Introducción a la creación de un conjunto de datos para *machine learning*
Introduction to constructing a dataset for machine learning

G. Santolària Rossell†, M. Olivera, C. Méndez Mangana, Antón Barraquer Kargacin

Resumen

Cualquier proyecto de *data science* (ciencia de los datos) necesita de un conjunto de datos o *dataset*. Por eso se dice que la piedra angular de todo proyecto de *machine learning* (aprendizaje automático) y *deep learning* (aprendizaje profundo) es el conjunto de datos. Su desarrollo ocupa aproximadamente el 70% del tiempo total que se dedica a todo un proyecto. Esto se debe a que una mala generación del conjunto de datos puede resultar en modelos ineficaces o sesgados. A través del ejemplo de nuestro grupo de investigación, se detalla la generación de un *dataset* oftalmológico con imágenes para el entrenamiento de un modelo de *machine learning* de clasificación. Finalmente, se presentan casos de la literatura a modo de ejemplo, para que el lector pueda extender su conocimiento con casos de éxito.

Palabras clave: *Dataset.* Conjunto de datos. *Deep learning. Machine learning.*

Resum

Tot projecte de *data science* necessita un conjunt de dades o *dataset*. Per això es diu que la pedra angular de tot projecte de *machine learning* i *deep learning* és el conjunt de dades. El seu desenvolupament ocupa aproximadament el 70% del temps total que es dedica a tot un projecte. Això es deu al fet que una mala generació del conjunt de dades pot resultar en models ineficaços o esbiaixats. A través de l'exemple del nostre grup de recerca es detalla la generació d'un *dataset* oftalmològic amb imatges per a l'entrenament d'un model de *machine learning* de classificació. Finalment, es presenten casos de la literatura a tall d'exemple, perquè el lector pugui estendre el seu coneixement amb casos d'èxit.

Paraules clau: *Dataset.* Conjunt de dades. *Deep learning. Machine learning.*

Abstract

Any data science project relies on its dataset. That is why it is said that the backbone of any machine learning/deep learning project is its dataset. Its development takes up approximately 70% of the total time spent on a data science project. Poor dataset generation can result in an inefficient or biased model. Through this example from our research team, we provide the generation of an ophthalmological dataset with classification images to train a machine learning model. To finish, we provide examples from literature so that the reader can extend their knowledge of the subject in question.

Key words: Dataset. Data set. Deep learning. Machine learning. Keratoplasty.

1.3. Introducción a la creación de un conjunto de datos para *machine learning*

Introduction to constructing a dataset for machine learning

G. Santolària Rossell†, M. Olivera[1], C. Méndez Mangana[2], Antón Barraquer Kargacin[3]

[1]Complejo Hospitalario Universitario Insular Materno Infantil. Instituto Canario de la Retina. Las Palmas de Gran Canaria. [2]Centro de Ojos de La Coruña. [3]Centro deOftalmología Barraquer. Barcelona.

Correspondencia:
Maximiliano Olivera
E-mail: mxolivera@gmail.com

En memoria del Ing. Gil Santolària Rossell (1993-2023), especialista en Inteligencia Artificial. Coautor de capítulos de la presente ponencia, publicaciones científicas y congresos internacionales del grupo de investigación en IA del Centro de Oftalmología Barraquer. Además de sus logros profesionales, y siempre dispuesto a ayudar a los demás, Gil era una persona maravillosa con quien tendremos imborrables recuerdos y viajes compartidos. Tu legado perdurará en cada uno de nosotros. Descansa en paz amigo.

El conjunto de datos

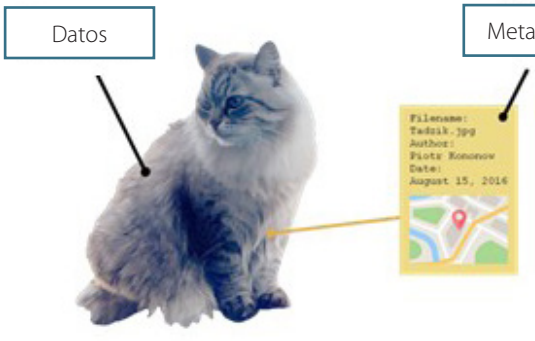
En cualquier proyecto de *machine learning*, el recurso más importante es el *dataset*. En este capítulo, introduciremos este concepto, su estructura, y analizaremos varios ejemplos de generación y uso de diferentes conjuntos de datos en casos reales (literatura).

