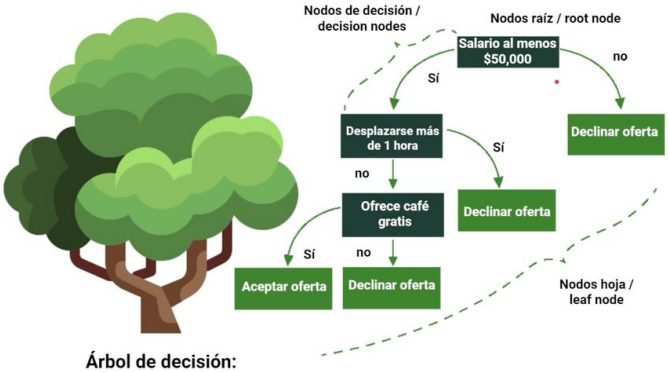


Árbol de decisión: ¿debería aceptar una nueva propuesta laboral?

Terminología

Nodo raíz	Poda (Pruning)
División	Rama / Subárbol
Nodo de decisión	Nodo madre/padre e hijo
Nodo de hoja o terminal	

Ventajas

- Algoritmo de caja blanca.
- Resultados fáciles de interpretar y de entender.
- Las combinaciones de los mismos pueden dar resultados muy certeros. Por ejemplo, *random forest*.

Desventajas

- Tienden al sobreajuste u overfitting.
- Se ven influenciadas por los outliers.
- Crear árboles demasiado complejos puede conllevar que no se adapten bien a los nuevos datos.
- Se pueden crear árboles sesgados si una de las clases es más numerosa.

¿Cuándo usar árboles de decisión?

- Sencillo y fácil de entender.
- Funcionan bastante bien con grandes conjuntos de datos.
- Relativamente robusto.
- Es un método muy útil para analizar datos cuantitativos.
- Aplica para clasificación y regresión.

Ejemplo:
https://economipedia.com/definiciones/arbol-de-decision-en-valoracion-de-inversiones.html

Como evaluar un modelo?

Matriz de confusión

- Permite visualizar el desempeño de un algoritmo de aprendizaje supervisado.
- Cada **columna** representa el número de predicciones de cada clase.
- Cada **fila** representa a las instancias en la clase real.

Matriz de confusión

- En términos prácticos nos permite ver **qué tipos de aciertos y errores** está teniendo nuestro modelo.

Interpretación de matriz de confusión

- ***Verdadero Positivo (TP):*** predije que era positivo y lo era.
- ***Verdadero Negativo (TN):*** predije que era falso y lo era.
- ***Falso Positivo (FP):*** predije que era positivo, pero resultó ser negativo.
- ***Falso Negativo (FN):*** predije que era negativo, pero resultó siendo positivo.

Interpretación de matriz de confusión

- **Verdaderos positivos** como **negativos** son **aciertos**.
- **Falsos negativos** como **positivos** son **errores**.

Exactitud o accuracy

- Cercanía al resultado de una medición del valor verdadero.
- En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación.

Exactitud o accuracy

- Proporción entre los positivos reales predichos y todos los casos positivos.
- En forma práctica, la exactitud es el **% total de elementos clasificados correctamente**.

Fórmula accuracy

$$\frac{(VP+VN)}{(VP+FP+FN+VN)} * 100$$

Precisión

- Dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud.
- Cuanto menor es la dispersión, mayor la precisión.
- Proporción entre el número de predicciones correctas (tanto positivas como negativas) y el total de predicciones.

Precisión

- En forma práctica, es el **porcentaje de casos positivos detectados**.
- Sirve para medir la **calidad del modelo** de machine learning en **clasificación**.

Fórmula precisión

$$\frac{(VP)}{(VP+FP)}$$

Sensibilidad

- *Recall*, *sensitivity* o **tasa de verdaderos positivos**.
- Proporción de **casos positivos** que fueron correctamente identificados.

