INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

SITUACIÓN PROFESIONAL 3

CLASE 5

Prof. Ricardo Piña

Ver en Video:

Título: [IA] Clase 6
https://www.youtube.com/watch?v=DPUwL8e145g (https://www.youtube.com/watch?v=DPUwL8e145g)
desde: inicio
hasta: fin

Out[6]:

A esta altura de la carrera Ud. todavía no sabe programar en Python, así que en este archivo hemos ocultado las celdas que contienen código para facilitar su lectura. Si Ud. quiere ver el código u ocultarlo, haga click aquí.

Out[7]:

ENTRENAMIENTO, TUNNING, Y EVALUACIÓN DEL MODELO

Dado un Conjunto de Datos (Dataset) que contiene todas las instancias u observaciones de que disponemos hemos utilizado un procedimiento en el cual definimos que teníamos que dividir el Dataset en dos partes:

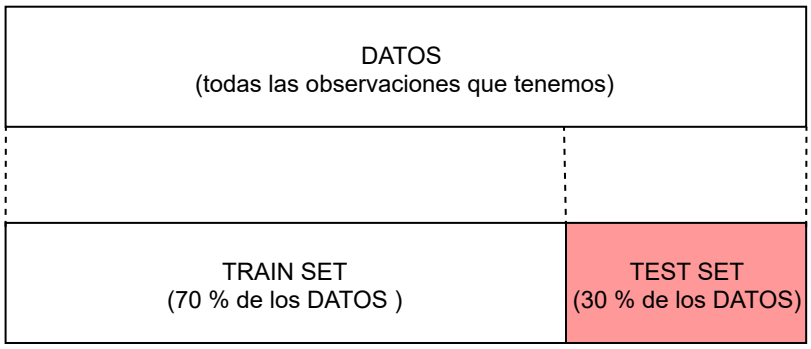
- **Train Set**
- **Test Set**

con el objetivo de que nuestro modelo **aprenda** del Train Set y sea **evaluado** en el Test Set.

La idea es que el Test Set represente en alguna medida a todas las observaciones que existen en el Universo y que no forman parte las que conocemos en el momento de entrenamiento, por lo tanto nuestra consigna es que

> **"no debemos permitir que nuestro modelo aprenda del Test Set"**,

sólo debe servir para evaluar de la mejor manera cómo se comportará nuestro modelo frente a las observaciones que no conocemos.

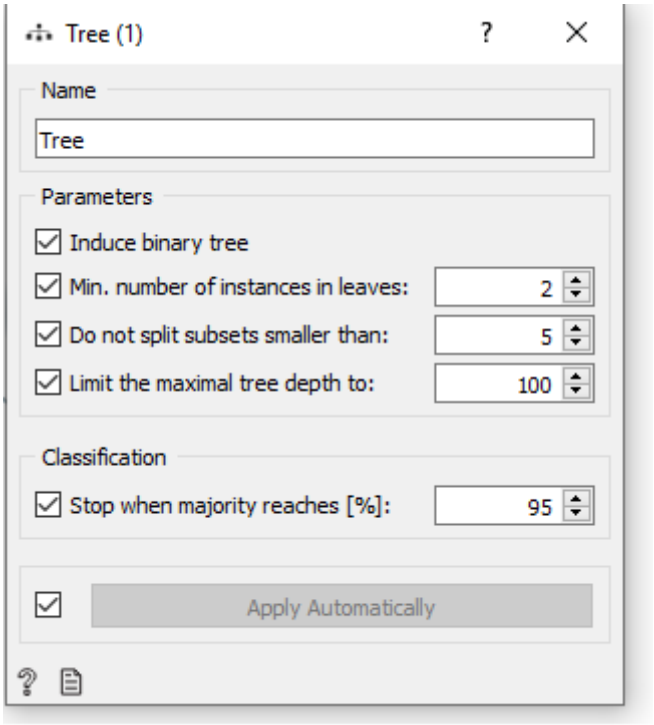


HIPERPARÁMETROS

Todos los modelos que aplicamos en machine learning tienen lo que se denominan como ***hiperparámetros***. En el caso de los árboles de decisión los hiperparámetros que se suelen considerar son:

- **profundidad** del árbol
- **cantidad mínima de observaciones** o instancias en cada hoja
- inlcusive el método que utilizamos para efectuar las divisiones en el árbol (nosotros conocemos sólo el método de la **entropía**, pero hay otros que veremos en una próxima materia)

También son hiperparámetros los otros parámetros que que figuran en el control Tree de Orange3 que hemos utilizado con anterioridad:



> Antes de ***entrenar*** a nuestro ***modelo*** tenemos que seleccionar qué valores de los hiperparámetros vamos a utilizar.

