## Clase N°2. Introducción a los árboles de decisión.

### Introducción

Un **árbol de decisión** representa un conjunto de reglas dicotómicas utilizadas para predecir el valor de una variable respuesta en función de un conjunto de variables observables o factores conocidos por el investigador. Su estructura está representada por un árbol invertido como se muestra a continuación:

La parte superior se llama **nodo raíz** y contiene la totalidad del conjunto de datos, los cuales luego serán particionados o repartidos a través de las ramas los cuales representan los **nodos de decisión**, para luego llegar al **nodo terminal** o **nodo hoja** el cual recoge las predicciones del algoritmo.

Entre las principales ventajas presentes en un árbol de decisión se pueden destacar que:

- Son fáciles de representar e interpretar.
- No requieren gran limpieza de los datos y no se ven afectados por la presencia de outliers.
- Pueden utilizar predictores numéricos como categóricos.
- Permiten identificar los predictores más representativos.
- Pueden ser utilizados tanto para problemas de regresión como de clasificación.
- Utiliza un algoritmo de división binaria recursiva o "*recursive binary splitting"*.

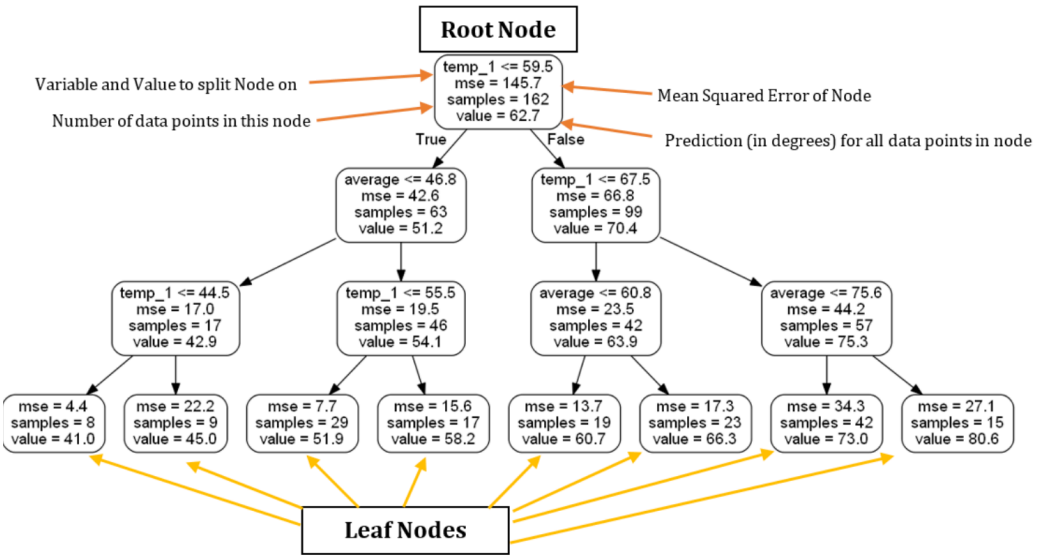
Por otra parte, entre sus principales desventajas, se pueden mencionar:

- Su capacidad de predicción puede ser a veces poco efectivo dado que utiliza un solo árbol de decisión.
- Es propenso a estar sobreajustado "overfitting", lo cual afecta el resultado de su performance.
- Puede perder información al categorizar variables continuas.

## Árboles de regresión

Los árboles de regresión son el subtipo de árboles de predicción que se aplica cuando la *variable respuesta es* ***continua***. En la etapa del entrenamiento de un árbol de regresión, los datos se van distribuyendo por bifurcaciones llamadas *nodos de decisión* generando la estructura del árbol hasta alcanzar un *nodo terminal*. Al momento de querer predecir un nuevo valor ingresado en las variables observables, el algoritmo recorre el árbol, pasando por los nodos secundarios hasta llegar al nodo hoja donde recoge el valor de la predicción.

Para poder comprender mejor los elementos de un árbol de regresión, se muestra a continuación el siguiente esquema:




Donde "Root Node" es el nodo raíz, "leaf Nodes" son los nodos hojas o terminales, "sample" indica la cantidad de elementos pertenecientes al nodo, "mse" el error cuadrático medio cometido por el nodo, "value" el valor de la predicción y "temp_1" y "average" son las variables observables en este caso.

## Etapas de un modelo predictivo

1. **Preprocesamiento de los datos.**
   Comprende la limpieza de los datos previo a utilizar un modelo predictivo. Ello comprende por ejemplo: eliminar datos duplicados, completar o eliminar datos nulos, estandarizar variables numéricas o transformar variables categóricas en numéricas.
2. **Partición.**
   Luego de la limpieza de los datos, se procede a dividir el conjunto de datos en dos subconjuntos llamados "training" y "test" utilizando el método: training_test o el método de validación cruzada "cross-validation".
3. **Entrenamiento.**

En esta etapa se enseña al modelo a que realice una operación de clasificación usando el subconjunto de entrenamiento.

