Modelos de Regresión con Árboles de Decisión

¿Qué es la Regresión?

Predecir Valores Numéricos

La regresión nos ayuda a predecir números específicos basándose en información que ya conocemos.

- Precio de un taxi según la distancia
- Costo de una casa según su tamaño
- Temperatura según la época del año

Clasificación vs Regresión

Clasificación

Responde: "¿A qué categoría pertenece?"

- ¿Es spam o no spam?
- ¿Gato o perro?
- ¿Aprobado o reprobado?

Regresión

Responde: "¿Cuánto vale?"

- ¿Cuánto cuesta el taxi?
- ¿Qué precio tiene la casa?
- ¿Cuántos grados hace?

¿Qué es un Árbol de Decisión en Regresión?

Modelo predictivo segmentado

Divide el espacio de variables para predecir valores continuos mediante reglas binarias

Predicción por media

Cada hoja del árbol predice la media de la variable respuesta de las muestras que contiene

Ventajas clave

Interpretación intuitiva, manejo natural de interacciones no lineales y robustez ante outliers

Árboles de Decisión

Haces preguntas de sí/no para llegar a una respuesta. Cada pregunta divide las opciones hasta encontrar el resultado más probable.

1

¿Es un viaje largo?

Primera pregunta que divide los datos

2

¿Es en hora pico?

