Modelos Basados en Árboles de Decisión

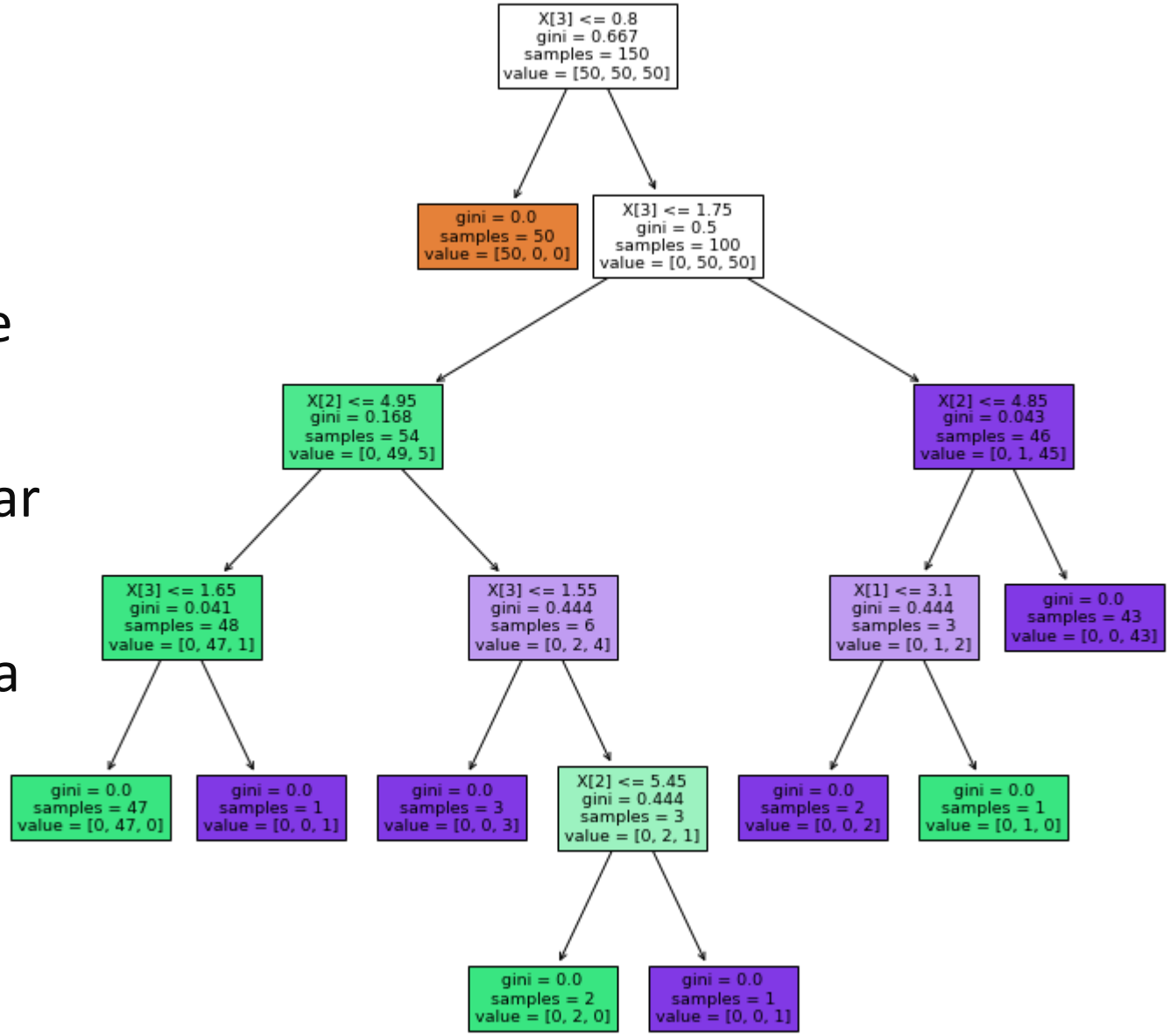
UNMSM – FISI – EP DE INGENIERÍA DE SOFTWARE