¿Qué se entiende por conjunto de datos?

Definición

El concepto de conjunto de datos se refiere a la agrupación de datos vinculados a una temática común. Podemos observar la implementación de esta definición por parte de la Unión Europea en la descripción de los *high-value datasets* (conjuntos de datos de valor elevado)[1], que son datasets de interés público. Para comprender mejor la idea de conjunto de datos, vamos a definir el concepto de dato y el de su complementario, el metadato.

Un dato es una representación simbólica de la información en lenguaje de ordenador[2,3]. Por otro lado, los metadatos son datos que forman una descripción estandarizada de las características de un conjunto de datos[4], que comúnmente se describen como datos que definen otros datos. Para dar un ejemplo, una imagen en formato ".jpeg", ".jpg", ".png", etc., sería un dato, y sus metadatos serían la geolocalización, el tamaño, sus dimensiones, etc. (Figura 1).



Fuente: Dataedo[5].

Figura 1. Dato y metadato.

La correcta generación de un conjunto de datos puede suponer el éxito o el fracaso de un proyecto de *machine learning,* de ahí que se considere el pilar de todo proyecto de *data science*[6].

El ciclo de vida de los datos

Para entender un conjunto de datos, primero debemos introducir y analizar el concepto de "ciclo de vida de los datos", que hace referencia a las diferentes etapas que atraviesan los datos, desde su generación y captura hasta el archivado o destrucción[7]. Poder entender y analizar las fases que han permitido la generación de un conjunto de datos puede aportar información/pistas sobre posibles sesgos y/o características intrínsecas en los datos. Por esta razón, documentar este proceso, ya sea generando una memoria del *dataset* o un simple diario, es una práctica muy recomendable. En la literatura, se hallan variaciones en el número o en la denominación de dichas fases, la representación más común[7] es la siguiente (Figura 2):

1. Captura.
2. Preprocesado.
3. Almacenamiento.
4. Análisis.
5. Visualización.
6. Publicación.
7. Archivado o destrucción.

En este apartado, nos centraremos en las tres primeras fases que modelan la generación de un conjunto de datos, mientras que en apartados posteriores, introduciremos las fases restantes del proceso:

- La fase de captura se caracteriza por la recopilación de todos los datos de interés. Esta puede darse por dos eventos: por la propia creación de los datos (integrando los datos, en la misma fuente de generación) o bien por extracción. Esto último acontece cuando no se tiene acceso a la fuente y debemos realizar una búsqueda manual de los datos (en el registro, historia clínica, banco de imágenes, etc.).
- La fase de preprocesado es la fase más importante en cuanto a la generación de un conjunto de datos. Trata de homogeneizar los datos tanto en estructura como en tipo. De su buena ejecución depende el correcto desempeño del modelo.
- La fase de almacenamiento, como su nombre indica, define el formato de almacenamiento de los datos. Se va a tener en cuenta la tipología de los datos y el uso posterior que se va a hacer de estos. Es importante recordar que el formato elegido debe facilitar su manipulación posterior.

El proceso descrito anteriormente se conoce como extraer-transformar-cargar (ETL, *extract-transform-load*).

Calidad de los datos

Un aprovechamiento óptimo de los datos recae en su calidad, siendo fundamental para que el desarrollo del proyecto sea fiable e imparcial. La causa más frecuente de desconfianza en un estudio de este tipo es la mala calidad de los datos[7]. El término "calidad de los datos" es comúnmente empleado por organizaciones para referirse al grado de confianza que debemos depositar sobre ellos. Así mismo, desde el ámbito empresarial, se percibe una tendencia en considerar los datos como activos de valor para la empresa. Tres características o dimensiones son los principales determinantes de la calidad de los datos[7]: precisión *(accuracy),* completitud *(completeness)* y atemporalidad *(timeliness)*:

