

clusterAI

ciencia de datos en ingeniería industrial

UTN BA

curso I5521

**clase_07: Ingeniería de
Atributos**

Agenda clase07: Ingeniería de atributos

- Que es y para que sirve la ingeniería de atributos?
- Normalización de variables
- Transformación de variables
- Discretización de variables
- Imputación de valores faltantes
- Codificación de variables categóricas
- Creación de nuevas variables
- Incorporación de información externa

Ingeniería de Atributos

Feature Engineering

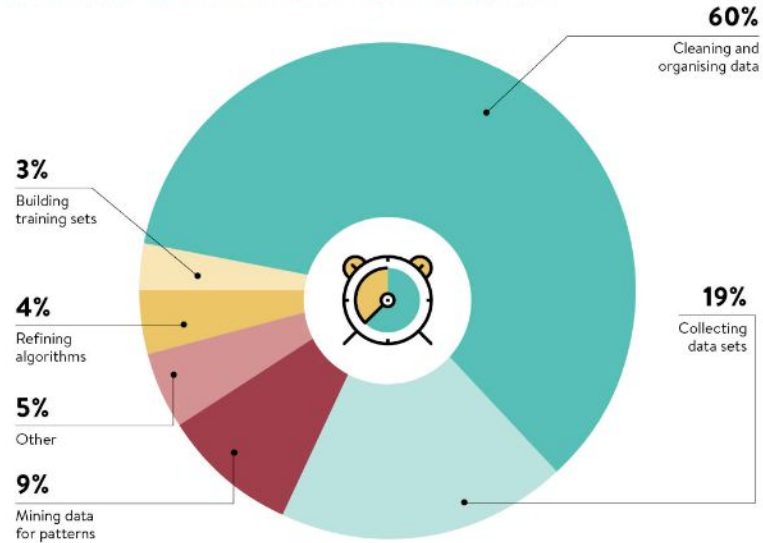
Se refiere al proceso de seleccionar, transformar y crear variables (características o atributos) a partir de los datos crudos para mejorar el rendimiento y precisión de los modelos de aprendizaje automático.

Puede incluir:

- Normalización de variables
- Transformación de variables
- Discretización de variables
- Manipulación de valores faltantes
- Tratamiento de valores atípicos
- Codificación de variables categóricas
- Creación de nuevas variables
- Incorporación de información externa

Ingenieria de Atributos

WHAT DATA SCIENTISTS SPEND THE MOST TIME DOING




Normalizacion Min-Max

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

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

- ❑ Valores normalizados van de 0 a 1.
- ❑ No modifica la distribución de los datos.
- ❑ Amplifica la magnitud de los outliers debido a que baja los desvíos standard


Normalizacion Z-Score (estandarización)


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

- ❑ Escala de los valores a partir de la media (μ) y el desvio (σ) de la distribución.
- ❑ Suaviza el efecto de los outliers.

Sepal.Length

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


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


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

Transformación de variables

La transformación es el reemplazo de la variable por una función de esa variable:

- Utilizar la raíz cuadrada
- Utilizar el logaritmo

En un enfoque más riguroso, una transformación implica un cambio en la distribución original, reemplazandola por una nueva forma.

Se emplea para mitigar sesgos en los datos o lograr dispersiones similares entre las variables.

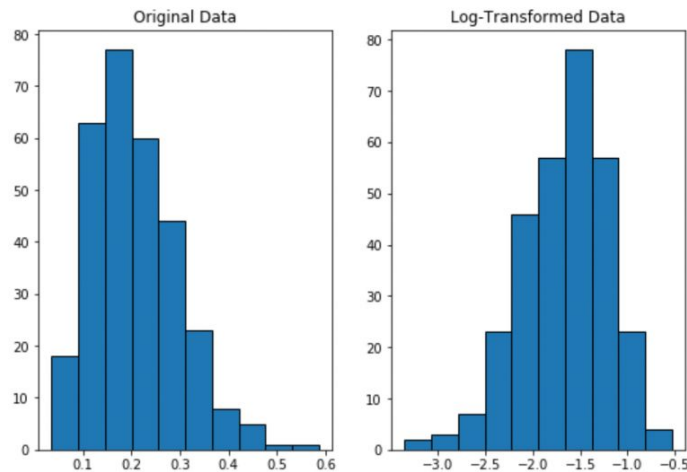
Transformación Logarítmica

Es una de las transformaciones más comunes para variable numéricas. Sus propiedades más importantes son:

- Aproxima a una normal una distribución sesgada
- Ayuda a estabilizar la varianza
- Si existe una relación no lineal entre las variables, la transformación logarítmica puede convertir esta relación en una relación lineal

Limitaciones:

- Solo se puede aplicar a valores positivos
- Su interpretación difiere de la interpretación de los valores originales



Discretizacion de variables

La estrategia consiste en agrupar los registros en bins según ciertos criterios, lo que resulta en una pérdida de variabilidad en las variables, pero a cambio, se logra una mayor robustez en el modelo general.