Entonces es cuando nos preguntamos si habrá algunos valores de estos hiperparámetros con los cuales nuestro modelo funcione mejor: el proceso para determinar los mejores valores para los hiperparámetros se suele denominar **TUNNING**.

Para determinar los mejores valores de los hipermatrámetros *podríamos pensar* en:

- seleccionar distintos valores para los hiperparámetros,
- entrenar con el Train Set y luego
- testear frente al Test Set.
- Repetir la operación con varias combinaciones de hiperparámetros
- elegir aquellos con los cuales obtuvimos mejores resultados en el Test Set.

El procedimiento anterior parece razonable, pero tiene un problema:

> Si estamos seleccionando los valores para los hiperparámetros al ver qué resultados obtenemos en el Test Set ... debemos admitir que nuestro modelo está **aprendiendo** del Test Set ... y anteriormente establecimos que el Test Set se utilizará **sólo** para testear nuestro modelo!

Cómo resolvemos este problema?

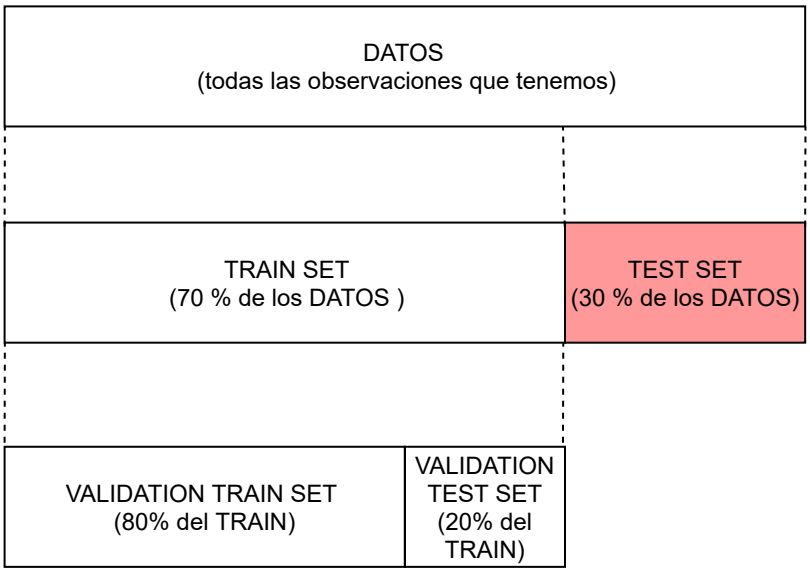
VALIDATION SET

Primero que todo, **dejemos absolutamente de lado al Test Set**, es intocable; cualquier cosa que querramos hacer para determinar el valor de los hiperparámetros deberá ser hecha con observaciones o instancias que no provengan de él.

Entonces sólo nos queda una fuente de datos disponibles ... el Train Set.

