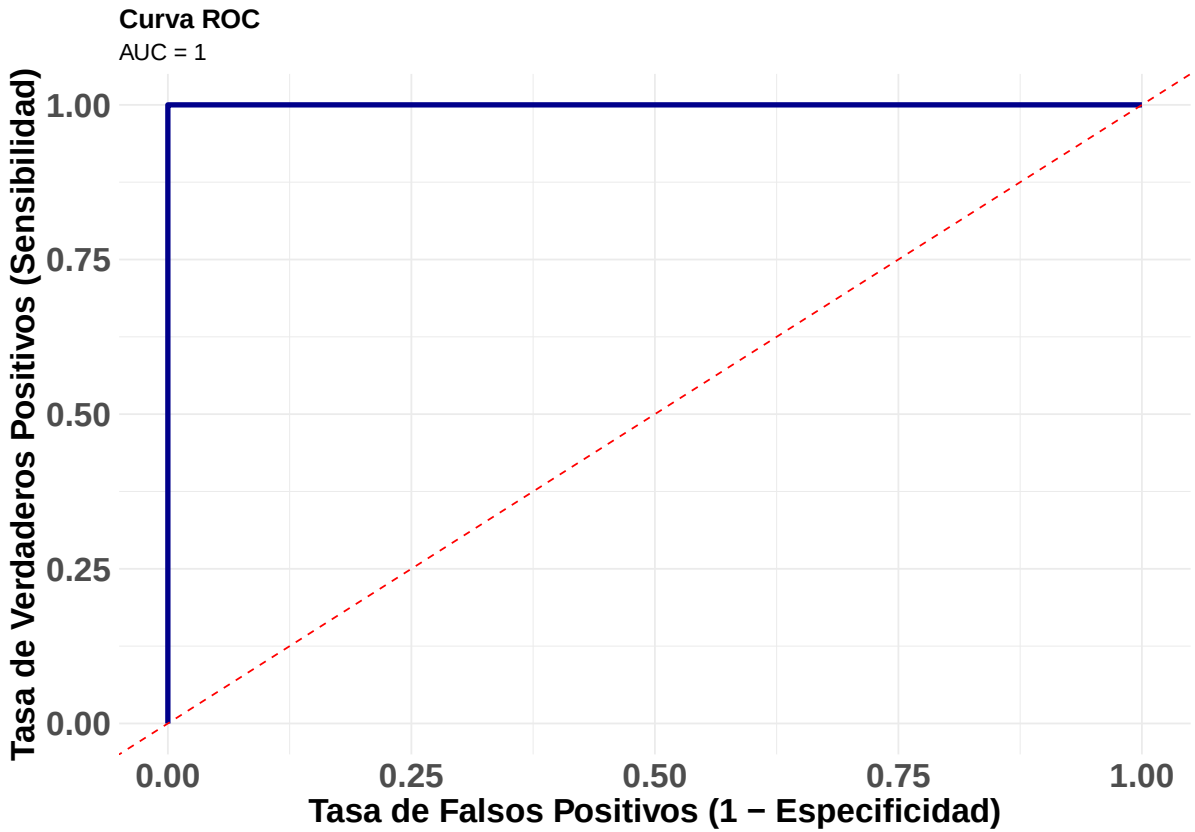
	Maduro	Inmaduro
Maduro	14	0
Inmaduro	0	12

$$ACC = \frac{(14+12)}{(14+0+0+12)} = 1$$

$$SEN = \frac{(14)}{(14+0)} = 1$$

$$SPE = \frac{(12)}{(12+0)} = 1$$

AREA BAJO LA CURVA ROC (AUC)



VENTAJAS Y DESVENTAJAS DE LOS ARBOLES DE DECISIÓN

Ventajas	Desventajas
- Fácil interpretación y visualización	- Sensibles al ruido y outliers
- Manejan datos numéricos y categóricos	- Tendencia al sobreajuste
- Identifican variables predictoras importantes	- Pueden ser inestables ante cambios
- Versátiles: clasificación y regresión	- Pueden generar árboles complejos
- No requieren suposiciones sobre los datos	- No capturan relaciones no lineales
	- Dificultad con problemas de alta dimensionalidad

RANDOM FOREST (RF)

