



ugr | Universidad
de Granada

TRABAJO FIN DE GRADO
GRADO EN INGENIERÍA INFORMÁTICA

Cuantificación de la incertidumbre de las
predicciones de modelos de aprendizaje
automático en problemas de estimación
del perfil biológico

Autor
David González Durán

Director
Pablo Mesejo Santiago

Mentor
Javier Venema Rodríguez



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE
TELECOMUNICACIÓN

—
Granada, mes de 2025

Cuantificación de la incertidumbre de las predicciones de modelos de aprendizaje automático en problemas de estimación del perfil biológico

David González Durán

Palabras clave: palabra_clave1, palabra_clave2, palabra_clave3, ...

Resumen

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Quantification of the uncertainty in machine learning model predictions for biological profile estimation problems

David González Durán

Keywords: Keyword1, Keyword2, Keyword3, ...

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Yo, **David González Durán**, alumno de la titulación **TITULACIÓN de la Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada**, con DNI 32071015E, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Grado en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: David González Durán

Granada, a X de mes de 202.

D. **Pablo Mesejo Santiago**, Profesor del Área de XXXX del Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada.

D. **Javier Vénema Rodríguez**, Esdudiente de Doctorado del programa de Tecnologías de la Información y de la Comunicación e investigador en Inteligencia Artificial en Panacea Cooperative Research.

Informan:

Que el presente trabajo, titulado *Cuantificación de la incertidumbre de las predicciones de modelos de aprendizaje automático en problemas de estimación del perfil biológico*, ha sido realizado bajo su supervisión por **David González Durán**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada a X de mes de 2025.

Los directores:

Pablo Mesejo Santiago

Javier Vénema Rodríguez

Agradecimientos

Poner aquí agradecimientos...

Índice general

1. Introducción	1
1.1. Descripción del problema	1
1.1.1. Limitaciones de la antropología forense	2
1.2. Motivación	5
1.3. Objetivos	8
1.4. Planificación económica y temporal del proyecto	9
2. Fundamentos teóricos	11
2.1. Machine Learning	11
2.1.1. Problemas de regresión	12
2.1.2. Problemas de clasificación	18
2.1.3. Selección de modelos y optimización	21
2.2. Deep Learning	22
2.2.1. El perceptrón multicapa	24
2.2.2. Entrenamiento y validación de la red	25
2.2.3. Redes Neuronales Convolucionales	29
2.2.4. Transfer Learning	35
2.3. Incertidumbre	37
2.3.1. Intervalos de valores razonables	38
2.3.2. Incertidumbre en <i>machine learning</i>	39
2.3.3. Cuantificación de la incertidumbre en <i>machine learning</i>	40
2.4. Conformal Prediction	42
2.4.1. Conformal Prediction en problemas de regresión	42
2.4.2. Conformal Prediction en problemas de clasificación	42

3. Estado del arte	45
3.1. Estimación de la edad en antropología forense	45
3.2. Estimación de la edad en antropología forense usando Machine Learning	45
3.3. Cuantificación de incertidumbre en antropología forense . . .	45
4. Materiales y métodos	47
4.1. Conjunto de datos disponibles	47
4.2. Métodos propuestos	49
4.2.1. Arquitectura empleada	49
4.2.2. Entrenamiento	51
4.2.3. <i>Split conformal regression</i>	51
4.2.4. Regresión cuantílica	51
4.2.5. <i>Conformalized Quantile Regression</i>	51
4.3.	51
5. Experimentación	53
5.1. Protocolo de validación experimental	53
5.2. Métricas	54
5.3. Experimentos realizados	54

Índice de figuras

1.1.	Procedimiento secuencial para la identificación forense basada en el esqueleto humano (<i>skeleton-based forensic identification</i>) [20].	4
1.2.	Procedimiento secuencial para el método propuesto en [32].	6
1.3.	Evolución de hallazgos/identificación de cadáveres en España (2010-2024) [42].	7
1.4.	Evolución del número de diligencias preprocesales de determinación de edad abiertas en España (2011–2023). Elaboración propia a partir de [44-47].	8
2.1.	Gráfica de puntos de valores de edad reales vs. predichos obtenidos por el modelo propuesto en [39].	15
2.2.	Gráfica de cajas de valores de edad reales vs. predichos obtenidos por el modelo propuesto en [38].	16
2.3.	Histograma de errores residuales del modelo de estimación de edad propuesto en [38].	17
2.4.	Gráfica de cajas de la distribución del error de estimación de edad en función de la edad real, obtenida en el modelo propuesto en [39].	17
2.5.	Matriz de confusión para la estimación de sexo según el modelo <i>random forest</i> propuesto en [59].	19
2.6.	Matrices de confusión para la estimación de mayoría/minoría de edad según el modelo de [60].	20
2.7.	Diagrama de división del dataset para la validación cruzada. Recuperado de [61].	23
2.8.	Esquema visual del funcionamiento de una unidad artificial. Adaptado de [67].	24

2.9.	Diagrama de obtención de probabilidad en problemas de clasificación. Adaptado de [69].	26
2.10.	Arquitectura simplificada de un MLP. Recuperado de [70].	27
2.11.	Esquema gráfico de la aplicación de un filtro convolucional sobre una región de una imagen.	30
2.12.	Esquema gráfico de <i>max pooling</i> con un filtro 2x2 y <i>stride</i> de 1. Recuperado de la Figura 14.12 de [66].	32
2.13.	Esquema gráfico de la arquitectura conocida como “AlexNet”, diseñada para resolver un problema de clasificación con 1000 clases. Recuperado de la Figura 5.39 de [71].	34
2.14.	Diagrama del funcionamiento de neuronas con <i>dropout</i> . Recuperado de la Figura 5.29 de [71].	35
2.15.	Diagrama de <i>fine-tuning</i> de un modelo en una nueva tarea. Recuperado de la Figura 19.2 de [66].	36
2.16.	Gráfico que ilustra las 3 predicciones que arroja un modelo de regresión cuantílica. Recuperado de [94].	41
2.17.	Ejemplo adverario mal clasificado por un modelo ML entrenado con datos textuales. Adaptado de la Figura 2 de [91], original de [99].	43
4.1.	Distribución de edad de los individuos del conjunto de datos disponible.	49
4.2.	Distribución de edad por sexo de los individuos del conjunto de datos disponible.	50
4.3.	Distribución de edad de los individuos del conjunto de datos disponible por sexo.	51
5.1.	Diagrama de división del <i>dataset</i> en <i>train</i> , <i>validation</i> y <i>test</i> . .	53
5.2.	Diagrama de división del <i>dataset</i> en <i>train</i> , <i>validation</i> , <i>calibration</i> y <i>test</i>	54

Índice de tablas

4.1. Instituciones participantes en la recolección de datos e imágenes	48
--	----

Convenciones

Caracteres en negrita

Los términos empleados para definir un concepto por primera vez están en negrita.

Caracteres en cursiva

Se empleará letra cursiva para palabras en otros idiomas, títulos de obras y palabras mencionadas como términos (no por su significado).

Citas

Las citas y referencias se realizarán en estilo IEEE. Al hacer clic en una cita, el lector será redirigido a la referencia correspondiente en la bibliografía. Desde allí, podrá regresar a la página original, ya que las referencias incluyen enlaces a los lugares donde han sido citadas. No obstante, dado que una misma referencia puede aparecer varias veces a lo largo del texto, se recomienda al lector tomar nota de la página en la que se encontraba antes de saltar a la bibliografía.

Introducción de conceptos

Por regla general, la primera vez que se introduce un nuevo concepto relevante se presentará su término en español y, entre paréntesis: su término en el idioma de origen (en inglés normalmente), y su acrónimo o sigla (cuando exista). Por ejemplo: “Red Neuronal Convolutacional (*Convolutional Neural Network*, CNN)”.

Algunas excepciones:

- Si se refiere a un concepto que en español no tiene una traducción estandarizada y/o se emplea habitualmente en su forma original (anglicismo técnico), se usará directamente el término en inglés, seguido

de sus siglas si las tiene. Por ejemplo: “*batch size* (en lugar de ‘tamaño de lote’)”.

- Si el término es ampliamente conocido en su forma abreviada (incluido en español), puede omitirse la explicación extendida. Por ejemplo: “ADN (en lugar de ‘ácido desoxirribonucleico, DNA’)”.

Todo lo anterior sin perjuicio del término que se emplee más tarde en el texto, que puede emplear cualquiera de los términos presentados, en base a su

Introducción de obras y entidades

Las obras se introducirán en

Comillas

Se emplearán comillas (“...”) para enmarcar citas literales.

- **Omisiones textuales:** La secuencia “[...]” dentro de una cita indica que se ha omitido una parte del texto original.
- **Adaptaciones textuales:** La secuencia “[<texto>]” dentro de una cita indica que se ha introducido una adaptación o paráfrasis del verbo original para ajustar la cita al contexto de la oración.

Signos decimales y millares

Dado que este trabajo está en español, se empleará la coma (”,”) como signo decimal, y el punto (”.”) como separador de millares.

Aclaraciones, incisos e información complementaria

Se han seguido las recomendaciones de la RAE para el uso de guion largo y paréntesis, si bien se ha optado por añadir también notas a pie de página:

- **Raya o guion largo (—):** Se emplea para enmarcar aclaratorios breves —especialmente cuando interrumpen el flujo de la oración— dentro del texto principal. Por ejemplo: “*El error es la diferencia entre el valor verdadero —asumiendo que existe— y el valor medido.*”

- **Paréntesis (“()”):** Se utilizan para añadir información complementaria concisa y no esencial para la comprensión del texto. Por ejemplo: “*El dataset (compuesto por 10.000 ejemplos) se divide en 5 subconjuntos:...*”
- **Notas a pie de página:** Se utilizan para incluir información adicional o aclaratoria que, por su extensión o nivel de detalle, interrumpiría el flujo del texto principal si se incorporara directamente. También resultan útiles para insertar comentarios dentro de textos ya delimitados por rayas o paréntesis.

Intervalo

Los intervalos pueden expresarse de dos formas:

- Con los extremos del intervalo: $[a, b]$
- Con el punto central y la mitad de la amplitud del intervalo: $x \pm \frac{b-a}{2}$

Consideraciones lingüísticas

Los términos “F” y “M” se refieren al sexo de las personas: femenino o masculino, respectivamente.

Capítulo 1

Introducción

1.1. Descripción del problema

La antropología es la ciencia que estudia la humanidad en todas sus dimensiones: biológica, cultural, lingüística o arqueológica [1], a lo largo del tiempo y en distintas partes del mundo. La antropología biológica o física se centra en el estudio de la anatomía, el crecimiento, la adaptación y la evolución del cuerpo humano [2].

Dentro de este campo, la **antropología forense (AF)** es el subcampo especializado que aplica métodos y técnicas antropológicas para resolver cuestiones médico-legales [2], empleando conocimientos de antropología física, aunque a veces también de la arqueología, para la correcta recuperación y análisis de la evidencia forense.

Tradicionalmente, los antropólogos forenses han tenido cinco principales objetivos en su trabajo [3]:

1. Determinar el **perfil biológico** de un individuo (es decir, sexo, edad, estatura y ascendencia) cuando los tejidos blandos se han deteriorado hasta el punto de que estas características no pueden determinarse mediante inspección visual.
2. Identificar la naturaleza de lesiones traumáticas (como heridas de bala, puñaladas o fracturas) en huesos humanos, así como sus causantes, con el objetivo de recopilar información sobre la causa y circunstancias de la muerte.
3. Estimar el intervalo *post mortem*, es decir, el tiempo transcurrido desde la muerte, gracias a su conocimiento sobre los procesos de descomposición corporal.

4. Asistir en la localización, recuperación y conservación de los restos (superficiales o enterrados) aplicando técnicas arqueológicas, garantizando la recolección de toda la evidencia forense relevante.
5. Proporcionar información clave para la **identificación** de los fallecidos, basándose en las características distintivas de los esqueletos.

Además de estos roles, en la actualidad los antropólogos desempeñan otros trabajos que no están relacionados con el ámbito criminalístico. Entre ellos, uno de sus campos de acción más relevantes es la **identificación de víctimas en contextos de catástrofes masivas** [4-6], como accidentes aéreos, ataques terroristas o desastres naturales, donde los restos suelen estar mutilados o desfigurados.

Su labor también es fundamental en la **recuperación e identificación de violaciones sistemáticas de derechos humanos**, como exterminios, persecuciones políticas y represiones dictatoriales [7]. Casos como la Guerra Civil Española y la Dictadura Franquista [8, 9], así como las múltiples dictaduras en el Cono Sur de América [10], han requerido la intervención de equipos forenses para esclarecer la verdad histórica y restituir la identidad de las víctimas a sus familiares, contribuyendo al proceso de memoria, justicia y reparación para las familias afectadas. Esta vinculación con la justicia trasciende lo nacional: la ciencia forense es clave en la **investigación de crímenes de guerra contra poblaciones civiles**. Organizaciones como Médicos por los Derechos Humanos y la ONU financian equipos especializados que documentan estos crímenes, proporcionando pruebas esenciales para tribunales internacionales [11].

Y por último, también son fundamentales para **estimar la edad de personas vivas en casos legales**, especialmente cuando no existen registros confiables. Esto ocurre, por ejemplo, en casos de solicitudes de asilo, adopciones internacionales o procesos judiciales donde es necesario determinar si una persona es menor o mayor de edad, lo cual puede tener importantes implicaciones legales. Según el tipo de procedimiento, se puede requerir tanto la estimación de la edad mínima como la edad más probable del individuo, con el fin de priorizar la protección de los menores, evitando que queden expuestos a violaciones de sus derechos.

1.1.1. Limitaciones de la antropología forense

Como hemos visto, la **identificación humana (ID)** es una de las principales tareas que aborda la AF. Consiste en la determinación y verificación de la identidad de una persona en base a [12]: evidencias circunstanciales (hora y lugar del descubrimiento del cuerpo, efectos personales, confirmación visual por parte de familiares y amigos); y evidencias físicas, obtenidas

a través de examinación externa de características como el sexo, color de piel, tatuajes, o huellas dactilares, o, cuando estas no estén disponibles, mediante examinación interna con técnicas médico-científicas, donde se aplican técnicas de antropología y genética forense.

Cabe destacar que, aunque los análisis dactilares y genéticos superan en precisión identificativa a los métodos antropológicos, su aplicabilidad enfrenta limitaciones técnicas significativas que condicionan su uso en ciertos contextos forenses [6]. Las huellas dactilares requieren de: tejido blando preservado, lo que es común en cadáveres frescos, pero se pierde con la descomposición o la carbonización; y una base de datos que incluya la huella del individuo en vida (registros *ante mortem*). Por otro lado, en cuanto al análisis genético, este puede verse comprometido por una mala conservación del ADN que puede deberse a su degradación o contaminación. La concentración presente en un cadáver se reduce drásticamente en los primeros 8 meses *post mortem* [13], y factores como las altas temperaturas, la exposición a humedad ambiental o la presencia de aguas subterráneas y entornos ricos en oxígeno, que fomentan la presencia microbiana, perjudican la conservación del ADN [14]. Y, aún extraída una secuencia válida de ADN, se necesita de muestras con las que compararla, a ser posible de familiares de primer grado, para establecer una identificación concluyente.

Por tanto, la AF contribuye al problema de identificación humana en dos escenarios [15]:

1. Cuando los otros métodos no son viables, dado que las pruebas no se puedan recoger o no sean válidas, o no haya registros con los que compararlas.
2. Como apoyo a otras técnicas de identificación. Por ejemplo, las técnicas de estimación del perfil biológico pueden reducir el grupo de posibles coincidencias en bases de datos genéticos, facilitando el cotejo de secuencias genéticas y reduciendo el coste del proceso.

La **estimación del perfil biológico (PB)** es, por tanto, un proceso fundamental de la AF, en el cual se determinan características biológicas clave de un individuo [3]:

- **sexo**, mediante el análisis morfológico y métrico de rasgos sexuales en el esqueleto, especialmente en la pelvis y el cráneo;
- **edad**, estimada a partir de cambios morfológicos y de desarrollo en el esqueleto, pudiendo referirse tanto a la **edad al momento de la**

muerte en restos óseos, como a la **edad cronológica**¹ en personas vivas en contextos forenses o humanitarios;

- **estatura**, mediante la estimación de la talla a partir de longitudes óseas, particularmente de los huesos largos; y
- **ascendencia o afinidad poblacional**, analizando variaciones craneométricas y morfológicas asociadas a poblaciones o grupos geográficos (actualmente en revisión [17-19]).

En los problemas de ID, cuando estas características biológicas coinciden con los registros *ante mortem*, se fortalece la hipótesis de identificación; en cambio, si existen una o más discrepancias —especialmente de alguna característica firme como múltiples epífisis no fusionadas, que no pueden ocurrir en un adulto mayor—, el individuo es excluido como posible coincidencia [3]. En la Figura 1.1 podemos observar que la estimación del PB es uno de los primeros pasos en el proceso de ID forense.

