Clase 1

Inteligencia Artificial

Es la inteligencia llevada a cabo por máquinas. Tiene 2 ramas

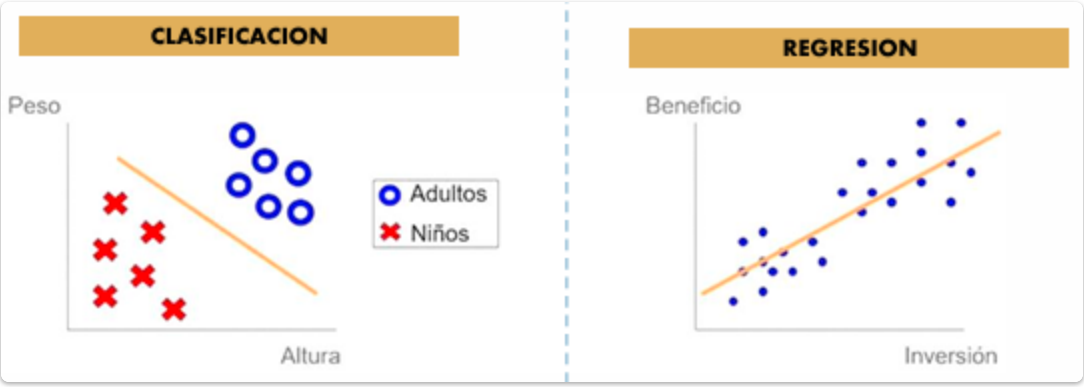
- *Deductiva (Lógica):* Sistemas Expertos
- *Inductiva (Ejemplos):* Redes Neuronales y Técnicas de Optimización.
    - En esta está el Aprendizaje Automático.

El Aprendizaje Automático es el subconjunto de técnicas que brindan a las computadoras la capacidad de aprender sin haber sido programadas explícitamente.

Tipos de Aprendizaje

Aprendizaje Supervisado

Según si la respuesta a predecir es discreta o continua se trata de un problema de clasificación o de regresión respectivamente.



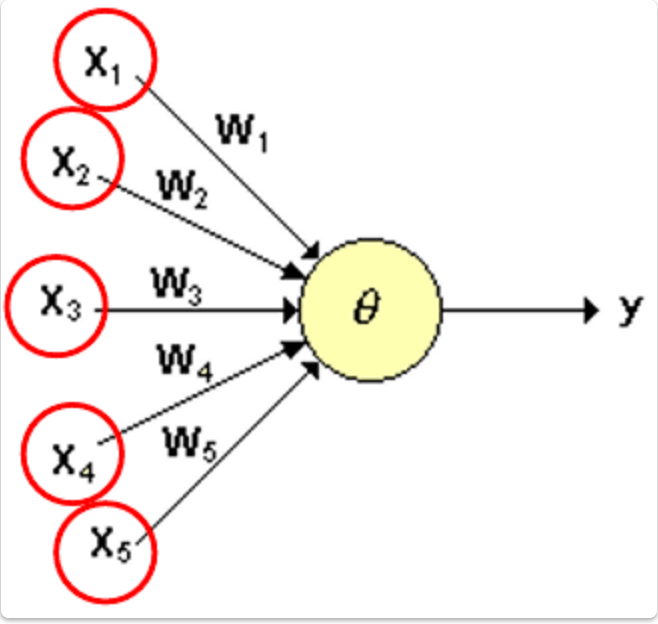
Con este Aprendizaje podemos dar una predicción de un resultado a futuro a partir de los daos disponibles.

Aprendizaje No Supervisado

Se centra más en el Agrupamiento según los datos propuestos. Podemos brindar con este Aprendizaje una segmentación de los datos en subgrupos con características similares.

