Aprendizaje automático

Clase 5

Martin Pustilnik, Iris Sattolo,
Maximiliano Beckel

UNAHUR

Preparación de los datos

Introducción

Datos: conjunto de entidades u objetos.

Atributos: características de un objeto.

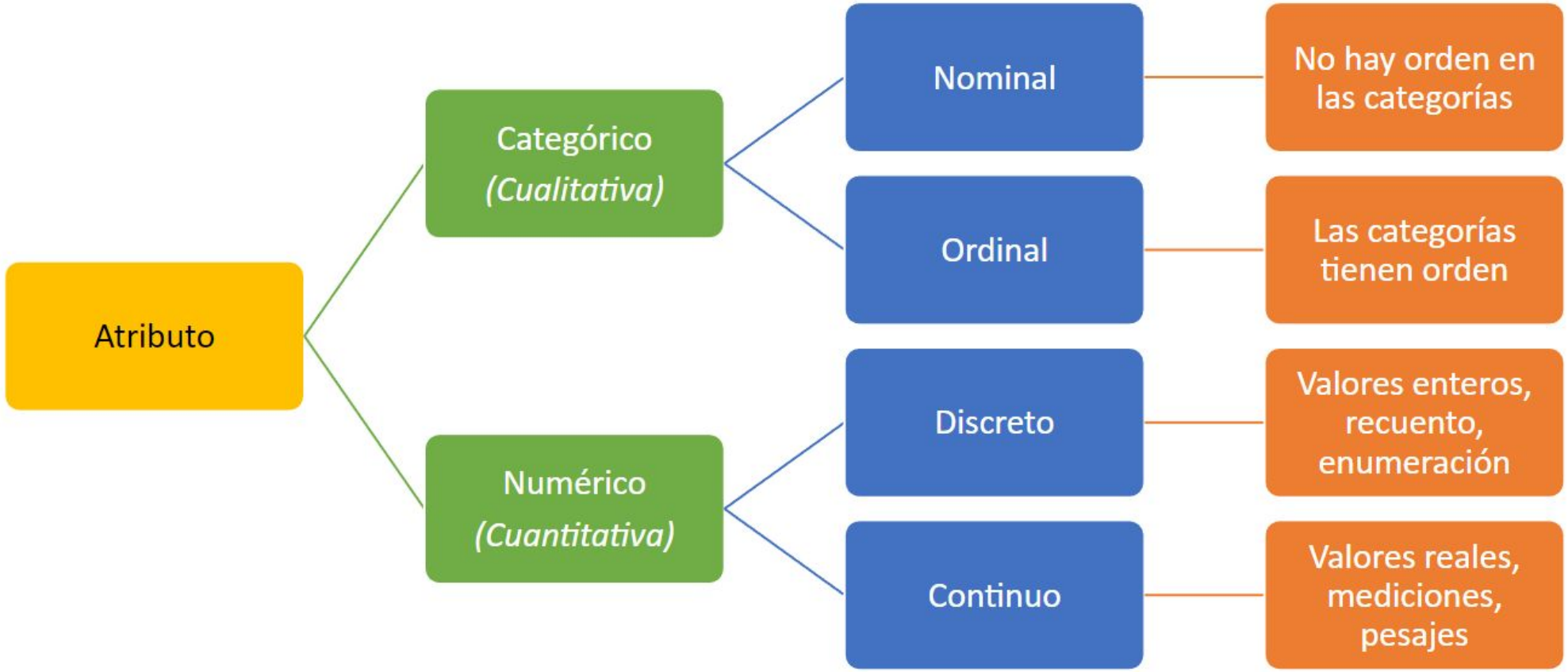
Marca	Modelo	Año	Km	...	Precio
Ford	Focus	2010	84000	...	192000
Chevrolet	S-10	2013	67000	...	409000
Volkswagen	Bora 2.0	2011	50300	...	196500
Ford	Focus ll	2012	75000	...	225000
Peugeot	308	2012	82000	...	270000
Audi	A3	2001	190000	...	135000
...	...	...	...		...

