UAM | Escuela Politécnica Superior

Datos personales y anonimización

Francisco Jurado, francisco.jurado@uam.es

Contenido

1. Marco legal
2. Anonimización
3. Técnicas de anonimización
4. Ejemplo paso a paso
5. Datos abiertos

Contenido

Reglamento General de Protección de Datos (RGPD)

- Es una **normativa europea** de referencia que se aplica a empresas que tratan datos de ciudadanos de la UE.

- En España:
  - Ley Orgánica 3/2018 de Protección de Datos y Garantía de los Derechos Digitales (LOPDGDD)
    - Ley española que **complementa y desarrolla el RGPD** dentro del territorio español
  - Modificación en mayo del 2023

- Se trata de uno de los marcos legales **más garantistas** del mundo para los ciudadanos.

- Establece los siguientes **derechos**:
  1. Acceso
  2. Rectificación
  3. Oposición
  4. Supresión (“derecho al olvido”)
  5. Limitación del tratamiento
  6. Portabilidad
  7. A no ser objeto de decisiones individualizadas automatizadas
  8. Derecho de información
  9. Derechos Schengen y Marco de Privacidad UE-EE.UU

Re

Derecho a no ser objeto de decisiones individuales automatizadas

Este derecho pretende garantizar que no seas objeto de una decisión basada únicamente en el tratamiento de tus datos, incluida la elaboración de perfiles, que produzca efectos jurídicos sobre ti o te afecte significativamente de forma similar.

Sobre esta elaboración de perfiles, se trata de cualquier forma de tratamiento de tus datos personales que evalúe aspectos personales, en particular analizar o predecir aspectos relacionados con tu rendimiento en el trabajo, situación económica, salud, las preferencias o intereses personales, fiabilidad o el comportamiento.

No obstante, este derecho no será aplicable cuando:

- **Sea necesario para la celebración o ejecución de un contrato entre tú y el responsable**
- **El tratamiento de tus datos se fundamente en tu consentimiento prestado previamente**

No obstante, en estos dos primeros supuestos, el responsable debe garantizar tu derecho a obtener la intervención humana, expresar tu punto de vista e impugnar la decisión.

- **Esté autorizado por el Derecho de la Unión o de los Estados miembros y se establezcan medidas adecuadas para salvaguardar los derechos y libertades e intereses legítimos del interesado.**

A su vez, estas excepciones no se aplicarán sobre las categorías especiales de datos (art.9.1), salvo que se aplique el artículo 9.2.letra a) o g) y se hayan tomado las medidas adecuadas citadas en el párrafo anterior.

E.

Reglamento general de protección de datos (RGPD)

- La "otra cara de la moneda", es que se trata de uno de los marcos legales más restrictivos para las empresas.
- *Aunque dispongamos de los datos*, si son datos personales, **NO** podemos tratarlos para fines distintos a aquel para el que fueron recogidos.
    - Salvo consentimiento explícito por parte de los interesados
        - Interesados = las personas a las que pertenecen los datos personales

Reglame

- La "otra s restricti
- *Aunque* ntarlos para fin
  - Sa
    -

Artículo 6. Licitud del tratamiento

1. El tratamiento solo será lícito si se cumple al menos una de las siguientes condiciones:

a) el interesado dio su consentimiento para el tratamiento de sus datos personales para uno o varios fines específicos;

b) el tratamiento es necesario para la ejecución de un contrato en el que el interesado es parte o para la aplicación a petición de este de medidas precontractuales;

c) el tratamiento es necesario para el cumplimiento de una obligación legal aplicable al responsable del tratamiento;

d) el tratamiento es necesario para proteger intereses vitales del interesado o de otra persona física;

e) el tratamiento es necesario para el cumplimiento de una misión realizada en interés público o en el ejercicio de poderes públicos conferidos al responsable del tratamiento;

f) el tratamiento es necesario para la satisfacción de intereses legítimos perseguidos por el responsable del tratamiento o por un tercero, siempre que sobre dichos intereses no prevalezcan los intereses o los derechos y libertades fundamentales del interesado que requieran la protección de datos personales, en particular cuando el interesado sea un niño.

¿Qué es un dato personal? (I)

- Reglamento General de Protección de Datos (RGPD):
  - *"Los datos personales son **cualquier información relativa a una persona física viva identificada o identificable**. Las distintas informaciones, que recopiladas pueden llevar a la identificación de una determinada persona, también constituyen datos de carácter personal."*
  - *"Los datos personales que hayan sido anonimizados, cifrados o presentados con un seudónimo, **pero que puedan utilizarse para volver a identificar a una persona**, siguen siendo datos personales y se inscriben en el ámbito de aplicación del RGPD."*

¿Son esto datos personales?

- "(...) *cualquier información relativa a una persona física viva* ***identificada***..."
- Sí, lo son: persona identificada por varios **identificadores directos**
- Esto se denomina **identificación**

Nombre	Apellido1	Apellido2	DNI	F. nacimiento	Localidad	Género	Teléfono	Nivel de estudios	Estado civil
José	López	López	01234567-Z	7/2/1930	Alcafrán	Hombre	640000000	Licenciatura/Ingeniería	Divorciado

¿Y ahora?

Nombre	Apellido1	Apellido2	DNI	F. nacimiento	Localidad	Género	Teléfono	Nivel de estudios	Estado civil
-	-	-	-	7/2/1930	Alcafrán	Hombre	640*********	Licenciatura/Ingeniería	Divorciado

- ¿Cuántos hombres nacidos el 7/2/30 hay en Alcafrán (10 habitantes)?
- "*Las distintas informaciones, que recopiladas pueden llevar a la identificación de una determinada persona, también constituyen datos de carácter personal*"
- Por tanto, sí, podrían serlo: persona identificable mediante **identificadores indirectos**
  - Por ejemplo, es áltamente probable que pueda vincular estos datos de forma unívoca a un determinado registro del censo electoral
- Esto se denomina **"reidentificación"**

Contenido

Consecuencia importante

- Que un conjunto de datos contenga o no datos personales **no depende únicamente del contenido de ese conjunto de datos**, sino que depende también de **factores externos y potencialmente cambiantes,** por ejemplo:
  - De las realidades a las que los datos hacen referencia
  - De la disponibilidad de otros conjuntos de datos publicados por terceros

- Esto implica la reevaluación periódica de la eficacia de los procesos de anonimización, teniendo en cuenta la evolución del contexto.