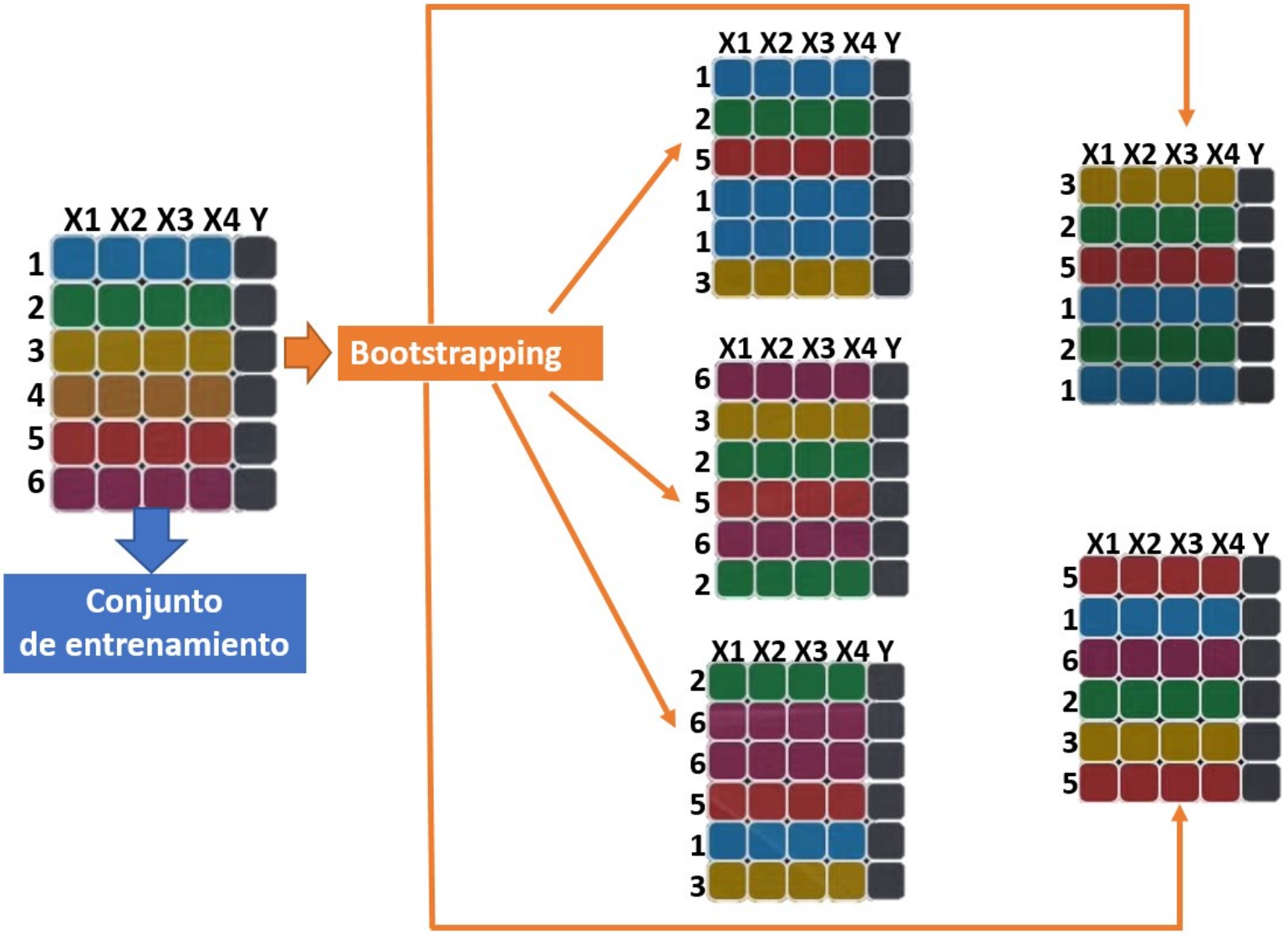
- El algoritmo Random Forest fue propuesto por Leo Breiman en 2001. Breiman fue un estadístico y científico de la computación reconocido, conocido por sus contribuciones significativas en el campo del aprendizaje automático y la estadística.
- RF es una técnica de aprendizaje supervisado que combina múltiples árboles de decisión para crear un modelo robusto y preciso y es usado tanto para problemas de clasificación como de regresión.

PASO 1: ¿QUÉ ES BOOSTRAPING?

