Ciencia de Datos

- **Módulo 2**

Limpieza de datos

UTNFra

GyEE Gestión y Evaluación Educativa

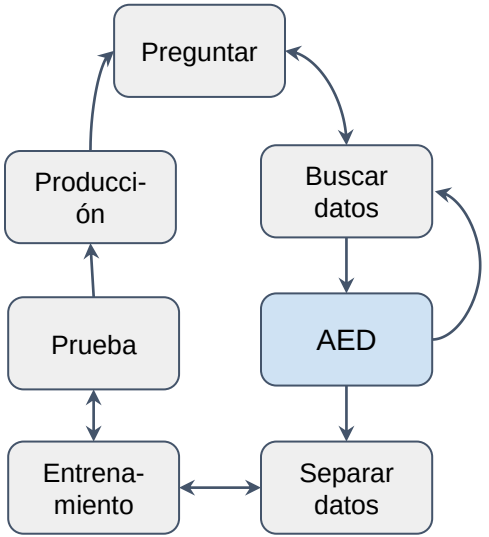
<aprendé programando Virtual>

SECUNDARIA del futuro

Aprendizaje Automático - Pasos

En todo modelo de Machine Learning hay algunos **pasos comunes** que nos sirven de guia:

- **Definición del problema**: Siempre que se va a realizar un proyecto de ML es para resolver un problema específico, por ende, primero debemos de realizarnos una pregunta la cual buscaremos resolver.
- **Buscar los datos**: Por lo general debemos definir qué datos necesitamos y de dónde los podemos encontrar. Hoy en día hay muchos datos libres, sean generados por organismos gubernamentales u ONG, o en algunos casos incluso empresas privadas.
- **Análisis exploratorio de datos**: Durante esta etapa se *limpian* los datos y se realiza una exploración de los mismos, buscando características y relaciones que nos ayuden a seleccionar qué variables usar o que modelo usar.

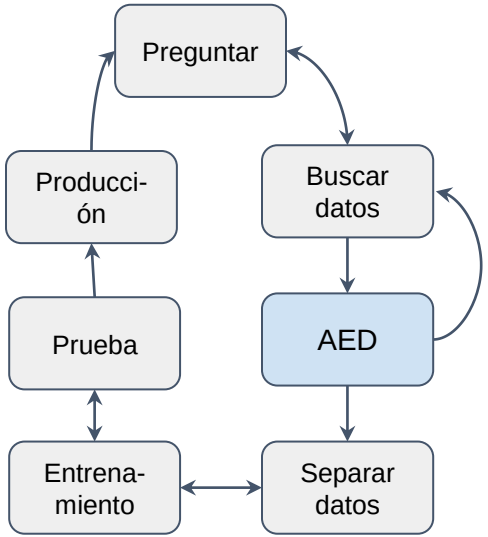


Aprendizaje Automático - Pasos

- **Dividir los datos de entrenamiento y prueba**. Esta división es de vital importancia para poder definir si el modelo tiene un buen rendimiento o no.

- **Entrenamiento del modelo**: En esta etapa se **entrenan** diferentes de modelos y se **prueban** tomando en cuenta alguna métrica previamente definida, seleccionando el modelo que mejor desempeño tenga.

- **Puesta en producción**: Finalmente se realiza la puesta en producción, este paso depende de qué clase de proyecto sea, puede que en algunos casos, colocar en producción solo signifique generar las predicciones para alguna colección de datos, o puede ser implementarlo en un sistema que reciba datos periódicamente.



Calidad de los Datos

Los datos son de calidad cuando representan correctamente la construcción del mundo real a la que se refieren.
Los principales aspectos a considerar son:

- Validez
- Precisión
- Integridad
- Consistencia
- Uniformidad

Validez

La validez de los datos, es el grado en que los datos se ajustan a las reglas o **restricciones**:

- **De tipo de dato:** los valores en una columna en particular deben ser de un tipo de datos.
- **De rango:** normalmente, los números o fechas deben estar dentro de un cierto rango.
- **Obligatorias:** determinadas columnas no pueden estar vacías.
- **Únicas:** un campo, debe ser único en un conjunto de datos.
- **De pertenencia al conjunto:** los valores de una columna provienen de un conjunto de valores discretos. Por ejemplo, el sexo biológico de una persona en general se marca como masculino o femenino.
- **Patrones de expresión regular:** campos de texto que deben seguir un patrón determinado. (Email)
- **Validación de campo cruzado:** deben cumplirse determinadas condiciones que abarcan varios campos. Por ejemplo, la fecha de alta de un paciente del hospital no puede ser anterior a la fecha de admisión.