¿Para qué sirve anonimizar?

- Para que un conjunto de datos **deje de estar sujeto al RGPD**
- "*Los datos personales que hayan sido **anonimizados**, de forma que la persona no sea identificable, **dejarán de considerarse datos personales.***"
- Esto posibilita tratamientos que **estarían prohibidos o serían extremadamente complejos** dentro del marco regulatorio del RGPD: análisis, explotación, publicación en abierto, etc...

¿Para qué sirve a

- Para que un conjunt
- "*Los datos personal* *a que la persona no sea identificable,* ***dejará***
- Esto posibilita tratar **extremadamente complejos** dentro d xplotación, publicación en abierto, etc...

Considerando 26

(26) Los principios de la protección de datos deben aplicarse a toda la información relativa a una persona física identificada o identificable.

Los datos personales seudonimizados, que cabría atribuir a una persona física mediante la utilización de información adicional, deben considerarse información sobre una persona física identificable.

Para determinar si una persona física es identificable, deben tenerse en cuenta todos los medios, como la singularización, que razonablemente pueda utilizar el responsable del tratamiento o cualquier otra persona para identificar directa o indirectamente a la persona física.

Para determinar si existe una probabilidad razonable de que se utilicen medios para identificar a una persona física, deben tenerse en cuenta todos los factores objetivos, como los costes y el tiempo necesarios para la identificación, teniendo en cuenta tanto la tecnología disponible en el momento del tratamiento como los avances tecnológicos.

Por lo tanto los principios de protección de datos no deben aplicarse a la información anónima, es decir información que no guarda relación con una persona física identificada o identificable, ni a los datos convertidos en anónimos de forma que el interesado no sea identificable, o deje de serlo.

En consecuencia, el presente Reglamento no afecta al tratamiento de dicha información anónima, inclusive con fines estadísticos o de investigación.

Contenido

Técnicas de anonimización

- Las técnicas de anonimización implican **transformaciones en los datos** cuya finalidad es armonizar dos **objetivos contrapuestos**:

1. Que los datos **no** puedan asociarse al individuo del que proceden

2. Pero que sigan siendo **útiles** para un determinado fin
   - Normalmente, un determinado tipo de análisis
   - En ocasiones, tal fin no se conoce (ej. datos abiertos)

Técnicas de anonimización

- Normalmente, las técnicas de anonimización implican alguna **pérdida de información** respecto a los datos originales.

- Esto se denomina **diferencial de privacidad**:
  - *El diferencial de privacidad es la "diferencia o distorsión resultante del análisis de la información personal anonimizada frente al análisis de los datos personales originales no anonimizados."*

Técnicas

- Aleatorización
  - Adición de ruido → pierde precisión
  - Permutación → pierde información relacional
- Supresión de registros o atributos
- Enmascaramiento de caracteres
- Generalización → pierde nivel de precisión
- Perturbación de datos
- K-anonimización

Métricas

- K-anonimidad
- L-diversidad
- T-proximidad

Aleatorización

- **Altera los datos originales** mediante procesos aleatorios **para evitar la identificación directa**.

- **Objetivo:** Romper la exactitud o la estructura original de los datos.

- Formas de hacerlo: adición de ruido y permutación.

Aleatorización: adición de ruido

- **Agrega pequeñas cantidades aleatorias** (ruido) a datos numéricos **para proteger la privacidad** sin perder utilidad.

- **Ejemplo:**
  - Modificar edades o ingresos con valores aleatorios mantiene la privacidad mientras se conserva el análisis estadístico.
  - **Edad original: 45 años** → **Edad anonimizada: 45 + (ruido aleatorio entre -3 y +3)** → **47 años**.

- Pierde precisión → Balance entre privacidad y precisión
  - El ruido excesivo puede distorsionar resultados, por lo que debe aplicarse cuidadosamente para preservar la calidad del análisis.

Ejemplo con código Python

```
data = {
    'Nombre': ['Ana', 'Luis', 'María', 'Pedro'],
    'Edad': [45, 32, 28, 50],
    'Ingresos': [2500, 1800, 2200, 3000]
}
df = pd.DataFrame(data)

# Añadir ruido aleatorio a la columna Edad e Ingresos, creando nuevas columnas con los
valores perturbados, manteniendo los originales para comparación.

np.random.seed(42)  # Para reproducibilidad

# ruido con media 0 y desviación 2
df['Edad_ruido'] = df['Edad'] + np.random.normal(0, 2, size=len(df))

# ruido con media 0 y desviación 100
df['Ingresos_ruido'] = df['Ingresos'] + np.random.normal(0, 100, size=len(df))
```

Aleatorización: permutación

- **Intercambia valores entre registros para romper relaciones** originales entre atributos.

- Protección de privacidad
  - **Evita inferir correlaciones** sensibles, dificultando la identificación de individuos.

- Impacto en análisis → Pierde información relacional
  - Puede afectar análisis que dependen de correlaciones entre atributos originales.

- **Ejemplo:**

Original:	
Nombre	Ciudad
Ana	Madrid
Luis	Sevilla

Permutado:	
Nombre	Ciudad
Ana	Sevilla
Luis	Madrid

- **Limitación:**
  - No altera valores, por eso se combina con técnicas como ruido o supresión para mayor protección.

Ejemplo con código Python

```
data = {
    'Nombre': ['Ana', 'Luis', 'María', 'Pedro'],
    'Ciudad': ['Madrid', 'Sevilla', 'Valencia', 'Bilbao'],
    'Ingresos': [2500, 1800, 2200, 3000]
}
df = pd.DataFrame(data)

# Permutar columnas Ciudad e Ingresos
np.random.seed(42)  # Para reproducibilidad

df['Ciudad_permutada'] = np.random.permutation(df['Ciudad'].values)
df['Ingresos_permutados'] = np.random.permutation(df['Ingresos'].values)
```

Ejemplo con código Python

```
def permutar_columnas(df, columnas):
    """
    Permuta aleatoriamente los valores de las columnas
    especificadas en el DataFrame.
    """

    np.random.seed(42)  # Para reproducibilidad

    for col in columnas:
        if col in df.columns:
            df[col] = np.random.permutation(df[col].values)
        else:
            print(f"Advertencia: La columna '{col}' no existe.")

    return df
```

Supresión de registros o atributos

- **Elimina registros o atributos** que contienen información que puede identificar directamente a una persona.

