09481: Inteligencia Artificial

Profesor del curso: Breyner Posso, Ing. M.Sc.
e-mail: breyner.posso1@u.icesi.edu.co

Programa de Ingeniería de Sistemas.
Departamento TIC.
Facultad de Ingeniería.
Universidad Icesi.
Cali, Colombia.

Agenda

- Conceptos básicos de limpieza de datos (data cleaning).
- Conceptos básicos de transformación de datos.

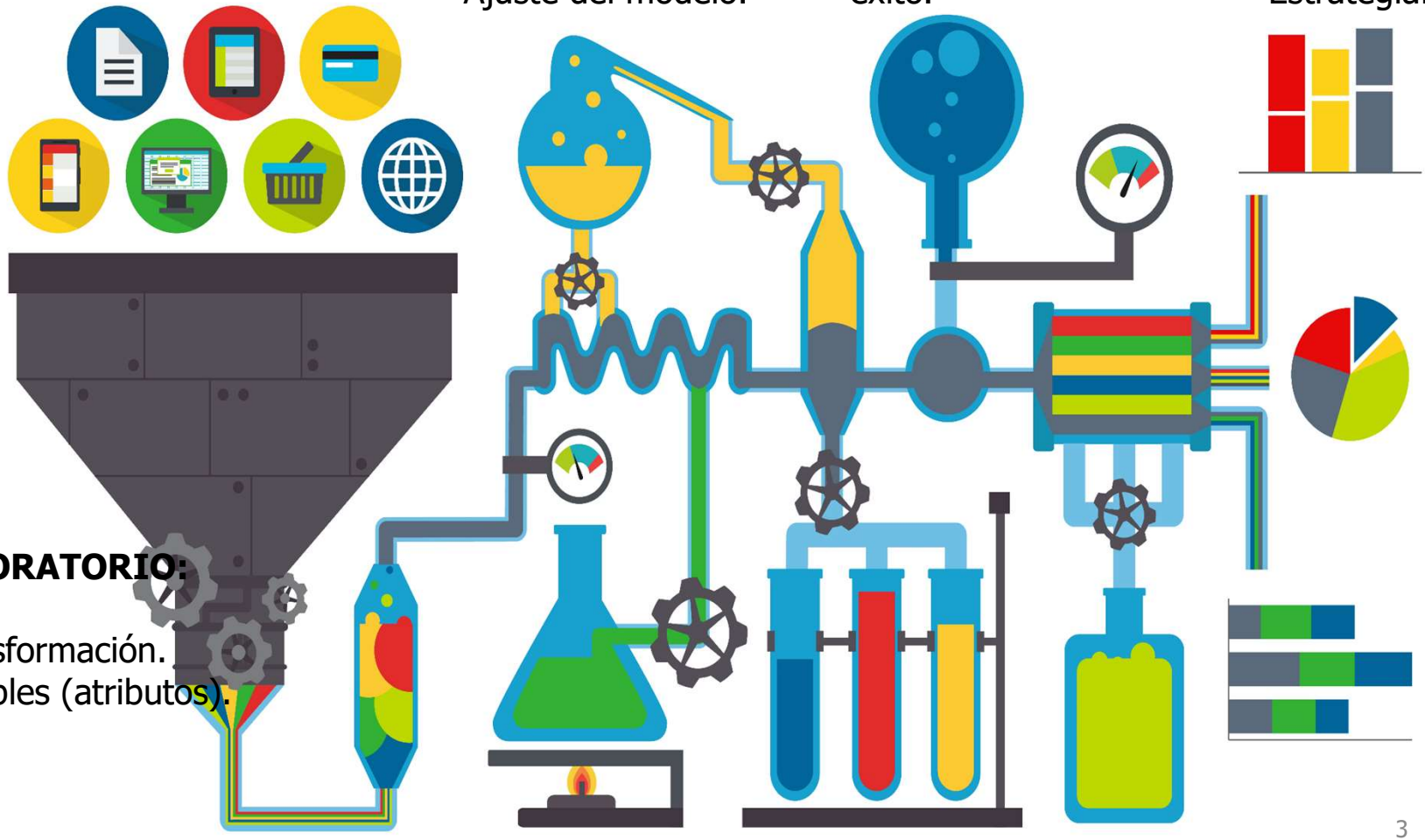
DATOS:
Materia prima.