Objetos (o instancias, ejemplos, registros, filas)

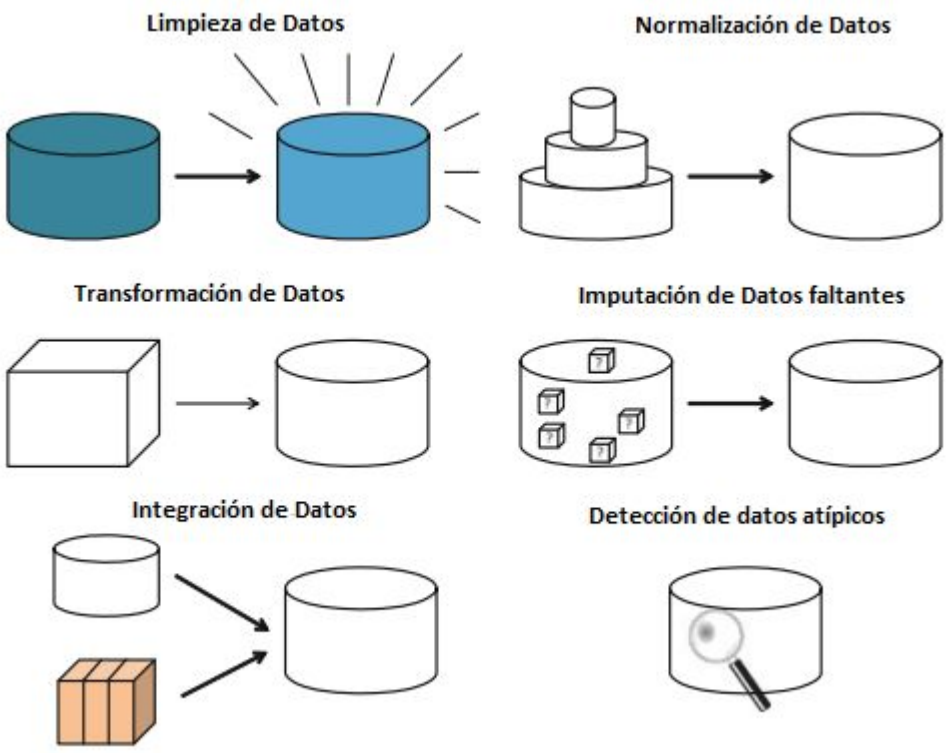
Atributos (o características, columnas, campos)

Introducción

Tipos de atributos (o variables)



Preprocesamiento de Datos

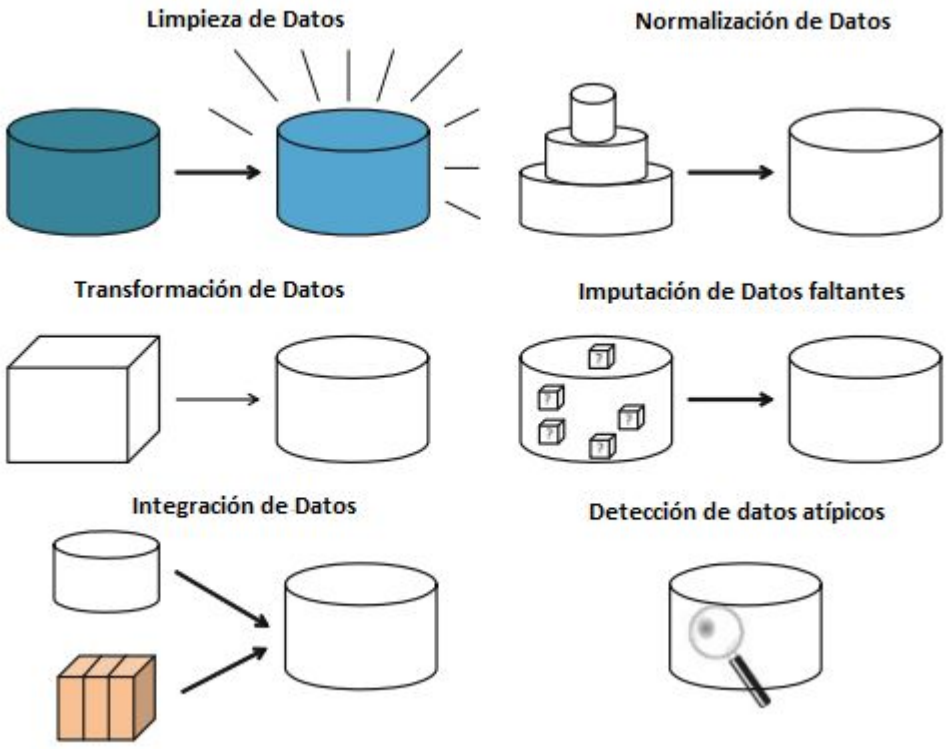


El preprocesamiento de datos implica una serie de pasos necesarios para poder extraer información de ellos y construir modelos de AA que sean de utilidad.