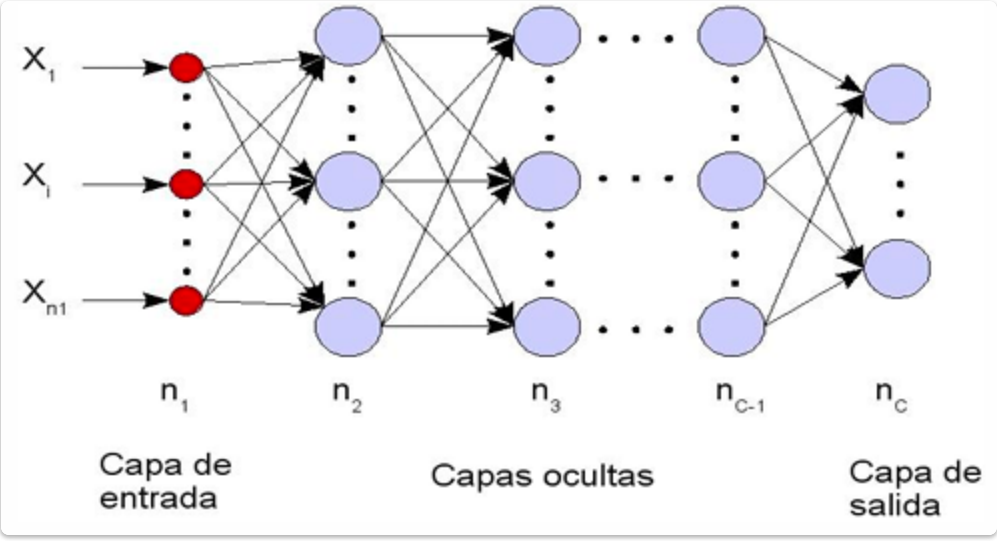
Redes Neuronales

Buscan emular el comportamiento del cerebro humano.



- Las entradas X_i representan las señales/entradas de información que debe de ser numérica.
- Los W_i son los pesos, estos son numéricos y representan el conocimiento real. Varían su valor en el aprendizaje, buscamos que estos se estabilicen.
- θ es la función umbral que la neurona tiene que pasar para activarse.
- Y es la respuesta que da la neurona.

Las neuronas se juntan formando una estructura de este estilo:



Tipos de Redes Neuronales

- Para el *análisis de imágenes* se presentan las Redes Neuronales Convolucionales
- También existen redes neuronales que generan datos, tenemos:
  - Redes Generativas Adversarias (GAN): Generan nuevos datos en situaciones en que éstos son limitados.

- Autoencoders Variacionales (VAE): Tienen por objetivo reconstruir los datos de entrada. Mejores que las GAN en la generación de caras.

Análisis de los datos disponibles

Tipos de Variables

Tenemos las Cuantitativas o Numéricas:

- *Discretas:* Cantidad de empleados, de alumnos, etc. Números enteros por así decirlo.
- *Continuas:* Sueldos, metros cuadrados, beneficios, etc. Números flotantes por así decirlo. También están las Cualitativas o Categóricas:
- *Nominales:* Nombran al objeto al que se refieren sin poder establecer un orden, por ejemplo estado civil, raza, idioma, género, etc.
- *Ordinales:* Se puede establecer un orden entre sus valores, por ejemplo alto, mediano, bajo, etc.

Descripciones estadísticas

Identifican propiedades de los datos y destacan qué valores deben tratarse como ruido o valores atípicos.

Medidas de Tendencia Central

Media/Promedio

Es el promedio de los valores del atributo, dicho atributo debe de ser numérico. Nosotros la truncamos.

$$\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}$$

N es la cantidad de valores a promediar

Con Numpy podemos sacar la Media de una Vector de 1D haciendo lo siguiente:

```
import numpy as np

vector = np.array([1, 2, 3, 4])
media = vector.mean() # 2.5
```

Mediana

Es literalmente el valor del medio, divide a los valores del atributo en dos partes iguales. Los valores tienen que estar ordenados para poder obtenerla.