MODELO:
Implementación del modelo de analítica.
Ajuste del modelo.

EVALUACION:
Viabilidad del negocio.
Validación criterios de éxito.

DESPLIEGUE:
Resultados.
Conocimiento.
Estrategia.

PREGUNTAS:
¿Cómo?
¿Cuáles?
¿Cuándo?
¿Por qué? *



ANALISIS EXPLORATORIO:
Limpieza de datos.
Preparación / transformación.
Selección de variables (atributos).

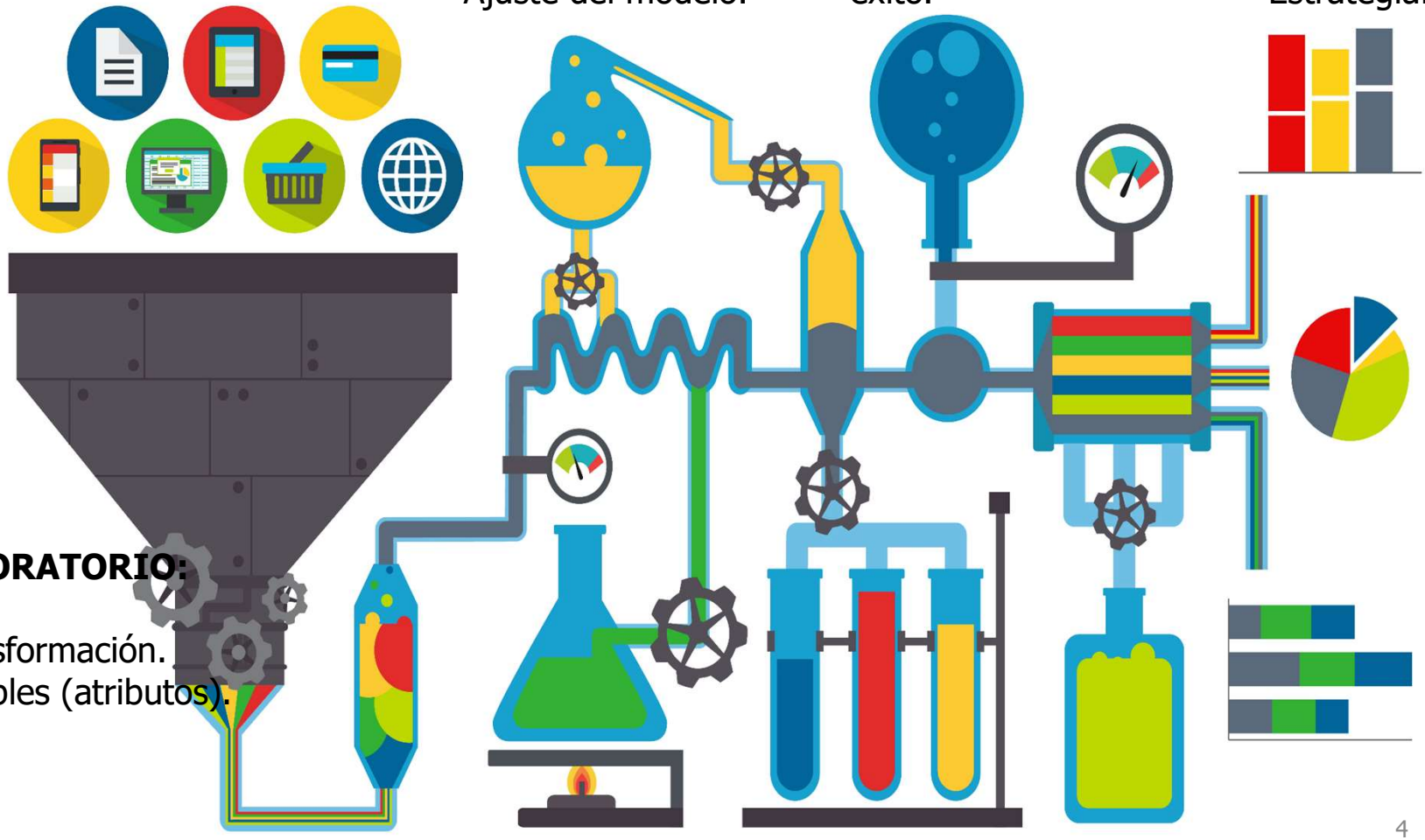
DATOS:
Materia prima.