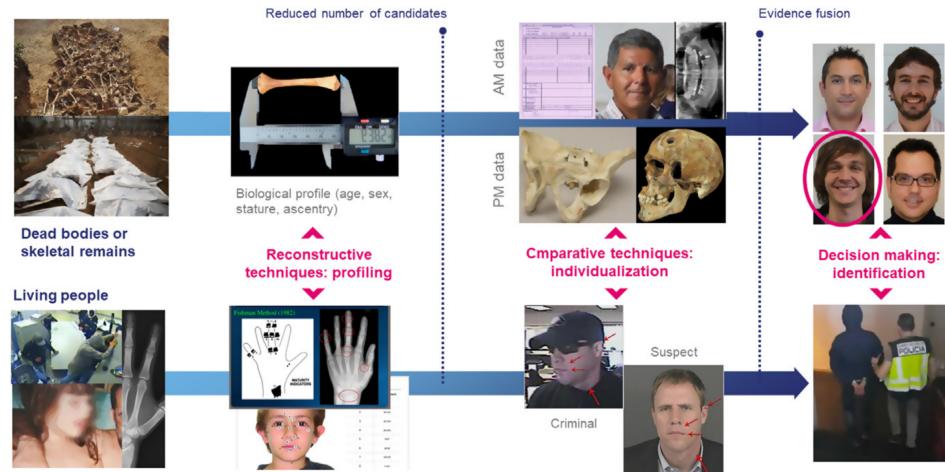


Figura 1.1: Procedimiento secuencial para la identificación forense basada en el esqueleto humano (*skeleton-based forensic identification*) [20].

La estimación del PB en restos humanos es una tarea compleja, especialmente cuando se estima la edad en el momento de la muerte, ya que hay diferentes métodos a aplicar dependiendo de la fase de desarrollo del individuo. Las variaciones en la morfología de los huesos son bien conocidas, pero estas no siempre ocurren al mismo tiempo en diferentes individuos, ya que no están expuestos a las mismas condiciones genéticas y del entorno.

¹La edad cronológica es la edad real de una persona desde su nacimiento, mientras que la edad biológica refleja la condición fisiológica del cuerpo [16].

Además, como se ha mencionado anteriormente, la estimación de edad también se realiza sobre personas vivas en casos legales donde la edad es un factor determinante [21], por ejemplo, con menores migrantes no acompañados. En estos casos no se tiene acceso a los huesos de la persona de forma directa, por lo que el análisis se realiza sobre imágenes médicas.

1.2. Motivación

Los métodos de estimación del PB se basan en la evaluación visual y en el análisis morfométrico de rasgos esqueléticos, que requieren de conocimiento especializado. Sin embargo, su aplicación puede presentar ambigüedades en su formulación que den lugar a intérpretes variables —muchas veces fruto de sesgos cognitivos [22, 23]— y están sujetos a posibles errores de medición [24]. Además, la gran variabilidad genética y ambiental entre individuos, que afecta la morfología del esqueleto y genera diferencias significativas entre poblaciones de distintas regiones [25], hace que muchos de estos métodos —basados en muestras de referencia limitadas o no representativas de la diversidad humana global— pierdan precisión. Esto puede introducir sesgos al estimar el PB de individuos de grupos poco estudiados o con características atípicas.

Frente a estas limitaciones, recientes avances en inteligencia artificial (IA) y machine learning (ML) han demostrado el potencial de mejorar la exactitud y objetividad de estimación del PB, tanto para la estimación de sexo [26-28] como de edad [29-31].

Estos modelos, que emplean imágenes médicas con algoritmos de visión por computador, siguen dos principales enfoques. En el primer enfoque, parten de un método de AF clásico e intentan automatizarlo y/o mejorarlo [32, 33] (véase la Figura 1.2). Para ello, es necesario especificar:

1. cómo extraer las características relevantes de las imágenes médicas, mediante técnicas de procesamiento de imágenes o morfometría tradicional; y
2. un modelo de clasificación o regresión (como SVM, redes neuronales simples o árboles de decisión) que opere sobre estas características predefinidas.

En cambio, en el enfoque más popular, el *end-to-end*, el modelo aprende automáticamente tanto la extracción de características como la clasificación/regresión a partir de los datos en bruto. Este enfoque es posible gracias a las redes neuronales convolucionales, que eliminan la dependencia de criterios antropológicos preestablecidos y permite al modelo extraer por sí mismo las características más relevantes para la estimación de sexo, edad, etc.

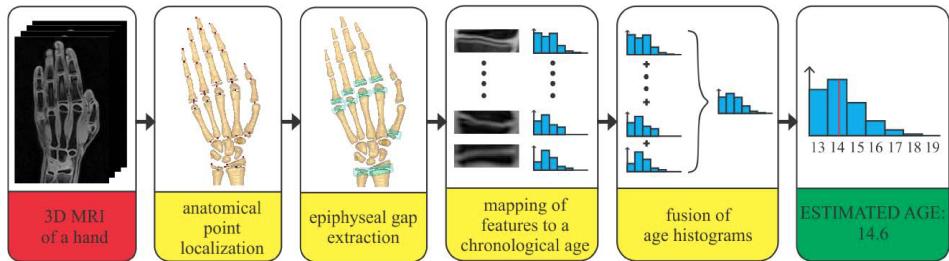


Figura 1.2: Procedimiento secuencial para el método propuesto en [32].

Este enfoque se ha visto potenciado por el auge del Deep Learning, permitiendo a las CNNs aprender patrones complejos que podrían pasar inadvertidos por el ser humano, y mejorando la precisión de las predicciones [34, 35]. Sin embargo, aún mejorando la exactitud de las predicciones, los modelos siguen mostrando carencias respecto a la cuantificación de incertidumbre, pues no todas las predicciones tienen el mismo nivel de confianza o fiabilidad. Ya en [36] se introducía no solo la necesidad de identificar el método adecuado para estimar la edad a partir de los elementos disponibles, sino también de evaluar su confiabilidad y realizar un estudio del error arrojado por las predicciones del método. Estos generalmente se han basado en la estadística frequentista [37-39]². Un ejemplo de este tipo de análisis se ilustra en la Figura ??, donde se examina la distribución probabilística del error residual arrojado por el modelo de regresión propuesto en [37].

Aunque existen métricas para evaluar el error cuando se dispone de *ground truth*, la mayoría de los modelos actuales se limitan a ofrecer predicciones puntuales en regresión [38, 40, 41] o etiquetas únicas en clasificación [35, 40], sin cuantificar la incertidumbre asociada a cada predicción.

Con lo anterior se expone la motivación de la aplicación de ML a la AF, así como de la necesidad de cuantificar la incertidumbre en las predicciones, para ofrecer garantías de confiabilidad estadística que aspiren a sustentar la validez legal en contextos judiciales. Algunos datos que magnifican la necesidad de técnicas de AF confiables actualmente son:

- En los últimos años, ha aumentado significativamente el número de cadáveres hallados en el territorio español, como podemos apreciar en la Figura 1.3 [42]. En 2024 se ha alcanzado una cifra record, —en gran parte debido a las inundaciones de la DANA Valencia—, de 531 cadáveres en 2024, de los cuales se pudo identificar a 323.

²La estadística frequentista es la corriente estadística que desarrolla a partir de los conceptos de probabilidad y que se centra en el cálculo de probabilidades y el contraste de hipótesis.

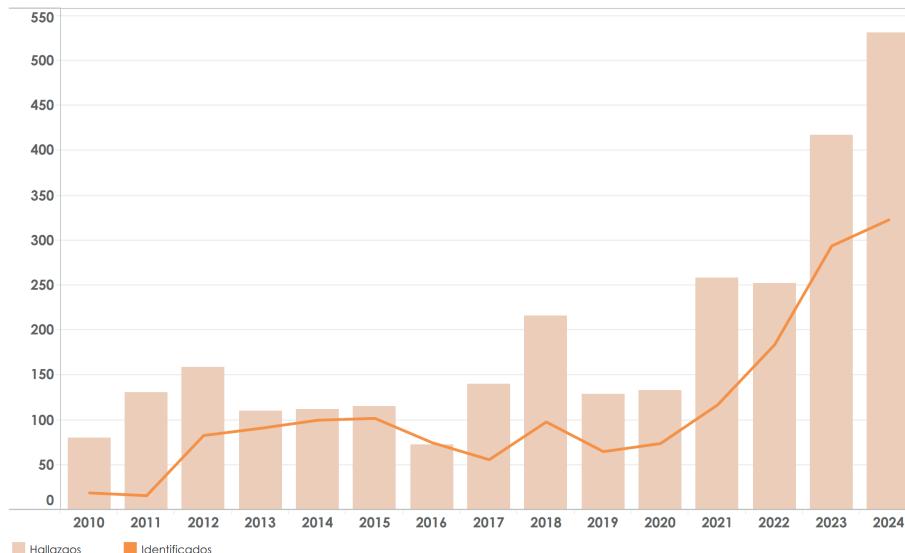


Figura 1.3: Evolución de hallazgos/identificación de cadáveres en España (2010-2024) [42].

- En 2020, de las 2.457 fosas totales documentadas de la Guerra Civil y el franquismo, aún 1.221 seguían sin ser intervenidas y se estimaba que “con una intervención oficial del Estado podrían recuperarse unos 20 a 25.000 individuos” e identificar “entre 5 y 7.000 de ellos”, estimándose necesario contar con unos 40-50 profesionales de la antropología forense [43].
- En España, se ha registrado en la última década (2013-2023) un aumento significativo en la llegada de Menores Extranjeros No Acompañados [44-47], que ha disparado consigo el número de diligencias abiertas para la determinación de su edad, como se ve reflejado en la Figura 1.4.
- La relevancia de la ciencia forense en la identificación de víctimas y la protección de la dignidad humana ha convertido su aplicación en un pilar fundamental de los derechos humanos y la justicia internacional, naciendo así la **acción forense humanitaria** [48]. Esta disciplina emplea la ciencia forense con un propósito exclusivamente humanitario, con los objetivos de: identificar a las personas fallecidas, gestionar dignamente sus restos y aliviar el sufrimiento de sus familias en situaciones de conflicto, migración y desastres naturales [49].

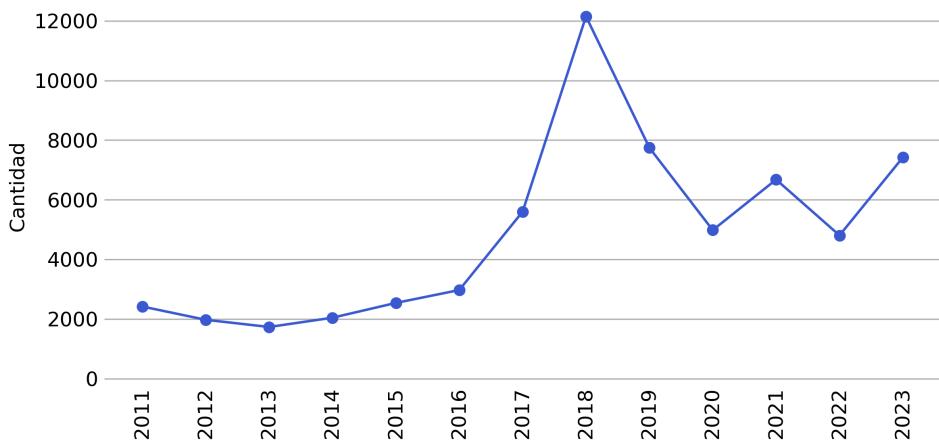


Figura 1.4: Evolución del número de diligencias preprocesales de determinación de edad abiertas en España (2011–2023). Elaboración propia a partir de [44–47].

1.3. Objetivos

La *Conformal Prediction* emerge como un marco teórico robusto para generar intervalos de predicción con garantías estadísticas sólidas, independientemente de la distribución subyacente de los datos. A diferencia de los enfoques tradicionales, este método no solo ofrece predicciones puntuales, sino que cuantifica la incertidumbre asociada a cada estimación mediante intervalos adaptativos o conjuntos de predicción que reflejan la confiabilidad de la predicción en cada caso particular.

Este proyecto tiene un doble objetivo: por un lado, desde un prisma teórico, estudiar las ventajas y costes asociados a las diversas técnicas de inferencia conformal actuales; y, por otro, aplicarlo a un contexto práctico como es el problema de estimación del PB, centrándonos en la estimación de edad y de sexo a partir de datos biológicos e imágenes médicas. De esta forma, cuando estemos ante datos biológicos ambiguos, la conformal prediction podrá devolver conjuntos de predicciones con más de una etiqueta predicha (p.ej., {masculino, femenino}) en problemas de clasificación, o intervalos de predicción más amplios (p.ej., edad \in [17,20]) en problemas de regresión, en ambos casos para un nivel de confianza determinado.

¿Es más correcto “técnicas de cuantificación de la incertidumbre”?

Por tanto, ponemos desgranar los objetivos en:

- Estudiar de forma exhaustiva la bibliografía sobre *conformal prediction* y sus diversas variantes, así como de la estimación de sexo y edad, centrándolo nuestra atención en el estado del arte.

- Implementar, entrenar y validar modelos de regresión —en problemas de estimación de edad— y clasificación —tanto en problema de estimación de sexo como edad legal— a los que aplicar la inferencia conformal.
- Comparar los intervalos y conjuntos de predicciones generados para evaluar su calibración empírica, robustez ante datos ambiguos y utilidad forense, contrastándolos con métodos tradicionales (p.ej., intervalos de confianza clásicos).
- Realiza una primera aproximación a un marco interpretable y con garantías estadísticas para la estimación del perfil biológico, donde la incertidumbre cuantificada pueda integrarse en informes periciales bajo estándares jurídicos.

En resumen, este trabajo pretende explorar la integración de marcos probabilísticos en la práctica forense que capturen la incertidumbre de los problemas, y facilitar el uso de la inferencia conformal en ellos. Este enfoque proporciona estimaciones calibradas de incertidumbre, con garantías estadísticas de cobertura válidas bajo supuestos mínimos, útiles para la toma de decisiones fundamentadas en contextos prácticos donde la interpretabilidad y robustez son críticas.

1.4. Planificación económica y temporal del proyecto

Falta escribir este apartado

Capítulo 2

Fundamentos teóricos

Este capítulo tiene el propósito de presentar y describir los fundamentos teóricos que sustentan los métodos utilizados en el trabajo, además de justificar su importtancia para abordar los problemas planteados.

2.1. Machine Learning

Frente a la idea de intentar crear un programa que simulara directamente el comportamiento inteligente de una “mente adulta”, Alan Turing ya vaticinó un enfoque alternativo [50]: que las máquinas pudieran aprender como lo hace un niño, mediante un “proceso educativo” con el cual se logra alcanzar progresivamente una “mente adulta”, obteniendo así comportamientos inteligentes complejos.

En los años 50, surgió el concepto de *machine learning* (ML) —o aprendizaje automático en español—, popularizado por Arthur L. Samuel [51], para designar una rama marginal de la IA, centrada en el desarrollo de modelos y algoritmos que permitiesen a las computadoras imitar la forma en la que los humanos aprenden, realizar tareas autónomas y mejorar su rendimiento a través de la experiencia y exposición a más datos. De esta forma, estos modelos podrían realizar predicciones o tomar decisiones sin ser programados para cada caso.

En las décadas de 1960, 1970 y 1980, surgieron algoritmos fundamentales como el perceptrón [52, 53] o los árboles de decisión [54], que sentaron los cimientos teóricos para el desarrollo posterior de técnicas más complejas. Sin embargo, el progreso fue lento debido a las limitaciones computacionales y el gran escepticismo académico.

Los años 90 y 2000 marcaron un punto de inflexión para el ML, gracias a los avances teóricos, el mayor poder computacional y la disponibilidad

de grandes volúmenes de datos. De 2010 en adelante, la evolución del ML ha sido exponencial, marcada por la consolidación del *deep learning*, la escalabilidad masiva y su integración en numerosas aplicaciones: de visión por computador, reconocimiento de lenguaje natural, robótica, diagnóstico médico y forense, finanzas o recomendación de contenidos, entre otros. De esta forma, el ML se ha convertido en un campo tan amplio y exitoso que ahora “eclipsa” al resto de campos de la IA [55].

El ML diferencia tres tipos de aprendizaje en base a tres tipos de retroalimentación [56]:

- **Aprendizaje supervisado**, en el que el agente (refiriéndose con este al modelo de ML y su algoritmo de aprendizaje) observa ejemplos de pares entrada-salida y aprende la función que mejor mapea las entradas (inputs) a las salidas (outputs) correspondientes. El objetivo es generalizar este aprendizaje para hacer predicciones precisas sobre datos nuevos y no vistos [57].
- **Aprendizaje por refuerzo**, en el que los datos de entrenamiento no contienen salida objetivo, sino que contiene posibles resultados junto con medidas de calidad de dicho resultado, es decir, una función de evaluación del estado. En este tipo de aprendizaje, el agente toma decisiones en un entorno y recibe recompensas o penalizaciones por las acciones que realiza, ajustando su comportamiento mediante prueba y error, maximizando la recompensa acumulada en el tiempo [58].
- **Aprendizaje no supervisado**, en el que el agente tampoco dispone de valores de salida, solo de entrada [57], y los objetivos pueden ser muy variados, centrándose en descubrir patrones, estructuras o relaciones ocultas en los datos. A diferencia de los otros enfoques, aquí no hay una “respuesta correcta” predefinida, sino que el modelo debe inferir conocimiento directamente desde la distribución de los datos.