Numéricas

- Deciles: se agrupan los datos en deciles y valores pertenecientes al mismo decil tomarán como valor el número del decil
- Ranqueo de la variable

Categóricas

- Baja frecuencia: aquellas categorías con baja frecuencia se las puede agrupar en una nueva categoría llamada "Otra"
- Según la variable: se puede buscar una agrupación más general de acuerdo a la variable.
Ej:
 - Provincias -> Países
 - Modelo de autos -> Marca de Autos

Manipulación de valores faltantes

El problema de los datos faltantes está presente en el análisis de datos desde los orígenes del almacenamiento. Es importante saber porqué se generan los valores faltantes para saber cómo manipular los.

Imputación de datos

- Se utilizan distintos criterios y técnicas para imputar los nulos.

Borrar casos seleccionados o variables

- Puede utilizarse cuando hay un patrón no aleatorio de datos faltantes.
- Si al eliminar un subconjunto de datos disminuye la utilidad de los datos, la eliminación del caso puede ser no efectiva.

Tratamientos de valores atípicos

Capping

- Se evalúa un determinado percentil (Ej: 99) y para valores por arriba o debajo de este se les asigna el valor del percentil definido.

Eliminación de registros

- Se opta por eliminar los registros por mayores o menores a determinado valor.

Codificación de variables categóricas

One-Hot Encoded

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29



One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

Codificación de variables categóricas

Label Encoded

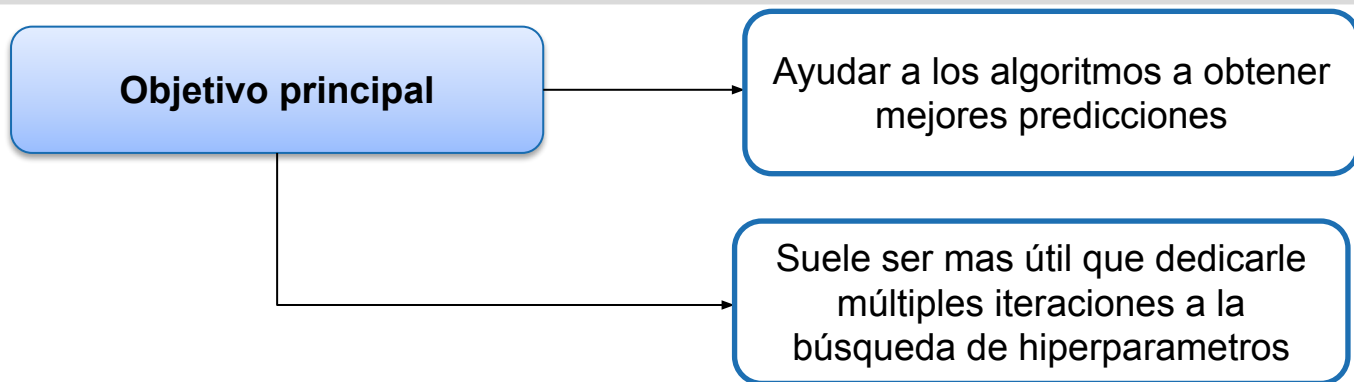
Crea una sola variable con un valor para cada una de las distintas categorías

Height
Tall
Medium
Short



Height
0
1
2

Creación de nuevas variables



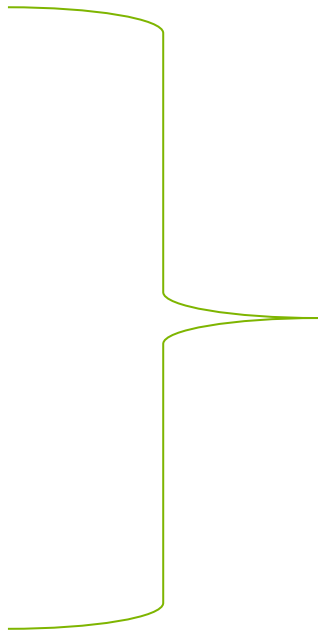
Variables a crear:

- Medias / Medianas por grupo
- Información relevante sobre fechas (Si es feriado, Navidad, día de la semana, diferencia de fechas, etc.)
- Búsqueda de keywords en variables de texto
- Interacción entre variables
- Incorporación de información externa que creamos relevante para el problema de negocio
- Con información geográfica se pueden crear variables en función a puntos de interés.

Agrupación de variables

Se pueden crear nuevas variables mediante la agrupación resumiendo la información de las observaciones y agrupándolas según el criterio de otra variable

Empleado	Puesto	Salario
1	Junior	\$45.000
2	Senior	\$60.000
3	Manager	\$105.000
4	Manager	\$125.000
5	Senior	\$75.000
6	Junior	\$35.000
7	Senior	\$55.000
8	Junior	\$40.000



Puesto	Media Salario
Junior	\$45.000
Senior	\$63.333
Manager	\$115.000

Interaccion entre variables

Se pueden crear nuevas variables a partir de la interacción de variables originales del dataset:

- Se relacionan 2 o mas variables de manera no aditiva

Sirve para:

- Mejorar la precisión del modelo
- Identificación de efectos no lineales

Ejemplos:

- $x_3 = x_1^2 \cdot x_2$