Fórmula sensibilidad

$$\frac{VP}{(VP + FN)}$$

Especificidad

- **Tasa de verdaderos negativos**.
- Proporción de **casos negativos** que fueron correctamente identificados.

Fórmula especificidad

$$\frac{VN}{VN + FP}$$

F1-score

- Resume la **precisión y sensibilidad** en una sola métrica.

Fórmula F1-score

$$2 * \frac{precision * recall}{precision + recall}$$

En resumen

$$Precision = \frac{TruePositive}{ActualResults} \text{ or } \frac{TruePositive}{TruePositive + FalsePositive}$$

$$Recall = \frac{TruePositive}{PredictedResults} \text{ or } \frac{TruePositive}{TruePositive + FalseNegative}$$

$$Accuracy = \frac{TruePositive + TrueNegative}{Total}$$

$$Specificity = \frac{TrueNegatives}{TrueNegatives + FalsePositives}$$

Random Forest

Random forest

- Bosques aleatorios.
- **Ensamble** en machine learning en donde se **combinan árboles de decisión**.

¿Qué es un ensamble?

- También conocidos como métodos combinados.
- Intentan ayudar a mejorar el **rendimiento** de los modelos de machine learning.

¿Qué es un ensamble?

- Proceso mediante el cual se construyen estratégicamente varios modelos de machine learning para resolver un problema particular.

Random forest

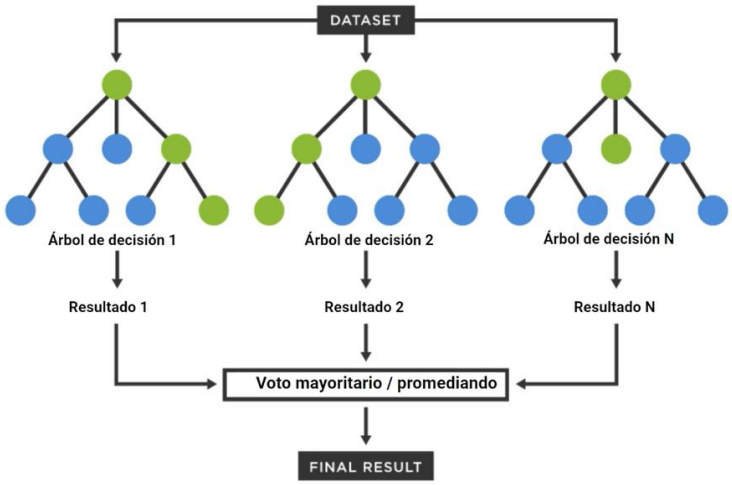
- Al igual que el árbol de decisión, es un algoritmo de aprendizaje supervisado.
- Utilizados en problemas de clasificación.
- También puede usarse para regresión.

Problemas de overfitting

Uno de los problemas con la creación de un árbol de decisión es que si le damos la **profundidad suficiente**, tiende a "memorizar" las soluciones en vez de generalizar.

Es decir, a tener **overfitting**. ❌

Ara solucionar el overfitting se recomienda hacer random forest, el cual sigue el sig. Algoritmo:



Ventajas

- Funciona bien aún sin ajuste de hiper parámetros.
- Al utilizar múltiples árboles se reduce considerablemente el riesgo de overfitting.
- Suele mantenerse estable frente a nuevas muestras de datos.

Desventajas

- Es mucho más “costoso” de crear y ejecutar que “un solo árbol” de decisión.
- No funciona bien con datasets pequeños.
- Puede requerir muchísimo tiempo de entrenamiento.
- Su interpretación a veces se vuelve compleja.

¿Cuándo usar random forest?

- Rápido y fácil de aplicar.
- En el caso de realizar técnicas de hypertuning de hiper parámetros.
- Para problemas de clasificación y también de regresión.
- Datasets grandes.
- Para evitar el overfitting mediante la aplicación de métodos de ensamble.

Ejemplo: https://www.iartificial.net/random-forest-bosque-aleatorio/