MODELO:
Implementación del modelo de analítica.
Ajuste del modelo.

EVALUACION:
Viabilidad del negocio.
Validación criterios de éxito.

DESPLIEGUE:
Resultados.
Conocimiento.
Estrategia.

PREGUNTAS:
¿Cómo?
¿Cuáles?
¿Cuándo?
¿Por qué? *



ANALISIS EXPLORATORIO:
Limpieza de datos.
Preparación / transformación.
Selección de variables (atributos).

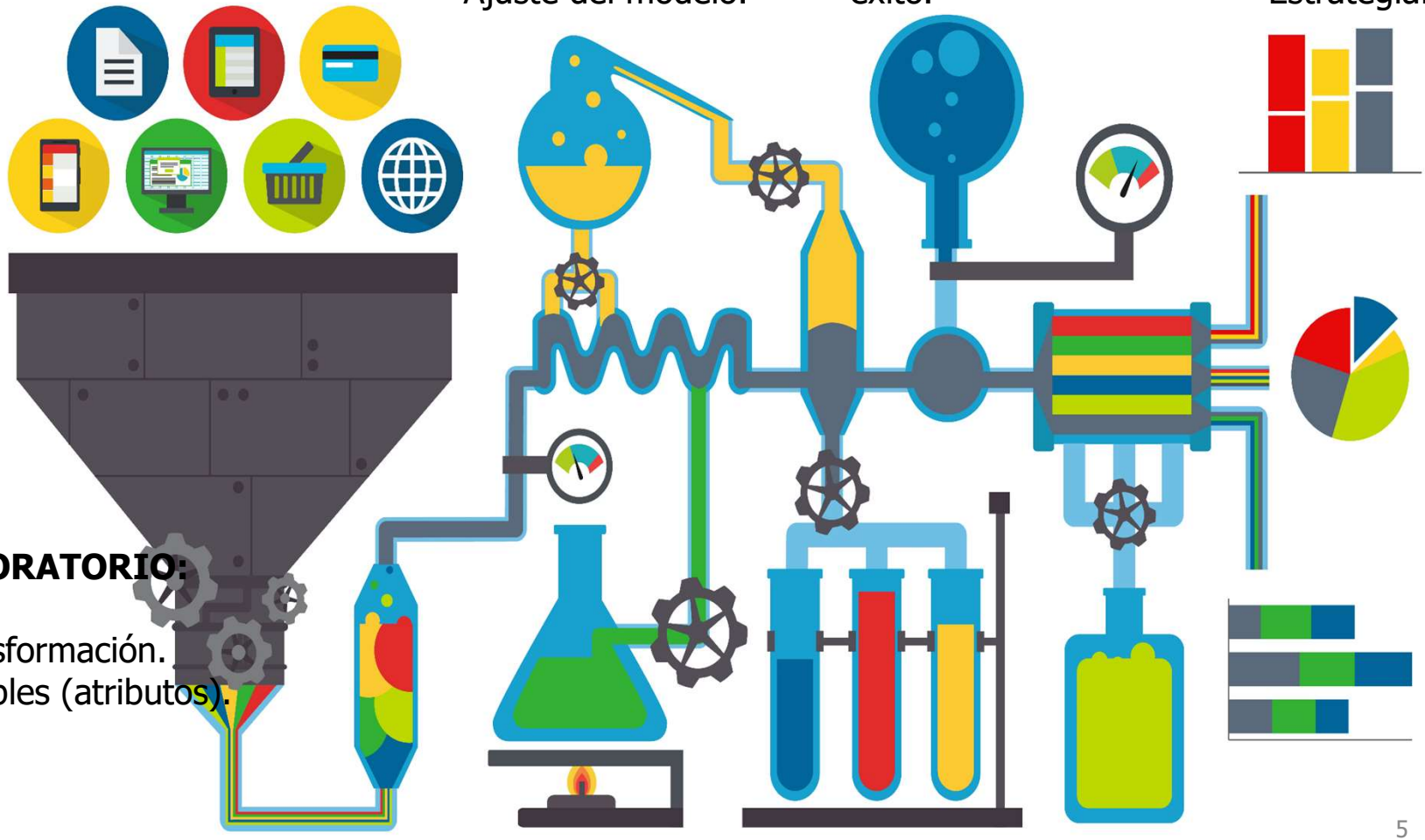
DATOS:
Materia prima.