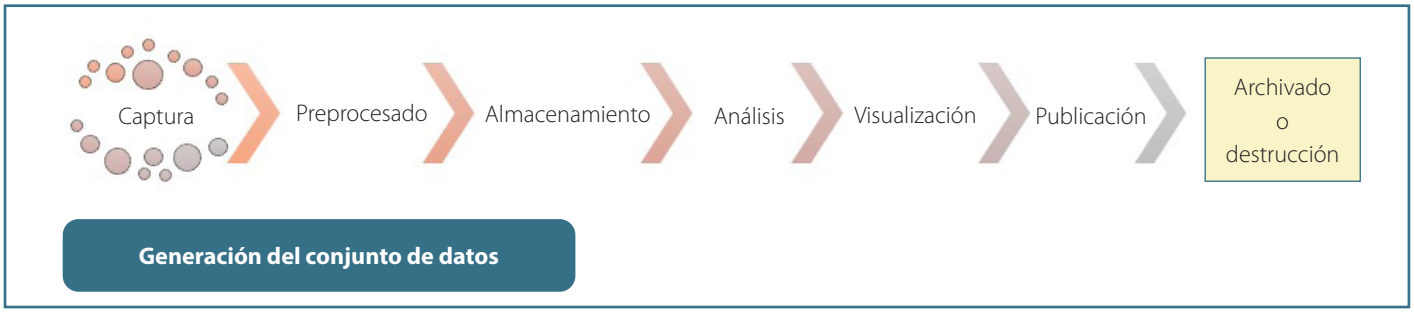


Figura 2. Ciclo de vida de los datos: generación del dataset. (Fuente: propia).

- Por precisión nos referimos a si la información recogida es o no correcta. Es decir, a modo de ejemplo, que no haya ninguna incongruencia en el uso del punto o la coma como marca decimal (en caso de datos numéricos), calidad o representatividad de la imagen, etc. De igual importancia es que los datos no contengan entradas repetidas. Así mismo, en el caso de imágenes, su etiquetado (asignación de categoría) ha de ser congruente en la totalidad del conjunto de datos.
- La completitud de los datos hace referencia a que no falta información. En otras palabras, debemos asegurar que todos los registros capturados estén completos. La escasez de información, de forma generalizada para una característica o atributo del conjunto de datos, suele traducirse en la imposibilidad de usar dicha característica en el desarrollo futuro de un modelo.
- Finalmente, la atemporalidad hace alusión al grado en que los datos representan la realidad en un momento preciso/específico y, también, que algunos de estos datos pueden quedar obsoletos. Pese a que esta característica no se aplica en la mayoría de los datos capturados en el sector médico (en el caso de imágenes), debemos ser conscientes de que los datos tienen una componente temporal. Por ejemplo, sin este componente temporal los datos relacionados con transacciones pierden su sentido.

Caso real: cómo constituir un conjunto de datos

En este apartado, se describen los pasos de la generación de un *dataset* para el entrenamiento de un modelo (supervisado) de *machine learning* para la clasificación binaria de imágenes, en base a nuestra experiencia con el conjunto de datos generado para el estudio: *"Detection of graft failure in post-keratoplasty patients by Automated Deep Learning"*[8].

Captura de los datos

La captura de los datos se hizo por extracción de dos fuentes distintas. Ambas utilizadas en la creación de un capítulo sobre seguimiento a largo plazo de los trasplantes corneales, incluido en la ponencia de la Sociedad Española de Oftalmología "Queratoplastias: nuevas técnicas para el siglo XXI"[9]. La generación del *dataset* requiere la unión de la información del paciente con imágenes del segmento anterior. La primera fuente de información son datos tabulares (presentados como tablas), donde cada fila representa una intervención de queratoplastia total y las columnas representan los atributos o características de dicha intervención (causa, injertos previos, edad, presencia de vascularización, etc.).

Por otra parte, la totalidad de las imágenes extraídas de la segunda fuente de información, el banco de imágenes del Centro de Oftalmología Barraquer, estaban en formato digital ".jpeg". En el hipotético caso de que algunas imágenes estuvieran en formato físico, estas deberán ser digitalizadas acorde con la resolución de las imágenes digitales disponibles. Siendo el objetivo del estudio la detección del fracaso del injerto (problema de clasificación), se utilizaron las notas clínicas para la determinación del estado final (etiqueta), siendo en este caso: fallo del injerto/injerto sano.

Los datos en formato de tabla fueron extraídos con facilidad (solicitud y recepción del archivo). Por lo que respecta a la extracción de las imágenes, se organizó por fases, debido a la presencia de pacientes con múltiples intervenciones. Las fases se organizaron según el orden de intervención, hasta recolectar todos los registros (imágenes) presentes en el archivo en formato tabular. Finalmente, las imágenes se almacenaron en carpetas según la fase y fueron comprobadas por dos oftalmólogos de segmento anterior.