Algunas de las tareas más comunes son:

- Normalización de Datos
- Transformación de Datos
- Imputación de Datos faltantes
- Detección de datos atípicos
- Integración de datos
- Reducción de la dimensionalidad

Preprocesamiento de Datos



El preprocesamiento de datos implica una serie de pasos necesarios para poder extraer información de ellos y construir modelos de AA que sean de utilidad.

Algunas de las tareas más comunes son:

- Normalización de Datos
- Transformación de Datos
- Imputación de Datos faltantes
- Detección de datos atípicos

→ PED

- Integración de datos
- Reducción de la dimensionalidad

Limpieza de datos

Los datos en la vida “real” suelen traer consigo una serie de problemas y errores que pueden deberse tanto a la forma en la que fueron tomados (problemas técnicos) o errores humanos.

- **Datos incompletos**

  ```
  ocupacion = []
  ```

- **Ruidosos**

  ```
  edad = 250; salario = -100; Fecha de hoy: 07/04/1987
  ```

- **Inconsistentes**

  ```
  mes = [‘Septiembre’, ‘Setiembre’, 9, ‘Sep’]
  ```

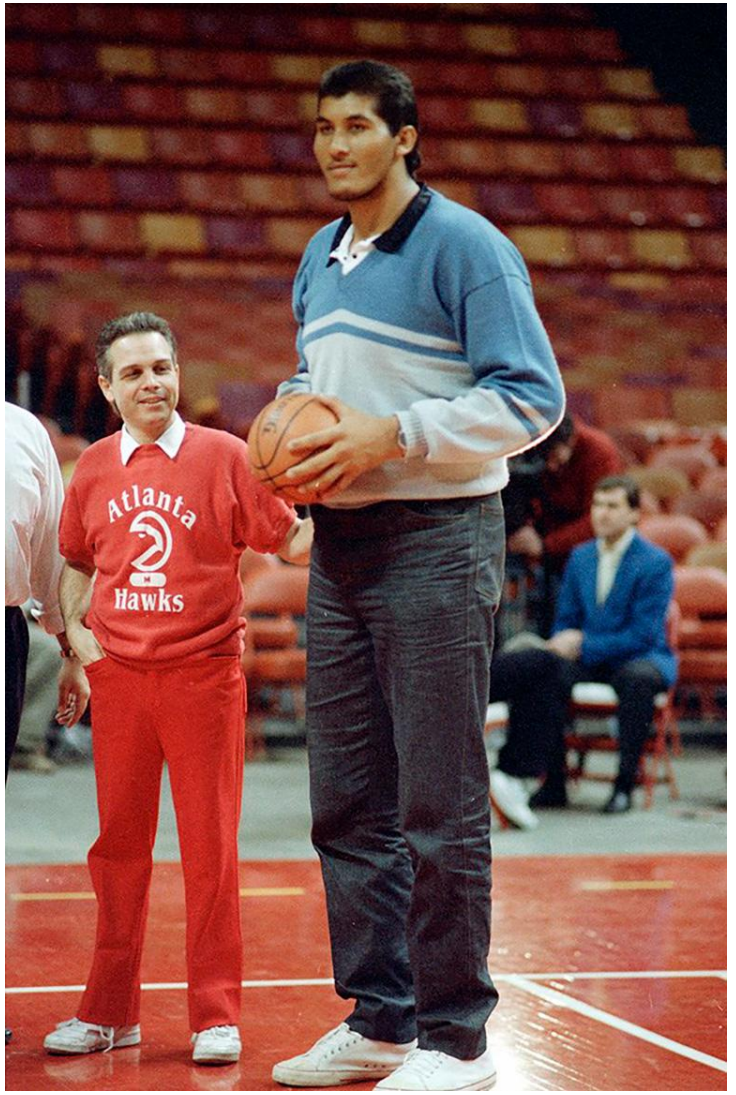
- **Errores sistemáticos/intencionales**

  ```
  fecha de nacimiento = 01/01/1900
  estado civil = ‘feliz’
  ```