MODELO:
Implementación del modelo de analítica.
Ajuste del modelo.

EVALUACION:
Viabilidad del negocio.
Validación criterios de éxito.

DESPLIEGUE:
Resultados.
Conocimiento.
Estrategia.

PREGUNTAS:
¿Cómo?
¿Cuáles?
¿Cuándo?
¿Por qué? *



ANALISIS EXPLORATORIO:
Limpieza de datos.
Preparación / transformación.
Selección de variables (atributos).

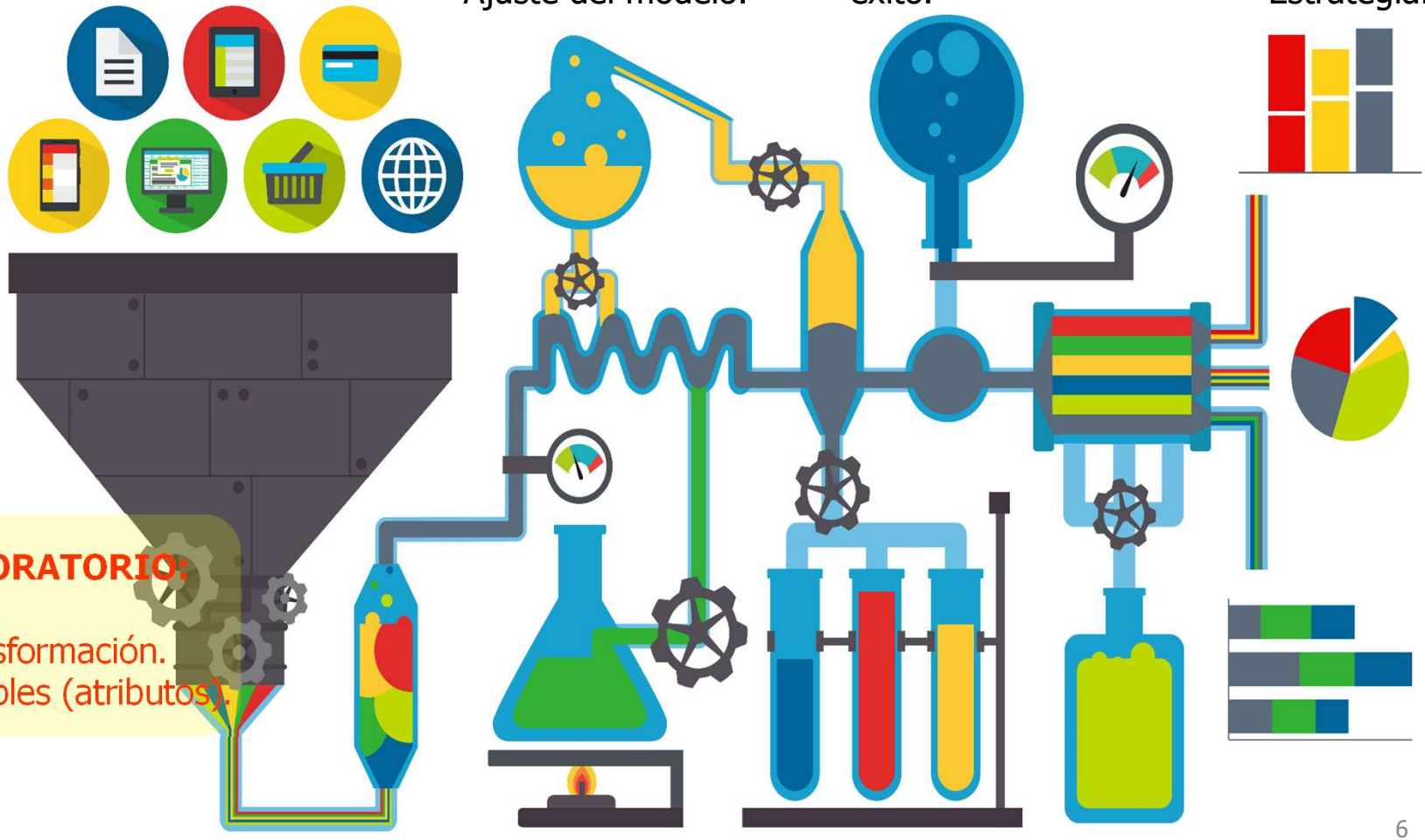
DATOS:
Materia prima.