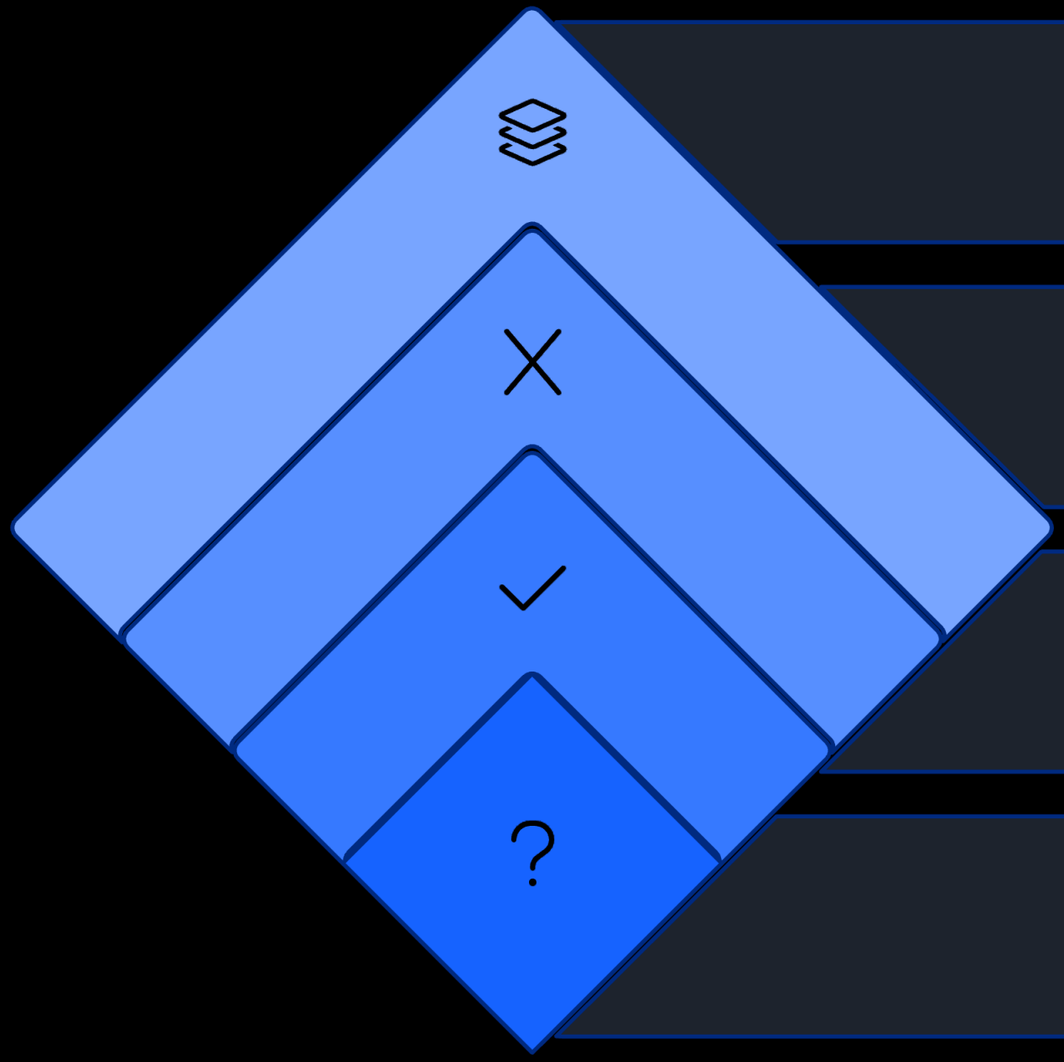
Segunda pregunta para refinar

3

Predicción final

El precio estimado del taxi

Cómo se Divide un Conjunto de Datos



Subdivisión

Refinar grupos similares

Rama No

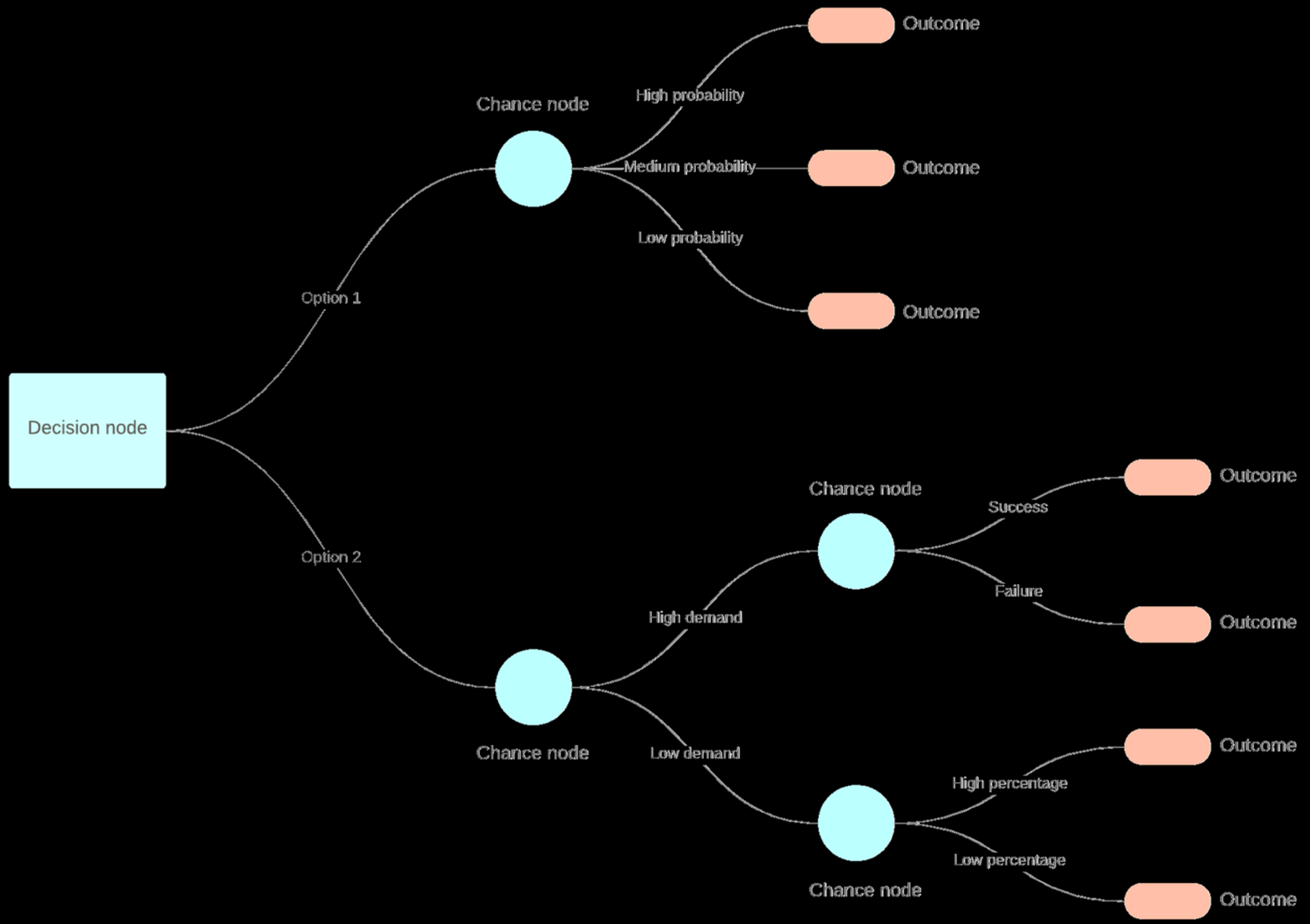
Puntos que no cumplen

Rama Sí

Puntos que cumplen condición

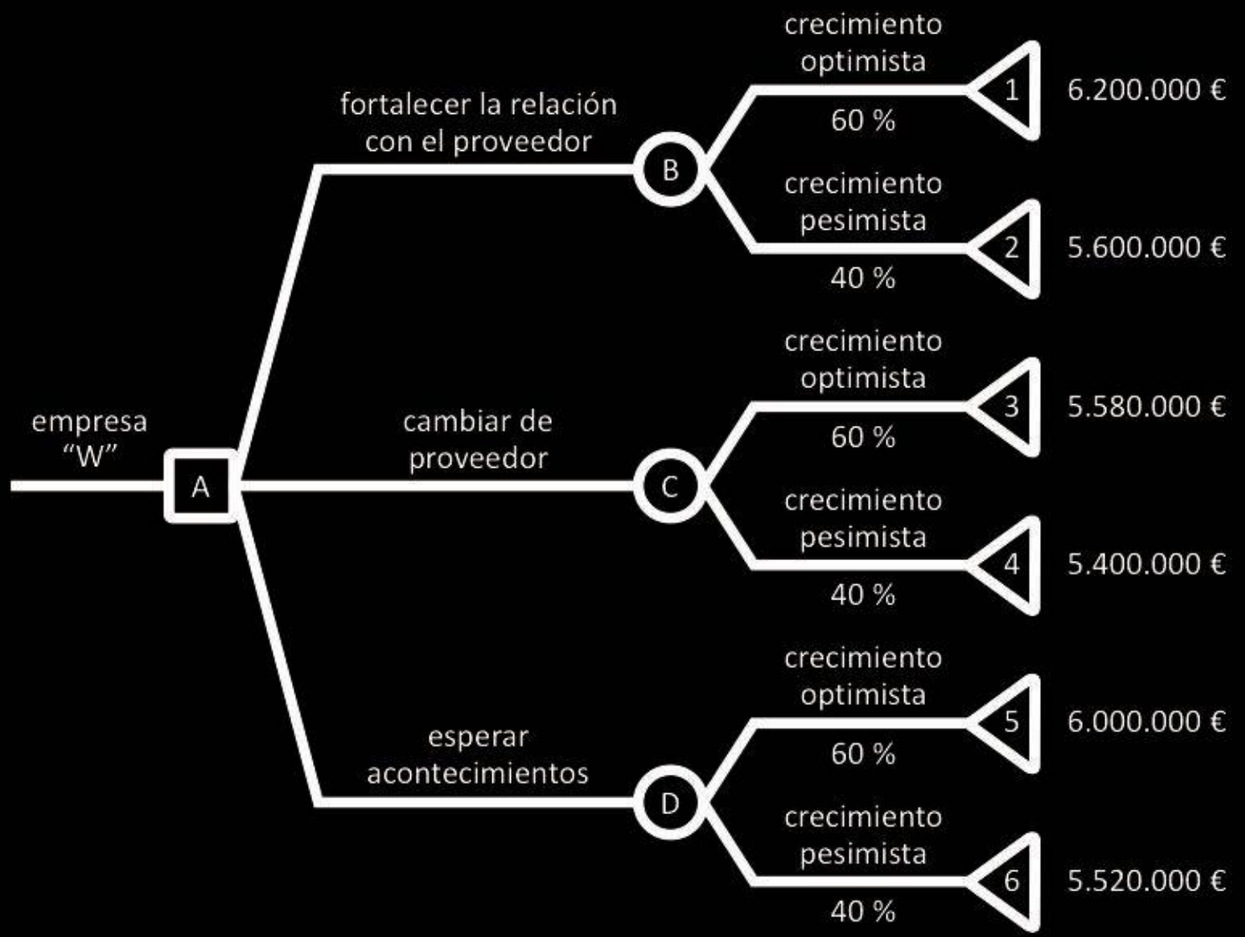
Pregunta Inicial

Dividir por criterio simple


Decision node
Option 1
Chance node
High probability
Outcome
Medium probability
Outcome
Low probability
Outcome
Option 2
Chance node
High demand
Chance node
Success
Outcome
Failure
Outcome
Low demand
Chance node
High percentage
Outcome
Low percentage
Outcome

árbol de decisión

ejemplo práctico



El problema del tamaño del árbol

Árbol muy pequeño

Underfitting: Muy pocas preguntas = predicciones imprecisas. Es como generalizar demasiado.

