De mirar el pasado a anticipar el futuro

Machine Learning con Python para la Toma de Decisiones Empresariales

Jorge Israel Frometa Moya

Ensembles: La fuerza de la inteligencia colectiva.

- Random Forest: ¿Por qué 100 árboles ven más que uno solo?

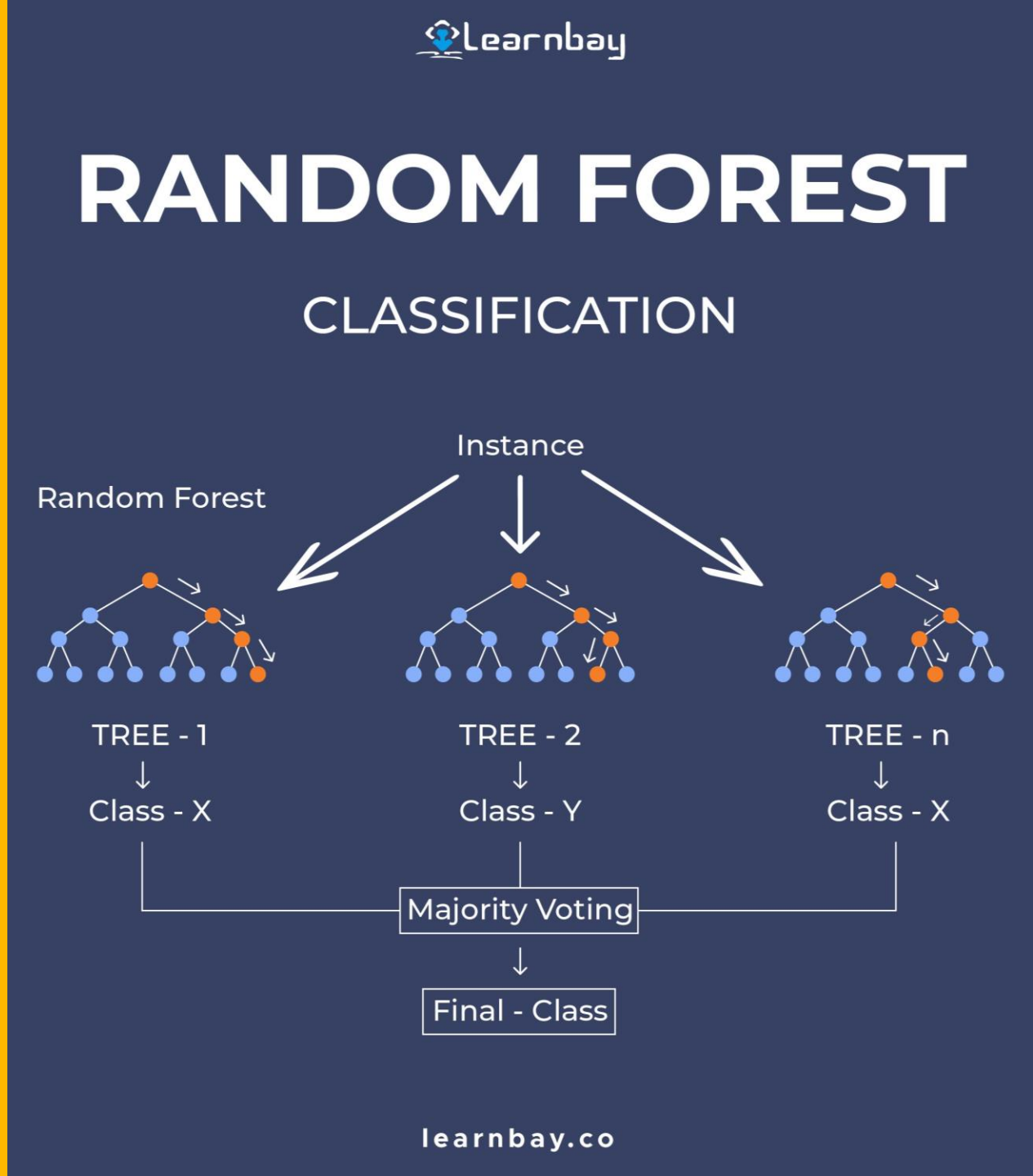
Concepto clave: Reducción del riesgo de error mediante la combinación de modelos

El concepto de "Sabiduría de la Multitud"

• Analogía de Negocio: Si quieres saber si una inversión es buena, ¿le preguntas a un solo asesor o a un comité de varios expertos independientes?

• Cómo funciona: Random Forest crea muchos árboles. Cada uno ve una parte distinta de los datos y de las variables.

•Al final, se vota: si 70 árboles dicen "Fuga" y 30 dicen "Se queda", la predicción final es "Fuga".