Este trabajo se centrará en el aprendizaje supervisado, pues es este tipo de aprendizaje el empleado en los problemas de clasificación y regresión que aplicaremos en el ámbito de la antropología forense.

2.1.1. Problemas de regresión

Como se ha mencionado antes, la regresión es un tipo de problema clásico en el aprendizaje supervisado, y consiste en predecir el valor de una o más **variables continuas** objetivo a partir de unos datos de entrada [57], utilizando un modelo entrenado con ejemplos ya con valores conocidos.

Matemáticamente, este proceso implica modelar la relación entre la variable dependiente Y y las variables independientes X , de modo que se pueda predecir o explicar el comportamiento de Y en función de los valores de X . El modelo aprende una función de predicción f que, dado un nuevo ejemplo i con características X_i , genera una estimación \hat{Y}_i :

$$f(X_i) = f(X_{i0}, X_{i1}, \dots, X_{in}) = \hat{Y}_i = Y_i + \varepsilon_i$$

donde

- $X_{i0}, X_{i1}, \dots, X_{in}$ son las características o atributos del ejemplo i ,
- Y_i es el valor real de la variable objetivo para ese ejemplo,
- \hat{Y}_i es la predicción generada por el modelo, y
- ε_i representa el error o residuo ¹, es decir, la diferencia entre la predicción y el valor real. Este término captura factores aleatorios o imprecisiones que el modelo no logra explicar perfectamente.

El análisis y la evaluación estadística del error son fundamentales para valorar la utilidad práctica del modelo y optimizar su capacidad predictiva mediante técnicas de ajuste y validación. Existen numerosas métricas para evaluar el rendimiento en problemas de regresión, pero tres destacan especialmente por ser *model-agnostic*, es decir, aplicables a cualquier modelo de regresión independientemente del algoritmo subyacente. Estas son:

- El **error absoluto medio (mean absolute error, MAE)** mide el promedio de las diferencias absolutas entre los valores reales (Y_i) y los valores predichos (\hat{Y}_i) por el modelo.

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i|$$

donde n es el número de ejemplos/instancias con las que se cuenta en los datos a evaluar.

La interpretación más inmediata de esta métrica es que representa cuánto se desvía en promedio la predicción del valor real sin considerar la dirección del error (positivo o negativo) y, por tanto, cuanto más se acerque a cero el valor, mejor es el ajuste del modelo.

Existe una variante denominada **error absoluto mediano (median absolute error, MedAE)**, que realiza la mediana de las diferencias

¹... a pesar de que en la literatura estos términos se distinguen, ...

absolutas, en vez de la media, aumentando la robustez frente a valores atípicos con errores extremos.

- El **error cuadrático medio** (*mean squared error*, **MSE**) mide el promedio de los errores al cuadrado entre valores reales (Y_i) y los valores predichos (\hat{Y}_i) por el modelo.

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Al igual que el MAE, cuantifica qué tan cerca están las predicciones de los valores reales, pero penaliza más los errores grandes, y es más sensible por tanto a valores atípicos.

Como veremos más tarde, esta métrica es muy útil en optimización mediante gradiente descendente, usado a la hora de entrenar modelos de regresión basados en redes neuronales.

Y también tiene una variante, la **raíz del error cuadrático medio** (*root mean square error*, **RMSE**), que se obtiene extrayendo la raíz cuadrada del MSE:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$$

Esta métrica conserva las mismas unidades que la variable objetivo, lo que facilita su interpretación práctica. Es comparable con el MAE en cuanto a escala, aunque sigue penalizando más los errores grandes.

- El **coeficiente de determinación**, o más conocido como **R²** o **bondad de ajuste**, mide la proporción de la variabilidad de la variable dependiente (Y) que es explicada por el modelo.

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

donde

- Y_i es el valor real de la variable dependiente para la instancia i ,
- \hat{Y}_i es la predicción generada por el modelo, y
- \bar{Y} es el promedio de los valores reales de la variable dependiente a lo largo de todas las instancias del conjunto de datos.

El valor de esta métrica varía entre $-\infty$ y 1, y su interpretación es la siguiente:

- $R^2 \leq 0$ significa que el modelo no explica ninguna variabilidad y que las predicciones del modelo no son mejores que simplemente predecir la media de los valores reales.
- $R^2 \in (0, 1)$ indica que el modelo está explicando una fracción de la variabilidad de los datos, y cuanto más cercano sea a 1, mejor será el ajuste del modelo.
- $R^2 = 1$ indica un ajuste perfecto y, por tanto, el modelo explica toda la variabilidad de los datos.

A diferencia de las anteriores, es una métrica relativa y adimensional, es decir, no depende de las unidades de la variable objetivo y evalúa qué tan bien se ajusta el modelo en comparación con un modelo base que siempre predice la media de los valores reales.

No obstante, el uso exclusivo de métricas numéricas resulta en un análisis pobre, ya que estas no permiten identificar patrones ocultos, detectar relaciones no lineales ni distinguir entre errores positivos o negativos. Por esta razón, se recomienda completar el análisis con representaciones gráficas, tales como:

- La **gráfica de puntos de valores reales vs. predichos**, que permite visualizar la relación entre las predicciones del modelo y los valores reales. Idealmente, los puntos deberían alinearse alrededor de la recta $Y = \hat{Y}$. Desviaciones sistemáticas indican sesgos o problemas de ajuste. Un ejemplo de esta gráfica lo encontramos en la Figura 2.1.

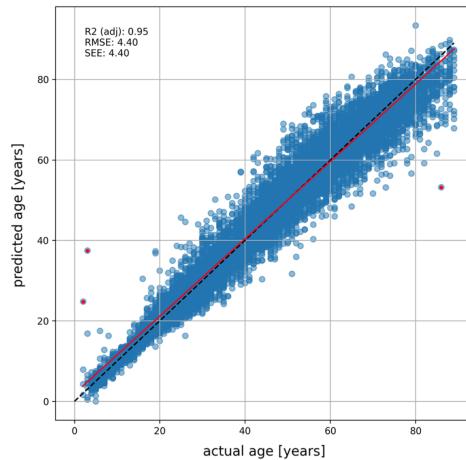
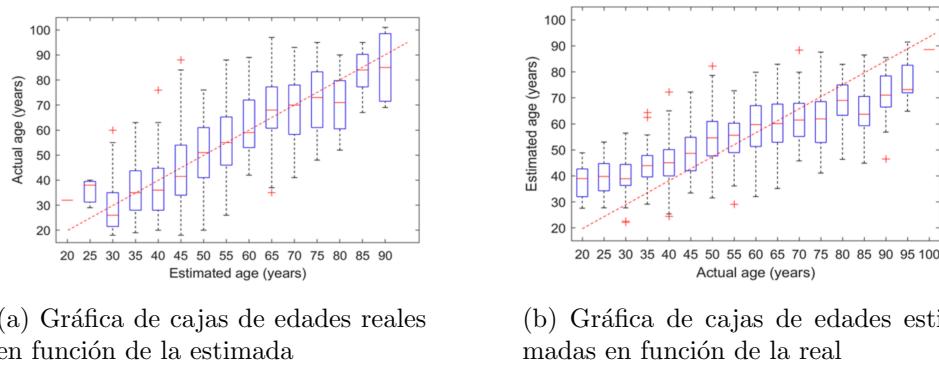


Figura 2.1: Gráfica de puntos de valores de edad reales vs. predichos obtenidos por el modelo propuesto en [39]. Se observan valores más dispersos en edades avanzadas.

- También existe una versión más refinada de presentar esta información, especialmente útil en casos en los que muchos datos sobrecargan la gráfica, en la **gráfica de cajas** (en inglés *boxplot*) de **valores reales vs. predichos**. Estos proporcionan una visión clara de la distribución de los datos, con mediana, cuartiles y valores atípicos, ya sea agrupando por valores reales o por valores predichos.

La Figura 2.2 muestra la distribución de las edades predichas en función de distintos grupos de edad real, y viceversa, lo que facilita la identificación de errores en el desempeño del modelo.



(a) Gráfica de cajas de edades reales en función de la estimada

(b) Gráfica de cajas de edades estimadas en función de la real

Figura 2.2: Gráfica de cajas de valores de edad reales vs. predichos obtenidos por el modelo propuesto en [38]. Se observa en b que se sobreestima la edad en personas jóvenes y se subestima en personas de edad avanzada.

- El **histograma de residuos** muestra la distribución de los errores ($Y_i - \hat{Y}_i$) del modelo. Una distribución simétrica y centrada en cero sugiere un buen ajuste, mientras que una distribución sesgada o asimétrica podría indicar que el modelo está subajustado o que hay algún patrón no capturado por el modelo.

Un ejemplo de este tipo de gráfica lo encontramos en la Figura 2.3, donde se analizaba el error obtenido con el modelo propuesto en [38].

Histograma de errores residuales del modelo de regresión propuesto en [37] que predice la estatura a partir de la longitud de la tibia.

- Y una versión más completa que este último es la **gráfica de cajas de la distribución del error en base a los valores reales o predichos**, que permite analizar cómo varía el error del modelo a lo largo de diferentes rangos de valores, ya sean reales o predichos. Estas visualizaciones nos permiten detectar fácilmente las fortalezas y debilidades en las predicciones del modelo, así como diagnosticar sesgo o insuficiencia de datos en ciertas categorías.

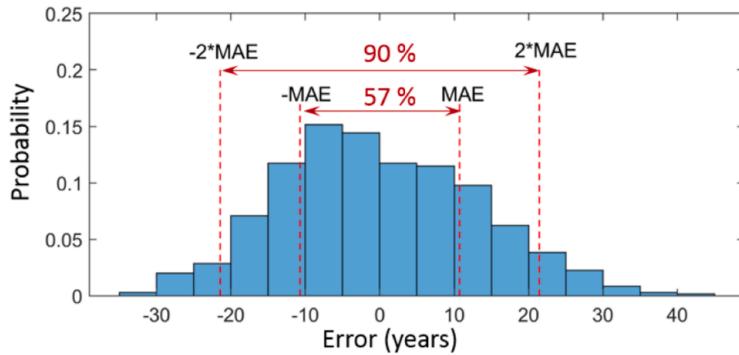


Figura 2.3: Histograma de errores residuales del modelo de estimación de edad propuesto en [38]. Se evidencia una mayor probabilidad de errores negativos (infraestimaciones) en comparación con los positivos. Además, se destaca que el 57 % de las predicciones presentan un error inferior al MAE, y que el 90 % se encuentra dentro de un margen de error menor a 2MAE.

Un ejemplo ilustrativo de esta gráfica se presenta en la Figura 2.4, donde se observa la variación del error del modelo propuesto en [39] a través de distintos rangos de edad real. En particular, se evidencia una tendencia a cometer errores mayores en los grupos de edad más avanzada.

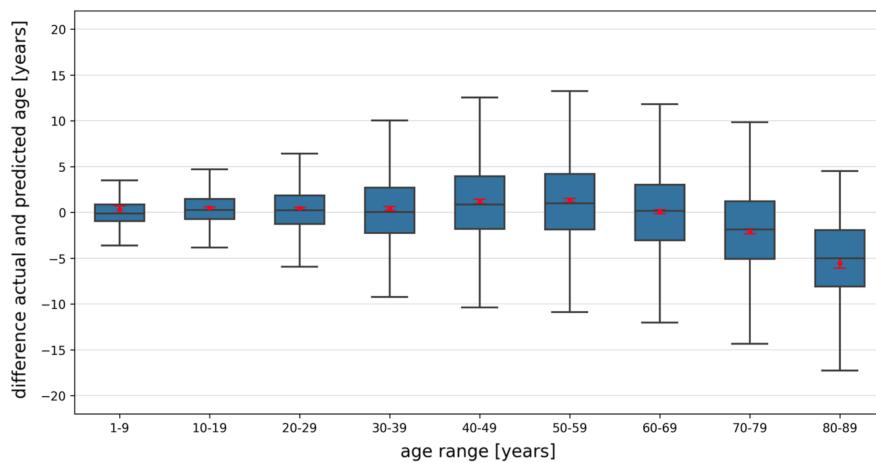


Figura 2.4: Gráfica de cajas de la distribución del error de estimación de edad en función de la edad real, obtenida en el modelo propuesto en [39].

2.1.2. Problemas de clasificación

En cambio, en los problemas de clasificación, los valores de salida son categóricos, denominados más comúnmente como **clases**, y a cada valor individual asignado a una instancia de datos se le conoce como **etiqueta** (*label* en inglés).

Existen multitud de variantes de clasificación, que pueden diferenciarse según diversos criterios:

- En base a la cardinalidad de las clases de salida: **clasificación binaria o multiclasa**, según si existen dos clases posibles o más de dos, respectivamente.
- En base al número de etiquetas asignadas a cada instancia: **clasificación con etiqueta única o multietiqueta**, según si cada instancia pertenece a una sola clase o a varias de forma simultánea.
- En base a la certeza de la asignación de clases: **clasificación con etiqueta precisa o difusa**, donde en el primer caso la asignación a una clase es determinista, y en el segundo caso se permite una pertenencia parcial a varias clases, con distintos grados de afinidad.

No obstante, la mayoría de los problemas estudiados en la literatura de ML, y concretamente en antropología forense, corresponden a clasificación binaria o multiclasa, con etiquetas únicas y asignación precisa [57], que serán el foco de este trabajo. La cardinalidad de las clases tiene implicaciones significativas en el diseño del modelo y la evaluación de su desempeño:

- **Clasificación binaria**, que es aquella en la que existen únicamente dos clases posibles para la variable objetivo, siendo común en problemas donde se desea discriminar entre dos estados mutuamente excluyentes (p.ej., “positivo” vs. “negativo”, “spam” vs. “no spam”, “fraude” vs. “no fraude”).

Se suele denominar a una de las clases como “positiva” y a otra como “negativa” para facilitar la interpretación de métricas como la precisión, la sensibilidad o la especificidad, si bien no tiene por qué existir una connotación valorativa entre ambas clases.

- **Clasificación multiclasa**: en este caso, la variable objetivo puede tomar más de dos valores posibles, pertenecientes a un conjunto finito. Un ejemplo de problema clásico es el de clasificar dígitos manuscritos (0-9).

En este tipo de problemas, el error ocurre cuando no se acierta al predecir la clase del ejemplo.

Una vez definido los tipos de problemas de clasificación, es fundamental establecer cómo medir la efectividad del modelo predictivo. A continuación, se detallan los principales criterios y elementos gráficos utilizados para evaluar y comparar modelos de clasificación:

- La **matriz de confusión** es una herramienta fundamental que permite visualizar el rendimiento de modelos de clasificación, tanto binarios como multiclas. Esta muestra una tabla con tantas columnas y filas como clases haya. En un eje, se representan las clases reales (etiquetas verdaderas), y en el otro eje, las clases predichas por el modelo. Cada celda de la matriz indica la cantidad de ejemplos que pertenecen a una clase real específica y que han sido clasificados como una clase predicha específica (véase la Figura 2.5).

Idealmente, los valores se concentrarían en la diagonal principal, lo que indicaría que las predicciones coinciden con los valores reales.

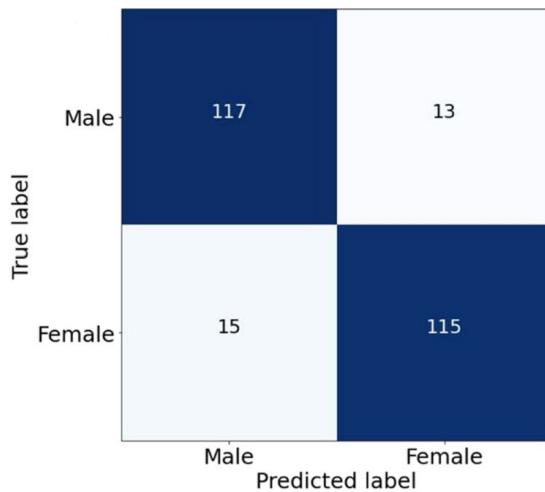


Figura 2.5: Matriz de confusión para la estimación de sexo según el modelo *random forest* propuesto en [59].

Esta visualización admite muchas variantes, por ejemplo, como la vista en la Figura 2.6.

Prácticamente todas las métricas y visualizaciones parten de la información ofrecida en esta matriz.