Si la cantidad de datos es impar, es el valor del medio.

$$\tilde{X} = x_{(N+1)/2}$$

Si la cantidad de datos es par, es el promedio de los 2 valores centrales.

$$\tilde{X} = \frac{x_{N/2} + x_{(N+1)/2}}{2}$$

Con Numpy la podemos calcular así:

```
import numpy as np

# Vector con cantidad impar
vector_impar = np.array([1, 3, 5])
print(np.median(vector_impar))  # 3

# Vector con cantidad par
vector_par = np.array([1, 3, 5, 7])
print(np.median(vector_par))  # (3 + 5)/2 = 4.0
```

También puede calcularse sobre atributos ordinales que estén ordenados. En tal caso, el resultado será o bien *el valor que divide al conjunto en dos partes iguales* o bien se dirá que *"la mediana está entre los valores ...".*

Moda

Es el valor que aparece con mayor frecuencia. Por lo tanto, puede determinarse para atributos cualitativos y cuantitativos.

Es posible que la mayor frecuencia corresponda a varios valores diferentes, lo que da lugar a más de una MODA (unimodal, bimodal, trimodal, etc.).

Un conjunto de datos con más de una moda se considera multimodal. Si cada valor de los datos tiene solo una ocurrencia, entonces no hay moda.

Podemos calcular la Moda de varias formas, se puede usar Numpy + SciPy, o Pandas

```
import numpy as np
from scipy import stats
import pandas as pd

vector = np.array([1, 2, 2, 3, 3, 4, 4, 4, 5])
moda = stats.mode(vector)
```

```
moda.mode[0] # valor de la moda
moda.count[0] # cuantas veces aparece

nuevo_vector = pd.Series([1, 2, 2, 3, 3, 4, 4, 4, 5])
nuevo_vector.mode() #valor de la moda
```

Rango Medio

Puede utilizarse para evaluar la tendencia central de un conjunto de datos numéricos. Es la media/promedio de los valores máximo y mínimo del conjunto.

$$rango\ medio = \frac{maximo + minimo}{2}$$

Con Numpy lo podemos calcular así:

```
import numpy as np

vector = np.array([1, 2, 3, 4])
rango_medio = (vector.max() - vector.min()) / 2
```

Medidas de Dispersión

Varianza y Desviación standard

Mide la dispersión de los datos con respecto a la media.

Valores **bajos** indican que las observaciones de los datos tienden a estar *muy cerca de la media*, mientras que valores **altos** indican que los datos *están muy dispersos*.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2$$

La **desviación standard** es la raíz cuadrada de la varianza.

Con Numpy las calculamos así:

```
import numpy as np

vector = np.array([1, 2, 3, 4])
varianza = vector.var()
desviacion = vector.std()
```

Rango

El rango de un conjunto de valores numéricos es la diferencia entre los valores máximo y mínimo de dicho conjunto.

Cuantiles, Cuartiles y Percentiles

Los cuantiles son valores que dividen un conjunto numérico ordenado en partes iguales. Es decir que determinan intervalos que comprenden el mismo número de valores. Los más usados son:

- *Cuartiles*: Dividen en 4 partes.
- *Deciles*: Dividen en 10 partes.
- *Centiles o Percentiles*: Dividen en 100 partes.

El percentil en estadística nos indica, una vez ordenados los datos de menor a mayor, el valor de la variable por debajo del cual se encuentra un porcentaje dado de observaciones en un grupo.

Se representan como $Q1, Q2, Q3$, tomando a N como la cantidad de valores:

- $Q1 = (N + 1)/4$
- El 2do, es decir, $Q2$ coincide con la MEDIANA
- $Q3 = 3(N + 1)/4$

Si no hay parte decimal en la posición, se toma directamente el elemento, es decir, justo nos quedó un número entero.