Al final del proceso, se obtuvieron un total de 593 imágenes, que representan 593 registros, con 42 atributos cada uno del archivo de datos tabulares.

Preprocesado de los datos

Unificación de los datos

Hasta este momento, no se ha hecho uso del concepto "conjunto de datos" para referirnos a los datos mencionados en el apartado anterior. La razón es que aún no se han unificado los datos y, por lo tanto, no se han relacionado las imágenes obtenidas con los registros presentes en el archivo de datos tabulares. El proceso de unificación se ha automatizado mediante un *script* (instrucciones escritas en lenguaje de programación) Python (lenguaje de programación); generando el *dataset* de queratoplastias penetrantes que contiene datos tabulares e imágenes. Cabe señalar que el proceso de unificación puede realizarse de manera manual, sin embargo, puede resultar un proceso largo y tedioso, incluso con pocos datos.

El paso siguiente a la unificación, dejando de lado las imágenes, es el análisis de los 42 atributos presentes en el dataset generado. Se observó que la mayoría de los atributos no aportaban información, porque consistían en agrupaciones de otros atributos. Vamos a ejemplificar el concepto de agrupaciones a partir de las agrupaciones halladas con respecto al atributo "edad":

- Se observa una agrupación por rangos de años (11-30, 31-50, 51-70, más de 70).
- Se observa una agrupación que elimina las edades menores a 10 años.

Estas agrupaciones, aunque pueden ser muy útiles en estudios futuros, dificultan la comprensión del conjunto de datos. De modo que debemos centrarnos en obtener un resultado lo más limpio y accesible posible, sin perder información. Con esta limpieza, obtuvimos un total de 593 registros con 22 atributos, reduciendo casi a la mitad las características sin perder información.

Anonimización de los datos

Este procedimiento puede permitir la publicación del *dataset* para uso público, aunque en este caso aún no se ha llevado a cabo. La finalidad del proceso de anonimización es reducir al mínimo la posibilidad de reidentificación de los datos anonimizados y, a su vez, mantener la veracidad de los resultados del análisis[10,11].

El proceso se inicia con la detección de esas variables que permiten una identificación, directa o indirecta, del individuo. Según la Agencia Española de Protección de Datos, en un conjunto de datos, se pueden hallar[10]:

- Identificadores directos: todas aquellas características que, por sí mismas, permiten la identificación del individuo.
- Identificadores indirectos: aunque por sí solos no identifican una persona, el cruce de varios identificadores podría permitir su identificación.
- Datos especialmente sensibles: tal como datos financieros, infracciones, etc.

La Figura 3, que se presenta a continuación, ilustra un caso habitual de anonimización de datos médicos.

Con lo que respecta a las imágenes, estas también deben ser anonimizadas. Para ello, se ha desarrollado un *script* Python automatizado. El proceso renombra todas las imágenes disponibles asignando un nombre aleatorio distinto a cada una de ellas. El formato del nombre que se ha seleccionado es el siguiente:

P-XXXX.jpg

Donde XXXX es un número entero mayor que cero y menor a 10.000.