- **Limitación:**
  - Reducción de utilidad del conjunto.
  - Eliminar datos clave puede disminuir la utilidad del conjunto de datos para análisis posteriores.

- Combinación con otras técnicas
  - Se usa junto a otras técnicas para mantener la mayor información posible sin perder privacidad.

Ejemplo con código Python

```
data = {
    'Nombre': ['Ana', 'Luis', 'María', 'Pedro'],
    'Edad': [45, 32, 28, 50],
    'Ciudad': ['Madrid', 'Sevilla', 'Valencia', 'Bilbao']
}
df = pd.DataFrame(data)

# Supresión de la columna 'Nombre'
df_sin_nombre = df.drop(columns=['Nombre’])

# Supresión de registros donde Edad > 45
df_sin_registros = df[df['Edad'] <= 45]
```

Enmascaramiento de caracteres

- **Oculta información sensible** usando símbolos o caracteres ficticios **para protección parcial**.
- Aplicación en Datos Sensibles
  - Como números de tarjetas, teléfonos e identificadores, etc. para mostrar sólo parte de la información.
- **Ejemplo:**
  - Tarjeta `'1234 5678 9012 3456'` puede mostrarse como `'**** **** **** 3456'`
    - Revela solo los últimos dígitos para fines de verificación.
  - Útil cuando se necesita **mostrar parcialmente** la información **para validación** sin exponerla completamente.
- **Limitación:**
  - Protege contra exposición directa pero no elimina la posibilidad de identificar patrones en conjuntos completos.

Ejemplo con código Python

```
data = {
    'Nombre': ['Ana', 'Luis', 'María', 'Pedro'],
    'DNI': ['12345678A', '87654321B', '11223344C', '44332211D'],
    'Tarjeta': ['1234-5678-9012-3456', '9876-5432-1098-7654’,
                '1111-2222-3333-4444', '5555-6666-7777-8888']
}
df = pd.DataFrame(data)

# Enmascarar DNI (mostrar solo los 3 últimos caracteres)
df['DNI_enmascarado'] = df['DNI'].apply(lambda x: '*' * (len(x)-3) + x[-3:])

# Enmascarar Tarjeta (mostrar solo los últimos 4 dígitos)
df['Tarjeta_enmascarada'] = df['Tarjeta'].apply(lambda x: '****-****-****-' + x[-4:])
```

Generalización

- **Sustituye valores específicos con rangos o categorías** más amplias **para simplificar datos**.

- **Ejemplo:**
  - en lugar de edad exacta (45 años) → indicar un rango ('40-50 años').
  - en lugar de una ciudad específica (Madrid) → generalizar al país (España).

- Aplicación en k-anonimato (hablaremos más tarde)
  - Se usa para que registros sean indistinguibles dentro de un grupo, mejorando la privacidad mediante agrupación de datos.

- Impacto en análisis → Generalizar datos protege privacidad, pero reduce la granularidad y detalle del análisis (el nivel de precisión).

- Ideal para situaciones donde la precisión individual no es crítica, pero se requiere mantener tendencias generales.

Ejemplo con código Python

```
data = {
    'Nombre': ['Ana', 'Luis', 'María', 'Pedro'],
    'Edad': [45, 32, 28, 50],
    'Ciudad': ['Madrid', 'Sevilla', 'Valencia', 'Bilbao']
}
df = pd.DataFrame(data)

# Generalización de Edad en rangos
def generalizar_edad(edad):
    if edad < 30:
        return 'Menos de 30'
    elif edad < 40:
        return '30-39'
    elif edad < 50:
        return '40-49'
    else:
        return '50 o más'


df['Edad_generalizada'] = df['Edad'].apply(generalizar_edad)


# Generalización de Ciudad a nivel de país
df['Ubicacion_generalizada'] = 'España'
```

Perturbación de datos

- **Modifica valores** originales para **conservar la distribución estadística** del conjunto de datos.

- **Ejemplo:**
  - Si un ingreso original es de 2500 €, se puede reemplazar por un valor ficticio como 2600 €, calculado para conservar la media y la desviación estándar del grupo.

- Útil en análisis estadísticos donde se necesita **proteger la privacidad sin alterar** significativamente los **resultados globales**.

- **Diferencia con la adición de ruido**
  - La perturbación **busca mantener la coherencia** del conjunto, lo que la hace adecuada para estudios agregados.

- **¡OJO!** Si se aplica mal, puede introducir sesgos y afectar la validez de los análisis estadísticos.

Ejemplo con código Python

```
data = {
    'Nombre': ['Ana', 'Luis', 'María', 'Pedro'],
    'Ingresos': [2500, 1800, 2200, 3000]
}
df = pd.DataFrame(data)

# Reemplazar ingresos por valores aleatorios con misma media y desviación std.
np.random.seed(42)  # Para reproducibilidad
media = df['Ingresos'].mean()
desviacion = df['Ingresos'].std()

# Generar nuevos valores con distribución similar
df['Ingresos_perturbados'] = np.random.normal(loc=media,
                                              scale=desviacion,
                                              size=len(df)).round(2)
```

K-anonimización

- Para cada registro en un conjunto de datos, debe haber **al menos k registros que sean indistinguibles** en términos de ciertos atributos identificativos.
  - Si alguien intenta identificar a una persona específica, no podrá hacerlo con certeza porque hay al menos **k** personas que comparten las mismas características.
  - Es decir, **cada individuo está oculto dentro de un grupo de al menos k personas con características similares**.

- Un conjunto de datos cumple **k-anonimidad** si, para cualquier combinación de valores en los atributos, existen al menos **k registros con esos mismos valores**.

- **Cuasi-identificadores**: atributos que pueden reidentificar a alguien, como edad, código postal, género.

K-anonimización

- **Ejemplo:**

  Dataset original

Edad	Código Postal	Diagnóstico
29	28001	Diabetes
30	28001	Hipertensión
31	28001	Asma

- Si aplicamos **k-anonimización con k=3**, podemos generalizar:

Edad: 29-31	Código Postal: 2800*	Diagnóstico: (sin cambio)

  - El atributo 'Edad' es un cuasi-identificador → aplicamos generalización
  - El atributo 'Código postal' es un cuasi-identificador → aplicamos enmascaramiento
  - El atributo 'Diagnóstico' no es un cuasi-identificador, sino un dato sensible que queremos conservar para análisis (p.ej., estudios médicos). Por eso no se modifica para cumplir k-anonimidad.