Lo dividiremos en dos conjuntos que aquí hemos denominado como

- **VALIDATION TRAIN SET**
- **VALIDATION TEST SET**

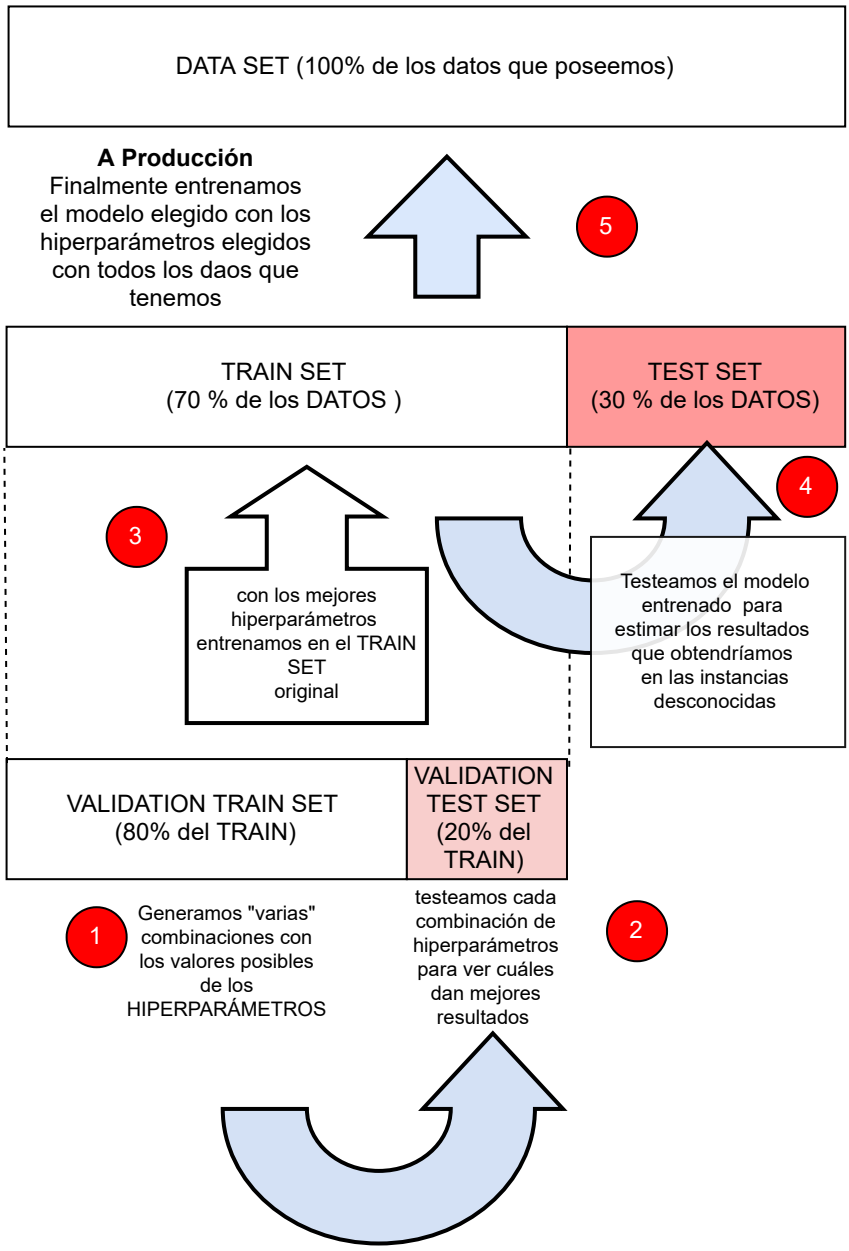


Con estos datos se determinan los Hiperparámetros del modelo.

También sirve para hacer SELECCIÓN DE MODELO
cuando estamos decidiendo qué modelo utilizar.

Los porcentajes propuestos son sólo indicativos

Cómo procederemos?



Cómo procederemos?

- Utilizaremos el **VALIDATION TRAIN SET para entrenar** a nuestro modelo con **distintas combinaciones de valores para los hiperparámetros**. Cuando conozcamos más modelos de machine learning también probaremos con diversos modelos y sus respectivos hiperparámetros para ver cuál nos conviene utilizar.
- a cada una de estas combinaciones la **testearemos frente al VALIDATION TEST SET**
  - calculando por ejemplo Accuracy (Exactitud) y
  - F1
  - otros tests de nuestro interés
- Elegiremos la combinación de valores de hiperparámetros que obtuvo mejores resultados y con ella **armaremos nuestro modelo**.
- Una vez que tenemos el modelo con los mejores valores de los hiperparámetros, lo **entrenaremos con *TODO* el TRAIN SET y lo evaluaremos con el TEST SET**.
- De esta manera al efectuar la selección de hiperparámetros **nuestro modelo no habrá visto al Test Set**, y por lo tanto al evaluarlo frente a él no habrá ningún ***sesgo*** en los resultados que obtengamos de Accuracy, F1 u otras formas de evlaución.
- Finalmente a la hora de poner nuestro modelo el **Producción** lo entrenaremos de nuevo, pero esta vez aprovecharemos **todos** los datos que tenemos en el Data Set original.

Para pensar

Si bien es cierto que siguiendo el procedimiento anterior nuestra estimación del modelo no estará sesgado por el conocimiento de las instancias u observaciones del Test Set , que es lo que pretendíamos, sin embargo suele aducirse con razón, que en algún paso del procedimiento sí se produce un sesgo de aprendizaje. Podría indicar cuál?

NOTA

Dado que durante el proceso de tunning o selección de hiperparámetros habrá que correr varias veces el modelo y luego evaluarlo, si el data set contiene muchas observaciones **esta tarea consumirá muchos recursos computacionales y podría demandar mucho tiempo**.

Antes de continuar, aclaremos que no existe uniformidad en cuanto a las denominaciones de estos conjuntos de datos, así que esté atento a la función que cumple cada conjunto de datos más que a los nombres propiamente dichos.

Resolución de la Situación Profesional

Resuelva con Orange3 el problema planteado en la Situación Profesional con Árbol de Decisión, pero ahora hágalo en forma **completa**:

- Divida el conjunto de datos originales en TRAIN Set y TEST Set
- Divida el TRAIN Set en TRAIN VALIDATION y TEST VALIDATION
- Seleccione el modelo (la profundidad del árbol en este caso) que mejor AC / F1 obtenga.
- Una vez elegido el modelo, entrénelo en el TRAIN SET y calcule en el TEST SET los valores que esperamos se obtengan al generalizar.
- Finalmente cree el modelo que utilizaríamos en Producción entrenando en todo el conjunto de datos original.

EJERCICIO

En clase se elegirá un problema y se decidirá qué hiperparámetros seleccionar para luego aplicar el procedimiento con Orange3.