- La **exactitud (accuracy)** es la proporción de predicciones correctas sobre el total. En clasificación binaria esto sería:

		predicted label	
		≥ 18	< 18
true label	≥ 18	85.40%	14.60%
	< 18	13.70%	86.30%

(a) Sin información de sexo

		predicted label	
		>= 18	< 18
true label	>= 18	83.10%	16.90%
	< 18	16.20%	83.80%

(b) Sexo femenino

		predicted label	
		>= 18	< 18
true label	>= 18	89.50%	10.50%
	< 18	9.33%	90.67%

(c) Sexo masculino

Figura 2.6: Matrices de confusión para la estimación de mayoría/minoría de edad según el modelo de [60]. Se representan los valores de cada celda en términos porcentuales de los ejemplos reales que hay de cada clase (< 18 y ≥ 18), lo que permite comparar la matriz de confusión general de todos los ejemplos (a) con la de ejemplos se sexo femenino (b) y sexo masculino (c), permitiendo identificar posibles sesgos en el modelo respecto al género, y así realizar una evaluación más precisa del rendimiento del modelo en diferentes subgrupos de la población.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

En el caso de clasificación multiclas, esta se generaliza como:

$$\text{Accuracy} = \frac{\text{Número de predicciones correctas}}{\text{Número total de ejemplos}} = \frac{1}{N} \sum_{i=1}^N 1(\hat{y}_i = y_i)$$

donde \hat{y}_i es la etiqueta predicha, y_i la etiqueta verdadera para el ejemplo i , N es el número de ejemplos, y 1 es la función indicadora, que vale 1 si la predicción es correcta y 0 si no lo es.

Los valores de esta métrica varían entre 0 y 1, donde 0 es el peor desempeño posible (todas las predicciones son incorrectas) y 1 es el mejor desempeño posible (todas las predicciones son correctas).

Es la medida más intuitiva, si bien puede dar una falsa impresión de buen desempeño si las clases mayoritarias dominan la métrica. Es por esto que el análisis debe completarse con otras métricas informativas.

- La **precisión** (*precision*) indica qué proporción de las predicciones positivas corresponde a casos realmente positivos. En clasificación multiclase, se interpreta como la proporción de ejemplos correctamente clasificados entre todos los que fueron asignados a una clase determinada.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Una alta precisión significa pocos falsos positivos. Esto puede interesar por ejemplo

La **exhaustividad (recall)** indica qué proporción de los casos positivos fueron correctamente detectados.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Un alto *recall* significa pocos falsos negativos.

Estas dos métricas complementarias se pueden calcular por cada clase, y hay varias formas de combinar sus valores:

- Macro:
- Micro:
- Weighted:

■ El **F1-Score**

Por completar

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Al igual que con *precision* y *recall*, también se puede calcular por cada clase.

2.1.3. Selección de modelos y optimización

El objetivo del ML es establecer una hipótesis que se ajuste de forma óptima a los ejemplos futuros. Para ello, suponemos que los ejemplos futuros mostrarán un comportamiento similar a los pasados. Bajo este supuesto, el ajuste óptimo de un modelo es, por tanto, la hipótesis que minimiza la tasa de error del problema [56].

Pero medir el error del modelo sobre los mismos datos empleados en el entrenamiento suele sesgar el resultado, ya que el modelo puede estar sobreajustado (*overfitting*) a los datos de entrenamiento, capturando no solo el patrón subyacente, sino también el ruido o las peculiaridades específicas de ese conjunto de datos. Para evitar esto, es fundamental evaluar el modelo en un conjunto de datos de prueba independiente, que simule cómo se comportaría con ejemplos futuros no vistos durante el entrenamiento. Por este motivo, es común dividir los datos disponibles en dos conjuntos distintos:

el **conjunto de entrenamiento (*training set*)** y **conjunto test (*test set*)**.

Aun así, incluso con esta división de conjuntos, puede persistir el riesgo de sobreajuste si se realizan múltiples ajustes y selecciones de hiperparámetros basados en el rendimiento en el conjunto test. Esto se debe a que, indirectamente, el modelo podría estar “aprendiendo” características específicas del conjunto de prueba, comprometiendo su capacidad de generalización. Para abordar este problema, se introduce un tercer subconjunto: el **conjunto de validación**. Este conjunto se utiliza para evaluar y ajustar los hiperparámetros del modelo durante el desarrollo, reservando el conjunto test únicamente para la evaluación final.

Además, técnicas como la **validación cruzada (*cross-validation*)** son ampliamente utilizadas para maximizar el uso de los datos disponibles, especialmente en conjuntos pequeños. En lugar de una única división entrenamiento-validación, este método:

1. Divide los datos en k particiones (*folds*) (véase la Figura 2.7).
2. En cada iteración, usa $k-1$ particiones para entrenamiento y la restante para validación, rotando sistemáticamente la partición de validación hasta que cada una de las k particiones haya sido utilizada exactamente una vez como conjunto de validación.
3. Promedia los resultados de todas las iteraciones para obtener una métrica robusta.

El modelo final se entrena con todos los datos de entrenamiento (incluyendo los usados en validación durante el ajuste). Si bien esta técnica proporciona estimaciones más confiables, su costo computacional es significativo, ya que requiere entrenar el modelo $k+1$ veces (k iteraciones de validación más el entrenamiento final), lo que puede ser prohibitivo para modelos complejos, como redes neuronales profundas.

2.2. Deep Learning

El **aprendizaje profundo (*deep learning, DL*)** es una familia de técnicas de ML que utilizan múltiples capas de procesamiento para aprender representaciones de datos con varios niveles de abstracción [62]. Las redes neuronales han demostrado ser especialmente eficaces para este propósito, al permitir la composición jerárquica de características que capturan patrones cada vez más complejos en los datos.

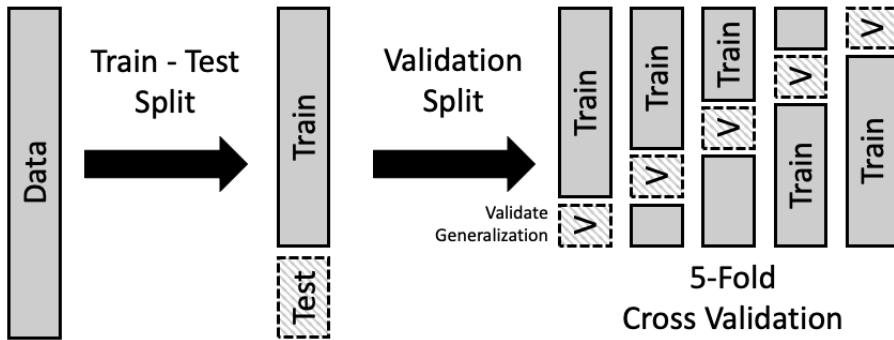


Figura 2.7: Diagrama de división del dataset para la validación cruzada. Recuperado de [61].

Las redes neuronales tienen su origen en el intento de modelar las redes de neuronas del cerebro humano [52]. Se requirió de numerosas contribuciones teóricas —como el perceptrón [53] o el algoritmo de *backpropagation* [63, 64], entre otras—, disponibilidad de datos estandarizados y un gran aumento en la capacidad computacional para poder escalar estas redes y obtener resultados sorprendentes en tareas complejas.

Las **redes neuronales profundas** (*deep neural networks*, DNNs) destacan por su capacidad para aprender representaciones jerárquicas: cada capa extrae características progresivamente más abstractas [62], desde líneas en imágenes hasta formas geométricas complejas, objetos completos e incluso escenas compuestas. Esta propiedad las hace excepcionalmente versátiles, ya que procesan datos de muy diversa naturaleza —datos tabulares, imágenes, audio, texto o señales temporales—, dados que ellas mismas aprenden los procesos de extracción de características de estos, hasta ahora realizados “a mano” (mediante procesos diseñados por la ingeniería de características) [56]². Gracias a ello, las DNNs han alcanzado rendimientos sobresalientes en dominios como visión por computador (clasificación de imágenes, detección de objetos, segmentación) o procesamiento de lenguaje natural (traducción, generación de texto) [65]. No obstante, su eficacia depende críticamente de grandes volúmenes de datos y recursos computacionales, lo que ha impulsado técnicas como el *transfer learning* y modelos eficientes para democratizar su uso.

²Este enfoque se denomina aprendizaje extremo a extremo (*end-to-end learning*), en el cual tanto la extracción de características como la clasificación son parte de un modelo integral que se entrena de manera conjunta, optimizando todos los componentes del sistema en un mismo proceso [56].

2.2.1. El perceptrón multicapa

El **perceptrón multicapa** (*multilayer perceptron*, MLP) forma la base del *deep learning*. Su diseño —con capas ocultas, funciones de activación no lineales y entrenamiento mediante *backpropagation*— sentó las bases conceptuales para arquitecturas más complejas, como las redes neuronales convolucionales o los *transformers* [66]. El MLP sigue siendo un referente teórico y la expresión más simple de cómo el aprendizaje jerárquico puede capturar patrones en los datos.

Cada nodo en la red es denominado **unidad o neurona artifical**. Siguiendo el diseño propuesto en [52, 53], cada unidad recibe señales de entrada —que o bien son las características de los datos o bien las salidas de las unidades de la anterior capa—, realiza una suma ponderada de estas con los pesos entrenables de cada conexión —más un término independiente o sesgo, también entrenable—, aplica una función no lineal sobre esta para producir una salida que propaga a las unidades de la siguiente capa (véase la Figura 2.8).

Matemáticamente, la operación de una unidad artifical se expresaría como:

$$y = f \left(\sum_{i=1}^n w_i x_i + b \right)$$

donde x_i son las entradas, w_i son los pesos entrenables (w_0 el sesgo)³, y f es la función de activación.

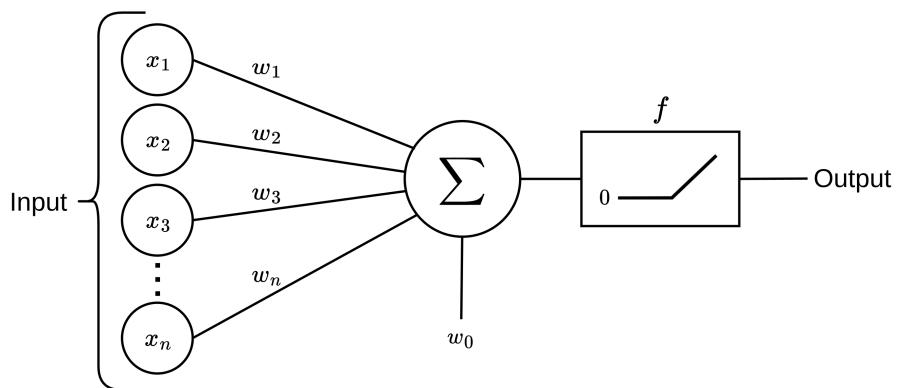


Figura 2.8: Esquema visual del funcionamiento de una unidad artificial. Adaptado de [67].

³El sesgo se considera un peso, puesto que, en la implementación, son un peso más conectado a una unidad de sesgo con valor constante unitario (1).

Esta **función de activación** a la salida de la unidad es un componente esencial que introduce no linealidad en el modelo, permitiendo a la red aprender relaciones complejas en los datos⁴. Existe multitud de funciones de activación, como la sigmoide, la tangente hiperbólica o ReLu —y sus múltiples variantes—, cada una con sus ventajas y limitaciones⁵.

La arquitectura de un MLP conecta estas unidades formando una red neuronal retroalimentada⁶, que consta de tres partes (véase la Figura 2.10):

- **Capa de entrada**, en las que el número de unidades debe coincidir con el formato de entrada de los datos, por ejemplo: en un problema con datos tabulares, debería haber una unidad por cada característica.
- **Capas ocultas**, donde se realizan las transformaciones no lineales de los datos. Es en estas donde el diseño puede variar en número de unidades y tipo de capas según la complejidad del problema y los datos.
- **Capa de salida**, que proporciona el resultado del modelo. Su forma depende del problema a resolver:
 - en problemas de regresión, esta capa tendrá tantas unidades como variables a predecir —sin función de activación, ya que esto limitaría el rango de valores posibles—;
 - en problemas de clasificación, esta capa tendrá una sola unidad —generalmente, con activación sigmoide— en clasificación binaria, o múltiples unidades —con activación *softmax*⁷— en clasificación multiclasa (véase la Figura 2.9).

2.2.2. Entrenamiento y validación de la red

En el caso de las redes neuronales, el conjunto de datos suele dividirse en tres subconjuntos: entrenamiento, validación y prueba. A diferencia de

⁴Sin ella, el MLP se reduciría a una simple combinación lineal de las entradas, incapaz de representar jerarquías de características [66].

⁵Si bien, actualmente, ReLU y sus variantes (*Leaky ReLU*, *Parametric ReLU* o *Swish*) se han convertido en el estándar *de facto* para las capas ocultas en DNNs, por su eficiencia computacional, y su eficacia empírica [68].

⁶Una red neuronal retroalimentada (*feed-forward neural network*) es aquella en la que las conexiones entre las unidades no forman un ciclo y, por tanto, la información solo se mueve en una dirección: adelante.

⁷La activación *softmax* no se aplica sobre la salida de una única unidad, sino que se aplica sobre un vector de salidas de múltiples unidades, transformándolas en una distribución de probabilidad, donde cada valor representa la probabilidad de pertenecer a una clase distinta y la suma de todas las salidas es igual a 1.

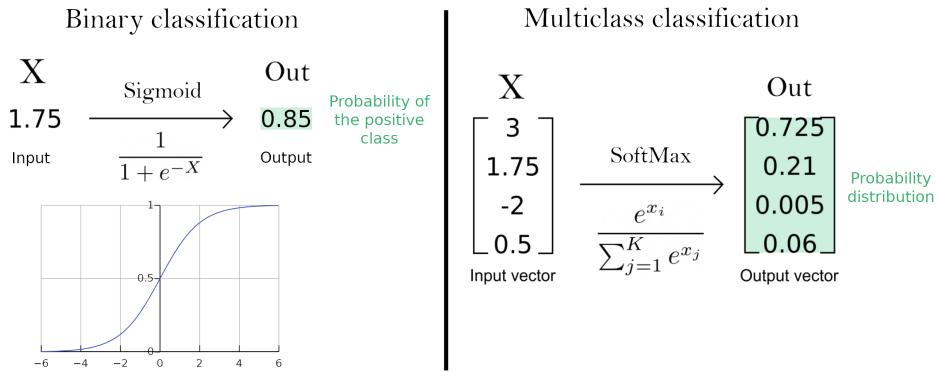


Figura 2.9: Diagrama de obtención de probabilidad en problemas de clasificación. Adaptado de [69].

métodos más tradicionales, no se utiliza validación cruzada, ya que entrenar redes profundas conlleva un elevado coste computacional.

Una vez hemos definido la arquitectura a emplear para resolver un problema, y definido los datos disponibles debemos entrenar la red con los datos de ejemplo. Este proceso implica ajustar los pesos del modelo para minimizar el error en las predicciones.

El método de entrenamiento estándar en redes neuronales es el **algoritmo de retropropagación (*backpropagation*)**, que funciona en dos fases clave [71]:

- **Propagación hacia adelante (*forward pass*):** Los datos de entrada se procesan a través de las capas de la red, generando una predicción.
- **Propagación del error hacia atrás (*backward pass*):** El error entre la predicción y el valor real se calcula y se propaga hacia atrás en la red, ajustando los pesos mediante el descenso de gradiente.

Sin entrar en demasiado detalle, esto consiste en calcular el gradiente de la función de pérdida con respecto a cada peso de la red, indicando cómo cada parámetro contribuye al error total. A mayor aporte al error de un peso, más se ajustará ese peso. Así, el algoritmo priorizará modificar significativamente los parámetros que más afectan al rendimiento de la red.

Este proceso explicado de manera vaga, tiene infinidad de detalles y variantes que influyen en su eficiencia y eficacia:

- El error obtenido entre la predicción y el valor real se calcula mediante la **función de pérdida (*loss function*)**. Esta función cuantifica el

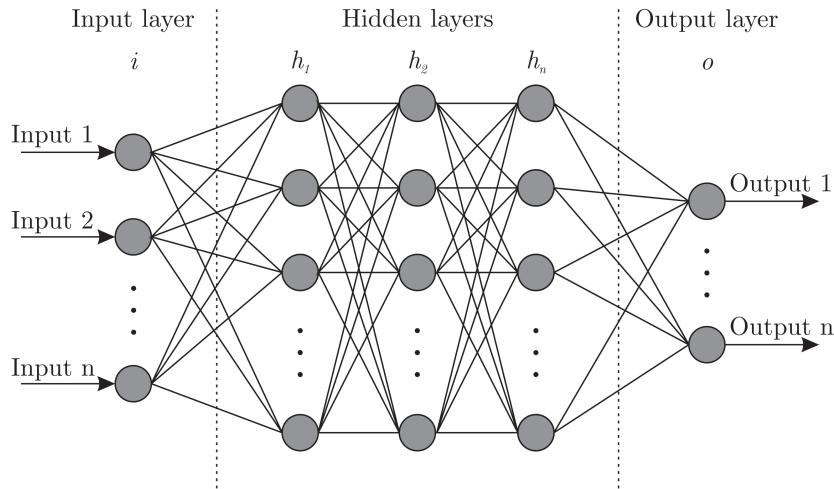


Figura 2.10: Arquitectura simplificada de un MLP. Recuperado de [70].

error del modelo durante el entrenamiento, midiendo la discrepancia entre las predicciones generadas y los valores o clases reales (*ground truth*).