Si la posición corresponde a un número con parte decimal entre el elemento i y el $i + 1$, se determinar un factor realizando una interpolación lineal. El cuartil será:

$$Q = X_i + (X_{i+1} - X_i) \cdot \text{factor}$$

Cálculo del Factor

Para calcular el factor primero tenemos que saber el valor de la Frecuencia Acumulada Relativa de las posiciones que se involucran según el cálculo que nos dio el primer cálculo del cuartil, el cálculo de la frecuencia acumulada relativa se define como:

$$F_i = \frac{i + 1}{N - 1}$$

Una vez tenemos los valores de las 2 frecuencias acumuladas relativas que se van a involucrar en el cálculo de nuestro factor ya podemos realizar correctamente el cálculo de la siguiente manera:

$$factor = \frac{\text{porcentaje que equivale el cuartil} - F_i}{F_{i+1} - F_i}$$

Tener en cuenta que el porcentaje que equivale el cuartil varía según cuál estemos calculando, por ejemplo, $Q1 = 0.25$ pero $Q3 = 0.75$.

Rango Intercuartil

Es la distancia que hay entre $Q1$ y $Q3$, esta distancia es una medida sencilla de dispersión que da el rango cubierto por la mitad de los datos.

Se calcula como: $Q3 - Q1$

Gráficos

Diagrama de Barras

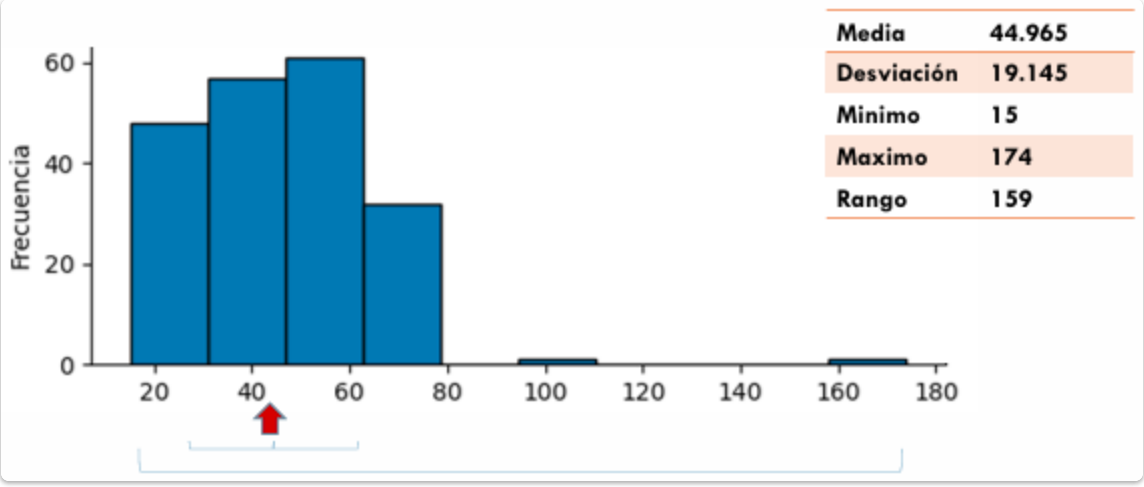
Sirve para mostrar comparaciones entre categorías, es mejor usarlo cuando se tienen datos cualitativos o discretos. Por ejemplo, mostrar la cantidad de estudiantes en cada facultad.

Diagrama de Torta

Sirve para ver la proporción que representa cada categoría respecto al total, es mejor usarlo cuando queremos resaltar la participación porcentual de cada grupo.

Histograma

Sirve para ver la distribución de frecuencias de datos continuos, es mejor usarlo cuando se quiere ver cómo se distribuyen los datos (de forma simétrica, sesgada, normal, etc.)