MODELO:
Implementación del modelo de analítica.
Ajuste del modelo.

EVALUACION:
Viabilidad del negocio.
Validación criterios de éxito.

DESPLIEGUE:
Resultados.
Conocimiento.
Estrategia.

PREGUNTAS:
¿Cómo?
¿Cuáles?
¿Cuándo?
¿Por qué? *



ANALISIS EXPLORATORIO:
Limpieza de datos.
Preparación / transformación.
Selección de variables (atributos).

Conceptos básicos de limpieza de datos (data cleaning)

Limpieza de datos

- La calidad de los análisis y descubrimientos (insights) va a depender de la calidad de los datos.
- La limpieza de datos es el proceso de identificar y arreglar o remover:
  - ❑Datos incompletos.
  - ❑Datos incorrectos o inexactos.
  - ❑Datos corruptos.
  - ❑Datos con formato incorrecto.
  - ❑Datos duplicados.
  - ❑Datos irrelevantes.

Limpieza de datos

- Los datos de baja calidad pueden provocar una gran variedad de problemas. Por ejemplo, una campaña de marketing puede estar mal orientada y, por tanto, fracasar.
- En el ámbito de la sanidad, unos datos deficientes pueden llevar a tratamientos inadecuados e incluso al fracaso en el desarrollo de medicamentos. Un estudio realizado por Accenture revela que la falta de datos limpios es la principal barrera para la adopción de la IA en este campo.
- En el ámbito de la logística, los datos pueden causar problemas de inventario, de planificación de las entregas y, por tanto, de satisfacción del cliente. En el ámbito de la fabricación, las fábricas que configuran los robots con datos erróneos pueden causar graves problemas.

Limpieza de datos

- Los pasos que se lleven cabo a para hacer la limpieza pueden variar dependiendo del dataset y del problema.
- Sin embargo, es muy importante establecer un protocolo para hacer la limpieza de los datos, una opción puede ser la siguiente:
  - ❑Paso 0: Hacer una inspección de los datos.
  - ❑Paso 1: Eliminar duplicados u observaciones irrelevantes.
  - ❑Paso 2: Solucionar errores estructurales.
  - ❑Paso 3: Filtrar outliers indeseados.
  - ❑Paso 4: Manejar datos faltantes.
  - ❑Paso 5: Validar.
  - ❑Paso 6: Reportar.

Limpieza de datos

- Paso 0: Inspeccionar los datos.

- Perfilamiento de los datos usando estadísticas.
- Visualizaciones.
- Software especializado.

Limpieza de datos

❑Paso 1: Eliminar duplicados u observaciones irrelevantes.

- Los duplicados se pueden presentar durante la fase de recolección de los datos o cuando se combinan datos de diferentes fuentes.

- Las observaciones irrelevantes son aquellas que no se ajustan al problema que se está intentando resolver. Por ejemplo, si se está haciendo un análisis de consumo de usuarios millenials, no es necesario incluir generaciones más antiguas.

- También se pueden eliminar atributos que no se vayan a usar en el análisis. Por ejemplo, en un estudio de salud puede no ser necesario el número telefónico del paciente.

Limpieza de datos

❑Paso 1: Eliminar duplicados u observaciones irrelevantes.

- Solo se deben eliminar observaciones o atributos si uno está completamente seguro que no son relevantes para el problema. De lo contrario, puede usar la información entregada por la matriz de correlación entre las variables.
- También, se puede consultar a un experto en el área del problema que se está intentado resolver. Puede suceder que una característica que parecía irrelevante, en realidad lo sea cuando se mira el problema desde una perspectiva diferente.