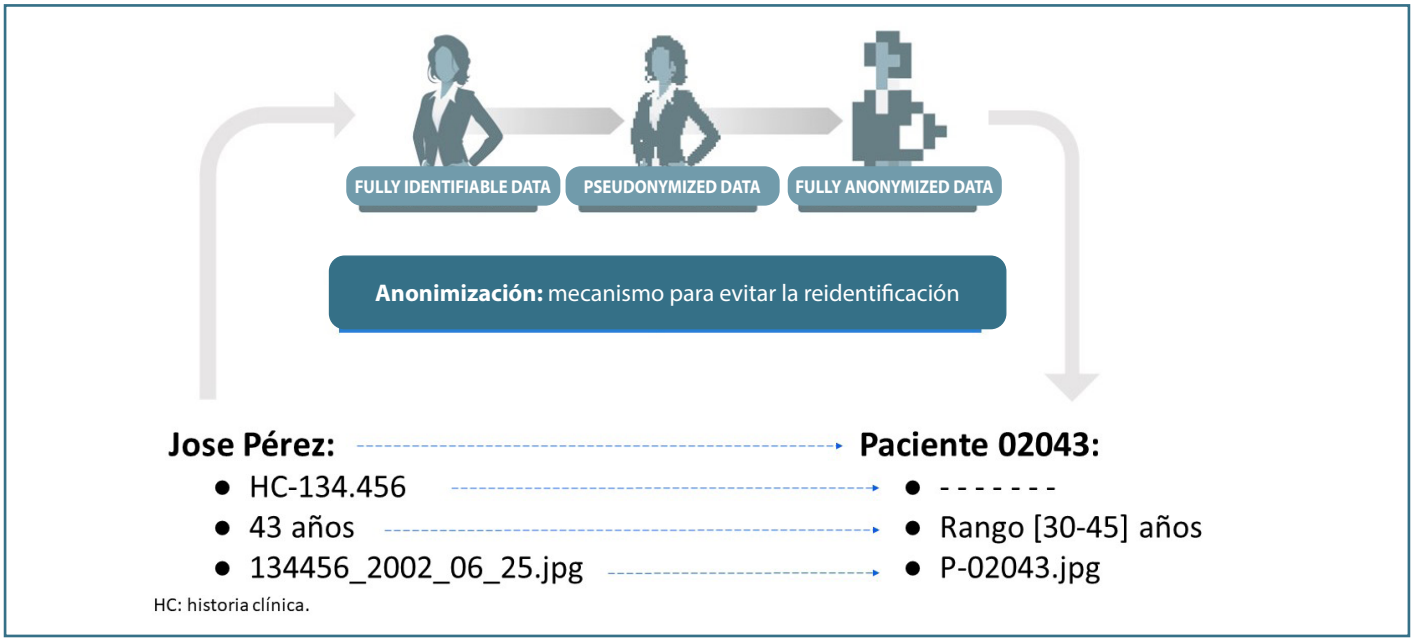


Figura 3. Esquema de la anonimización de los datos. (Fuente: Kisić12, con modificaciones).

Del *dataset* al análisis exploratorio de datos, una introducción al análisis exploratorio de datos

El análisis exploratorio de datos (AED), comprende el análisis y el desarrollo de visualizaciones de los datos. El AED se realiza a *posteriori* de la generación del conjunto de datos, siendo un paso imprescindible previo a la generación de cualquier modelo de *machine learning* (Figura 4).

El AED recoge un conjunto de técnicas estadísticas que permiten explorar, describir y resumir la naturaleza de los datos, con el objetivo de garantizar resultados posteriores objetivos y un intercambio de información veraz[13] (Figura 5).

La aplicación de este conjunto de técnicas permite:

- Detectar valores atípicos *(outliers)*.
- Normalizar atributos.
- Revelar posibles errores.
- Estudiar las correlaciones (la relación entre variables).
- Realizar representaciones gráficas para facilitar el análisis descriptivo.
- Resumir los aspectos más relevantes del conjunto de datos.

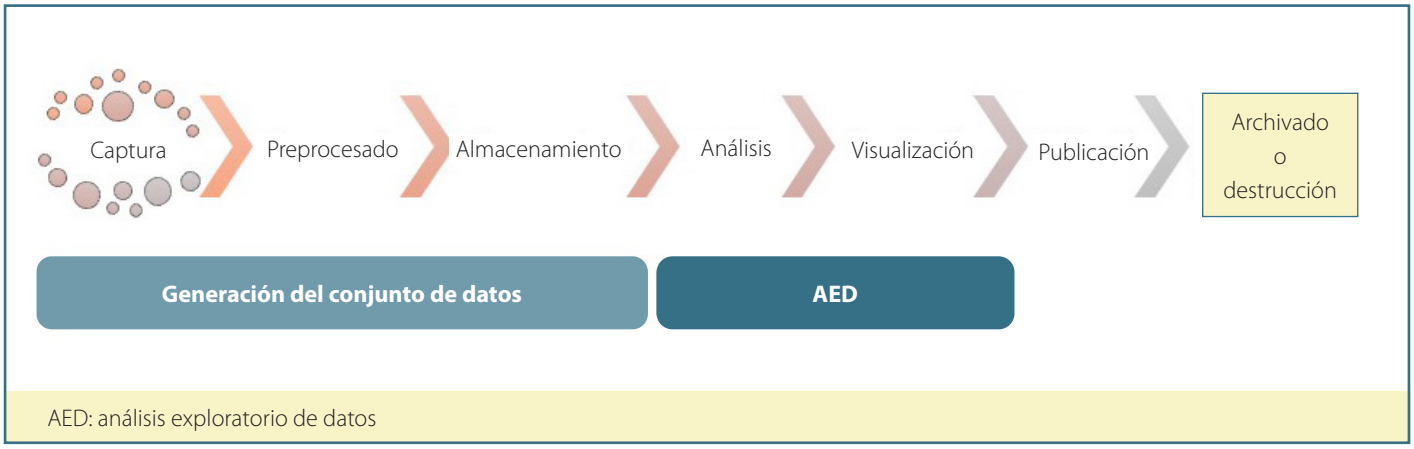


Figura 4. Ciclo de vida de los datos: análisis exploratorio de datos. (Fuente: propia).