- **Cada registro comparte los mismos valores en los cuasi-identificadores con al menos 2 más**, formando un grupo de 3.

Dataset Original

Edad	Código Postal	Diagnóstico
29	28001	Diabetes
30	28001	Hipertensión
31	28001	Asma
45	28002	Migraña
46	28002	Diabetes
47	28002	Hipertensión

Tras aplicar k-anonimización (k=3)

Edad	Código Postal	Diagnóstico
29-31	2800*	Diabetes
29-31	2800*	Hipertensión
29-31	2800*	Asma
45-47	2800*	Migraña
45-47	2800*	Diabetes
45-47	2800*	Hipertensión

K-anonimización

- Características:
  - **Indistinguibilidad**: Los registros deben ser similares en atributos clave (como edad, género, ubicación, etc.) para que no se pueda identificar a un individuo específico.
  - **Privacidad**: Al aplicar k-anonimización, se reduce el riesgo de reidentificación de individuos a partir de datos que podrían ser considerados sensibles.
  - El valor de **k** determina el **nivel de anonimato**.
    - Valor más alto de k implica un mayor nivel de protección, pero también puede afectar la utilidad de los datos.

K-anonimización

- Un conjunto de datos sobre empleados:
  - si k = 3: para cada registro de empleado, debe haber al menos otros dos empleados con características similares
  - si un registro tiene un nombre, edad y departamento, debe haber al menos dos registros más con la misma combinación de esos atributos

- Claramente **k=1 y k=n (generalmente) inútiles** para un conjunto de datos de tamaño n.
  - **K=1** no proporciona anonimato
  - **K=n** no tiene ninguna utilidad (ej. información básica como tamaño del conjunto de datos).

K-anonimización

K=2 en general			
	Código postal	Edad	Serie Favorita
k=2	22xxxx	21-25	La Casa de Papel
	22xxxx	21-25	La Casa de Papel
k=4	10xxxx	41-45	Peaky Blinders
	10xxxx	41-45	Peaky Blinders
	10xxxx	41-45	Peaky Blinders
	10xxxx	41-45	Peaky Blinders
K=3	58xxxx	56-60	Juego de Tronos
	58xxxx	56-60	Juego de Tronos
	58xxxx	56-60	Juego de Tronos

K-anonimización

- Con k-anonimidad, cada grupo tiene al menos k registros con los mismos cuasi-identificadores.

- **Problema → riesgo de inferencia**
  - Si todos los registros del grupo comparten el mismo valor sensible (p.ej., el mismo diagnóstico), entonces **saber el grupo implica saber el dato sensible**.

- **Ejemplo:**
  - Grupo k-anónimo:

    `Edad: 30-39 | CP: 2800* | Diagnóstico: Diabetes, Diabetes, Diabetes`
  - Aunque hay 3 personas (k=3), el diagnóstico es no cambiará, por lo que podrá inferirse el dato sensible (el diagnóstico).

L-diversidad

- La l-diversidad es un **concepto complementario a la k-anonimidad**, diseñado **para reducir un problema importante**: la homogeneidad dentro de los grupos k-anónimos.

- La l-diversidad busca asegurar que haya suficiente variabilidad en los atributos para dificultar inferencias.

- Necesidad de que, dentro de cada grupo de equivalencia (grupos de registros que son indistinguibles entre sí), haya al menos (l) valores diferentes para un atributo sensible.
  - (l): Cantidad de valores diferentes que hay en un grupo de registros indistinguibles entre sí.

- Ayuda a prevenir inferencia, donde se podría deducir información sobre un individuo basándose en los datos disponibles.

L-diversidad

- Un **grupo** cumple l-diversidad si, para cada conjunto de cuasi-identificadores, hay al menos l valores bien representados para el atributo sensible.
  - l ≥ 2 significa que hay al menos dos diagnósticos diferentes en cada grupo.
  - Cuanto mayor sea l, más diversidad y menos riesgo.

- Ej. grupo de registros de pacientes con la misma edad y código postal.
  - Si todos tienen el mismo diagnóstico ("diabetes") → no hay diversidad.
  - Si hay diferentes diagnósticos ("diabetes", "hipertensión" y "asma") → hay diversidad.
  - Para proteger la privacidad, no debe haber un solo valor que se repita en exceso.
    - Aunque sepamos el grupo, **no podremos inferir el diagnóstico con certeza**.

Ejemplo con código Python

```
data = {
    'Edad': [29, 30, 31, 45, 46, 47, 33, 34, 35],
    'CodigoPostal': ['28001', '28001', '28001', '28002', '28002', '28002', '28003', '28003', '28003'],
    'Diagnostico': ['Diabetes', 'Hipertensión', 'Asma', 'Migraña', 'Diabetes', 'Hipertensión', 'Asma',
                    'Diabetes', 'Hipertensión']
}
df = pd.DataFrame(data)


# Paso 1: Generalización para k-anonimidad
# Agrupar Edad en rangos y truncar Código Postal

def generalizar_edad(edad):
    if edad < 32:
        return '29-31'
    elif edad < 40:
        return '33-35'
    else:
        return '45-47'

df['Edad_generalizada'] = df['Edad'].apply(generalizar_edad)
df['CP_generalizado'] = df['CodigoPostal'].apply(lambda x: x[:4] + '*')
```

Ejemplo con código Python

```
# Paso 2: Verificar grupos para l-diversidad
# Agrupar por cuasi-identificadores y comprobar diversidad en Diagnóstico
agrupado = df.groupby(['Edad_generalizada', 'CP_generalizado'])

# Crear lista para almacenar grupos válidos
resultado = []
for (edad_g, cp_g), grupo in agrupado:
    diagnosticos_unicos = grupo['Diagnostico'].nunique()
    if diagnosticos_unicos >= 2:  # Cumple l-diversidad con l=2 -> incluir grupo en resultado
        for _, fila in grupo.iterrows():
            resultado.append({
                'Edad_generalizada': edad_g,
                'CP_generalizado': cp_g,
                'Diagnostico': fila['Diagnostico']
            })

# Convertir el resultado a DataFrame final
resultado_df = pd.DataFrame(resultado)
```

T-proximidad

- Con l-diversidad, garantizamos que haya al menos l valores distintos en el atributo sensible.
- **Problema** → esos valores pueden ser muy similares (por ejemplo, tres tipos de cáncer diferentes), lo que sigue permitiendo inferencias.
- Necesitamos medir qué tan cerca está la distribución de valores sensibles en un grupo respecto a la distribución global.