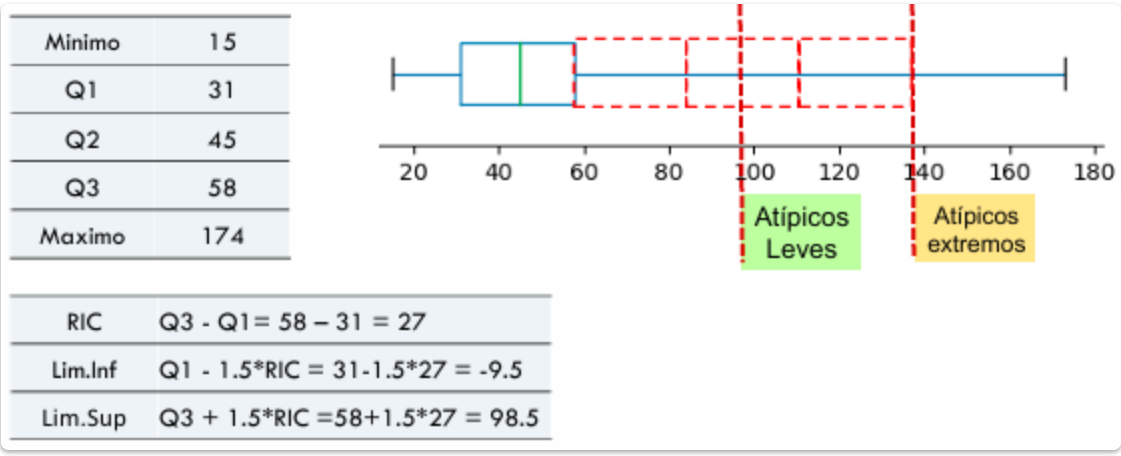
- La Media nos representa el centro de gravedad de los datos, si la distribución de los datos fuera simétrica, la media va a estar cerca del centro, si está desplazada a un lado (es este caso), indica una asimetría.
- La Desviación Standard mide cuánto se dispersan los datos respecto a la media hacia la izquierda y derecha.

- Los Mínimos y Máximos son los valores más extremos observados, nos dicen el rango total de valores observados.
- El Rango nos dice qué tan amplia es la distribución, se interpreta como la anchura total del eje X que ocupan las barras.

Diagrama de Caja

Sirve para ver los cuartiles, la mediana y los valores atípicos, es mejor usarlo cuando se quiere analizar la dispersión y simetría de los datos, o comparar distribuciones entre grupos.

- Consideramos valores atípicos leves a los que se encuentran a: $1.5 \cdot RIC$ (RIC Rango Intercuartil) más allá de los límites de la caja hacia la izquierda o derecha.
- Consideramos valores atípicos extremos a los que se encuentran a: $3 \cdot RIC$ más allá de los límites de la caja hacia la izquierda o derecha.



- Los límites teóricos son los que se calculan en la imagen como $Lim.Inf$ y $Lim.Sup$, el rango comprendido por estos, es decir, $[Lim.Inf,\ LimSup]$ son todos los valores no atípicos.
- Los bigotes en el gráfico (Valores $Mínimo$ y $Máximo$) indican el rango de los valores de la muestra comprendidos en el intervalo.

Diagrama de Dispersión

Sirve para ver la relación entre 2 variables numéricas, es mejor usarlo para analizar correlaciones o tendencias entre dos variables.

Relación entre atributos numéricos

Cuando hagamos un modelo de IA nos interesa saber si dos atributos numéricos se encuentran linealmente relacionados o no. Para ver eso se usa el coeficiente de correlación lineal.

Dados dos atributos X e Y, el coeficiente de correlación lineal entre ellos se calcula de esta forma:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_x \cdot \sigma_y}$$

Teniendo en cuenta que:

- $Cov(X,Y)$ es la covarianza entre X e Y
  - Este valor nos indica el grado de variación conjunta de dos variables aleatorias respecto a sus medidas. Puede ser Positiva, Negativa o Cercana a cero.
- σ_x y σ_y son los desvíos de cada variable.

$$Cov(X,Y) = \left[\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)\right]/N$$

$$\sigma_X = \sqrt{\left[\sum_{I=1}^{N}(x_i - \mu_X)^2\right]/N}$$

Interpretación del Coeficiente

- Si $0.5 \leq |Corr(A,B)| < 0.8$ se dice que A y B tienen una correlación lineal débil.
- Si $|Corr(A,B)| \geq 0.8$ se dice que A y B tienen una correlación lineal fuerte.
- Si $|Corr(A,B)| < 0.5$ se dice que A y B no están correlacionados linealmente. Eso no implica que son independientes.