4. **Evaluación.**
   Una vez que el modelo ha sido entrenado se genera una predicción y se compara con los datos de la variable respuesta del conjunto de testing.
5. **Validación.**
   Finalmente, se utilizan métricas con el fin de poder evaluar el rendimiento del modelo y de esta forma comparar el desempeño de diferentes modelos o rechazarlo en caso de bajo performance.

**Entrenamiento del árbol**

En esta etapa se realizan divisiones sucesivas del conjunto de datos en regiones independientes entre sí llamadas nodos. En cada nodo se realiza una predicción y se vuelve a realizar una división dicotómica hasta llegar al nodo hoja.

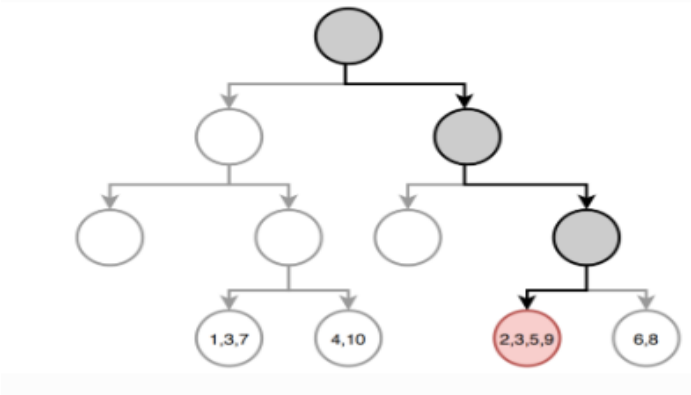
**Predicción**

Luego de la creación del árbol de entrenamiento se realizan predicciones partiendo de un conjunto de datos observables conocidos por el usuario.

Por ejemplo, al considerar 10 observaciones, Y la variable respuesta y X la variable observable, resulta:

| id | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 10 | 18 | 24 | 8 | 2 | 9 | 16 | 10 | 20 | 14 |
| X | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

En base a la tabla al aplicar el algoritmo, se puede observar en color gris el camino de decisión que recorre el dato hasta llegar al nodo terminal en color rojo donde se muestra la predicción.



El valor predicho por el árbol en nodo terminal es la media de la variable respuesta Y correspondiente a las observaciones del id= 2, 3, 5, 9, cuyos valores, se calculan a continuación:

$$\hat{\mu} = \frac{18 + 24 + 2 + 20}{4} = 16$$

Aunque la media aritmética es el método más utilizado para estimar la predicción de un árbol e regresión, existen otros métodos basados en la ponderación o peso de cada observación en la etapa de entrenamiento.

**Evaluación del modelo**

En los árboles de regresión, los criterios más empleados para identificar la eficiencia del modelo son:

- (**MAE**): Error Absoluto Promedio.
- (**MSE**): Error Cuadrático Promedio.
- (**RMSE**): Raíz Cuadrada del Error Cuadrático Promedio.

El indicador RMSE se calcula mediante la sig. ecuación:

$$\text{RMSE}=\sqrt{\frac{\sum(pi-Oi)^2}{n}}$$

- Pi: es el valor de la variable predictoria.
- Oi: es el valor observable.
- n: el tamaño de la muestra.

Los valores próximos a cero indican menor pérdida de información y mayor exactitud en la predicción. Por otra parte los valores muy grandes, reflejan que existe una gran diferencia entre los valores de la variable respuesta del conjunto de evaluación y la variable predicha, por lo tanto el modelo es poco eficiente. En este caso, el modelo se puede ajustar o rechazar y seleccionar otro modelo predictivo.

**Arboles de decisión en Python**

Para utilizar arboles de decisión en Python se utiliza la librería scikit-learn y la sublibreria DecisionTreeRegressor.

**ACTIVIDAD**

Se quiere estimar el nivel de pobreza en relación a la deserción escolar y a la mortalidad infantil en las distintas provincias de la Argentina. Para ello se crea un árbol de regresión. En base al algoritmo desarrollado en la clase, responder las sig. preguntas:

1. ¿Qué información refleja el nodo raíz?
2. ¿Qué tipo de variables se muestran en los nodos de decisión?
3. ¿Qué variable se representa en los nodos hojas?
4. ¿Cuál es el nivel de árbol sin podar? ¿Qué nivel tiene el árbol podado?
5. ¿Al comparar el árbol podado y el árbol sin podar cual utilizarían para realizar la predicción del nivel de pobreza? Justificar.