Limpieza de datos

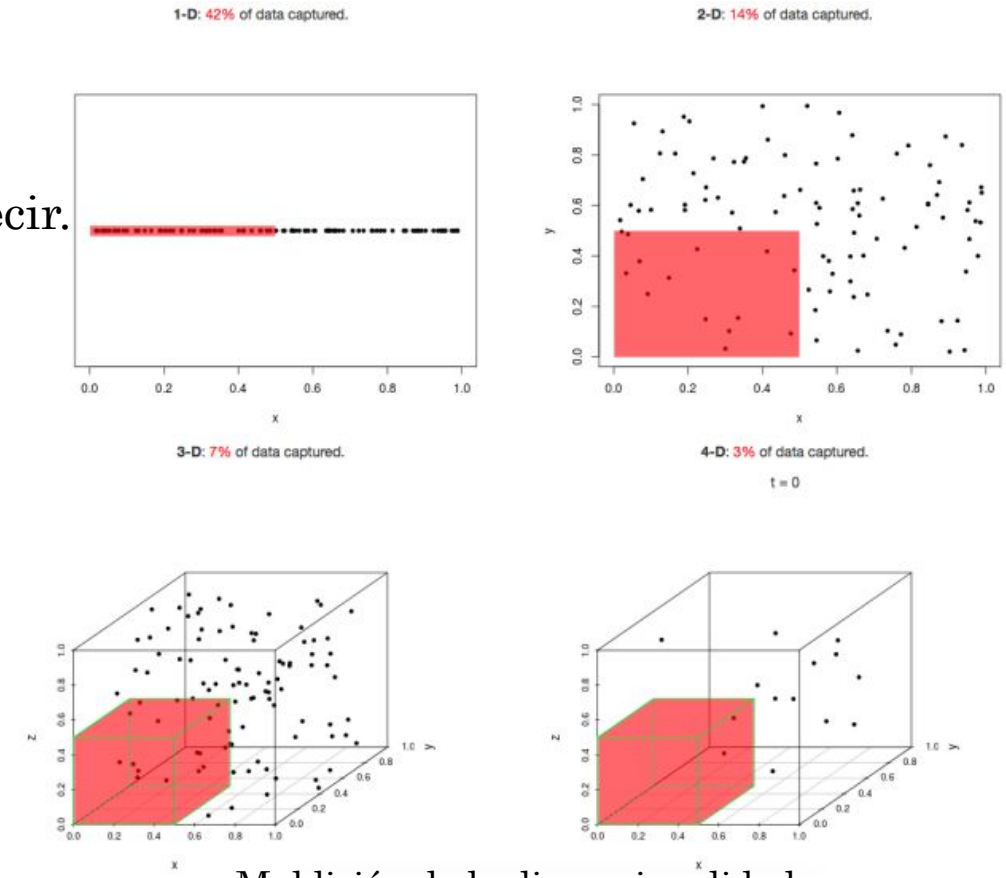
- **<u>Datos faltantes:</u>** no siempre vamos a contar con datos en todos los atributos de nuestros datos. La ausencia de un dato puede deberse al azar o estar vinculada con alguna otra variable o hecho que no conocemos.
  - Estrategias:
    - Eliminar el dato/atributo
    - Imputar el dato faltante (intentar estimar cuál es el valor que falta a partir del resto de los datos que tengo)
    - No hacer nada (algunos modelos admiten datos faltantes)

- **<u>Datos atípicos:</u>** pueden deberse a error en la toma de los datos o en su procesamiento. El desafío es poder distinguir los casos en los que un dato atípico representa verdaderamente un error y los que no.
  - Estrategias:
    - A los fines prácticos, se pueden tratar de la misma manera que a los datos faltantes.



Limpieza de datos

- **<u>Reducción de la dimensionalidad;</u>** muchas veces en nuestras bases de datos tenemos mucha redundancia que afectan negativamente la performance de los modelos que queramos hacer.
  - Presencia de registros/filas repetidas.
  - Atributos altamente correlacionados.
  - Atributos con muy poca varianza.
  - Atributos poco relacionados con la variable a predecir.