Fase de Preparación de los Datos

La información que está almacenada siempre va a presentar:

- Datos faltantes.
- Valores extremos.
- Inconsistencias.
- Ruido

Nosotros tenemos que detectar estas anomalías/fallas para hacer una Limpieza y una Transformación de la información.

Integración

Es un proceso que se realiza durante la recopilación de datos y, si se realiza un almacén de datos, durante el proceso de carga, puede, en muchos casos, detectar y solucionar problemas de datos no resueltos durante la integración, como los valores anómalos y faltantes.

Problemas que se pueden dar

1. El primer problema a la hora de realizar una integración de distintas fuentes de datos es identificar los objetos, es decir, conseguir que datos sobre el mismo objeto se unifiquen y datos de diferentes objetos permanezcan separados. Este problema se conoce como el problema del esclarecimiento de la identidad. Pueden pasar 2 errores:
   - Dos o más objetos que son *diferentes* terminan unificados.
   - Dos o más fuentes de objetos *iguales* terminan separadas.
     - Este es el más común.
2. Cuando se integran las fuentes de información se realiza un proceso de descomposición de claves para entender mejor la composición de las mismas, si las claves están mal diseñadas pueden entrañar información no normalizada.

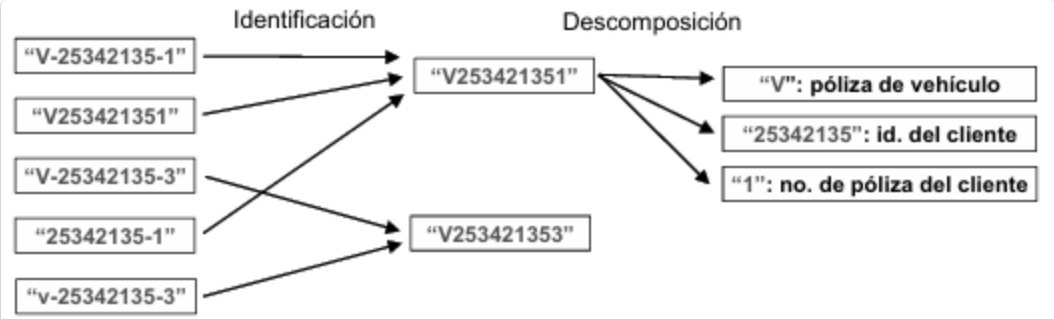


Figura 4.1. Ejemplos de integración: identificación y descomposición.

3. Cuando se integran (correctamente) dos fuentes diferentes de datos de distintos objetos suele suceder que puedan aparecer datos faltantes (el dato se registra en una fuente pero no en la otra) o datos inconsistentes (el dato es diferente en una fuente y otra).
4. La integración de formatos diferentes, que se produce si tenemos codificaciones diferentes (casado / matrimonio), idiomas diferentes, medidas diferentes, etc. requiere de una unificación.