Estabilidad vs. Interpretabilidad

- **El "Trade-off"**: Un árbol es fácil de dibujar; un bosque de 100 árboles no.
- **Ganancia**: Perdemos la capacidad de ver el "dibujo" del árbol, pero ganamos mucha estabilidad (menos Overfitting) y precisión.
- **Solución para el Analista**: Usaremos el ranking de Importancia de Variables para seguir explicando el "por qué".

PILARES DE RANDOM FOREST

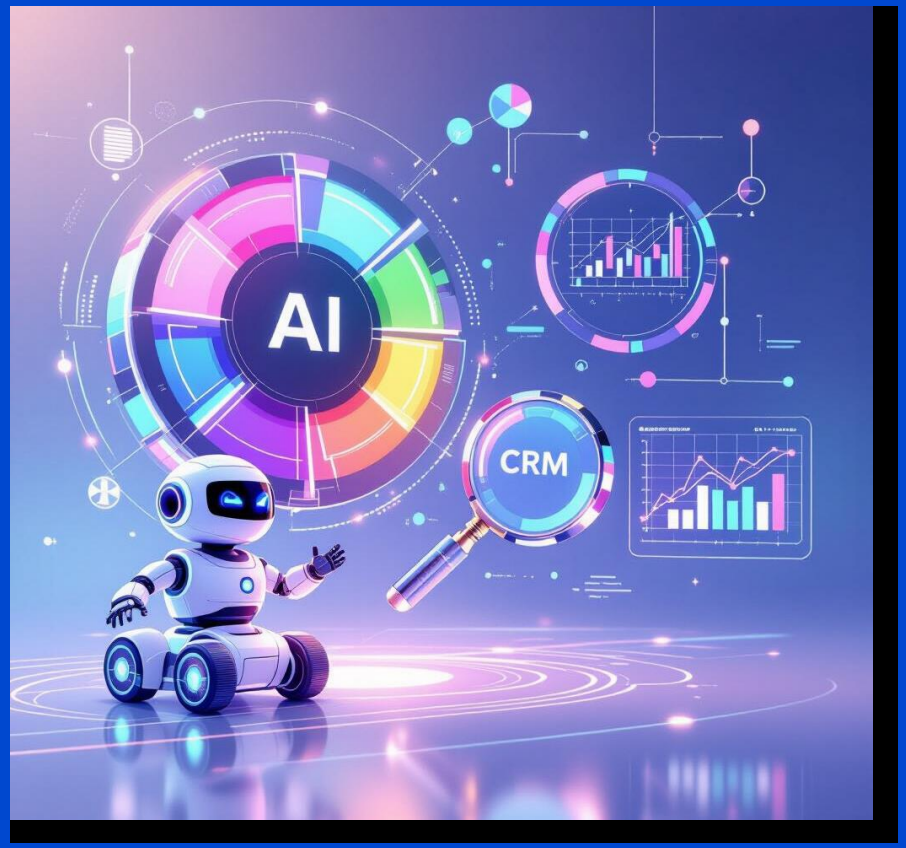
BAGGING (Bootstrap Aggregating)

- Crea 100 versiones diferentes del dataset
- Cada árbol ve una muestra aleatoria (con reemplazo)
- Ejemplo: 1,000 clientes → Cada árbol ve ~630 únicos

ALEATORIEDAD DE CARACTERÍSTICAS

- En cada división, considera solo un subconjunto de variables
- Ejemplo: 20 variables → Considera solo $\sqrt{20} \approx 4$ en cada división
- Fuerza diversidad: Árboles no se copian entre sí

VENTAJAS Y DESVENTAJAS



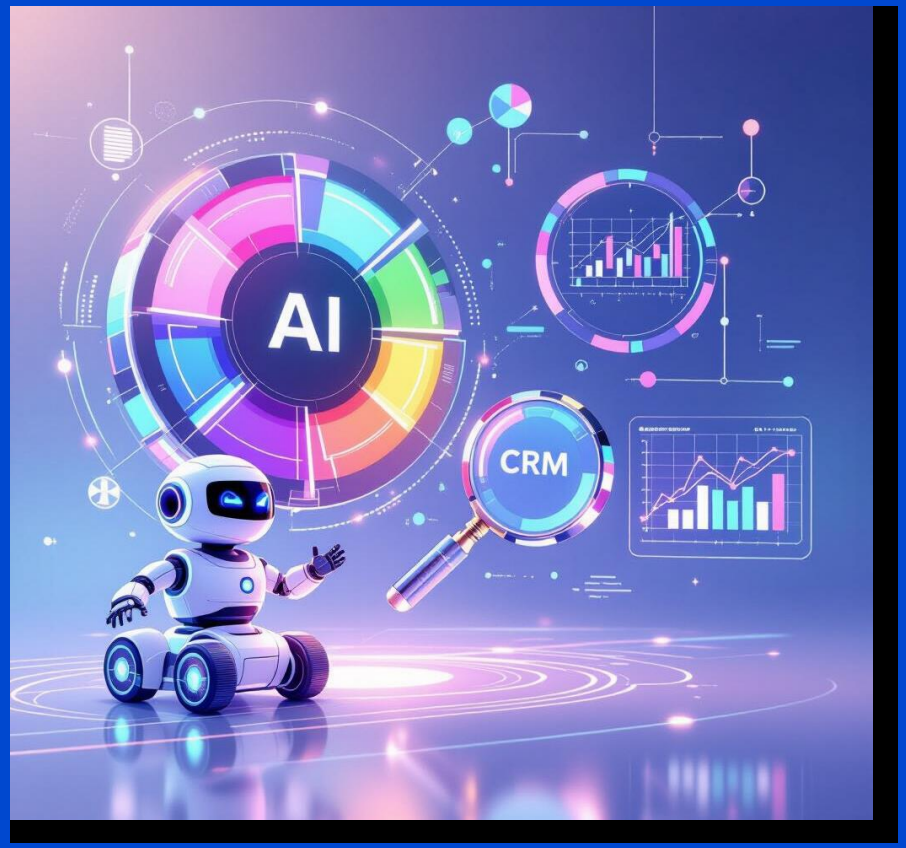
✅ VENTAJAS:

- Estable: Pequeños cambios → Resultados similares
- Robusto: Maneja outliers mejor
- Precisa: Mejor performance en test
- Versátil: Funciona bien "out of the box“

❌ DESVENTAJAS:

- Caja negra: No hay un árbol único para mostrar
- Computacionalmente costoso: 100× más lento
- Memoria: Consume más guardar 100 árboles

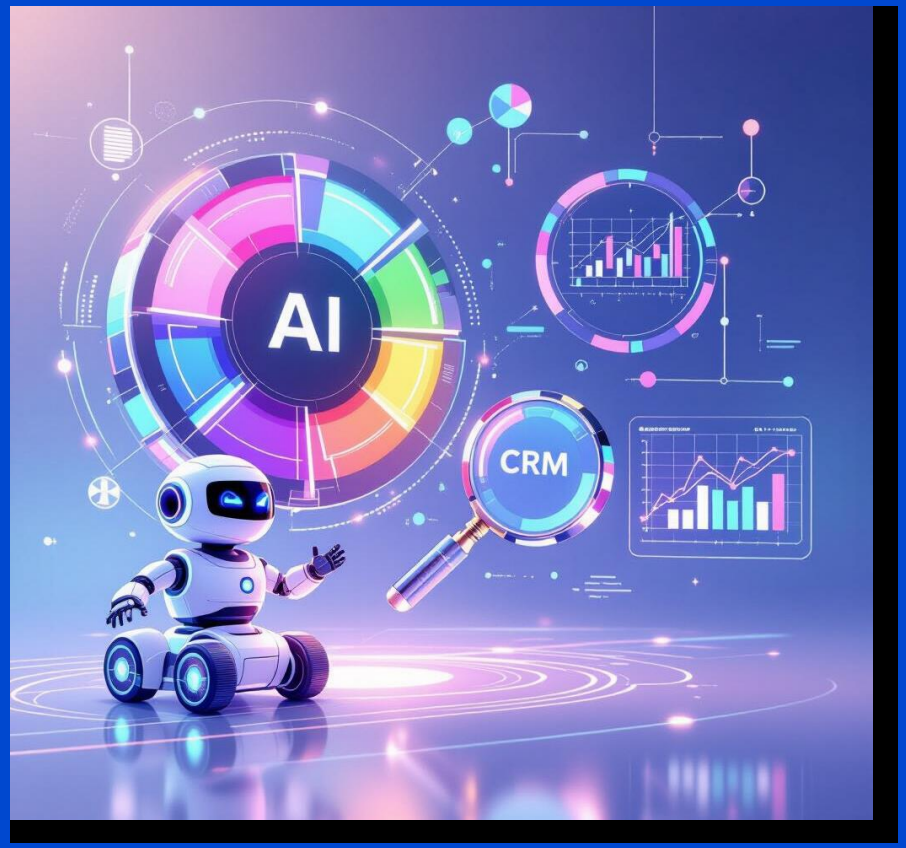
IMPORTANCIA DE VARIABLES - EL "POR QUÉ" APROXIMADO:



Aunque no podemos ver el árbol completo, podemos ver:

- Qué variables son más importantes en promedio
- Cómo contribuyen a la decisión colectiva
- Qué características diferencian a los grupos

Más Allá de Random Forest - Gradient Boosting



GRADIENT BOOSTING (BOOSTING):

- Filosofía: "Equipo que aprende de errores anteriores"
- Entrenamiento: Árboles en serie, cada uno corrige al anterior
- Fortaleza: Máxima precisión posible
- Debilidad: Más propenso a overfitting, más lento