Precisión

La precisión, es el grado en que los datos se acercan a los valores reales.

- Que un dato sea válido, no significa que sea preciso.
  - El color de ojos válido para una persona puede ser azul, pero no necesariamente es verdadero (no representa la realidad).
- Otra cosa a tener en cuenta es la diferencia entre exactitud y precisión. Decir que vivís en la tierra es exacto, pero no es preciso. ¿Dónde en la Tierra?

Integridad

La integridad de un dataset es el grado en el que se conocen todos los datos necesarios.

Cuando analizamos Data Sets, es muy común encontrar valores faltantes, y generalmente es imposible conseguirlos.

Dado el hecho de que los valores perdidos son inevitables, nos queda la pregunta de qué hacer cuando los encontramos.

Ignorar los datos faltantes es lo mismo que cavar agujeros en un barco; Se hundirá.

A continuación veremos tres formas de lidiar con este problema.

Integridad - Valores faltantes

Algunas estrategias para resolver este problema son:

Descartar

Si los valores perdidos en una columna rara vez ocurren y ocurren al azar, entonces la solución más fácil y directa es eliminar las observaciones (filas) que tienen valores perdidos.

Si faltan la mayoría de los valores de la columna y se producen al azar, una decisión típica es descartar toda la columna.

Imputar

Significa calcular el valor faltante en base a otras observaciones. Algunos métodos para resolverlo:

Por valores estadísticos - media, mediana, etc

Regresión linear - en relación a otras variables

Copiando valores de otras observaciones similares - solo cuando existen mucha cantidad de datos

Marcar

Significa indicar que el dato está faltando. Algunos argumentan que completar los valores faltantes conduce a una pérdida de información, independientemente del método de imputación que usemos. Esto se debe a que decir que faltan datos es informativo en sí mismo y el algoritmo debería saberlo.

Consistencia

La consistencia es el grado en que los datos son consistentes, dentro del mismo conjunto de datos o en varios conjuntos de datos.

La inconsistencia ocurre cuando dos valores en el conjunto de datos se contradicen.

- Una edad válida, digamos 10, podría no coincidir con el estado civil, digamos divorciado.
- Un cliente se registra en dos tablas diferentes con dos direcciones diferentes. ¿Cual es la verdadera?

Uniformidad

La uniformidad es el grado en el que se especifican los datos utilizando la misma unidad de medida.

- El peso se puede registrar en libras o en kilos.
- La fecha puede seguir el formato de Argentina o de EE. UU.

Para poder procesarlos, los datos deben convertirse a una única unidad de medida.

Metodología

El flujo de trabajo es una secuencia de cuatro pasos cuyo objetivo es producir datos de alta calidad y tener en cuenta todos los criterios de los que hemos hablado.

- Inspección: detecta datos inesperados, incorrectos e inconsistentes.
- Limpieza: Corrija o elimine las anomalías descubiertas.
- Verificación: después de la limpieza, los resultados se inspeccionan para verificar que sean correctos.
- Informes: se registra un informe sobre los cambios realizados y la calidad de los datos almacenados actualmente.

Pregunta evaluadoras

Estas preguntas ayudan a evaluar y mejorar la calidad de los datos:

¿Cómo se recopilan los datos y en qué condiciones ?
El entorno donde se recopilaron los datos es importante. El entorno incluye, entre otros, la ubicación, el tiempo, las condiciones climáticas, etc.
No es lo mismo cuestionar a los sujetos sobre su opinión sobre lo que sea mientras están de camino al trabajo que mientras están en casa.
¿Qué representan los datos ?
¿Incluye a todos? ¿Solo la gente de la ciudad ?. O, quizás, solo aquellos que optaron por contestar porque tenían una fuerte opinión sobre el tema.
¿Cuáles son los métodos utilizados para limpiar los datos y por qué ?
Diferentes métodos pueden ser mejores en diferentes situaciones o con diferentes tipos de datos.