Maldición de la dimensionalidad

Limpieza de datos

- **<u>Transformación de datos:</u>** muchas veces en nuestros modelos utilizamos atributos que son de distinta naturaleza (ej: variables categóricas y numéricas) o que varían en escalas muy distintas (ej: edad y altura de las personas). En estos casos, nos gustaría que nuestros atributos se "parezcan más entre sí".
  - Estandarización/Normalización:
    - Nos permite llevar a todas nuestras variables numéricas a la misma escala de variación, haciendo que sean más comparables entre sí.
    - Ayuda a evitar que atributos con mayores magnitudes tengan a su vez un mayor peso en los modelos que el resto de los atributos.
    - Métodos más usados: Min-Max, Z-Score y Decimal Scaling.

Limpieza de datos

- **<u>Transformación de datos:</u>** muchas veces en nuestros modelos utilizamos atributos que son de distinta naturaleza (ej: variables categóricas y numéricas) o que varían en escalas muy distintas (ej: edad y altura de las personas). En estos casos, nos gustaría que nuestros atributos se "parezcan más entre sí".
  - Estandarización/Normalización:
    - Min-Max:

$$X^*_{\text{mm}} = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

- Valores normalizados van de 0 a 1.
- Dominada por los valores atípicos

```
Sepal.Length

Min.    4.300
1st Qu. 5.100
Median  5.800
Mean    5.843
3rd Qu. 6.400
Max.    7.900
```

$$X_{mm} = \frac{X - 4.3}{7.9 - 4.3}$$

Para los valores extremos es 0 y 1

Limpieza de datos

- **<u>Transformación de datos:</u>** muchas veces en nuestros modelos utilizamos atributos que son de distinta naturaleza (ej: variables categóricas y numéricas) o que varían en escalas muy distintas (ej: edad y altura de las personas). En estos casos, nos gustaría que nuestros atributos se "parezcan más entre sí".
  - Estandarización/Normalización:
    - Z-score:

$$Z-score = \frac{X-mean(X)}{sd(X)}$$

Útil cuando…

- ❏ el verdadero mínimo y máximo son desconocidos
- ❏ hay valores atípicos que dominan la normalización min-max. Acá se puede usar la mediana

```
Sepal.Length

Min.    4.300
1st Qu. 5.100
Median  5.800
Mean    5.843
3rd Qu. 6.400
Max.    7.900
```

$$Z-score = \frac{4.3-5.843}{0.828} = -1,863$$

$$Z-score = \frac{7.9-5.843}{0.828} = 2,484$$

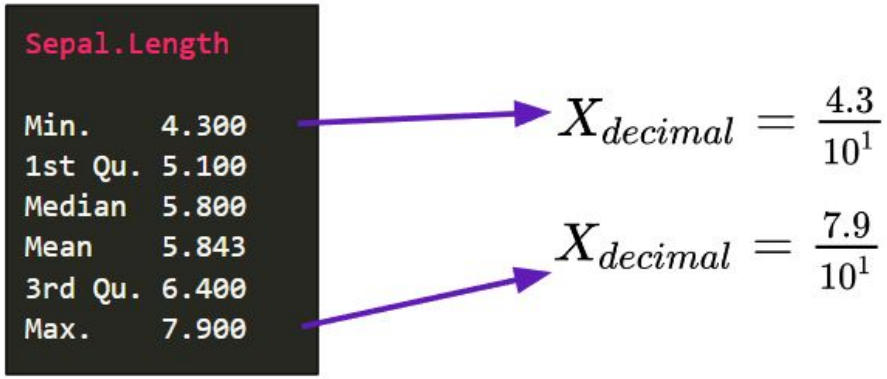
Limpieza de datos

- **<u>Transformación de datos:</u>** muchas veces en nuestros modelos utilizamos atributos que son de distinta naturaleza (ej: variables categóricas y numéricas) o que varían en escalas muy distintas (ej: edad y altura de las personas). En estos casos, nos gustaría que nuestros atributos se "parezcan más entre sí".
  - Estandarización/Normalización:
    - Decimal Scaling:

Decimal Scaling asegura que cada valor normalizado se encuentra entre - 1 y 1.

$$X_{decimal} = \frac{X}{10^d}$$

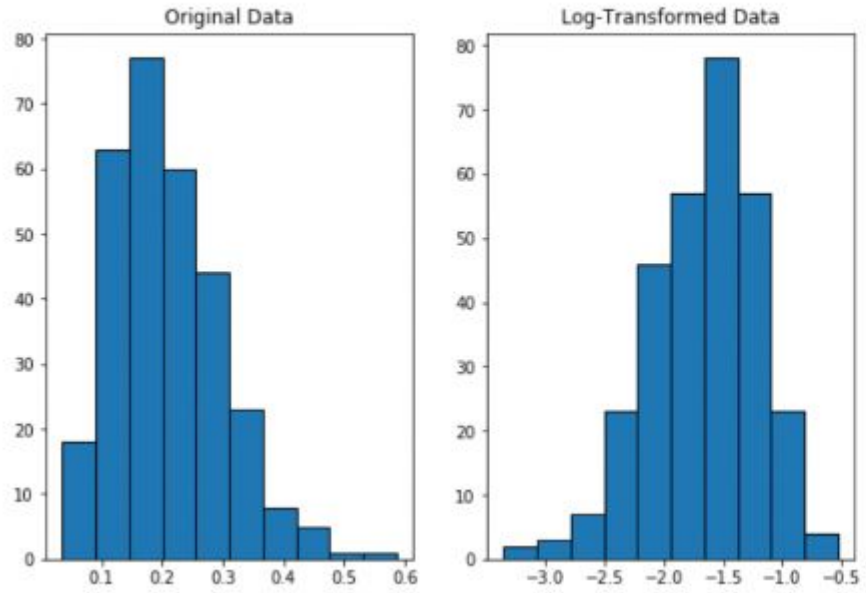
donde **d** es el número de dígitos en los valores de la variable con el valor absoluto más grande.



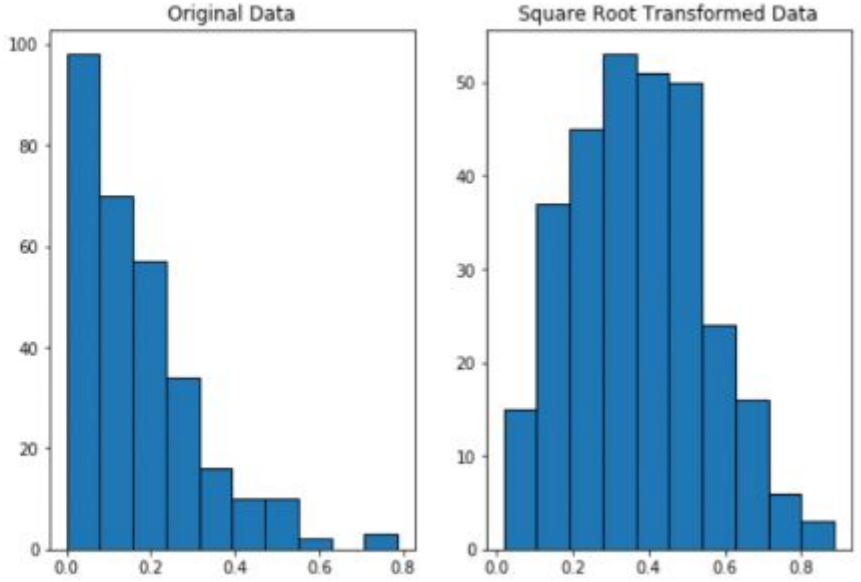
Limpieza de datos

- **<u>Transformación de datos:</u>** muchas veces en nuestros modelos utilizamos atributos que son de distinta naturaleza (ej: variables categóricas y numéricas) o que varían en escalas muy distintas (ej: edad y altura de las personas). En estos casos, nos gustaría que nuestros atributos se "parezcan más entre sí".
  - <u>Transformaciones en la distribución de los datos:</u>

Logaritmo (base 10 o natural)