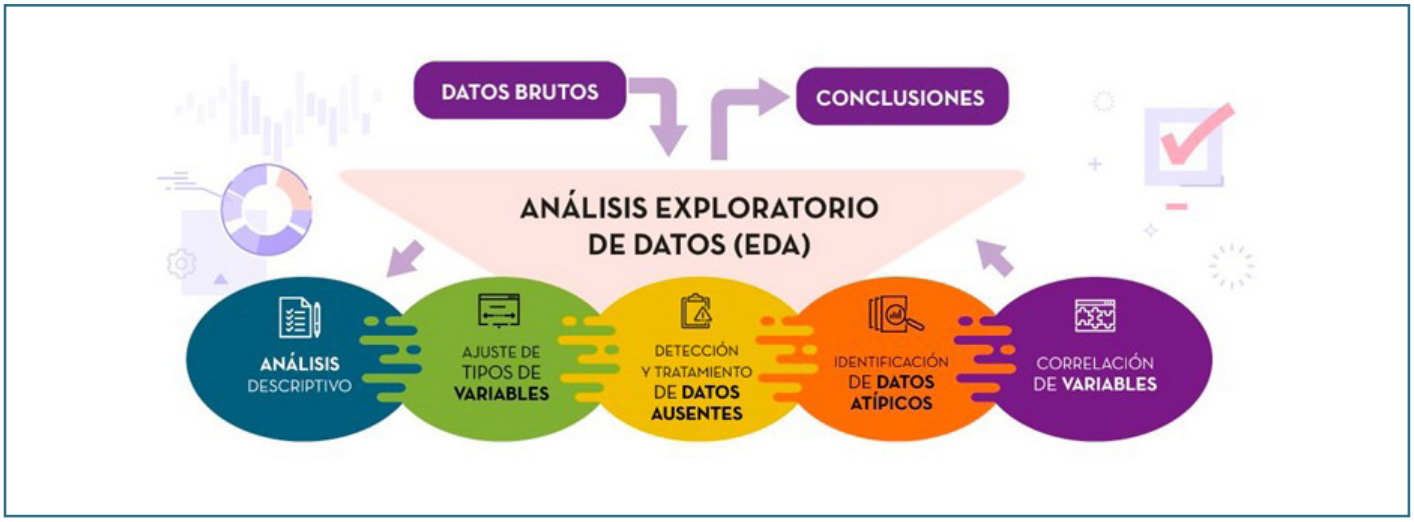


Figura 5. Esquema del análisis exploratorio de datos. (Fuente: *Guía Práctica de Introducción al Análisis Exploratorio de Datos*[13]).

Tendencias en el campo de la salud

Federated Learning

El mercado mundial de *big data* (macrodatos o análisis masivo de datos) relacionado con la sanidad se calcula que alcanzará los 70.000 millones de dólares estadounidenses en 2025, lo cual representa un crecimiento del 508,7% respecto al año 2016[14]. Este crecimiento esperado se debe, en parte, al desarrollo de tecnologías como el *federated learning* (aprendizaje federado). Dicha tecnología permite la generación de visualizaciones y el entrenamiento de herramientas de *machine learning* de forma remota. Su principal ventaja es que permite usar datos entre localizaciones remotas sin una transferencia manual de datos entre instituciones. Posibilita la colaboración directa entre investigadores pertenecientes a distintas instituciones, sin comprometer la anonimidad de los datos y evitando los costes asociados de computación[15].

Literatura: creación de un dataset al detalle

A continuación, se van a presentar diferentes proyectos de creación de conjuntos de datos que podemos hallar en la literatura que engloban diferentes tipos de datos: datos en formato tabular, imagen y vídeo. Los procesos se describen con detalle, desde la captura de los datos hasta la publicación del *dataset*.