Árbol muy grande

Overfitting: Demasiadas preguntas = memoriza los datos pero no predice bien casos nuevos.

Tamaño ideal

El equilibrio perfecto entre simplicidad y precisión para nuevas predicciones.

Elegir la profundidad del Árbol: ¿Por qué es clave?

Control de complejidad

La profundidad determina cuántas divisiones puede hacer el árbol. Muy profundo = overfitting, muy superficial = underfitting

Balance Sesgo-Varianza

La profundidad óptima minimiza el error total balanceando la capacidad de capturar patrones sin memorizar ruido

Métodos de Selección

Validación cruzada y análisis de MSE en conjunto de prueba revelan la profundidad ideal

Controlando el crecimiento del Árbol

Profundidad Máxima

Limita cuántas preguntas puede hacer el árbol consecutivamente

Muestras Mínimas

Exige un número mínimo de ejemplos antes de hacer una nueva división

Otros parámetros configurables en Árboles de Decisión

min_samples_split

Número mínimo de muestras requeridas para dividir un nodo interno. Controla cuándo el árbol deja de crecer.

min_samples_leaf

Mínimo de muestras que debe contener una hoja. Previene hojas con muy pocas observaciones.

max_features

Máximo número de variables consideradas en cada división. Introduce aleatoriedad y reduce overfitting.

Poda

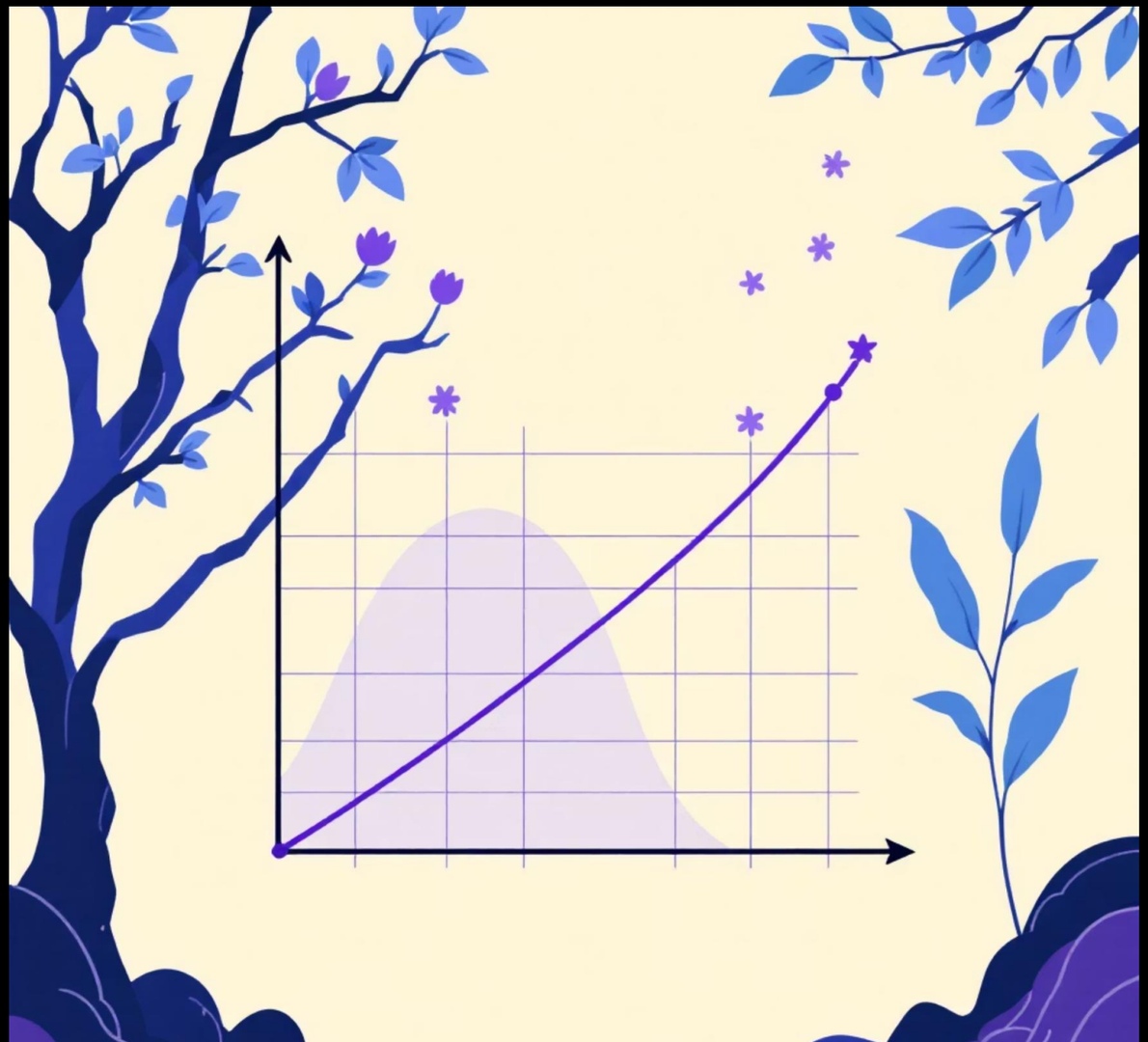
Técnica post-construcción para simplificar árboles eliminando ramas que no mejoran la validación.