Minería de Datos

PROFESOR: JUAN GAMARRA MORENO

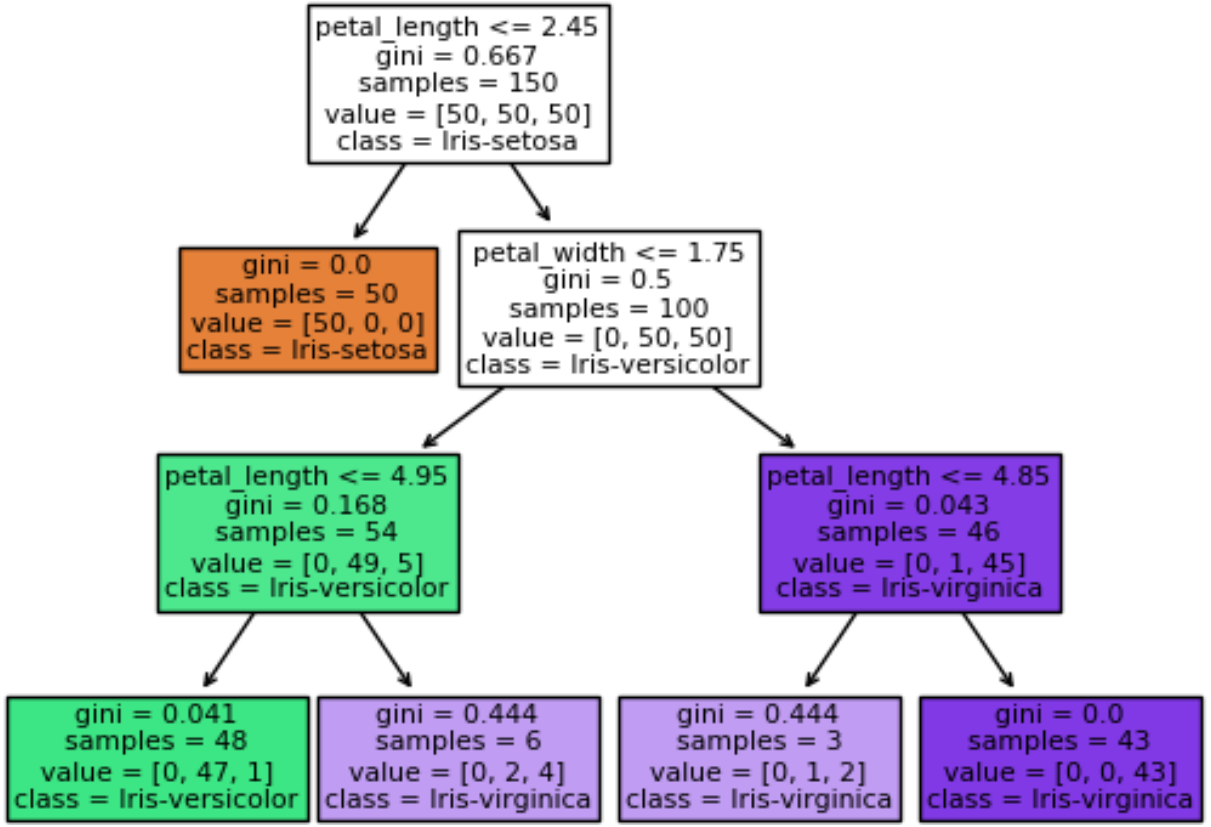
Introducción (1/3)

Los modelos basados en árboles de decisión son un tipo de modelo de aprendizaje automático que utiliza una estructura de árbol para realizar decisiones. En estos modelos, cada nodo interno del árbol representa una característica o atributo, y cada rama representa una posible salida basada en el valor de esa característica. Los nodos hoja representan las etiquetas de clasificación o los valores de predicción.



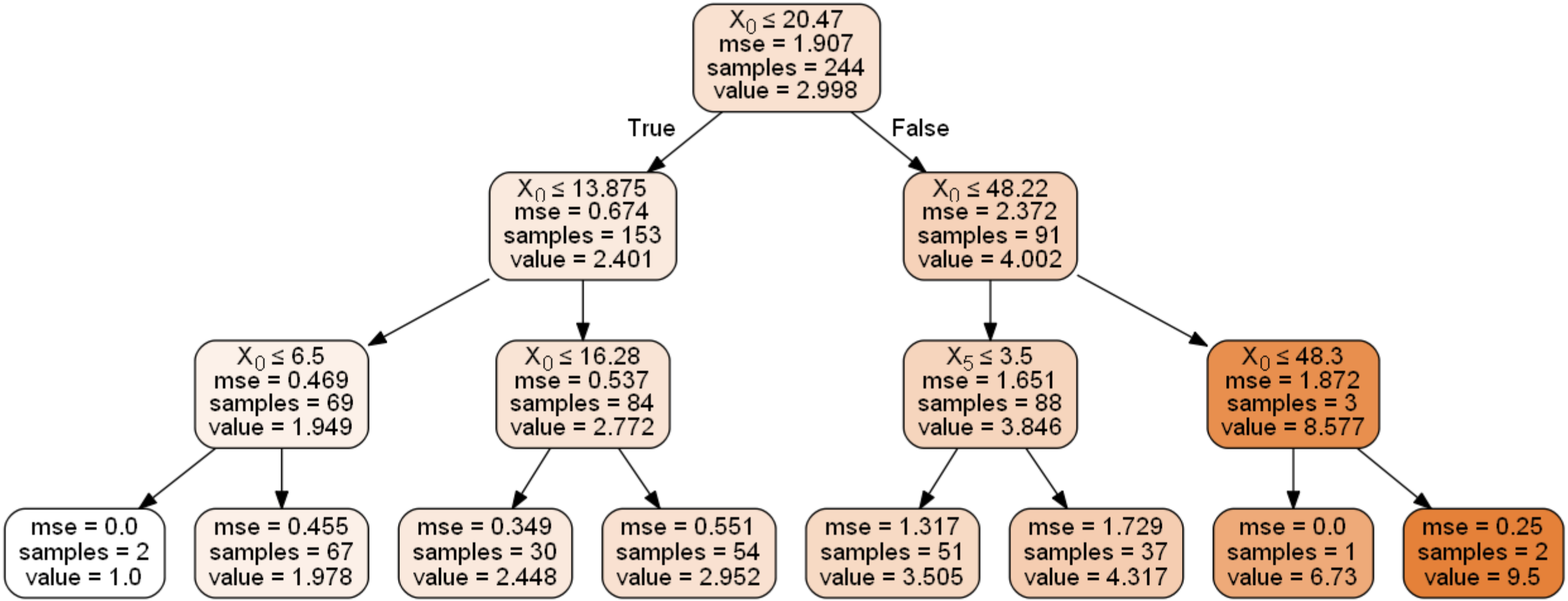
Introducción (2/3)