T-proximidad

- Extensión de la l-diversidad donde además de tener al menos l valores diferentes, estos deben **reflejar la distribución de los datos originales**.
  - La idea es que los datos anonimizados sean lo más cercanos posible a los datos originales en términos de distribución, lo que puede ser útil para mantener la utilidad de los datos mientras se protege la privacidad.

- Un grupo cumple **t-proximidad** si la **distancia** entre la distribución de valores sensibles en ese grupo y la distribución global de esos valores **no excede un umbral t**.
  - Se suele usar la distancia de Earth Mover (EMD) o alguna métrica similar para comparar distribuciones.
  - t es un valor entre 0 y 1 (cuanto menor, más estricta la privacidad).

T-proximidad

- **Ejemplo:**
- Supongamos que en todo el dataset:

  `Diagnóstico global: Diabetes (40%), Hipertensión (30%), Asma (30%)`

- En un grupo k-anónimo:

  `Diagnóstico: Diabetes (90%), Hipertensión (10%)`

  - La distancia entre estas distribuciones es alta → **NO cumple t-proximidad**.
- Si otro grupo tiene:

  `Diagnóstico: Diabetes (35%), Hipertensión (35%), Asma (30%)`

  - La distancia es baja → **SÍ cumple t-proximidad**.

k-anonimidad vs. l-diversidad vs. t-proximidad

- **k-anonimidad** protege **contra identificación directa**.
- **l-diversidad** protege **contra inferencia del atributo sensible**.
- **t-proximidad** protege **contra inferencia por distribución sesgada**.

Contenido

Paso 0: ¿Puede haber reidentificación?

- Dataset de ejemplo:
  - Datos de estudiantes de una Universidad
  - Se han eliminado los identificadores directos

TITULACIÓN	CENTRO	AÑO NACIMIENTO	GÉNERO	CRÉDITOS MATRICULADOS	CRÉDITOS PRESENTADOS	CRÉDITOS SUPERADOS
GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1980	HOMBRE	36	24	24
GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1995	MUJER	48	48	48
GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1990	HOMBRE	16	0	0
GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1995	HOMBRE	60	36	36
GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1980	MUJER	16	0	0
GRADO EN XENOMORFOLOGÍA	ESCUELA DE ASTROBIOLOGÍA	1990	MUJER	24	18	12
GRADO EN NIGROMANCIA	ESCUELA DE CIENCIAS FICTICIAS	1990	MUJER	36	36	30
GRADO EN NIGROMANCIA	ESCUELA DE CIENCIAS FICTICIAS	1995	MUJER	12	12	6

Paso 0: ¿Puede haber reidentificación?

- ¿Puede haber **reidentificaciones** en el dataset?
  - Sí, por ejemplo:
    - sólo hay una matrícula en el "Grado en Xenomorfología"
    - sólo hay un caso de **hombre** nacido en **1990**
- (Vamos a "colorear" los registros para poder hacer un mejor seguimiento del proceso)

TITULACIÓN	CENTRO	AÑO NACIMIENTO	GÉNERO	CRÉDITOS MATRICULADOS	CRÉDITOS PRESENTADOS	CRÉDITOS SUPERADOS
GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1980	HOMBRE	36	24	24
GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1995	MUJER	48	48	48
GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1990	HOMBRE	16	0	0
GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1995	HOMBRE	60	36	36
GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1980	MUJER	16	0	0
GRADO EN XENOMORFOLOGÍA	ESCUELA DE ASTROBIOLOGÍA	1990	MUJER	24	18	12
GRADO EN NIGROMANCIA	ESCUELA DE CIENCIAS FICTICIAS	1990	MUJER	36	36	30
GRADO EN NIGROMANCIA	ESCUELA DE CIENCIAS FICTICIAS	1995	MUJER	12	12	6

Paso 1: seleccionar variables "pivote" y agrupar

- Vamos a elegir un subconjunto de variables que denominaremos "variables pivote"
- Esas variables tienen un cometido y una propiedad esenciales:
  - **Cometido**: generar agrupaciones distintas de registros, dentro de las cuales se aplicará una determinada técnica de aleatorización.
  - **Propiedad**: las variables "pivote" no pierden información relacional respecto a ninguna otra variable del *dataset*.
- Por tanto, interesa elegir variables relevantes, **que suelan usarse como criterios de clasificación o agregación habituales** en el ese *dataset* concreto

	VARIABLES PIVOTE						
GRUPO	**TITULACIÓN**	**CENTRO**	**AÑO NACIMIENTO**	**GÉNERO**	**CRÉDITOS MATRICULADOS**	**CRÉDITOS PRESENTADOS**	**CRÉDITOS SUPERADOS**
1	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1980	HOMBRE	36	24	24
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1995	MUJER	48	48	48
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1990	HOMBRE	16	0	0
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1995	HOMBRE	60	36	36
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1980	MUJER	16	0	0
2	GRADO EN XENOMORFOLOGÍA	ESCUELA DE ASTROBIOLOGÍA	1990	MUJER	24	18	12
3	GRADO EN NIGROMANCIA	ESCUELA DE CIENCIAS FICTICIAS	1990	MUJER	36	36	30
	GRADO EN NIGROMANCIA	ESCUELA DE CIENCIAS FICTICIAS	1995	MUJER	12	12	6

Paso 2: definir los "bloques de coherencia"

- Los bloques de coherencia son subconjuntos de variables con una **interdependencia relacional o semánticas fuerte**
- Queremos **conservar** esa información relacional en el proceso de anonimización
- Casos habituales:
  - Jerarquías (localidad, provincia, comunidad autónoma)
  - Pares codificación-descripciones (código-provincia, nombre-provincia)
  - Datos numéricos semánticamente relacionados (salario, importe retención IRPF)
    - Suelen utilizarse en fórmulas para calcular magnitudes derivadas
- Cuantos más bloques de coherencia, menos riesgo de reidentificación…
  - …pero más información relacional se pierde
- Cuantas menos variables contengan los bloques de coherencia, menos riesgo de reidentificación…
  - …pero más información relacional se pierde