No se debe confundir con las métricas de evaluación de un modelo: aunque en algunos casos se pueden usar métricas como funciones de pérdida y viceversa, las métricas destacan por ser fáciles de interpretar y suele utilizarse más de una. En cambio, debe existir una única función de pérdida durante el entrenamiento de una red neuronal, que debe cumplir tres requisitos clave:

1. Reflejar el objetivo del aprendizaje: Debe capturar adecuadamente qué significa “éxito” para el modelo (p.ej., minimizar el error en regresión o maximizar la probabilidad de clasificación correcta).
2. Ser diferenciable: Es esencial para aplicar técnicas de descenso por gradiente, ya que el optimizador necesita calcular derivadas.
3. Ser eficiente computacionalmente: Dado que se evalúa en cada iteración del entrenamiento, su cálculo debe ser rápido incluso con grandes volúmenes de datos.

Mientras las métricas ayudan a entender el modelo, la función de pérdida es la que lo entrena.

En problemas de regresión se emplean funciones de pérdida como el error cuadrático medio (*mean squared error*, MSE), que mide la diferencia promedio al cuadrado entre las predicciones y los valores reales, o el error absoluto medio (*mean absolute error*, MAE), que calcula la

diferencia promedio en valor absoluto ⁸.

En clasificación, las funciones de pérdida más comunes son la entropía cruzada (*cross-entropy loss*) para problemas de clasificación binaria y multiclase, que penaliza fuertemente las predicciones incorrectas y ayuda a optimizar las probabilidades predichas para cada clase.

- Existen multitud de **algoritmos de optimización de parámetros**, como SGD, Adam o RMSProp. Estos algoritmos determinan cómo actualizar los pesos del modelo durante el entrenamiento para minimizar la función de pérdida. Están basados en el descenso de gradiente, que ajusta los pesos en dirección opuesta al gradiente de la función de pérdida respecto a los pesos, multiplicado por un factor escalar llamado **tasa de aprendizaje** (*learning rate*). Este hiperparámetro controla la magnitud de los pasos de actualización: un valor demasiado alto puede hacer que el entrenamiento diverja, mientras que uno demasiado bajo ralentiza la convergencia o estanca el modelo en mínimos locales.

Existen estrategias avanzadas para ajustar el *learning rate* de manera más eficiente durante el entrenamiento, como la búsqueda de un *learning rate* de punto de partida

- Si bien existen métodos de entrenamiento de redes ejemplo a ejemplo —como el Gradiente Descendente Estocástico (SGD) puro[72]—, estas se suelen entrenar por lotes (*minibatches*) ⁹ debido a ventajas clave, como el aprovechamiento de la paralelización de operaciones en GPU y una mayor estabilidad en la función de pérdida al promediarse el error entre varios ejemplos. Aún así, establecer un tamaño de lote óptimo no es una tarea trivial que requiere de encontrar un equilibrio entre generalización y velocidad: los lotes grandes aceleran el entrenamiento pero pueden reducir la generalización del modelo, mientras que los lotes pequeños puede presentar una gran varianza que introduzca ruido en el modelo [73], si bien esto puede ayudar a escapar de mínimos locales, y puede paliarse con un bajo *learning rate* (aunque esto aumentaría todavía más los tiempos de entrenamiento).
- Tras el uso de *minibatches* en el entrenamiento, surge el concepto de **época** (*epoch*), que hace referencia a un ciclo completo de presentación de todos los datos de entrenamiento a la red neuronal [56]. Durante una época, los *minibatches* se procesan secuencialmente, actualizando los pesos del modelo en cada iteración (o *step*) con el gradiente calculado sobre un lote. Por ejemplo, si un conjunto de entrenamiento

⁸Aunque esta no es derivable en $x = 0$, se define la derivada en ese punto como 0.

⁹Se denomina *batch* al *dataset* completo, y *minibatch* a los subconjuntos de este cuyo tamaño está determinado por el hiperparámetro *batch size*.

tiene 4096 ejemplos y el tamaño de lote es 32, una época constará de 128 iteraciones ($4096/32$).

El número de épocas es un hiperparámetro crítico: demasiadas pueden llevar a sobreajuste (*overfitting*), donde el modelo memoriza los datos de entrenamiento pero no generaliza bien; demasiado pocas pueden resultar en infraajuste (*underfitting*), donde el modelo no captura los patrones subyacentes. Además, la combinación de tamaño de lote y épocas influye en la dinámica de optimización, ya que lotes más pequeños requieren más pasos por época, introduciendo más ruido pero potencialmente mejorando la exploración del espacio de pesos.

En la práctica, se suele establecer un número muy alto de épocas, y monitorizar el error en un conjunto de validación para determinar cuándo detener el entrenamiento, evitando así el sobreajuste cuando el error de validación comienza a aumentar. A esta técnica se le denomina *early stopping* [74].

2.2.3. Redes Neuronales Convolucionales

Como ya se venía anticipando, la arquitectura MLP es especialmente adecuada para trabajar con datos estructurados o tabulares, donde la información se organiza en una matriz en la que cada columna representa una característica concreta (como sexo, altura o peso). Sin embargo, su diseño presenta limitaciones clave: al manejar vectores de entrada de tamaño fijo y carecer de mecanismos para aprovechar relaciones espaciales o secuenciales, no es óptima para datos no estructurados, como imágenes o texto, donde cada elemento individual (un píxel o una palabra) carece de significado por sí mismo [66].

Por ejemplo, los patrones aprendidos en una posición de una imagen podrían no ser reconocidos en otra ubicación, ya que las entradas tienen un recorrido distinto dentro de la red. Por tanto, el modelo carecería de **invarianza traslacional**, puesto que los pesos no se comparten entre distintas posiciones, a lo que se suma una marcada inefficiencia por el elevado número de parámetros requeridos [71].

Precisamente para estos casos, otras arquitecturas profundas resultan más apropiadas. Las **redes neuronales convolucionales** (*Convolutional Neural Network* en inglés, CNNs) son un tipo de DNN que, aprovechando las ventajas de las operaciones convolucionales, explotan los principios de localidad y correlación espacial. Esto les permite procesar imágenes (en 1D, 2D o 3D) de manera eficiente, interpretando patrones visuales jerárquicos que un MLP no podría capturar, y con significativamente menos parámetros.

Capas convolucionales

Como se ha introducido antes, el operador de **convolución** es la base de las CNNs. Este operador matemático aplica un **filtro** (también denominado *kernel*)¹⁰ a regiones locales de una imagen de entrada, realizando un producto punto¹¹ entre los valores del filtro y los píxeles correspondientes de la imagen, y sustituyendo el valor del pixel central por el resultado del producto (véase la Figura 2.11).

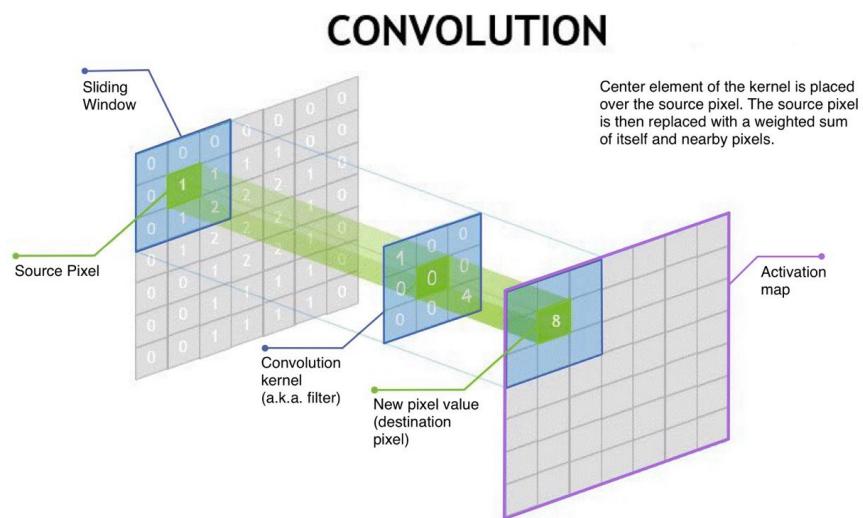


Figura 2.11: Esquema gráfico de la aplicación de un filtro convolucional sobre una región de una imagen. Adaptado de [75].

Este proceso se repite al desplazar el filtro por toda la imagen mediante una **ventana deslizante**, generando un **mapa de activación**, que permite destacar líneas, curvas o texturas simples. Este mapa de activación preserva la información de la localización de las características, si bien estas pueden ser detectadas en cualquier parte de la imagen. Esta propiedad se conoce como **equivarianza**.

Las CNNs aprovechan la convolución mediante **capas convolucionales**. Cada capa convolucional está compuesta por un conjunto de filtros convolucionales, donde cada uno a su vez tiene tantos *kernels* como canales de

¹⁰Aunque, como veremos después, a la hora de hablar de capas convolucionales, no son lo mismo.

¹¹El producto punto o producto escalar de dos vectores, se define como la suma de los productos componente a componente.

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}_1 \cdot \mathbf{v}_1 + \mathbf{u}_2 \cdot \mathbf{v}_2 + \dots + \mathbf{u}_n \cdot \mathbf{v}_n$$

entrada de la imagen haya en la capa (si es la primera capa convolucional, habrá 1 canal en imágenes de escala de grises, o 3 en imágenes RGB). El número de filtros en cada capa, su tamaño y la forma en que se deslizan sobre la entrada ¹² se determinan durante el diseño de la red, mientras que los valores de los *kernels* son parámetros entrenables.

Cada filtro convolucional realiza la operación convolucional sobre cada canal con el *kernel* que le corresponde. Después, se suman los mapas de activación de cada canal (pixel a pixel) añadiendo un sesgo (un mismo valor a todos los píxeles ¹³), generando lo que denominamos como **mapa de características** (ya que idealmente extrae características relevantes). Los mapas de características generados con cada uno de los filtros son los nuevos canales, que conforman la salida de la capa convolucional. Esta salida puede ser posteriormente procesada por otras capas, permitiendo a la red aprender representaciones jerárquicas cada vez más abstractas de los datos de entrada: las primeras capas convolucionales detectarán bordes, cambios de color o texturas básicas; a medida que avanzamos en las capas de la red, las combinaciones de estas características simples permite identificar formas más complejas, como objetos e incluso composiciones.

Añadir una imagen explicando stride y padding

Sin embargo, hemos pasado por alto algo fundamental: ¿cómo reunimos la información de dos regiones distantes de una imagen en un mismo sitio? Una primera aproximación intuitiva nos diría que los filtros convolucionales deben ser progresivamente más grandes, para capturar patrones de mayor tamaño y contexto. No obstante, esto incrementaría considerablemente el número de parámetros y, por tanto, aumentaría el coste computacional y aumentaría el riesgo de sobreajuste del modelo (ya que un modelo con más parámetros puede memorizar mejor los datos de entrenamiento). Es por esto que, en aquellos problemas en los que no es necesario preservar la información de localización de las características, —como en los que nos enfocamos en este trabajo: clasificación y regresión—, y, por tanto, el modelo sea invariante a la ubicación, se emplean técnicas de submuestreo (*downsampling*) [66], como usar *stride* mayor de 1 en los filtros de las capas convolucionales o realizar *pooling* ¹⁴.

Capas de pooling

Las **capas de agrupación** (*pooling layers*) tienen como objetivo principal comprimir la información de la imagen, reduciendo sus dimensiones (alto y ancho) mientras se preservan los datos más relevantes para la tarea.

¹²Definidos mediante los parámetros de *stride* y *padding*, que controlan el desplazamiento del filtro y la cantidad de relleno alrededor de la entrada, respectivamente.

¹³Es por ello que no rompe la propiedad de equivarianza.

¹⁴Nos centraremos en el último dado su amplio uso y fácil comprensión, además de su demostrada efectividad empírica.

Esta reducción del tamaño espacial de los mapas de características disminuye el número de parámetros y operaciones en las fases posteriores, lo que reduce el coste computacional. Además, tiene un beneficio adicional: ayuda a prevenir el sobreajuste, ya que al limitar la cantidad de parámetros, el modelo evita memorizar ruido o detalles irrelevantes de los datos de entrenamiento, favoreciendo así el aprendizaje de patrones generalizables.

Hay diversos métodos de *pooling*, entre los que destacan:

- **Max pooling**, que calcula el máximo valor de regiones del mapa de características, y lo usa para crear un mapa de características reducido (véase la Figura 2.12).
- **Average pooling**, que reemplaza el valor máximo del *max pooling* por el cálculo de la media entre los valores de la región.

La región de aplicación del *pooling*, al igual que en la convolución, viene determinada por ciertos parámetros, definidos por el diseñador, como el tamaño de filtro (que suele ser de 2x2), el *stride* y el *padding*, si bien también existen variantes adaptativas (*adaptive*), que ajustan automáticamente su cobertura para producir una salida con dimensiones específicas, independientemente del tamaño de la imagen de entrada. Esta funcionalidad es especialmente útil cuando se necesita adaptar los mapas de características para conectarlos a una capa *fully-connected*.

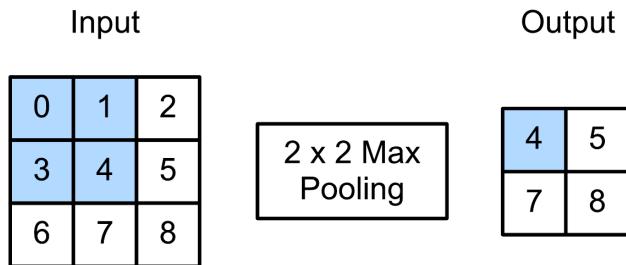


Figura 2.12: Esquema gráfico de *max pooling* con un filtro 2x2 y *stride* de 1. Recuperado de la Figura 14.12 de [66].

Capas *Fully-Connected*

Como hemos visto hasta ahora, en las CNNs, las primeras capas están diseñadas para extraer características espaciales a través de filtros convolucionales y de *pooling*. Sin embargo, una vez que se ha reducido la dimensionalidad y se han obtenido representaciones abstractas de alto nivel, es necesario realizar una predicción (en problemas de clasificación y regresión).

Aquí es donde las **capas completamente conectadas (fully-connected, FC)** juegan un papel crucial. Se utilizan en las últimas etapas de la red convolucional para combinar todas las características extraídas y producir una predicción final. Es decir, actúan como el clasificador/regresor ¹⁵ que toma todas las señales procesadas por las capas anteriores y predice la clase a la que pertenece la imagen o el valor objetivo.

La arquitectura de esta capa sigue la estructura del MLP, con neuronas organizadas en una o más capas densas, donde cada neurona está conectada con todas las salidas de la capa anterior. Para que esto sea posible, primero se aplica una operación de **flattening** que transforma el mapa de características multidimensional en un vector unidimensional. A partir de ahí, el procesamiento es equivalente al de una red neuronal tradicional: cada neurona calcula una combinación lineal de sus entradas seguida de una función de activación no lineal.

Diseño de la CNN para problemas de clasificación y regresión

Un patrón común de diseño de CNNs para la resolución de problemas de clasificación y regresión consta de dos componentes principales:

- el *backbone* o extracto de características, que alterna capas convolucionales con capas de *pooling*, cuya función es extraer representaciones jerárquicas y cada vez más abstractas de los datos de entrada; y
- el *classifier*, generalmente implementado mediante una o más capas totalmente conectadas, toma estas representaciones para realizar la tarea específica de salida, ya sea clasificación o regresión.

Regularización y normalización

Como en otras arquitecturas de redes neuronales, existen numerosas técnicas de regularización para evitar el sobreajuste. Veamos algunas de las técnicas empleadas en CNNs:

- **Data augmentation** [76, 77]: Consiste en añadir o modificar dinámicamente ejemplos a partir de los que se tienen originalmente, de forma que se entrene la red con un conjunto de datos más diverso y robusto, evitando el sobreajuste y mejorando la generalización.

¹⁵Si bien, independientemente de la tarea —regresión o clasificación—, a esta parte de la red se le denomina clasificador

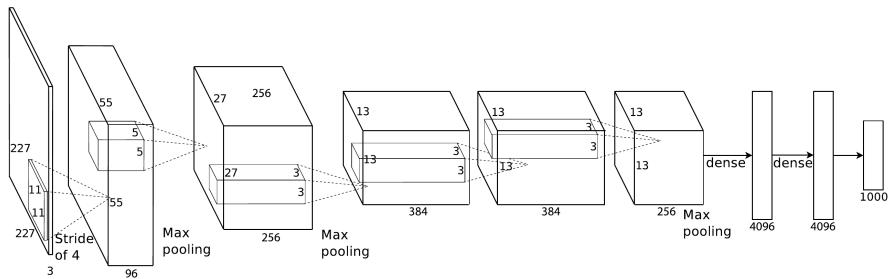


Figura 2.13: Esquema gráfico de la arquitectura conocida como “AlexNet”, diseñada para resolver un problema de clasificación con 1000 clases. Recuperado de la Figura 5.39 de [71]. Esta arquitectura presenta una serie de capas convolucionales con funciones de activación no lineales ReLU, max pooling, algunas capas totalmente conectadas y una capa final *softmax*, la cual se alimenta a una función de pérdida de entropía cruzada multiclas.

