# 5.**Bias y GIGO en datos**

***Bias*** es un anglicismo que significa **sesgo**. Este fenómeno sucede cuando deseamos estudiar una población y, para ello, partimos de una muestra que no es estadísticamente significativa.

De otro modo, la expresión GIGO significa “**basura entra, basura sale**”. Esta frase se fundamenta en el principio que dice “**la calidad del resultado (output) depende de la calidad de la entrada (input)**”.

Para que los datos representen las verdaderas características de la población estudiada se debe prestar especial atención a estos dos aspectos: **Bias** y **GiGO**.

De igual modo para que los resultados de un análisis de datos sean efectivos y permitan una correcta interpretación de la realidad que pretenden explicar, es muy importante la calidad de estos y que su recolección obedezca a criterios científicos. Por consiguiente, es indispensable que estos **datos realmente recojan las características de la población que se desea estudiar**, es decir, que realmente sean **representativos**.

## ***Bias* o sesgo**

En otras palabras, el **sesgo** se origina cuando se selecciona la información de manera errónea y se considera una muestra que no representa la totalidad de la población. En este caso, **la muestra aleatoria no es representativa**.

### **Ejemplo de estudio**

Supongamos que queremos estudiar el comportamiento de la población de Medellín en determinados aspectos. Para ello, seleccionamos **100 personas**, distribuidas en **80 hombres** y **20 mujeres**.

Ahora bien. según los datos del censo 2018, el Departamento Administrativo Nacional de Estadística -DANE- de Colombia, indica que esta ciudad tiene una población cercana a los 2.5 millones de habitantes, distribuidos en 47% hombres y 53% mujeres, lo que la hace la segunda ciudad más poblada de Colombia.

Surge la siguiente pregunta ¿Si Medellín tiene aproximadamente 2.5 millones de habitantes de los cuales 47% son hombres y 53% son mujeres, piensas que los datos de 100 personas seleccionadas al azar donde 80% son hombres y 20% son mujeres realmente son representativos? Seguramente tu respuesta a esta interrogante es que **no es representativa**.

Por esta razón, debemos asegurarnos de que el número de datos que poseemos sea significativamente grande y correctamente distribuida para sacar conclusiones de cómo se comporta la población en Medellín.

Esta situación ha sucedido en ejemplos prácticos sobre todo en reconocimiento facial. Históricamente las empresas que se han dedicado al reconocimiento facial han tenido más muestras de personas blancas que de personas de piel oscura, por tal motivo se ha creado un sesgo. Eso pone en riesgo la información, porque no podemos identificar a una población importante dentro del mundo.

## **Garbage in Garbage out (GIGO)**

Así pues, debemos asegurarnos de tener calidad en los datos y en la fuente de acceso, cantidad de datos suficientes, tipología de datos, entre otros, para que nuestros resultados sean de calidad.

Si no tenemos una fuente confiable para obtener nuestros datos, estos no van a permitir una conclusión relevante o fidedigna. Por lo tanto, **se precisa recolectar un buen input para tener un buen output**.

## **Reto**

Además de estos dos fenómenos que se han comentado, te invito a que pienses en otros tipos de sesgos y errores que incurren las empresas y el sector público. No olvides compartir tus conclusiones y también escríbelo en tu guía de retos.

## **Resumen de la clase Advertencias de uso de marketing.**

### **Advertencias.**

Se puede crear publicidad personalizada pero debe considerarse lo siguiente en cuanto lo que respecta advertencias… (Es necesario usar la data con responsabilidad y ética)

* Evitar Sesgos discriminatorios ⇒ Como evitar trato distinto dependiendo del tipo de usuarios en caso de no ser realmente necesario.
* Evitar la manipulación de precios ⇒ Suele suceder cuando las empresas conocen datos sobre el poder adquisitivo del cliente.
* Evitar la manipulación de campañas políticas.
* Evitar fomentar conductas violentas, adictivas o ataques emocionales