Los criterios de división como MSE y MAE determinan cómo se optimizan las particiones en cada nodo.

Análisis de Errores en Árboles de Decisión

Comportamiento del error

El error de entrenamiento siempre decrece con mayor profundidad, pero el error de validación puede aumentar cuando aparece sobreajuste.



Métricas de Evaluación

MSE
Error cuadrático medio - penaliza errores grandes

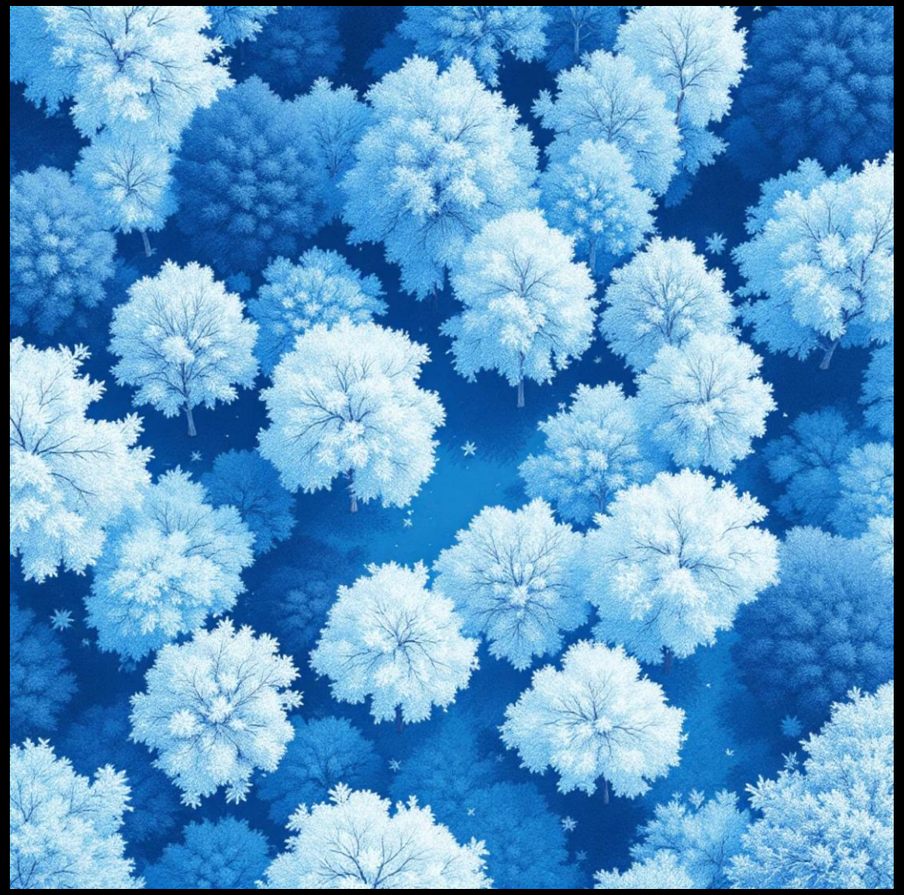
MAE
Error absoluto medio - más robusto a outliers