Algunas alteraciones realizadas pueden ser cambios en el nivel de brillo y contraste, rotaciones, traslaciones, escalados o volteos de imágenes, entre otras. No existe configuración óptima, y su configuración depende mucho del problema y las imágenes disponibles.

Esta técnica sirve especialmente para problemas como clasificación o regresión, donde las clases o valores predichos no suelen variar bajo pequeñas perturbaciones locales.

- **Dropout** [78]: Técnica que, durante el entrenamiento, “apaga” (pone a cero) aleatoriamente un porcentaje de neuronas en cada iteración, evitando así que la red dependa demasiado de determinadas unidades individuales (véase la Figura 2.14). En CNNs suele aplicarse a capas *fully-connected*, aunque existen variantes como *Spatial Dropout* [79] que elimina canales completos en capas convolucionales, forzando una distribución más robusta de características.
- **Batch normalization** [80]: Esta se introduce como una capa nueva a añadir en el diseño de las redes, con nuevos parámetros entrenables: *scale* y *shift*. Normaliza los valores de cada canal (media cero y desviación 1), y los reescaliza y desplaza en base a los valores de *scale* y *shift*. Esto suaviza significativamente el espacio de valores de optimización [81] y reduce la sensibilidad a la tasa de aprendizaje [82], permitiendo establecer valores más altos. En CNNs se aplica típicamente después de las capas convolucionales y antes de la función de activación

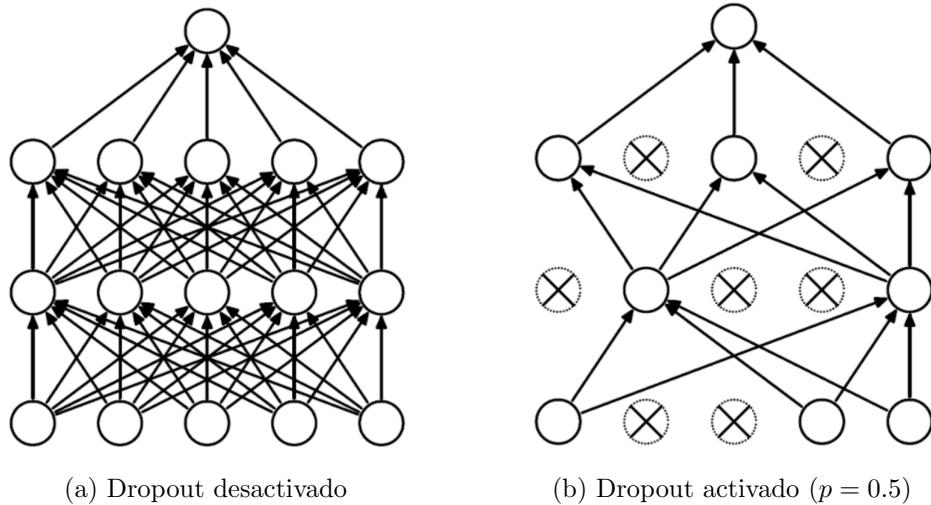


Figura 2.14: Diagrama del funcionamiento de neuronas con *dropout*. Recuperado de la Figura 5.29 de [71]. Cuando se evalúa el modelo, todas las unidades funcionan correctamente (a). Durante el entrenamiento, algunas son “apagadas” (b).

Conexiones residuales

Uno de los principales problemas que no permite aumentar mucho la profundidad de las redes convolucionales es el desvanecimiento de gradiente (*vanishing gradient problem*), que consiste en la disminución exponencial de los gradientes durante el proceso de *backpropagation* a medida que se retrocede hacia las capas iniciales de la red. Algunas de las soluciones a este problema han sido: utilizar funciones de activación ReLU, ya que evita gradientes pequeños para valores positivos; inicializar adecuadamente los pesos de la red; o *batch normalization*, que estabiliza la distribución de las activaciones. Sin embargo, las conexiones residuales han sido la contribución más significativa para resolver este problema.

Las **redes residuales** (*residual nets*, ResNet)

2.2.4. Transfer Learning

El **aprendizaje por transferencia** (*transfer learning*) es una técnica que consiste en aprovechar el conocimiento aprendido por un modelo entrenado en una tarea como punto de partida para mejorar el rendimiento y acelerar el entrenamiento en una nueva tarea relacionada [56].

En redes neuronales, el aprendizaje consiste en ajustar pesos, y en el caso del *transfer learning*, estos pesos se inicializan con valores previamente opti-

mizados para una tarea fuente, en lugar de comenzar con valores aleatorios (véase la Figura 2.15).

Se conoce como **fine-tuning** a la técnica de inicialización de los pesos de aquellas partes del modelo (como capas convolucionales) con los pesos previamente aprendidos, y que continúa el entrenamiento con los datos específicos de la nueva tarea. En este contexto, se denomina *head* a las capas finales del modelo que se sustituyen para adaptarse a la nueva tarea.

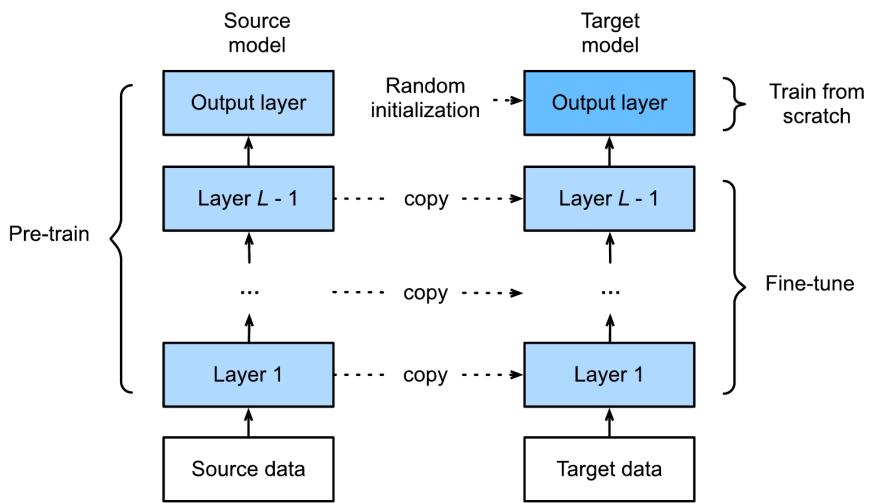


Figura 2.15: Diagrama de *fine-tuning* de un modelo en una nueva tarea. Recuperado de la Figura 19.2 de [66]. La capa final de salida es entrenada desde cero para la nueva tarea. El resto de capas son inicializadas con los pesos previos.

Por ejemplo, en [35] se utilizan dos modelos de CNN preentrenados en clasificación con ImageNet (que contiene imágenes de 1000 clases): VGG16 y ResNet50. Estos modelos se ajustan (*fine-tuning*) para estimar el sexo de una persona a partir de radiografías de húmero. Aunque ambas tareas parecen muy diferentes, las primeras capas de la red, especializadas en detectar características generales como bordes y texturas, pueden ser útiles en los dos casos, lo que permite una transferencia efectiva del conocimiento.

El *fine-tuning* puede aplicarse de forma gradual: primero se entrena solo el *head* (manteniendo el resto del modelo congelado) y luego, si es necesario, se afinan también algunas capas preentrenadas para mejorar el rendimiento en la tarea específica.

2.3. Incertidumbre

La metrología¹⁶, y la estadística comparten un papel fundamental en el análisis del error y la incertidumbre en campos como el ML. Mientras la metrología establece los fundamentos conceptuales de error e incertidumbre, la estadística proporciona métodos para cuantificar, modelar y reducir estos factores durante el desarrollo y validación de modelos.

El Comité Conjunto de Guías en Metrología (*Joint Committee for Guides in Metrology*)¹⁷ define el **error** como una “medición imperfecta” de la magnitud observada, que puede estar causada por efectos aleatorios (componente aleatoria del error) y por efectos sistemáticos (componente sistemática del error, más conocida como **sesgo**). Por otro lado, define a la **incertidumbre** como “parámetro, asociado con el resultado de una medición, que caracteriza la dispersión de los valores que podrían atribuirse razonablemente al **mensurado**, que es como se denomina a la magnitud a ser medida. [...] El parámetro puede ser, por ejemplo, una desviación estándar, o la anchura de un intervalo con un nivel de confianza establecido” [84].

Partiendo de estas definiciones generales, veamos las diferencias entre los dos enfoques principales en la evaluación de mediciones: el enfoque basado en el error y el enfoque basado en la incertidumbre.

El **enfoque basado en el error** o enfoque tradicional parte de la premisa de que existe un valor verdadero. En consecuencia, el propósito de la medición es aproximarse lo más posible a dicho valor, minimizando las distintas componentes del error [84]:

- para el error aleatorio, esto se logra aumentando el número de observaciones, ya que su distribución tiene una media igual a cero; y
- para el error sistemático, es necesario identificarlo y cuantificar su magnitud, lo que permite aplicar factores de corrección que compensen su efecto.

Se asume que el resultado de la medición ha sido corregido por todos los efectos sistemáticos identificados como significativos, de modo que la esperanza matemática de esta componente sea igual a cero.

¹⁶Ciencia de las mediciones y sus aplicaciones [83].

¹⁷Este Comité está formado por numerosas organizaciones internacionales de metrología y normalización: BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML e ILAC. Su objetivo principal es mantener y promover las guías internacionales clave en metrología, como la Guía para la Expresión de la Incertidumbre en la Medición (*Guide to the Expression of Uncertainty in Measurement*, GUM) [84] y el Vocabulario Internacional de Metrología (*Vocabulaire international de métrologie*, VIM) [83].

Sin embargo, en la práctica no existen reglas claras para distinguir las componentes del error ni cómo estas se combinan en el error total que permitan diferenciar claramente las componentes del error ni cómo estas se combinan en el error total. En general, solo es posible estimar un límite superior del valor absoluto del error total estimado, al que se denomina de forma inapropiada “incertidumbre”.

Frente a enfoque anterior, se presenta el **enfoque basado en la incertidumbre** [84], cuyo propósito no es hallar el mejor valor posible, sino establecer un intervalo de valores razonables para el mensurando, el cual puede refinarse con información adicional. Así, la medición misma se convierte en una herramienta para determinar el error del instrumento.

Creo que es mejor eliminar este apartado e incluirlo en el anexo pero lo dejo por ahora para que vea la diferencia entre el intervalo de confianza, el de credibilidad y el de predicción

2.3.1. Intervalos de valores razonables

Veamos qué tipos de intervalos de valores nos permiten cuantificar la variabilidad de los resultados y, por tanto, la incertidumbre de la medición realizada.

- El **intervalo de confianza (IC)** es una herramienta común de la estadística frecuentista¹⁸, que permite estimar un rango de valores tal que podamos confiar en que contiene al valor verdadero de un parámetro poblacional desconocido θ (p.ej., la media) [85].

Los métodos del cálculo del IC dependen de la distribución del estimador (p.ej., la distribución de la media muestral) y los parámetros conocidos.

Es importante aclarar un malentendido común: un intervalo de confianza con nivel 95 % para un parámetro θ no significa que exista un 95 % de probabilidad de que θ esté dentro del intervalo calculado a partir de una muestra específica. En realidad, el 95 % se refiere a la frecuencia con la que, si muestreásemos muchas veces los datos, los intervalos construidos a partir de esas muestras incluirían al valor verdadero de θ en aproximadamente el 95 % [66].

- El **intervalo de credibilidad o región creíble (RC)** es, de hecho, la que determina que el parámetro θ está contenido en el rango de sus valores con una probabilidad determinada por la confianza. Este intervalo es la aproximación bayesiana equivalente al intervalo de confianza, y, como este, requiere conocer la distribución a priori de los datos.

¹⁸La estadística frecuentista ... ¿Anexo explicando las diferencias entre frecuentista y bayesiana? (para agosto)

La diferencia radica en que, a diferencia del intervalo de confianza, que parte de que θ es un parámetro fijo desconocido y los datos son tratados como aleatorios, el enfoque bayesiano fija los datos (ya que son conocidos) y el parámetro θ lo trata como aleatorio (ya que es desconocido) [66].

Esta interpretación resulta más intuitiva y directa en comparación con la interpretación frecuentista del intervalo de confianza. En particular, una región creíble del 95 % sí puede interpretarse como que hay un 95 % de probabilidad de que el parámetro θ se encuentre dentro de ese intervalo, dado el conjunto de datos observado y la distribución a priori asumida.

- El **intervalo de predicción (prediction interval)** es radicalmente diferente a los intervalos previos, pues trata de predecir un valor futuro de una observación, no determinar un parámetro poblacional. Existen numerosos métodos, con mayores y menores garantías estadísticas, con y sin necesidad de conocer la distribución de los datos. El enfoque más prometedor es la predicción conformal, que han demostrado ser eficaz en contextos donde los supuestos clásicos (normalidad, homocedasticidad) no se cumplen [86], y es actualmente el enfoque más robusto para la construcción de intervalos de predicción en aplicaciones modernas de ML [86-90].

Incluir imagen comparando intervalo de confianza y credibilidad

Como podemos esperar, a más estrecho sea el intervalo que manejemos, más se puede confiar en las predicciones, pero no todos los tipos de intervalos revelan la misma información sobre incertidumbre.

2.3.2. Incertidumbre en *machine learning*

El enfoque basado en incertidumbre se puede extrapolar a modelos de ML, incorporando técnicas de **cuantificación de la incertidumbre (Uncertainty Quantification, UQ)** para evaluar y comunicar la confianza del modelo en sus predicciones. En problemas de regresión, la analogía es directa: es deseable que los modelos no solo proporcionen una predicción puntual, sino también un intervalo que indique el grado de incertidumbre asociado a cada predicción, conocido como **intervalo de predicción (prediction interval)**, el cual puede derivarse de métodos de UQ como *bootstrapping* o predicción conformal. En caso de los problemas de clasificación, el concepto equivalente al de intervalo de predicción se denomina **conjunto de predicción (prediction set)**, que puede construirse mediante técnicas como estimaciones de probabilidad calibrada o predicción conformal.

Las fuentes de incertidumbre pueden ser muy variadas, y su identificación requiere en muchos casos de conocimiento específico en el problema. Si bien

se suelen considerar dos tipos de incertidumbre en las predicciones realizadas en ML [91]:

- La **incertidumbre aleatoria** es la relativa al dato individual. Esta incertidumbre se debe a la variabilidad inherente del fenómeno observado y no puede reducirse, aunque se disponga de más datos. Por ejemplo, en un entorno médico, puede reflejar la variabilidad entre pacientes con condiciones similares.
- La **incertidumbre epistémica** es la causada por falta de conocimiento o precisión del modelo. Se relaciona con aspectos como la escasez de datos, la calidad de la información disponible o la capacidad limitada del modelo para generalizar. A diferencia de la incertidumbre aleatoria, la epistémica es reducible: puede disminuirse con más datos, mejores modelos o mayor comprensión del problema.

A estos, se le puede añadir un tercero: el ***drift*** [92], que procede de cambios en la distribución de los datos a lo largo del tiempo, ya sea en la distribución de las variables de entrada, en la distribución de las variables de salida, o en la relación entre las dos previas. Por ejemplo: una imagen de entrada a un modelo de clasificación que no corresponde a ninguna clase con la que se haya entrenado anteriormente, un cambio en la población objetivo de una aplicación médica —p.ej., debido a un cambio demográfico o a la aparición de una nueva enfermedad—, entre otros.

2.3.3. Cuantificación de la incertidumbre en *machine learning*

La cuantificación de la incertidumbre puede abordarse desde dos perspectivas:

- sobre nuevos datos (no vistos), donde el objetivo es estimar la confianza del modelo en situaciones reales de predicción, donde no se conoce el valor verdadero; o
- sobre datos ya observados, como los del conjunto de entrenamiento o validación. En este caso el propósito es evaluar si el modelo representa adecuadamente la variabilidad de los datos, detectar sobreajuste o validar la calibración de las predicciones.

Nos centraremos en el estudio de la cuantificación de la incertidumbre sobre datos nuevos, pues es en este contexto donde resulta más relevante para aplicaciones prácticas, como las de estimación del PB. Esta provee

s posible haber de métodos cuantificación incertidumbre sobre datos nuevos sin mencionar métodos sobre los previos? En la calificación, la calificación del modelo es importante para la cuantificación de incertidumbre. Hecho, es recomendable realizarlaivamente a la calificación de conmial prediction.

información clave para la toma de decisiones tanto del valor predicho como de la certeza asociada a la predicción.

Además, nos centraremos en aquellos métodos que no requieran asumir distribuciones específicas en los datos ni dependan fuertemente de supuestos paramétricos, ya que en muchos problemas reales (como la estimación del perfil biológico) los datos pueden presentar heterocedasticidad, asimetría o comportamientos complejos que dificultan su modelización con enfoques clásicos.