Limpieza de datos

❑Paso 2: Solucionar errores estructurales.

- Los errores estructurales se presentan cuando hay incongruencias en el contenido de algunos atributos. Por ejemplo, nombres incongruentes debido a errores tipográficos, uso inadecuado de mayúsculas o incumplimiento de algunas convenciones definidas previamente.

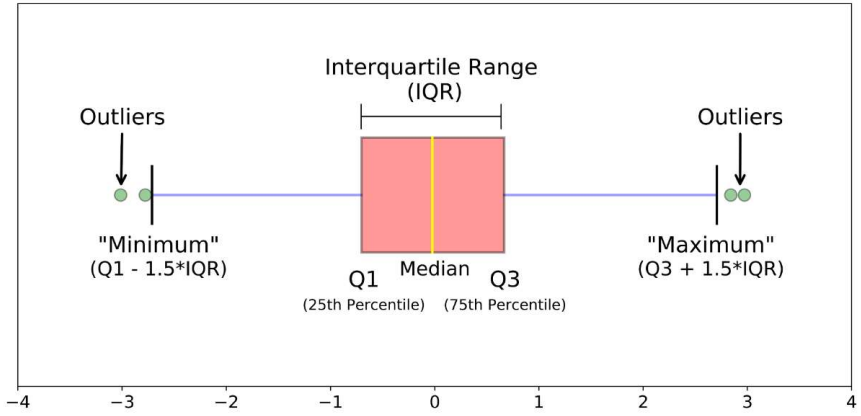
- Los valores se deben almacenar en el tipo de dato adecuado.

- Este tipo de error puede conducir a errores a la hora de procesar la información de ese atributo en particular.

Limpieza de datos

❑Paso 3: Filtrar outliers indeseados.

- Con frecuencia se presentan valores de atributos en alguna de las observaciones que parecen no encajar en los datos que se están analizando.
- Se puede considerar outlier cualquier valor que esté por fuera de Q1-(1.5*IQR) y Q3+(1.5*IQR).



Limpieza de datos

❑Paso 3: Filtrar outliers indeseados.

- Si se tiene una razón legítima para eliminar un valor atípico, como una entrada de datos incorrecta, hacerlo ayudará al rendimiento de los datos con los que está trabajando. Sin embargo, a veces es la aparición de un valor atípico lo que probará una teoría en la que está trabajando.
- Recuerde: el hecho de que exista un valor atípico no significa que sea incorrecto. Este paso es necesario para determinar la validez de ese número. Si un valor atípico resulta ser irrelevante para el análisis o es un error, considere eliminarlo.

Limpieza de datos

❑Paso 4: Manejar datos faltantes.

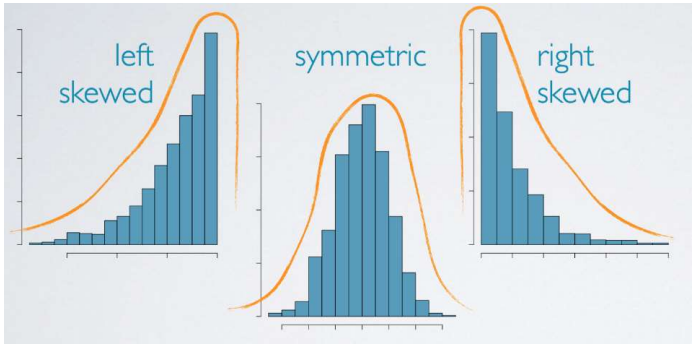
- En algunas ocasiones, faltan valores para algún atributo dentro de una observación y esto no se puede ignorar ya que muchos algoritmos no aceptan valores faltantes.
- Hay un algunas maneras de tratar con los datos que faltan, y la que se utilice va a depender del tamaño del dataset y del tipo de análisis que se esté realizando:

    - Como primera opción, se pueden eliminar las observaciones a las que les faltan valores, pero al hacerlo eliminará o perderá información, así que tenga esto en cuenta antes de hacerlo.

Limpieza de datos

❑Paso 4: Manejar datos faltantes.