R^2

Coeficiente de determinación – proporción de varianza explicada

Extensiones: Bosques y Boosting

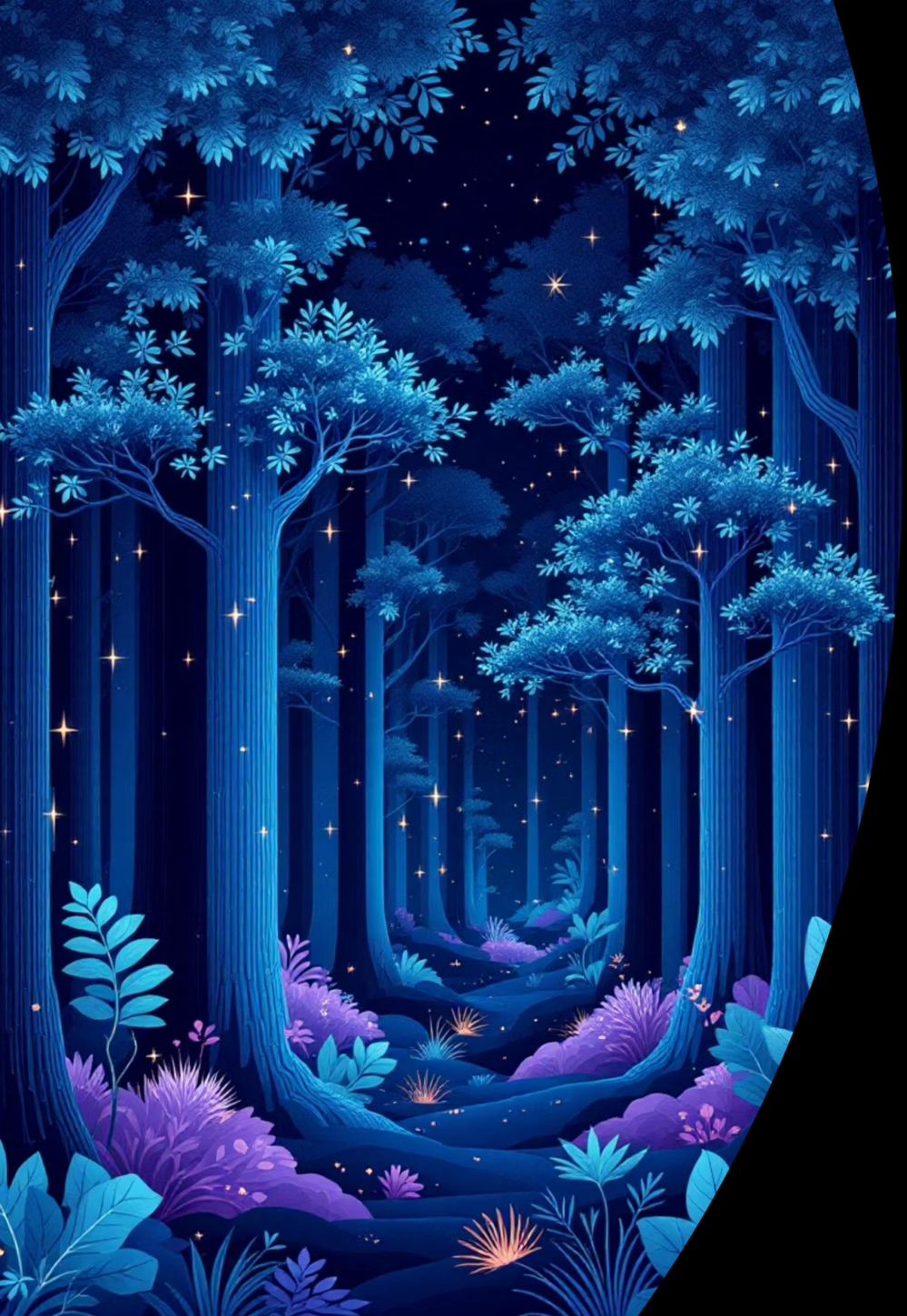
Random Forest



Múltiples árboles trabajando en equipo. Cada árbol vota y se toma el promedio de todas las predicciones.