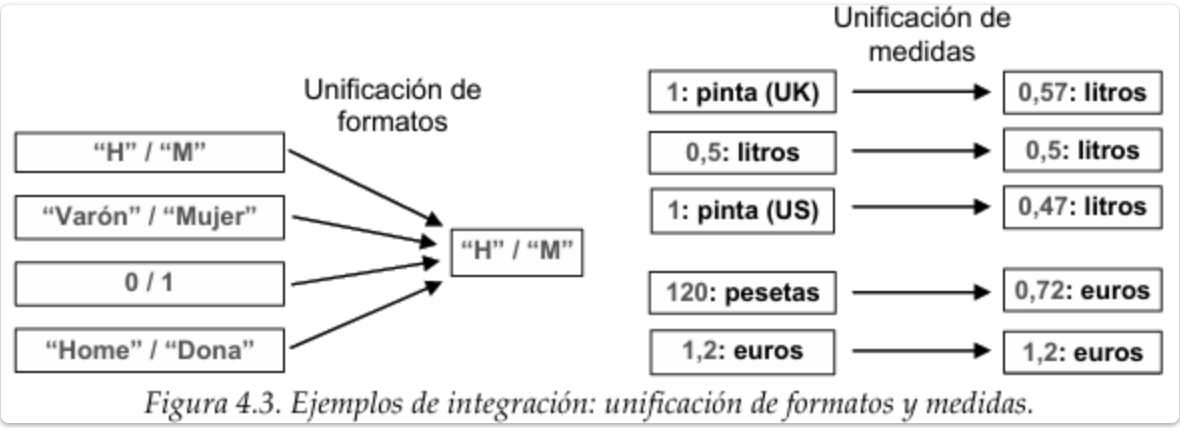


Figura 4.3. Ejemplos de integración: unificación de formatos y medidas.

Reconocimiento

Cuando tenemos integrados todos los datos lo primero que podemos realizar es un resumen de las características (o informe de estado) de atributos (ya sea tabla a tabla o para toda la

base o almacén de datos).

En este tipo de tablas se muestran las características generales de los atributos (medias, mínimos, máximos, posibles valores).

Atributo	Tabla	Tipo	# total	# nulos	# dists	Media	Desv.e.	Moda	Min	Max
Código postal	Cliente	Nominal	10320	150	1672	-	-	"46003"	"01001"	"50312"
Sexo	Cliente	Nominal	10320	23	6	-	-	"V"	"E"	"M"
Estado civil	Cliente	Nominal	10320	317	8	-	-	Casado	"Casado"	"Viudo"
Edad	Cliente	Numérico	10320	4	66	42,3	12,5	37	18	87
Total póliza p/a	Póliza	Numérico	17523	1325	142	737,24€	327€	680€	375€	6200€
Asegurados	Póliza	Numérico	17523	0	7	1,31	0,25	1	0	10
Matrícula	Vehículo	Nominal	16324	0	16324	-	-	-	"A-0003-BF"	"Z-9835-AF"
Modelo	Vehículo	Nominal	16324	1321	2429	-	-	"O. Astra"	"Audi A3"	"VW Polo"
...	...	...	...	...	...	...	...	...	...	...

Tabla 4.1. Tabla resumen de atributos.

Una vez tenemos esta información, podemos realizar más que nada Histogramas, Diagramas de Caja o Diagramas de Dispersión para poder reconocer y comprender más las anomalías que se presentan.

Valores Faltantes

Hay que estar atentos a que un valor faltante no necesariamente es un valor que está representado como NULO, puede pasar por ejemplo que en una columna que sea "dirección" nos digan "no tiene", eso se vuelve un valor faltante que no es NULO.

Para estos valores ya sean nulos o no podemos hacer muchas cosas:

- Ignorarlos.
- Rellenarlos manualmente.
- Usar una constante global para rellenarlos.
- Usar el valor de la Media u otra medida de centralidad para rellenarlos.
- Usar el valor de la Media u otra medida de centralidad de los objetos que pertenecen la misma clase.
- Usar alguna técnica de Aprendizaje Automático para calcular el valor más probable.

Problemas que surgen con estas medidas

Si sustituimos un dato faltante por un dato estimado, hemos de tener en cuenta que:

1. Perdemos información, ya que ya no se sabe que el dato era faltante y,
2. Inventamos información, con los riesgos que pueda tener de que sea errónea.

Solución

Creamos un nuevo atributo lógico (booleano) indicando si el atributo original era nulo o no. Esto permite al proceso y al método de minería de datos, si son bastante perspicaces, saber que el dato era faltante y, por tanto, que el valor hay que tomarlo con cautela.

En el caso en el que el atributo original sea nominal no es necesario crear un nuevo atributo, basta con añadir un valor adicional, denominado "faltante".

Transformación de Atributos

La transformación de datos engloba, en realidad, cualquier proceso que modifique la forma de los datos. Prácticamente todos los procesos de preparación de datos entrañan algún tipo de transformación.