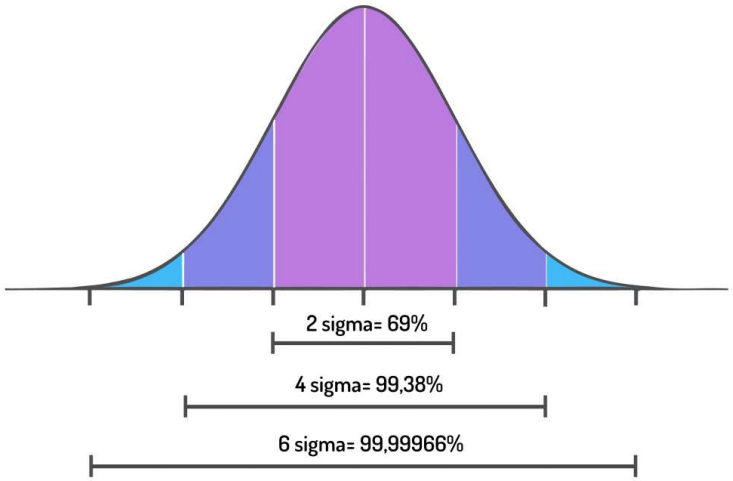
- Como segunda opción, se pueden reemplazar los valores faltantes basados en otras observaciones; nuevamente, existe la posibilidad de perder la integridad de los datos porque se puede estar operando a partir de suposiciones y no de observaciones reales.
    - Se pueden usar medidas estadísticas como la media o la mediana. La media es más usada cuando los datos son simétricos, mientras que la mediana es más robusta cuando los datos no son simétricos o hay presencia de outliers.



Limpieza de datos

❑Paso 4: Manejar datos faltantes.

- Si los datos tienen una distribución normal, se puede extraer la media y la desviación estándar y los datos faltantes se pueden generar de forma aleatoria dentro del intervalo de dos desviaciones estándar a cada lado de la media.



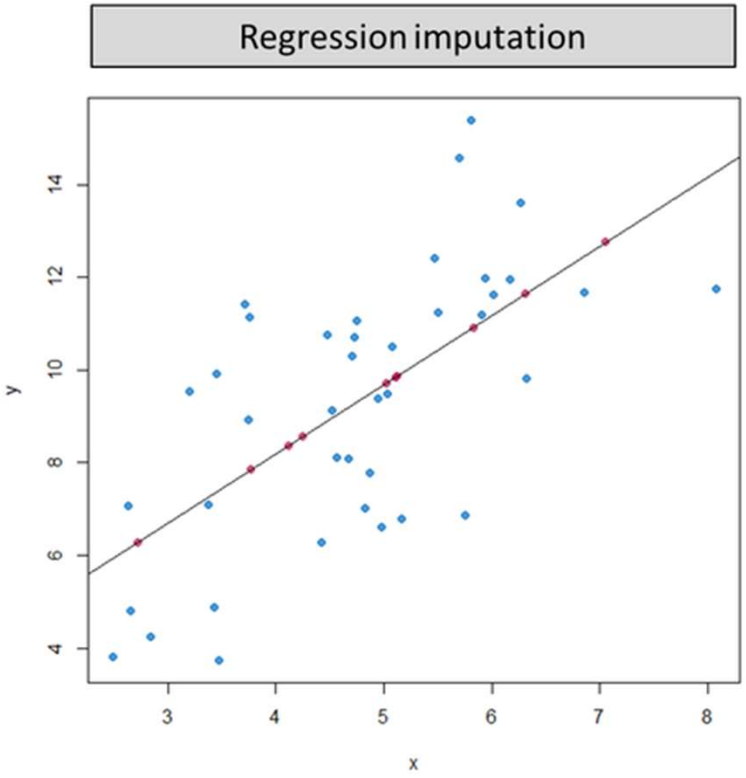

```
rand = np.random.randint(average_age - 2*std_age, average_age +
2*std_age, size = count_nan_age)

dataframe["age"][np.isnan(dataframe["age"])] = rand
```

Limpieza de datos

- Paso 4: Manejar datos faltantes.