La idea principal detrás de los árboles de decisión es dividir el conjunto de datos en subconjuntos cada vez más pequeños mientras se hace una serie de decisiones basadas en características específicas. Estas divisiones se realizan de manera que se maximice la pureza de los subconjuntos resultantes en términos de la variable objetivo (en el caso de clasificación) o se minimice el error de predicción (en el caso de regresión).



Introducción (3/3)

Los árboles de decisión son fáciles de entender e interpretar, lo que los hace muy populares en el análisis de datos. Además, pueden manejar datos numéricos y categóricos, y son robustos frente a valores faltantes en los datos. Sin embargo, pueden tender a sobreajustarse si no se controlan correctamente, lo que puede reducir su capacidad de generalización a nuevos datos


X0 ≤ 20.47
mse = 1.907
samples = 244
value = 2.998
True
False
X0 ≤ 13.875
mse = 0.674
samples = 153
value = 2.401
X0 ≤ 48.22
mse = 2.372
samples = 91
value = 4.002
X0 ≤ 6.5
mse = 0.469
samples = 69
value = 1.949
X0 ≤ 16.28
mse = 0.537
samples = 84
value = 2.772
X5 ≤ 3.5
mse = 1.651
samples = 88
value = 3.846
X0 ≤ 48.3
mse = 1.872
samples = 3
value = 8.577
mse = 0.0
samples = 2
value = 1.0
mse = 0.455
samples = 67
value = 1.978
mse = 0.349
samples = 30
value = 2.448
mse = 0.551
samples = 54
value = 2.952
mse = 1.317
samples = 51
value = 3.505
mse = 1.729
samples = 37
value = 4.317
mse = 0.0
samples = 1
value = 6.73
mse = 0.25
samples = 2
value = 9.5

Modelos Basados en Árboles

Algunos de los modelos más comunes basados en árboles incluyen:

1. **Árboles de decisión**: Es el tipo más básico de modelo basado en árboles. Divide el conjunto de datos en subconjuntos cada vez más pequeños basados en características específicas, con el objetivo de maximizar la pureza de los subconjuntos resultantes.
2. **Random Forests (Bosques Aleatorios)**: Este modelo construye múltiples árboles de decisión durante el entrenamiento y combina sus predicciones para obtener una predicción final. Cada árbol se entrena en un subconjunto aleatorio de características y observaciones, lo que ayuda a reducir el sobreajuste y mejorar la generalización.

Modelos Basados en Árboles

4. **Gradient Boosting Machines (GBM)**: Esta es otra técnica de ensamblaje que construye árboles de decisión de forma secuencial, donde cada árbol se enfoca en corregir los errores de predicción cometidos por los árboles anteriores. Esto se logra mediante la optimización de una función de pérdida específica.
5. **XGBoost (Extreme Gradient Boosting)**: Es una implementación optimizada y eficiente de Gradient Boosting, que utiliza técnicas como el muestreo por gradiente y la poda de árboles para mejorar el rendimiento y la velocidad de entrenamiento.

Modelos Basados en Árboles

5. **LightGBM**: Otra implementación eficiente de Gradient Boosting que utiliza una estrategia de crecimiento de árboles basada en histogramas, lo que lo hace aún más rápido que XGBoost en muchos casos.
6. **CatBoost**: Es una biblioteca de Gradient Boosting desarrollada por Yandex que está diseñada para manejar automáticamente variables categóricas, lo que simplifica el preprocesamiento de datos.

Estos son solo algunos ejemplos de modelos basados en árboles, y hay muchas otras variantes y extensiones disponibles en la comunidad de aprendizaje automático. Cada uno tiene sus propias características, fortalezas y debilidades, y la elección del modelo adecuado depende del problema específico y los requisitos del proyecto