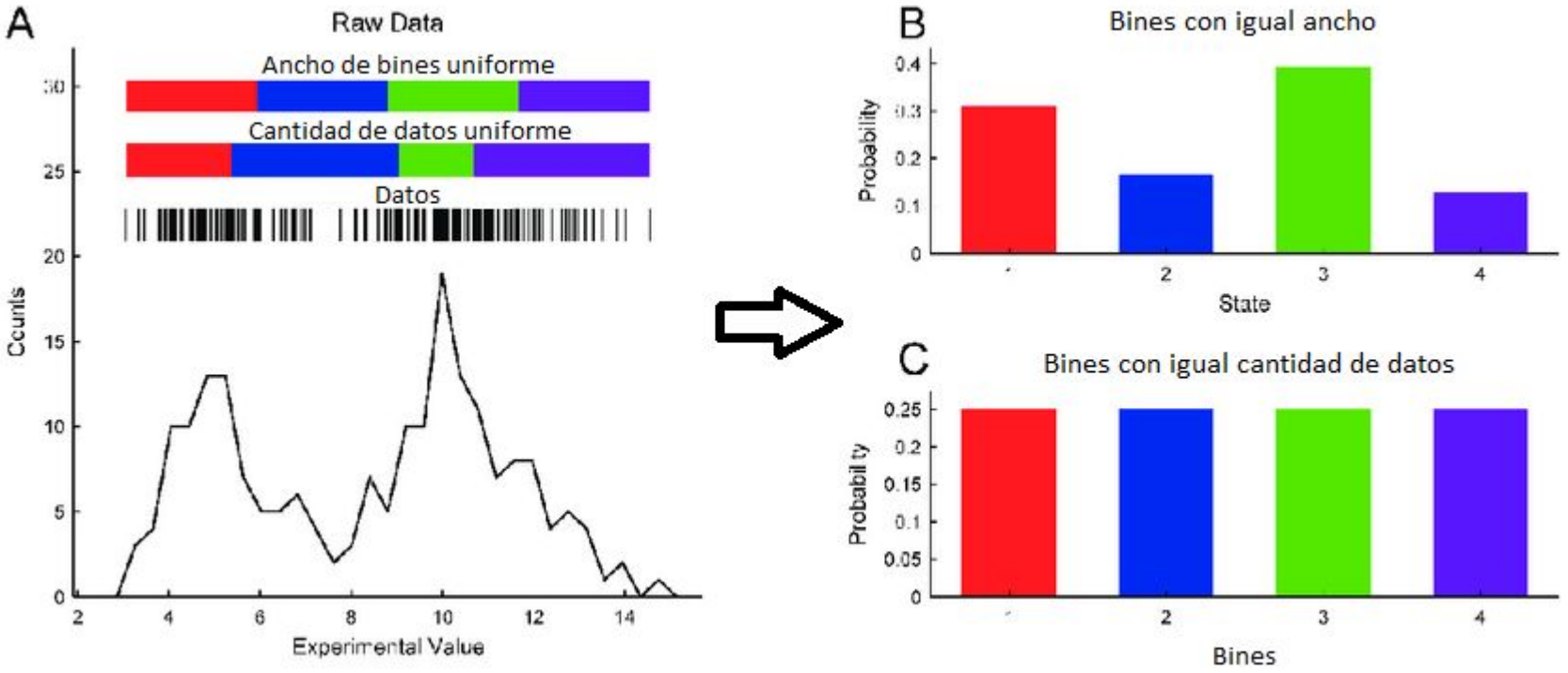


Raíz cuadrada



Limpieza de datos

- **<u>Transformación de datos:</u>** muchas veces en nuestros modelos utilizamos atributos que son de distinta naturaleza (ej: variables categóricas y numéricas) o que varían en escalas muy distintas (ej: edad y altura de las personas). En estos casos, nos gustaría que nuestros atributos se "parezcan más entre sí".
  - <u>Discretización de variables numéricas continuas:</u>



Limpieza de datos

- **Transformación de datos:** muchas veces en nuestros modelos utilizamos atributos que son de distinta naturaleza (ej: variables categóricas y numéricas) o que varían en escalas muy distintas (ej: edad y altura de las personas). En estos casos, nos gustaría que nuestros atributos se "parezcan más entre sí".
  - Codificación de variables categóricas:

Variable

Pet
Cat
Dog
Turtle
Fish
Cat

Variables Dummies

Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0