Paso 2: definir los "bloques de coherencia"

- Los bloques de coherencia son subconjuntos de variables con una **interdependencia relacional o semánticas fuerte**
- Queremos **conservar** esa información relacional en el proceso de anonimización
- Casos habituales:
  - Jerarquías (localidad, provincia, comunidad autónoma)
  - Pares codificación-descripciones (código-provincia, nombre-provincia)
  - Datos numéricos semánticamente relacionados (salario, importe retención IRPF)
    - Suelen utilizarse en fórmulas para calcular magnitudes derivadas

	VARIABLES PIVOTE		*BLOQUE 1*	*BLOQUE 2*	*BLOQUE 3*		
GRUPO	**TITULACIÓN**	**CENTRO**	**AÑO NACIMIENTO**	**GÉNERO**	**CRÉDITOS MATRICULADOS**	**CRÉDITOS PRESENTADOS**	**CRÉDITOS SUPERADOS**
1	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1980	HOMBRE	36	24	24
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1995	MUJER	48	48	48
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1990	HOMBRE	16	0	0
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1995	HOMBRE	60	36	36
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1980	MUJER	16	0	0
2	GRADO EN XENOMORFOLOGÍA	ESCUELA DE ASTROBIOLOGÍA	1990	MUJER	24	18	12
3	GRADO EN NIGROMANCIA	ESCUELA DE CIENCIAS FICTICIAS	1990	MUJER	36	36	30
	GRADO EN NIGROMANCIA	ESCUELA DE CIENCIAS FICTICIAS	1995	MUJER	12	12	6

Paso 3: Permutación

- Los datos de cada bloque de coherencia se permutan verticalmente de manera aleatoria **dentro de su grupo**.
- Tras este paso, en general ya no se puede afirmar que los datos de un sólo registro pertenecen al mismo individuo (ni lo contrario), **excepto:**
  - dentro de cada bloque de coherencia
  - Dentro del subconjunto formado por las variables pivote y cada bloque de coherencia

	VARIABLES PIVOTE		*BLOQUE 1*	*BLOQUE 2*	*BLOQUE 3*		
GRUPO	**TITULACIÓN**	**CENTRO**	**AÑO NACIMIENTO**	**GÉNERO**	**CRÉDITOS MATRICULADOS**	**CRÉDITOS PRESENTADOS**	**CRÉDITOS SUPERADOS**
1	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1990	HOMBRE	48	48	48
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1995	HOMBRE	16	0	0
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1980	MUJER	60	36	36
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1980	MUJER	16	0	0
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1995	HOMBRE	36	24	24
2	GRADO EN XENOMORFOLOGÍA	ESCUELA DE ASTROBIOLOGÍA	1990	MUJER	24	18	12
3	GRADO EN NIGROMANCIA	ESCUELA DE CIENCIAS FICTICIAS	1995	MUJER	12	12	6
	GRADO EN NIGROMANCIA	ESCUELA DE CIENCIAS FICTICIAS	1990	MUJER	36	36	30

Paso 4: Eliminación de grupos pequeños

- Las agrupaciones "pequeñas" con pocos registros, tienen un riesgo alto de permitir realizar inferencias estadísticas con niveles altos de certeza
- En el caso extremo (grupos "raros" con 1 solo registro) el proceso de anonimización no tiene absolutamente ningún efecto
- **Solución**: generalizamos, ocultando el valor de las variables pivote parcial o totalmente
  - Es decir, creamos un nuevo grupo "suma" de todos los grupos pequeños, entre los que se "diluyen" los datos de esos grupos.

	VARIABLES PIVOTE		*BLOQUE 1*	*BLOQUE 2*	*BLOQUE 3*		
GRUPO	TITULACIÓN	CENTRO	AÑO NACIMIENTO	GÉNERO	CRÉDITOS MATRICULADOS	CRÉDITOS PRESENTADOS	CRÉDITOS SUPERADOS
1	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1990	HOMBRE	48	48	48
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1995	HOMBRE	16	0	0
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1980	MUJER	60	36	36
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1980	MUJER	16	0	0
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1995	HOMBRE	36	24	24
2	GRADO EN <OCULTO>	ESCUELA <OCULTO>	1990	MUJER	24	18	12
	GRADO EN <OCULTO>	ESCUELA <OCULTO>	1995	MUJER	12	12	6
	GRADO EN <OCULTO>	ESCUELA <OCULTO>	1990	MUJER	36	36	30

¿Da soporte el dataset anonimizado a los siguientes análisis?

- Nº de estudiantes matriculados en la Universidad
- % de hombres y mujeres en la población estudiantil de la Universidad
- Edad media de los estudiantes de la Universidad
- Nº medio de créditos matriculados
- Tasa de presentación global (créditos presentados vs. créditos matriculados)
- Tasa de superación global (créditos superados vs. créditos presentados)

¿Da soporte el dataset anonimizado a los siguientes análisis?

- Nº de estudiantes matriculados en la Universidad
- % de hombres y mujeres en la población estudiantil de la Universidad

	VARIABLES PIVOTE		BLOQUE 1	BLOQUE 2	BLOQUE 3		
	TITULACIÓN	CENTRO	AÑO NACIMIENTO	GÉNERO	CRÉDITOS MATRICULADOS	CRÉDITOS PRESENTADOS	CRÉDITOS SUPERADOS
DATOS ANONIMIZADOS	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1990	HOMBRE	48	48	48
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1995	HOMBRE	16	0	0
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1980	MUJER	60	36	36
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1980	MUJER	16	0	0
	GRADO EN INFORMÁTICA	ESCUELA POLITÉCNICA	1995	HOMBRE	36	24	24
	GRADO EN <OCULTO>	ESCUELA <OCULTO>	1990	MUJER	24	18	12
	GRADO EN <OCULTO>	ESCUELA <OCULTO>	1995	MUJER	12	12	6
	GRADO EN <OCULTO>	ESCUELA <OCULTO>	1990	MUJER	36	36	30