Cuantificación de la incertidumbre en problemas de regresión

Existe multitud de maneras de cuantificar la incertidumbre en problemas de regresión, más y menos rigurosas.

Una primera aproximación es la **regresión cuantílica (*quantile regression*)**, que añade dos nuevas salidas en el modelo de regresión, que serán los límites inferior y superior de un intervalo de predicción, correspondientes a cuantiles específicos (por ejemplo, los percentiles 10 y 90)¹⁹. Esto permite estimar no solo la tendencia central (como en la regresión tradicional), sino también la incertidumbre de las predicciones (véase la Figura 2.16).

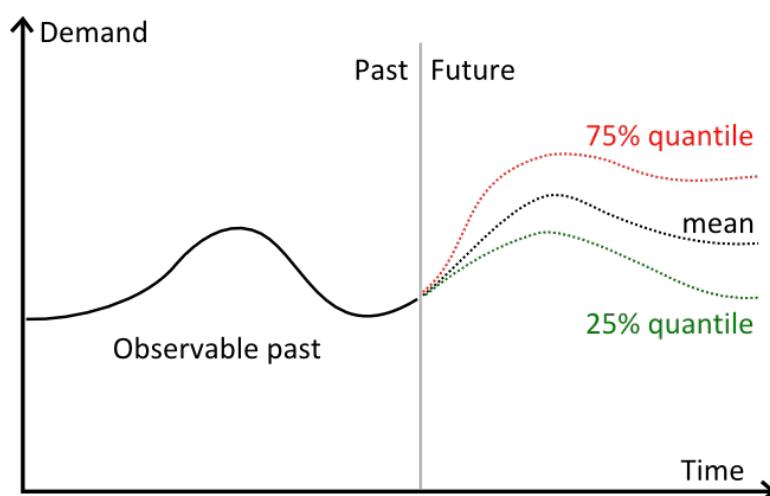


Figura 2.16: Gráfico que ilustra las 3 predicciones que arroja un modelo de regresión cuantílica. Recuperado de [94]. Se observa los límites superior e inferior los percentiles 75 y 25, respectivamente.

Sin embargo, este método presenta varios problemas:

¹⁹El entrenamiento del modelo para estas dos salidas se realiza con la función de pérdida *pinball*, [93] que combina los errores de predicción para múltiples cuantiles, penalizando asimétricamente las desviaciones según el cuantil objetivo.

- Sensibilidad a datos escasos o ruidosos: La estimación puede ser muy sensible a *outliers*²⁰ y a regiones con pocas observaciones, llevando a intervalos poco fiables. De hecho, se puede dar el fenómeno indeseable de que cuantiles mayores tengan valores predichos más bajos que cuantiles menores —conocido como cruzamiento de cuantiles [95]—.
- Falta de cobertura probabilística garantizada: No existe ningún tipo de garantía estadística que asegure que el cuantil estimado cubra la proporción deseada de observaciones en muestras futuras.
- No captura la incertidumbre epistémica: Solo captura la incertidumbre aleatoria [96], dado que modela cómo los valores de salida varían dado un conjunto de entradas. No captura automáticamente incertidumbre epistémica, a menos que se combine con otros enfoques, como métodos bayesianos [97].
- Muestran gran incertidumbre en cuantiles extremos (cerca de cero o uno) [98] y, por tanto, los intervalos de predicción obtenidos son extremadamente grandes, lo que logra una gran cobertura pero poca utilidad práctica.

Introducir la predicción conformal

Otra técnica que no ...

¿Introducir la calibración de modelos y centrarse en el Platt Scaling y Temperature Scaling?

Cuantificación de la incertidumbre en problemas de clasificación

2.4. Conformal Prediction

La predicción conformal (*conformal prediction*, CP)

Introducir la predicción conformal

2.4.1. Conformal Prediction en problemas de regresión

Conformalized Quantile Regression (CQR)

2.4.2. Conformal Prediction en problemas de clasificación

...

²⁰Los valores atípicos o *outliers* son instancias que se desvian significativamente del resto de instancias del conjunto de datos. Un *outlier* podría indicar un comportamiento anormal del sistema (variabilidad natural del fenómeno observado, eventos excepcionales que refleja comportamientos reales pero poco frecuentes, ...) o un error de recolección y registro de los datos [58].

Original Sentence	Adversarial Example
There is really but one thing to say about this sorry movie It should never have been made The first one one of my favourites An American Werewolf in London is a great movie with a good plot good actors and good FX But this one It stinks to heaven with a cry of helplessness	There is really but one thing to say about that sorry movie It should never have been made The first one one of my favourites An American Werewolf in London is a great movie with a good plot good actors and good FX But this one It stinks to heaven with a cry of helplessness
Negative sentiment	Positive sentiment

Figura 2.17: Ejemplo adverario mal clasificado por un modelo ML entrenado con datos textuales. Adaptado de la Figura 2 de [91], original de [99]. Se observa que el cambio de una sola palabra —y aparentemente sin mucha relevancia— (destacada en negrita) basta para cambiar la predicción de “sentimiento negativo” a “sentimiento positivo”.

Least-Ambiguous Set-Valued Classifiers

...

Adaptive Prediction Sets

El *Adaptive Prediction Sets* (APS) [89]

Regularized Adaptive Prediction Sets

Regularized Adaptive Prediction Sets (RAPS) [90] es una variante del método APS, que

añade una penalización a conjuntos de predicción demasiado grandes, realizando esto a través de la suma de un componente

Capítulo 3

Estado del arte

- 3.1. Estimación de la edad en antropología forense**
- 3.2. Estimación de la edad en antropología forense usando Machine Learning**
- 3.3. Cuantificación de incertidumbre en antropología forense**

Capítulo 4

Materiales y métodos

4.1. Conjunto de datos disponibles

Disponemos de un conjunto de datos compuesto por radiografías panorámicas maxilofaciales de individuos de diversos países y continentes (véase en la tabla 4.1), obtenidas con distintos modelos de máquinas de rayos X¹. Este conjunto de datos ha sido proporcionados por Panacea Cooperative Research, empresa *spin-off* de la Universidad de Granada.

Este *dataset* incluye:

- datos tabulares (en formato CSV), donde cada fila representa un ejemplo (un individuo), con los siguientes campos: un identificador único, sexo del individuo, edad del individuo y “muestra” (clasificación según el origen geográfico de la radiografía), e
- imágenes bidimensionales de radiografías panorámicas maxilofaciales, con una imagen asociada a cada individuo y se identifica mediante su ID único.

Se proporcionan los datos ya preprocesados, por lo que no es necesario realizar tareas adicionales de limpieza o transformación previa antes de su análisis.

Se ignora el campo de “muestra”, dado que se trata de una asignación sesgada y no representa necesariamente una clasificación fiable del origen poblacional de los individuos. Por tanto, este campo no se emplea en el

¹Los modelos empleados fueron: *Planmeca Promax Digital Panoramic*; *Sirona ORTHOPHOS-XG*, *ORTHOPHOS-DS*, y *SIDEXIS*. Las constantes radiológicas usadas fueron de 66 a a 70 kV, 7 a 11 mA, y 15 s.

País	Instituciones	Nº de ejemplos
Bosnia y Herzegovina	Universidad de Sarajevo	882
Botsuana	Dos clínicas dentales privadas en Garobone	1242
Chile	Dos clínicas dentales privadas en Santiago y Rancagua	1016
República Dominicana	Tres clínicas dentales privadas en Santo Domingo, La Vega y Santiago	541
Japón	Department of Forensic Sciences, Iwate Medical University, Iwate	1045
Corea	Catholic University of Korea, Seoul	500
Malasia	Faculty of Dentistry Universiti Teknologi MARA Selangor Branch, Selangor	667
Turquía	Department of Dentomaxillofacial Radiology, Baskent University, Turkey	2323
Uganda	Department of Dental Morphology with the Université Claude Bernard Lyon 1, Faculté d'odontologie, Lyon	283
Italia	Department of Surgical Sciences, University of Cagliari	173
Kosovo	University Dentistry Clinical Center, Pristina	1397
Líbano	Clínica dental privada en Beirut	690

Tabla 4.1: Lista de instituciones participantes en la recolección de los datos e imágenes dentales utilizados en el trabajo.

análisis ni en el entrenamiento de los modelos, centrándose exclusivamente en las variables de edad, sexo e imagen.

En el *dataset* hay un total de 10.739 ejemplos, de los que 5.756 son de individuos de sexo femenino y 4.983 de sexo masculino. Las edades mínima y máxima son 14 y 26 años, respectivamente, y la media son 19,13 años. En la Figura 4.1 se observa que el número de ejemplos por edad se mantiene relativamente constante desde los 14 hasta los 21 años, a partir de los cuales disminuye progresivamente, con una representación notablemente menor en los grupos de 24, 25 y 26 años.

En la Figura 4.2 podemos comprobar cómo en términos relativos la distribución de edad por sexo es muy similar, compartiendo ambas prácticamente el mismo rango de edades y patrones de dispersión, sin observarse diferencias sustanciales en la mediana ni en la forma general de las distribuciones.

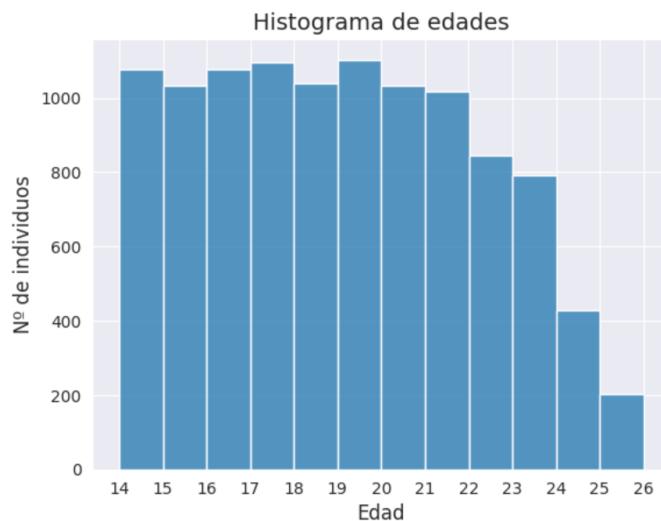


Figura 4.1: Distribución de edad de los individuos del conjunto de datos disponible. Elaboración propia.

En conclusión, el dataset presenta en general un buen balance entre clases y edades, lo que permite un análisis representativo de la población incluida. No obstante, será necesario examinar con mayor detalle la infrarepresentación de los grupos de mayor edad, especialmente a partir de los 22 años, para evaluar su posible impacto en el rendimiento y generalización de los modelos entrenados.

Se proporcionan los datos ya divididos en *train* —con un 80 % de los individuos— y *test* —con el 20 % restante—, con la intención de que puedan ser utilizados para entrenar y evaluar modelos de predicción. En la Figura 4.3 se puede observar cómo existe una distribución edad-sexo similar en los datos de ambos subconjuntos, por lo que se puede asumir que la partición respeta la representatividad de la población original, favoreciendo una evaluación más realista del rendimiento de los modelos en datos no vistos.

4.2. Métodos propuestos

4.2.1. Arquitectura empleada

El primer problema propuesto es el de estimación de edad. Partiremos de un planteamiento muy simple: imágenes bidimensionales de las radiografías panorámicas maxilofaciales como entrada, y estimación de edad a la salida. No incorporaremos el sexo del individuo en esta primera aproximación, si bien se explorará en el Anexo X.

¿Finalmente incluimos o no el sexo el sexo del individuo?

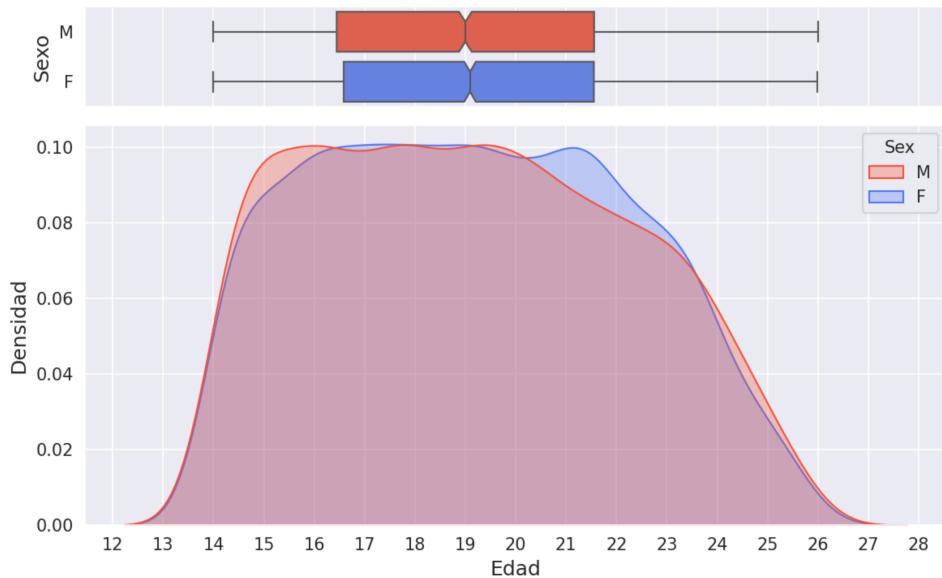
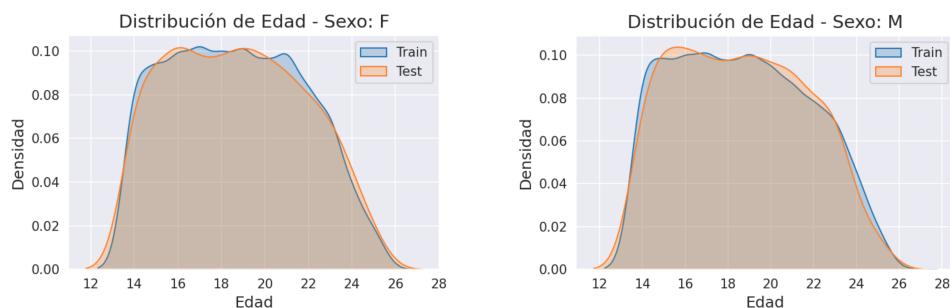


Figura 4.2: Distribución de edad por sexo de los individuos del conjunto de datos disponible. Elaboración propia.

Como modelo, empleamos una CNN, dado su buen desempeño en tareas de visión por computador. Específicamente, implementamos la arquitectura ResNeXt50 [100], utilizando un modelo entrenado con el *dataset* ImageNet ² [101] como punto de partida. Este modelo preentrenado es accesible a través de Pytorch.

Aunque ResNeXt50 fuera diseñado originalmente para un problema de clasificación de imágenes y entrenado con un dominio distinto al de nuestro problema, su adaptación a una tarea de estimación de edad es sencilla: reemplazar su cabecera de clasificación por una de regresión. Además, el uso de peso preentrenados proporciona una inicialización más robusta que el entrenamiento desde cero, ya que el modelo ya ha aprendido filtros genéricos para detectar características visuales básicas, como bordes o texturas.

²El dataset Imagenet contiene 1.000 clases de objetos. Estas clases abarcan una amplia variedad de categorías, como animales (*tiger, koala, zebra, ...*), vehículos (*ambulance, airliner, mountain bike, ...*), alimentos (*strawberry, pizza, bagel, ...*), entre otras.



(a) Distribución de edad de individuos de sexo femenino.

(b) Distribución de edad de individuos de sexo masculino.

Figura 4.3: Distribución de edad de los individuos del conjunto de datos disponible por sexo. Elaboración propia.

4.2.2. Entrenamiento

4.2.3. *Split conformal regression*

4.2.4. Regresión cuantílica

4.2.5. *Conformalized Quantile Regression*

4.3.

Capítulo 5

Experimentación

5.1. Protocolo de validación experimental

Como se ha comentado anteriormente, se han proporcionado los datos ya divididos en conjunto de entrenamiento (*train*) y de test, para evitar problemas asociados al *data snooping*. El ***data snooping*** ocurre cuando información del conjunto de test se filtra, directa o indirectamente, en el proceso de entrenamiento del modelo, lo que puede llevar a una sobreestimación del rendimiento y a modelos que generalizan pobremente en datos nuevos.

Debemos ser cuidadosos a la hora de tratar los datos en test, no debemos ... la variable de edad de los individuos, ya que es el target en nuestro problema de regresión, y cualquier ... puede ... Esto se conoce como data snooping

Para valorar los resultados obtenidos en los experimentos realizados se han dividido los datos de entrenamiento en *train* y *validation*.