Gradient Boosting

Árboles que aprenden de los errores de los anteriores, corrigiéndose uno tras otro para mejorar.



Random Forest

01

Bootstrap Sampling

Cada árbol se entrena con una muestra aleatoria con reemplazo del dataset original

ćč

Selección Aleatoria de Variables

En cada división, solo se considera un subconjunto aleatorio de las variables disponibles

03

Predicción por Promedio

La predicción final es el promedio de todas las predicciones individuales de los árboles

Ventajas clave: Mayor robustez, excelente manejo de grandes datasets, y significativa reducción del riesgo de sobreajuste comparado con árboles individuales.

¿Cómo Funciona Gradient Boosting?

1

Modelo Base

Inicia con un árbol simple que hace predicciones básicas

2

Cálculo de Residuales

Identifica errores del modelo actual (diferencia entre predicción y valor real)

3

Nuevo Árbol

Entrena un árbol para predecir estos errores residuales

4

Actualización

Suma la predicción del nuevo árbol (escalada por learning rate) al modelo

Resultado: Modelo aditivo que mejora gradualmente con cada iteración

Comparación: Árbol Simple vs Random Forest vs Gradient Boosting

Modelo	Tiempo Entrenamiento	Interpretabilidad
Árbol Simple	Rápido	Alta
Random Forest	Medio	Media
Gradient Boosting	Lento	Baja

Trade-off principal: Mayor precisión requiere mayor complejidad computacional y menor interpretabilidad del modelo resultante.

Modelo	¿Cómo funciona?	Cuándo usarlo	Cuándo no usarlo	Pros	Contras
Árbol de decisión	Es un conjunto de reglas tipo if/else que terminan en un valor	- Dataset pequeño o mediano.- Cuando quieres **interpretabilidad** (ver reglas claras).- Explicar resultados a alguien no técnico.	- Cuando los datos son muy complejos y no lineales.- Cuando necesitas máxima precisión.- Si hay riesgo de **sobreajuste** (aprende demasiado los datos).	- Fácil de entender y visualizar.- Rápido de entrenar.- Funciona con datos mixtos (numéricos y categóricos).	- Poco preciso solo.- Muy sensible a cambios en los datos.- Puede sobreajustar si no se poda.
Random Forest	Te quedas con lo que diga la mayoría (o el promedio si es un número).Esto hace que el resultado sea más estable y confiable que un único árbol.	- Datasets medianos y grandes.- Cuando quieres buena precisión sin mucho tuning.- Cuando los datos son ruidosos o desbalanceados.	- Si necesitas explicar reglas simples (es un “caja negra”).- Si necesitas predicciones en tiempo real con latencia muy baja (porque son muchos árboles).	- Mucho más preciso y robusto que un solo árbol.- Reduce sobreajuste.- Funciona bien “out of the box”.	- Más lento de entrenar y predecir.- Difícil de visualizar.- No capta relaciones muy complejas comparado con Gradient Boosting.
Gradient Boosting	Cada nuevo arbol aprende a corregir los errores de los anteriores, y juntos logran un resultado mucho más preciso.	- Cuando quieres **máxima precisión**.- Competencias tipo Kaggle.- Datos tabulares medianos/grandes.- Si las relaciones entre variables son complejas.	- Si el dataset es muy chico (puede sobreajustar).- Si no puedes dedicar tiempo a ajustar hiperparámetros.- Si necesitas interpretabilidad simple.	- Alta precisión.- Captura relaciones complejas.- Flexibilidad enorme en parámetros.	- Entrenamiento más lento.- Necesita tuning (learning rate, n_estimators, max_depth).- Menos interpretable.

Puntos Clave para Recordar

Predicción intuitiva

Los árboles toman decisiones como nosotros: pregunta tras pregunta

Balance es clave

Ni muy simple ni muy complejo - el tamaño perfecto es crucial

La unión hace la fuerza

Múltiples árboles juntos son más precisos que uno solo

Ejemplo práctico: Prediciendo tarifas del titanic

Variables Importantes

ćĉ

Clase del pasajero

Primera, segunda o tercera clase

02

Puerto de embarque

Southampton, Cherbourg o Queenstown

03

Edad del pasajero

Adultos vs niños