Estudio *"A novel dataset and deep learning-based approach for marker-less motion capture during gait"*

El estudio[16] presenta una visión novedosa en la generación de un *dataset* de secuencias de vídeo, con el fin de aplicar el aprendizaje profundo. Tiene como objetivo el desarrollo de métodos de estimación de la postura humana. Aunque se utilice un número reducido de sujetos (19 hombres y 12 mujeres), la metodología de generación del *dataset* se describe de forma clara, precisa y extensa.

Los datos se han capturado en formato vídeo a través de cuatro cámaras de acción (GoPro Hero 7 Black), obteniendo vídeos con una resolución de 1.920 × 1.080 pixels y 100 *frames* por segundo. También se ha utilizado un sistema basado en marcadores, en concreto, 12 cámaras VICON System, que han permitido registrar las posiciones del sujeto.

Trabajo *"Toward a Comprehensive Domestic Dirt Dataset Curation for Cleaning Auditing Applications"*

El trabajo[17] expone un conjunto de datos de imágenes de suciedad doméstica, con el objetivo de usarlo como auditoría de limpieza, incluyendo el análisis de la suciedad basado en inteligencia artificial y la inspección de limpieza asistida por robots.

Se presenta desde la creación del sistema robótico, que permite la adquisición de las imágenes, hasta el proceso de etiquetado de estas, con el objetivo final de entrenar un modelo de *machine learning* que permita revelar la composición de la suciedad presente en las imágenes. Dicho modelo consigue distinguir entre nueve tipos de suciedad. El *dataset* generado contiene 3.000 imágenes capturadas con un microscopio.

Conjunto de datos *"Data Resource Profile: National Cancer Registration Dataset in England"*

El *dataset* presentado[18] agrupa datos desde 1971, concretamente, sobre todas las personas residentes en Inglaterra diagnosticadas de neoplasias malignas y premalignas. Este conjunto de datos ha sido realizado por el *National Cancer Registration and Analysis Service* (NCRAS), que forma parte del *Public Health England* (PHE).

En este conjunto de datos, se agregan datos en formato tabular. Por otra parte, se asegura la calidad y el análisis de los datos mediante un enfoque poblacional. Además, se detallan las fases del "ciclo de vida de los datos", desde la captura hasta su publicación, y se tratan de forma extensiva los conceptos derivados de calidad de los datos, aportando información única al *dataset*.

Conjunto de datos *"Data Resource Profile: The Systemic Anti-Cancer Therapy (SACT) dataset"*

Nos encontramos con un segundo conjunto de datos poblacionales[19]. Se recoge la actividad de la terapia sistémica contra el cáncer reportada por el *National Health Service* (NHS) en Inglaterra, con el objetivo de proporcionar datos para apoyar y mejorar la toma de decisiones clínicas.

El estudio presenta un ejemplo de recolección de datos, detallando su estructura, y se enfoca en la realización del AED.

Conclusiones

Como se ha intentado exponer anteriormente, la creación de un conjunto de datos es un proceso largo y meticuloso que engloba

muchos procedimientos, desde la captura inicial de los datos hasta su momento de uso real. El éxito del proyecto de inteligencia artificial no solo radica en la forma en la que empleamos los datos de los que disponemos (historias clínicas y resultados de pruebas diagnósticas-terapéuticas), sino también en la planificación y razonamientos previos que llevan a la creación del conjunto de datos.