Discretización

Algunos algoritmos de minería de datos sólo operan con atributos cualitativos. La discretización convierte los atributos numéricos en ordinales ordenados que se representan con un intervalo.

Ayuda a que los valores anómalos caigan en alguno de los extremos sin problema.

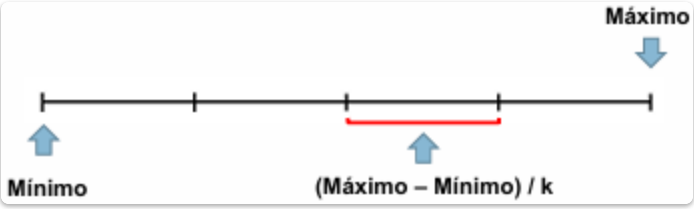
Ejemplo: La edad de la persona en categorías: $[0, 12]$ niño, $(12 - 21)$ joven, $[21, 65]$ adulto y > 65 anciano.

Se puede discretizar en un número fijo de intervalos, el ancho del intervalo se calcula:

- Dividiendo el rango en partes iguales.
- Dividiendo la cantidad de ejemplos en partes iguales (igual frecuencia).
- Indicando los límites de cada intervalo en forma manual.

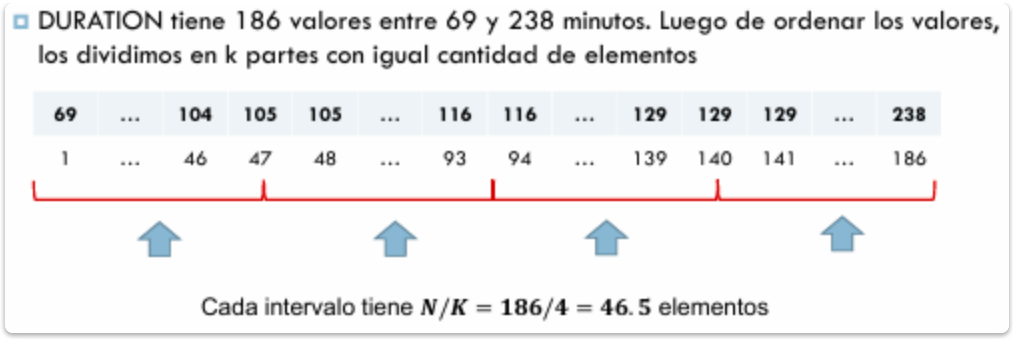
Discretización por Rangos

El objetivo es dividir el rango del atributo (intervalo entre el máximo y el mínimo) en una cierta cantidad k de partes iguales.



Discretización por Frecuencia

El objetivo es dividir los valores del atributo numérico en k partes con la misma cantidad de valores en cada una de ellas. El atributo debe tener al menos k valores diferentes.



Numerización

Es el proceso contrario a la discretización. Convierte atributos cualitativos en numéricos.

Para los nominales suele utilizarse una representación binaria y para los ordinales suele utilizarse una representación entera.

- Si se numeran en forma correlativa los valores de un *atributo nominal* se agrega un *orden* que originalmente *no está presente* en la información disponible.

Numerización Binaria (dummy)

Reemplaza al atributo nominal por tantos atributos numéricos binarios como valores distintos pueda tomar.

Las denominaciones de estos nuevos atributos surgen de igualar el nombre original con cada uno de los posibles valores. Para un mismo ejemplo sólo uno de estos nuevos atributos tendrá valor 1 y el resto 0.

Normalización

Permite expresar los valores de los atributos sin utilizar las unidades de medida originales facilitando su comparación y uso conjunto.

Si bien se aplica según el modelo a construir, la más común es la normalización lineal uniforme:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- Es muy sensible a los valores outliers.
- Si recortamos los extremos obtenemos valor negativos y/o mayores a 1.

Si los datos tienen una distribución normal, podemos tipificar:

$$X' = \frac{X - media(X)}{desviacion(X)}$$

- De esta forma los datos se distribuyen normalmente alrededor de 0 con desviación 1.