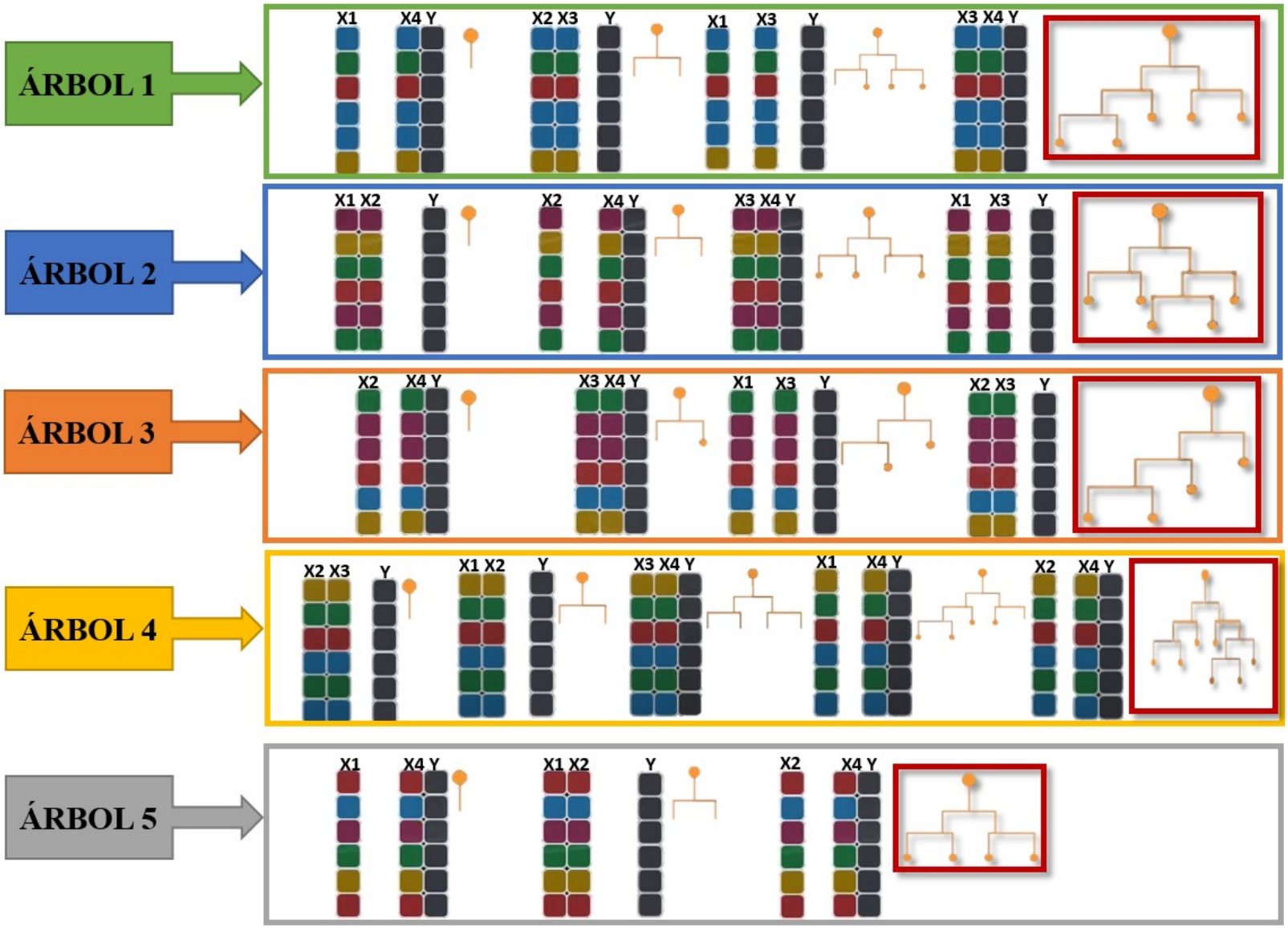
Bootstraping se refiere a una técnica en la que se generan muestras aleatorias con reemplazamiento a partir del conjunto de datos de entrenamiento.

- Algunas observaciones pueden aparecer múltiples veces en la muestra boostrap, mientras que otras pueden no ser seleccionadas.
- Esta técnica introduce variabilidad y diversidad en cada árbol del bosque.
- El bootstraping se realiza exclusivamente en los datos de entrenamiento y no involucra los datos de prueba o validación.
- El objetivo del bootstraping es generar diversidad y reducir la correlación entre los árboles del bosque.
- Al utilizar muestras bootstrap diferentes para cada árbol, se evita el sobreajuste en los datos de entrenamiento.

PASO 1: BOOSTRAPING



PASO 2: GENERACIÓN DEL ÁRBOL



PASO 3: AGREGAR LOS RESULTADOS DEL ÁRBOL DE RF

- El árbol final para cada muestra bootstrap en un bosque aleatorio es un árbol individual que representa una combinación de todas las divisiones realizadas durante su construcción.
- Cada árbol se construye de forma independiente utilizando una muestra bootstrap y un subconjunto de características, y luego se combinan las predicciones de todos los árboles para obtener la predicción final del bosque.
- La agregación en arboles de clasificación es el valor más frecuente (ej. si la hoja tiene 5 maduros y 4 no maduros la predicción será maduro), la agregación en arboles de regresión es el promedio de las observaciones (ej. 5 mg/l, 4 mg/l, 3 mg/l el promedio es 4 mg/l).

PARÁMETROS CLAVE EN RF

- **mtry:** Representa el número de variables predictoras (características) que se seleccionan aleatoriamente en cada árbol del bosque. Es decir, determina la cantidad de características que se consideran en cada división del nodo durante la construcción de cada árbol individual en el bosque.
- **importance:** Indica la importancia relativa de cada variable predictora en el modelo de Random Forest. Proporciona una medida de la influencia o contribución de cada característica (**Var. predictora**) en la predicción.
- **ntree:** Número de árboles que se construyen en el Random Forest. Cuanto mayor sea el número de árboles, más robusto y generalizado será el modelo.