Bibliografía

1. Unión Europea. EUR-Lex Document 32023R0138: Commission Implementing Regulation (EU) 2023/138 of 21 December 2022 laying down a list of specific high-value datasets and the arrangements for their publication and re-use (Text with EEA relevance). [Internet]. *DOUE*. 2023;19:43-75. [Citado 31 May 2023]. Disponible en: http://data.europa.eu/eli/reg_impl/2023/138/oj
2. Real Academia Española. Dato. En: [Internet]. *Diccionario de la lengua española*. 23ª ed. Espasa Calpe; 2014. [Citado 6 Jun 2023]. Disponible en: https://dle.rae.es/dato
3. Wikipedia contributors. Dato (informática) [Internet]. En: Wikipedia, The Free Encyclopedia. [Citado 6 Jun 2023]. Disponible en: https://es.wikipedia.org/wiki/Dato_(informática)
4. Ministerio de la Presidencia. Real Decreto 1708/2011, de 18 de noviembre, por el que se establece el Sistema Español de Archivos y se regula el Sistema de Archivos de la Administración General del Estado y de sus Organismos Públicos y su régimen de acceso. [Internet]. *BOE*. 2011;284. [Actualizado 16 Feb 2022]. [Citado 31 May 2023]. Disponible en: https://www.boe.es/eli/es/rd/2011/11/18/1708/con
5. What is Metadata (with examples) - Data terminology. [Internet]. En: Dataedo.com. Dataedo. [Citado 6 Jun 2023]. Disponible en: https://dataedo.com/kb/data-glossary/what-is-metadata
6. State of Data Science 2022: Paving the Way for Innovation. [Internet]. Anaconda; 2022. [Citado 31 May 2023]. Disponible en: https://know.anaconda.com/rs/387-XNW-688/images/ANA_2022SODSReport.pdf?mkt_tok=Mzg3LVhOVy02ODgAAAGK5xO1gHPIVLxSNz2R_BSRTV5F-OxSDZ0O6X3rLbt_4dTSYfP4NPZI7TKFAbd8FDQReOlSBFSgU6VqNVatNCGyLGztbm6X7S7G0K4EUSe9DDHI
7. Eryurek E, Gilad U, Lakshmanan V, Kibunguchy A, Ashdown J. *Data Governance: The Definitive Guide: People, Processes, and Tools to Operationalize Data Trustworthiness*. O'Reilly Media, Inc.; 2021. pp. 251.
8. Méndez Mangana C, Barraquer Kargacin A, Fernández-Engroba J, Taña P, Santolaria G, Olivera M, *et al*. Detection of graft failure in post-keratoplasty patients by Automated Deep Learning. *Invest Ophthalmol Vis Sci*. 2022;63(7):2330.
9. Barraquer R, De Toledo J. Queratoplastias: nuevas técnicas para el siglo XXI: 92 ponencia oficial de la Sociedad Española de Oftalmología de 2009. SEO; 2016. pp. 623.
10. ¿Cómo afecta GDPR a los datos personales abiertos? [Internet]. En: Datos.gob.es. Ministerio de Asuntos Económicos y Transformación Digital. 11 Abr 2018. [Citado 7 Jun 2023]. Disponible en: https://datos.gob.es/ca/noticia/como-afecta-gdpr-los-datos-personales-abiertos
11. Agencia Española de Protección de Datos. [Internet]. [Citado 7 Jun 2023]. Disponible en: https://www.aepd.es/es
12. Kisić M. Imamo li pravo na (pseudo)anonimnost? [Internet]. En: Securitysee.com. Security SEE - Magazin za bezbednost; 2022. [Citado 7 Jun 2023]. Disponible en: https://www.securitysee.com/2022/08/imamo-li-pravo-na-pseudoanonimnost/
13. *Guía Práctica de Introducción al Análisis Exploratorio de Datos*. [Internet]. En: Datos.gob.es. Ministerio de Asuntos Económicos y Transformación Digital. 22 Sep 2021. [Citado 7 Jun 2023]. https://datos.gob.es/es/documentacion/guia-practica-de-introduccion-al-analisis-exploratorio-de-datos
14. Global healthcare big data market size in 2016 and a forecast for 2025. [Internet]. En: Statista.com. *Statista*. [Citado 6 Jun 2023]. Disponible en: https://www.statista.com/statistics/909654/global-big-data-in-healthcare-market-size/
15. Nguyen A. *Hands-On Healthcare Data: Taming the Complexity of Real-World Data*. O'Reilly Media, Inc; 2022. pp. 242.
16. Vafadar S, Skalli W, Bonnet-Lebrun A, Khalifé M, Renaudin M, Hamza A, *et al*. A novel dataset and deep learning-based approach for marker-less motion capture during gait. *Gait Posture*. 2021;86:70-6. [Citado 31 May 2023].
17. Pathmakumar T, Elara MR, Soundararajan SV, Ramalingam B. Toward a Comprehensive Domestic Dirt Dataset Curation for Cleaning Auditing Applications. *Sensors*. 2022;22(14):5201.
18. Henson KE, Elliss-Brookes L, Coupland VH, Payne E, Vernon S, Rous B, *et al*. Data Resource Profile: National Cancer Registration Dataset in England. *Int J Epidemiol*. 2020;49(1):16.
19. Bright CJ, Lawton S, Benson S, Bomb M, Dodwell D, Henson KE, *et al*. Data Resource Profile: The Systemic Anti-Cancer Therapy (SACT) dataset. *Int J Epidemiol*. 2020;49(1):15.