	TITULACIÓN	CENTRO	AÑO NACIMIENTO	GÉNERO	CRÉDITOS MATRICULADOS	CRÉDITOS PRESENTADOS	CRÉDITOS SUPERADOS
DATOS ORIGINALES	*GRADO EN INFORMÁTICA*	*ESCUELA POLITÉCNICA*	*1980*	*HOMBRE*	*36*	*24*	*24*
	GRADO EN INFORMÁTICA	*ESCUELA POLITÉCNICA*	*1995*	*MUJER*	*48*	*48*	*48*
	GRADO EN INFORMÁTICA	*ESCUELA POLITÉCNICA*	*1990*	*HOMBRE*	*16*	*0*	*0*
	GRADO EN INFORMÁTICA	*ESCUELA POLITÉCNICA*	*1995*	*HOMBRE*	*60*	*36*	*36*
	GRADO EN INFORMÁTICA	*ESCUELA POLITÉCNICA*	*1980*	*MUJER*	*16*	*0*	*0*
	GRADO EN XENOMORFOLOGÍA	*ESCUELA DE ASTROBIOLOGÍA*	*1990*	*MUJER*	*24*	*18*	*12*
	GRADO EN NIGROMANCIA	*ESCUELA DE CIENCIAS FICTICIAS*	*1990*	*MUJER*	*36*	*36*	*30*
	GRADO EN NIGROMANCIA	*ESCUELA DE CIENCIAS FICTICIAS*	*1995*	*MUJER*	*12*	*12*	*6*

Inciso: seudonimización

- "*El tratamiento de datos personales de manera tal que **ya no puedan atribuirse a un interesado sin utilizar información adicional**, siempre que dicha información adicional figure por separado y esté sujeta a medidas técnicas y organizativas destinadas a garantizar que los datos personales no se atribuyan a una persona física identificada o Identificable*"

- **Seudonimizar**: sustituir los identificadores directos "naturales" por otro identificador directo "artificial", y no revelar (pero mantener) la correspondencia entre unos y otros.

Inciso: seudonimización

- Los datos de la tabla inferior son datos **seudonimizados**:

Nombre	Apellido1	Apellido2	DNI	F. nacimiento	Estado civil	Código
Juan	Pérez	López	00000001-Z	7/2/1930	Divorciado	**16384**

Código	Localidad	Género	Teléfono	Nivel de estudios
16384	Madrid	Hombre	640********	Licenciatura/Ingeniería

- Hemos extraído los identificadores directos a otra tabla, y los hemos sustituido por un código
- Los datos de la tabla inferior no pueden atribuirse a un individuo concreto sin conocer la correspondencia entre el código artificial y los identificadores directos (tabla superior)

¿Exime la seudonimización del cumplimiento del RGPD?

- ¡NO! Los datos seudonimizados se siguen considerando **datos personales:**
  - "*Los datos personales que hayan sido anonimizados, cifrados o* ***presentados con un seudónimo****, pero que puedan utilizarse para volver a identificar a una persona,* ***siguen siendo datos personales*** *y se inscriben en el ámbito de aplicación del RGPD.*"
- Cuestión clave: la "desidentificación" es **reversible** a través del código**.**

¿Entonces para qué sirve seudonimizar?

- Para **reducir riesgos** en determinados tipos de tratamientos de datos, por ejemplo:
  1. Encargo a otro departamento de mi organización que haga un determinado cálculo sobre un subconjunto de datos personales cuyos titulares no tiene por qué conocer
  2. Necesito alojar datos personales en *la nube,* pero quiero proteger a los titulares de los datos ante accesos no autorizados
- En general, casos en los que el responsable del tratamiento necesita poder volver a vincular los datos seudonimizados con los individuos de los que proceden:
  1. Vincular un resultado con el individuo adecuado
  2. Recuperar un dato almacenado externamente y asociarlo al individuo adecuado

Repetimos...

- ¿Los datos pseudonimizados se siguen considerando datos personales?

  ¡SÍ!

- ¿La seudonimización exime del cumplimiento del RGPD?

  ¡NO!

Contenido

1. Marco legal
2. Anonimización
3. Técnicas de anonimización
4. Ejemplo paso a paso
5. **Datos abiertos**

¿Qué son los datos abiertos?

- "***Datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona***, *y que se encuentran* ***sujetos, como mucho, al requerimiento de atribución*** *y de* ***compartirse de la misma manera en que se entregan***."

Open Knowledge Foundation

https://opendatahandbook.org/guide/es/what-is-open-data/

Principio de Accesibilidad universal:
Son accesibles para todos, sin restricciones de uso comercial o personal

Principio de Libre uso y redistribución:
Se puede pedir que se mencione la fuente original (atribución), pero no se pueden imponer restricciones fuertes.

Principio de Licencias abiertas:
Si se redistribuyen, debe mantenerse la misma licencia abierta. Ej. Creative Commons o similares.

Reutilización de la información del sector público (RISP)

- "***El uso por parte de personas físicas o jurídicas**, **de los datos generados y custodiados por los organismos del sector público**, **con fines comerciales o no**.”*

https://administracionelectronica.gob.es/pae_Home/pae_Estrategias/pae_Gobierno_Abierto_Inicio/pae_Reutilizacion_de_la_informacion_en_el_sector_publico.html

Cualquier persona (individual) o entidad (empresa, organización) puede acceder y usar esos datos.

Son datos producidos por administraciones públicas (ministerios, ayuntamientos, organismos oficiales) y que están bajo su responsabilidad.

Se pueden usar para cualquier propósito:

- Comercial: crear productos, servicios, aplicaciones (por ejemplo, apps de transporte usando datos abiertos).
- No comercial: investigación, periodismo, proyectos educativos.

Reutilización de la información del sector público (RISP)

- **Ejemplo práctico**

- Datos meteorológicos publicados por AEMET:
  - Una empresa puede usarlos para crear una app de predicción del tiempo (fines comerciales).
  - Un investigador puede analizarlos para estudiar el cambio climático (fines no comerciales).
  - Puedes descargarlo, crear gráficos, usarlo en una app.
  - Solo debes citar la fuente y mantener la misma licencia si lo redistribuyes.