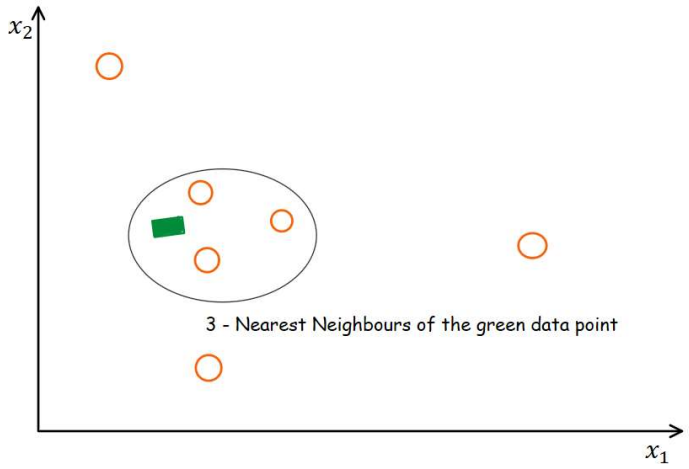
  - También se puede utilizar regresión linear para estimar los valores faltantes. Para poder aplicar esta técnica es necesario identificar que existe una correlación entre el atributo que se quiere completar y otro del que si se tienen los valores. De esta manera, el atributo que se desea completar se usa como variable dependiente y el otro atributo será la variable independiente. Cuando se aplica este método, se debe tener en cuenta que los modelos de regresión lineal son sensibles a los outliers.



Limpieza de datos

❑Paso 4: Manejar datos faltantes.

- Copiar valores de otros registros u observaciones similares. Esto se puede realizar tanto con valores numéricos como categóricos. También se puede usar KNN para encontrar las observaciones más cercanas y de esa manera generar los valores faltantes.



Limpieza de datos

- Paso 4: Manejar datos faltantes.

  - Algunos expertos consideran que rellenar los valores faltantes conduce a una pérdida de información, sin importar el método que se utilice.
  - Hay situaciones en que puede ser más adecuado usar una bandera (flag) en el lugar de los valores faltantes. Por ejemplo, si se trata de valores numéricos rellenar con 0, pero estos 0's deben ser ignorados al momento de generar estadísticas o hacer visualizaciones. Si son valores categóricos se puede crear una nueva categoría "missing", lo cual indica que ese valor falta.

Limpieza de datos

❑Paso 5: Validar.

- Al final del proceso de limpieza de datos vale la pena responder las siguientes preguntas:
  - ¿Tienen sentido los datos?
  - ¿Los datos siguen las reglas apropiadas para su campo?
  - ¿Prueba o refuta su teoría de trabajo, o saca a la luz alguna idea?

Limpieza de datos

- Paso 5: Validar.

  - ¿Puedes encontrar tendencias en los datos que te ayuden a formar tu próxima teoría?

  - Si no es así, ¿se debe a un problema de calidad de los datos?

  - Las conclusiones falsas debido a datos incorrectos pueden llevar a una toma de decisiones deficientes o erróneas. Las conclusiones falsas pueden llevar a un momento vergonzoso en una reunión cuando uno se da cuenta de que sus datos no resisten el escrutinio.

Limpieza de datos

❑Paso 5: Reportar.

- Hacer un reporte de los cambios realizados al dataset original, el estado final del dataset y las herramientas utilizadas.

Conceptos básicos de transformación de datos

Transformación de datos

- La transformación de datos es el proceso de convertir los datos originales a un formato o estructura que pueda ser utilizado para alimentar los modelos de machine learning.
- Las transformaciones van a depender del dataset, del modelo y del problema.
- Algunas de las posibles transformaciones son:
  - Escalado: normalización, estandarización, etc.
  - Simetría: aplicar algunas operaciones para que la distribución de los datos sea más simétrica.
  - Renombrar atributos.
  - Cambiar las unidades de algunas características.
  - Crear nuevas características (feature engineering): crear nuevos atributos a partir de los existentes.

Lecturas recomendadas

- Statistics and Probability - EDA

https://medium.com/omarelgabrys-blog/statistics-probability-exploratory-data-analysis-714f361b43d1#92d4

- Data Cleaning

https://www.kaggle.com/learn/data-cleaning

- Data Transformation and Feature Engineering in Python

https://www.visual-design.net/post/data-transformation-and-feature-engineering-in-python