EJERCICIO MADURACIÓN TEMPRANA CON RF

Table 6: Matriz de Confusión RF.

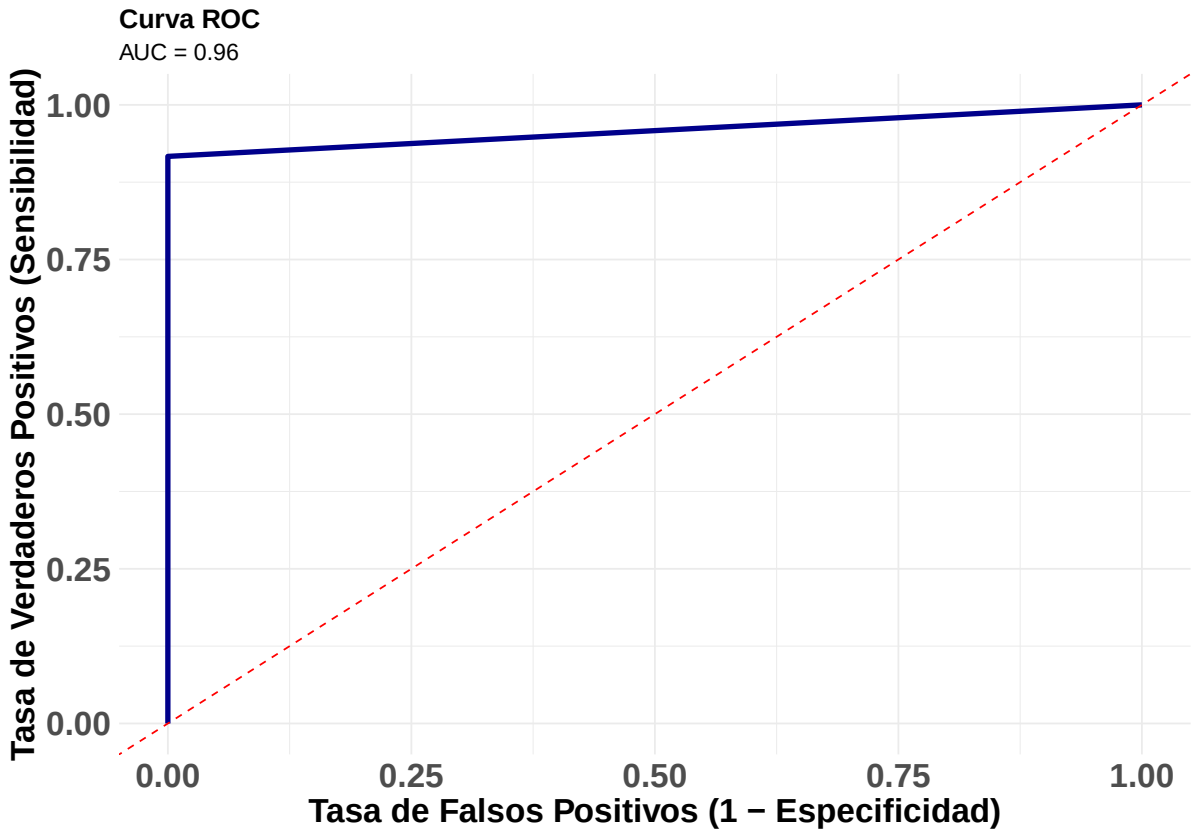
	Maduro	Inmaduro
Maduro	14	1
Inmaduro	0	11

$$ACC = \frac{(14+11)}{(14+0+1+11)} = 0.96$$

$$SEN = \frac{(14)}{(14+0)} = 1$$

$$SPE = \frac{(11)}{(11+1)} = 0.92$$

AREA BAJO LA CURVA ROC (AUC) PARA RF



VENTAJAS Y DESVENTAJAS DE RANDOM FOREST

Ventajas	Desventajas
- Combina múltiples árboles independientes	- Mayor capacidad de generalización
- Reduce el sobreajuste y la varianza	- Aumenta la complejidad computacional
- Maneja problemas no lineales	- Realiza promedios o votación de resultados de múltiples árboles
- Mejora la precisión y robustez	- Difícil de interpretar y visualizar
- Identifica importancia de características	- Requiere ajuste de hiperparámetros

PRÁCTICA ANÁLISIS DE DATOS

- El trabajo práctico se realiza en Rstudio.cloud.

RESUMEN DE LA CLASE

- Obtener los mejores parámetros para el modelo de RF.
- Realizar predicciones con árboles de decisión y RF.
- Métricas de evaluación (AUC, Accuracy).