Sector infomediario

- Empresas y organizaciones que toman, en gran medida, **datos públicos (abiertos) y los transforman en productos, servicios o aplicaciones con valor añadido** para terceros.
  - También pueden hacer uso de datos de fuentes cerradas (comerciales)
  - Sus productos o servicios suelen estar dirigidos a otras empresas, pero también hacia la ciudadanía en general

- **Ejemplos** de empresas y organismos que han creado soluciones novedosas en base al uso, entre otros, de datos generados por las administraciones públicas.
  - https://datos.gob.es/es/empresas

Sector infomediario. Ejemplos

- Apps de movilidad que reutilizan datos abiertos
- Utilizan datos abiertos publicados por administraciones, como horarios de transporte, ubicación de paradas.

  - **Moovit –** Plataforma de movilidad urbana que usa datos abiertos de transporte público.
    - https://moovitapp.com
  - **Citymapper –** App que integra datos abiertos para planificar rutas multimodales.
    - https://citymapper.com
  - **Blinkay –** Startup española para gestión inteligente de aparcamiento.
    - https://blinkay.com

Sector infomediario. Ejemplos

- Empresas que ofrecen informes económicos combinando datos públicos oficiales (Registro Mercantil, BOE) con datos privados y análisis propios.

  - **Infoempresa** – Información financiera y mercantil sobre empresas.
    - https://www.infoempresa.com
  - **Iberinform** – Informes comerciales y financieros basados en datos públicos y privados.
    - https://www.iberinform.
  - **Informa D&B** – Información empresarial y análisis de riesgo.
    - https://www.informa.es
  - **eInforma** – Informes de empresas y análisis sectoriales.
    - https://www.einforma.com

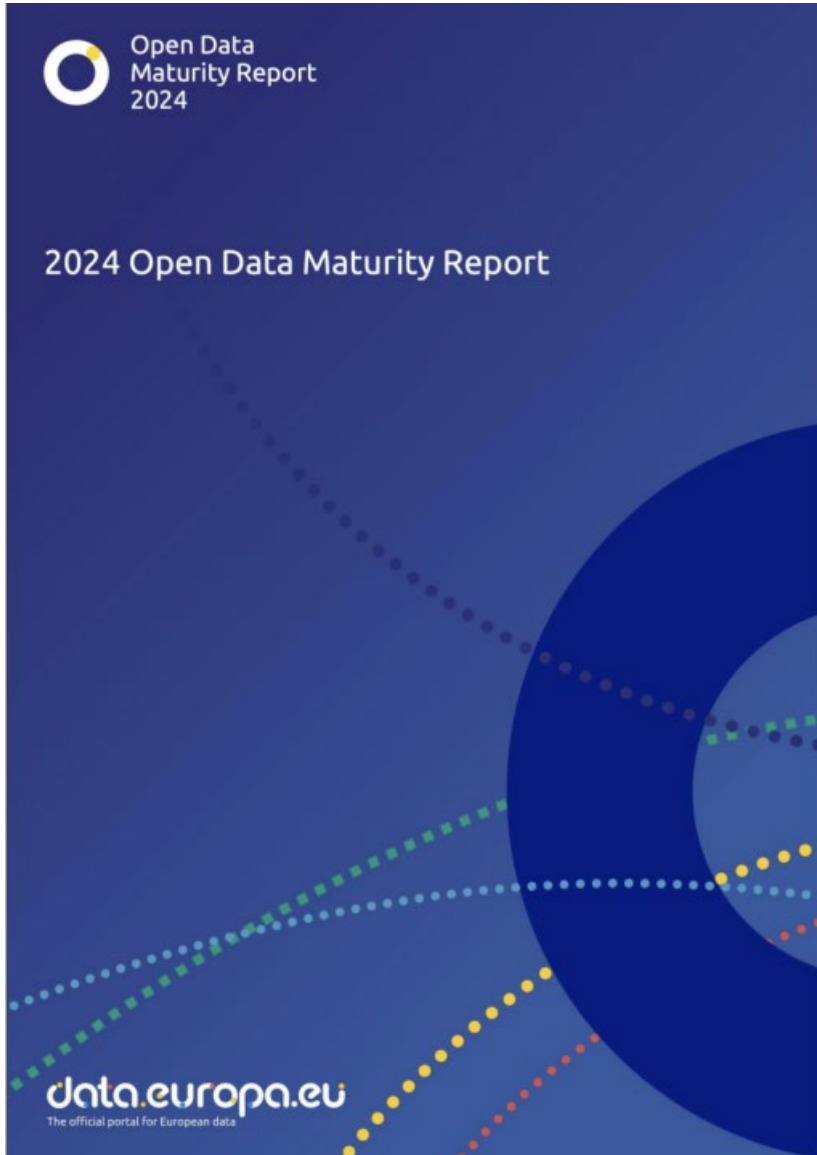
Sector infomediario. Ejemplos

- Empresas que usan datos meteorológicos (AEMET, Copernicus) junto con datos de sensores privados para agricultura o logística

  - **Cordulus** – Datos meteorológicos hiperlocales para agricultura de precisión.
    - https://www.cordulus.com/es
  - **Sencrop** – Red de estaciones meteorológicas conectadas para cultivos.
    - https://sencrop.com/es

España, potencia en datos abiertos

- Desde 2015, la UE publica anualmente el **Open Data Maturity Report** sobre madurez en datos abiertos para los Estados miembros
  - https://data.europa.eu
- España tiene una posición destacada desde hace algunos años
  - 10 March 2025: https://data.europa.eu/en/open-data-maturity/2024
  - **Sexto puesto global** (quinto entre los países de la UE) **en 2024**
    - En el grupo de países *"trendsetters"* (creadores de tendencias)
  - **Primera** posición en la dimensión *Open Data Impact*



España, potencia en datos abiertos

- Portal Nacional de Datos Abiertos: datos.gob.es

datos.gob.es
reutiliza la información pública

Ejemplo: *Location Intelligence*

- *¿Cuál es la zona más idónea en Madrid para abrir una nueva escuela infantil*?
- Sería estupendo que tuviéramos acceso a...
  - Las pirámides de población actuales y proyectadas en cada barrio
  - La ubicación de la oferta pública de escuelas infantiles
  - La ubicación de la oferta privada de escuelas infantiles
  - La ubicación de parques y zonas de juego infantil
  - La capacidad adquisitiva de cada barrio
  - Cómo se mueve la gente entre barrios para ir a trabajar
    - ...¡pues **lo tenemos!** ¡esos datos **están abiertos** por parte de distintas administraciones!

UAM | Escuela Politécnica Superior

Datos personales y anonimización

Francisco Jurado, francisco.jurado@uam.es