Se consideró la validación cruzada (*cross-validation*), pero *data split*

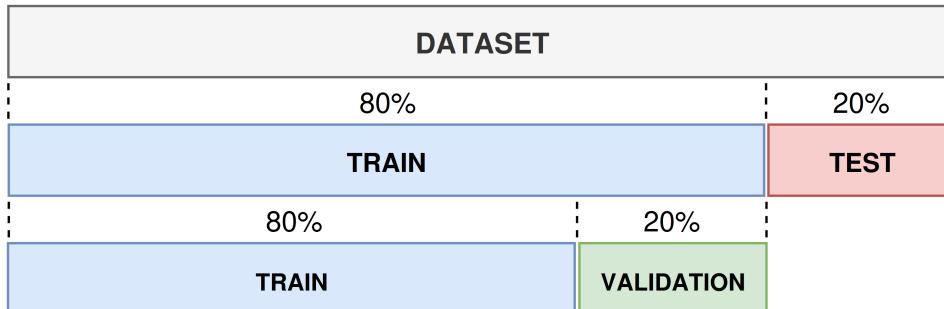


Figura 5.1: Diagrama de división del *dataset* en *train*, *validation* y *test*.
Elaboración propia.

Sin embargo, aquellos métodos de predicción conformal requieren de una fracción

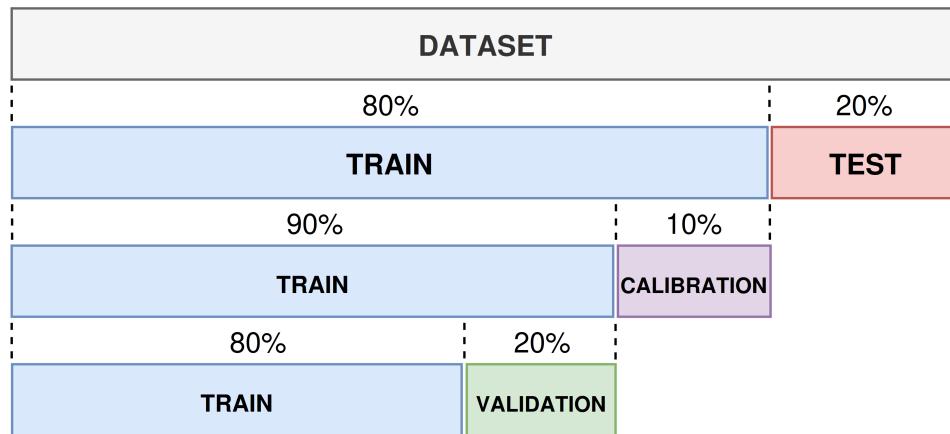


Figura 5.2: Diagrama de división del *dataset* en *train*, *validation*, *calibration* y *test*. Elaboración propia.

5.2. Métricas

5.3. Experimentos realizados

Bibliografía

- [1] American Anthropological Association. “What is Anthropology?” Consultado el 01/04/2025, American Anthropological Association. URL: <https://americananthro.org/learn-teach/what-is-anthropology/>. [Citado en pág. 1].
- [2] S. P. Nawrocki. “An Outline Of Forensic Anthropology.” Archivado del original (PDF) el 15 de junio de 2015. Consultado el 30 de abril de 2025. URL: <https://web.archive.org/web/20110615005707/>. [Citado en pág. 1].
- [3] S. N. Byers y C. A. Juarez, *Introduction to Forensic Anthropology*, 6.^a ed. Routledge, 2023. [Citado en págs. 1, 3, 4].
- [4] H. H. de Boer, S. Blau, T. Delabarre y L. H. and, “The role of forensic anthropology in disaster victim identification (DVI): recent developments and future prospects,” *Forensic Sciences Research*, vol. 4, n.^o 4, págs. 303-315, 2019. [Citado en pág. 2].
- [5] M. Prinz, A. Carracedo, W. Mayr, N. Morling, T. Parsons, A. Sajantila, R. Scheithauer, H. Schmitter y P. Schneider, “DNA Commission of the International Society for Forensic Genetics (ISFG): Recommendations regarding the role of forensic genetics for disaster victim identification (DVI),” *Forensic Science International: Genetics*, vol. 1, n.^o 1, págs. 3-12, 2007. [Citado en pág. 2].
- [6] J.-P. Beauthier, E. De Valck, P. Lefèvre y J. De Winne, “Mass Disaster Victim Identification: The Tsunami Experience,” *The Open Forensic Science Journal*, vol. 2, n.^o 1, págs. 54-62, 2009. [Citado en págs. 2, 3].
- [7] M. Skinner, D. Alempijevic y M. Djuric-Srejic, “Guidelines for International Forensic Bio-archaeology Monitors of Mass Grave Exhumations,” *Forensic Science International*, vol. 134, n.^o 2, págs. 81-92, 2003. [Citado en pág. 2].

- [8] J. A. Sanchis-Gimeno, J. Iglesias-Bexiga, M. E. Schwab, G. López-García, E. Ariza, A. Calpe, M. Mezquida, S. Nalla e I. Ercan, “Identification success rates in the post-Spanish Civil War mass graves located in the cemetery of Paterna, Spain: Meta-research on 15 mass graves with 933 subjects,” *Forensic Science International*, vol. 361, págs. 112-122, ago. de 2024. [Citado en pág. 2].
- [9] M. Baeta, C. Núñez, S. Cardoso, L. Palencia-Madrid, L. Herrasti, F. Etxeberria y M. M. de Pancorbo, “Digging up the recent Spanish memory: genetic identification of human remains from mass graves of the Spanish Civil War and posterior dictatorship,” *Forensic Science International: Genetics*, vol. 19, págs. 272-279, 2015. [Citado en pág. 2].
- [10] V. Ataliva, N. F. Bahamondes, C. M. Suárez y B. Rosignoli, “Arqueología Forense y prácticas genocidas del Cono Sur americano: reflexionando desde los confines,” *Revista de Arqueología Americana*, vol. 41, págs. 403-441, jun. de 2024. [Citado en pág. 2].
- [11] T. Tanaka, “International Humanitarian Law (IHL) and Forensic Document Examination,” *Journal of the American Society of Questioned Document Examiners*, vol. 23, n.º 1, 2020. [Citado en pág. 2].
- [12] T. Thompson y S. Black, *Forensic Human Identification: An Introduction*, 1.^a ed. Taylor & Francis, 2006. [Citado en pág. 2].
- [13] D. Higgins, A. B. Rohrlach, J. Kaidonis, G. Townsend y J. J. Austin, “Differential Nuclear and Mitochondrial DNA Preservation in Post-Mortem Teeth with Implications for Forensic and Ancient DNA Studies,” *PLoS One*, vol. 10, n.º 5, págs. 1-17, 2015. [Citado en pág. 3].
- [14] K. E. Latham y J. J. Miller, “DNA Recovery and Analysis from Skeletal Material in Modern Forensic Contexts,” *Forensic Sciences Research*, vol. 4, n.º 1, págs. 51-59, 2018. [Citado en pág. 3].
- [15] Scientific Working Group for Forensic Anthropology (SWGANTH). “Personal Identification.” Consultado el 25 de abril de 2025. URL: https://www.nist.gov/system/files/documents/2018/03/13/swganth_personal_identification.pdf. [Citado en pág. 3].
- [16] B. Marcante, L. Marino, N. E. Cattaneo, A. Delicati, P. Tozzo y L. Caenazzo, “Advancing Forensic Human Chronological Age Estimation: Biochemical, Genetic, and Epigenetic Approaches from the Last 15 Years: A Systematic Review,” *International Journal of Molecular Sciences*, vol. 26, n.º 7, 2025. [Citado en pág. 4].
- [17] A. Ross y S. Williams, “Ancestry Studies in Forensic Anthropology: Back on the Frontier of Racism,” *Biology*, vol. 10, n.º 7, pág. 602, 2021. [Citado en pág. 4].

- [18] A. Ross y M. Pilloud, “The need to incorporate human variation and evolutionary theory in forensic anthropology: A call for reform,” *American Journal of Physical Anthropology*, vol. 176, n.º 4, págs. 672-683, 2021. [Citado en pág. 4].
- [19] D. Flouri, A. Alifragki, J. Gómez García-Donas y E. Kranioti, “Ancestry Estimation: Advances and Limitations in Forensic Applications,” *Research and Reports in Forensic Medical Science*, vol. 12, págs. 13-24, 2022. [Citado en pág. 4].
- [20] P. Mesejo, R. Martos, Ó. Ibáñez, J. Novo y M. Ortega, “A Survey on Artificial Intelligence Techniques for Biomedical Image Analysis in Skeleton-Based Forensic Human Identification,” *Applied Sciences*, vol. 10, n.º 14, pág. 4703, 2020. [Citado en pág. 4].
- [21] A. Schmeling, R. B. Dettmeyer, E. Rudolf, V. Vieth y G. Gescrick, “Forensic Age Estimation,” *Deutsches Arzteblatt international*, vol. 113(4), págs. 44-50, 2016. [Citado en pág. 5].
- [22] S. Nakhaeizadeh, I. E. Dror y R. M. Morgan, “Cognitive bias in forensic anthropology: Visual assessment of skeletal remains is susceptible to confirmation bias,” *Science & Justice*, vol. 54, n.º 3, págs. 208-214, 2014. [Citado en pág. 5].
- [23] G. S. Cooper y V. Meterko, “Cognitive bias research in forensic science: A systematic review,” *Forensic Science International*, vol. 297, págs. 35-46, 2019. [Citado en pág. 5].
- [24] N. R. Langley, L. M. Jantz, S. McNulty, H. Maijanen, S. D. Ousley y R. L. Jantz, “Error quantification of osteometric data in forensic anthropology,” *Forensic Science International*, vol. 287, págs. 183-189, 2018. [Citado en pág. 5].
- [25] D. H. Ubelaker y C. M. DeGaglia, “Population variation in skeletal sexual dimorphism,” *Forensic Science International*, vol. 278, 407.e1-407.e7, 2017. [Citado en pág. 5].
- [26] F. Curate, C. Umbelino, A. Perinha, C. Nogueira, A. Silva y E. Cunha, “Sex determination from the femur in Portuguese populations with classical and machine-learning classifiers,” *Journal of Forensic and Legal Medicine*, vol. 52, págs. 75-81, 2017. [Citado en pág. 5].
- [27] M. F. Darmawan, S. M. Yusuf, M. A. Rozi y H. Haron, “Hybrid PSO-ANN for sex estimation based on length of left hand bone,” en *2015 IEEE Student Conference on Research and Development (SCORed)*, IEEE, 2015, págs. 478-483. [Citado en pág. 5].

- [28] S. C. D. Pinto, P. Urbanová y R. M. Cesar-Jr, “Two-Dimensional Wavelet Analysis of Supraorbital Margins of the Human Skull for Characterizing Sexual Dimorphism,” *IEEE Transactions on Information Forensics and Security*, vol. 11, n.º 7, págs. 1542-1548, 2016. [Citado en pág. 5].
- [29] J. R. Kim, W. H. Shim, H. M. Yoon, S. H. Hong, J. S. Lee, Y. A. Choy y S. Kim, “Computerized Bone Age Estimation Using Deep Learning Based Program: Evaluation of the Accuracy and Efficiency,” *American Journal of Roentgenology*, vol. 209, n.º 6, págs. 1374-1380, 2017. [Citado en pág. 5].
- [30] D. Larson, M. Chen, M. Lungren, S. Halabi, N. Stence y C. Langlotz, “Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs,” *Radiology*, vol. 287, págs. 313-322, 2018. [Citado en pág. 5].
- [31] H. Lee, S. Tajmir, M. Zissen, B. Yeshiwas, T. Alkasab, G. Choy y S. Do, “Fully Automated Deep Learning System for Bone Age Assessment,” *Journal of digital imaging*, vol. 30, págs. 427-441, 2017. [Citado en pág. 5].
- [32] D. Stern, T. Ebner, H. Bischof, S. Grassegger, T. Ehamer y M. Urschler, “Fully automatic bone age estimation from left hand MR images,” en *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2014: 17th International Conference, Boston, MA, USA*, Springer, vol. 17(Pt II), 2014, págs. 220-227. [Citado en págs. 5, 6].
- [33] S. Aja-Fernández, R. de Luis-García, M. Martín-Fernández y C. Alberola-López, “A computational TW3 classifier for skeletal maturity assessment. A Computing with Words approach,” *Journal of Biomedical Informatics*, vol. 37, n.º 2, págs. 99-107, 2004. [Citado en pág. 5].
- [34] D. Štern, C. Payer y M. Urschler, “Automated age estimation from MRI volumes of the hand,” *Medical Image Analysis*, vol. 58, págs. 101-138, 2019. [Citado en pág. 6].
- [35] J. Venema, D. Peula, J. Irurita y P. Mesejo, “Employing deep learning for sex estimation of adult individuals using 2D images of the humerus,” *Neural Comput & Applic*, vol. 35, págs. 5987-5998, 2022. [Citado en págs. 6, 36].
- [36] L. Ferrante y R. Cameriere, “Statistical methods to assess the reliability of measurements in procedures for forensic age estimation,” *International Journal of Legal Medicine*, vol. 123, n.º 4, págs. 277-283, 2009. [Citado en pág. 6].

- [37] R. Verma, K. Krishan, D. Rani, A. Kumar y V. Sharma, “Stature estimation in forensic examinations using regression analysis: A likelihood ratio perspective,” *Forensic Science International: Reports*, vol. 2, pág. 100 069, 2020. [Citado en págs. 6, 16].
- [38] M. Štepanovský, Z. Buk, A. Pilmann Kotěrová, J. Brůžek, Š. Bejdová, N. Techataweewan y J. Velemínská, “Application of machine-learning methods in age-at-death estimation from 3D surface scans of the adult acetabulum,” *Forensic science international*, vol. 365, pág. 112 272, 2024. [Citado en págs. 6, 16, 17].
- [39] A. Heinrich, “Accelerating computer vision-based human identification through the integration of deep learning-based age estimation from 2 to 89 years,” *Sci Rep*, vol. 14, pág. 4195, 2024. [Citado en págs. 6, 15, 17].
- [40] S. Park, S. Yang, J. Kim, J. Kang, J. Kim, K. Huh, S. Lee, W. Yi y M. Heo, “Automatic and robust estimation of sex and chronological age from panoramic radiographs using a multi-task deep learning network: a study on a South Korean population,” *Int J Legal Med*, vol. 138, págs. 1741-1757, 2024. [Citado en págs. 6].
- [41] K. Imaizumi, S. Usui, K. Taniguchi, Y. Ogawa, T. Nagata, K. Kaga, H. Hayakawa y S. Shiotani, “Development of an age estimation method for bones based on machine learning using post-mortem computed tomography images of bones,” *Forensic Imaging*, vol. 26, pág. 200 477, 2021. [Citado en págs. 6].
- [42] Ministerio del Interior de España, “Informe anual sobre personas desaparecidas 2025,” Ministerio del Interior, inf. téc., 2025. [Citado en págs. 6, 7].
- [43] F. Etxeberria, *Las exhumaciones de la Guerra Civil y la dictadura franquista 2000-2019: Estado actual y recomendaciones de futuro*. Madrid, España: Secretaría de Estado de Memoria Democrática, 2020, ISBN: 978-84-7471-146-2. URL: https://www.mpr.gob.es/servicios/publicaciones/Documents/Exhumaciones_Guerra_Civil_accesible_BAJA.pdf. [Citado en págs. 7].
- [44] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2024,” Fiscalía General del Estado, Madrid, España, inf. téc., 2024. [Citado en págs. 7, 8].
- [45] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2019,” Fiscalía General del Estado, Madrid, España, inf. téc., 2019. [Citado en págs. 7, 8].
- [46] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2016,” Fiscalía General del Estado, Madrid, España, inf. téc., 2016. [Citado en págs. 7, 8].

- [47] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2013,” Fiscalía General del Estado, Madrid, España, inf. téc., 2013. [Citado en págs. 7, 8].
- [48] S. Cordner y M. Tidball-Binz, “Humanitarian forensic action — Its origins and future,” *Forensic Science International*, vol. 279, págs. 65-71, 2017. [Citado en pág. 7].
- [49] M. V. Tidball-Binz y S. M. Cordner, “Humanitarian forensic action: A new forensic discipline helping to implement international law and construct peace,” *WIREs Forensic Science*, 2021. [Citado en pág. 7].
- [50] A. Turing, “I.—COMPUTING MACHINERY and INTELLIGENCE,” *Mind*, vol. LIX, n.º 236, págs. 433-460, 1950. [Citado en pág. 11].
- [51] A. L. Samuel, “Some Studies in Machine Learning Using the Game of Checkers,” *IBM Journal of Research and Development*, vol. 3, n.º 3, págs. 210-229, 1959. [Citado en pág. 11].
- [52] W. S. McCulloch y W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics*, vol. 5, n.º 4, págs. 115-133, 1943. [Citado en págs. 11, 23, 24].
- [53] F. Rosenblatt, “The perceptron: a probabilistic model for information storage and organization in the brain,” *Psychological Review*, vol. 65(6), págs. 386-408, 1958. [Citado en págs. 11, 23, 24].
- [54] J. R. Quinlan, “Induction of decision trees,” *Machine learning*, vol. 1, págs. 81-106, 1986. [Citado en pág. 11].
- [55] P. Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. USA: Penguin Books Limited, 2015. [Citado en pág. 12].
- [56] S. Russell y P. Norvig, *Artificial Intelligence: A Modern Approach*, 4rd. Prentice Hall Press, 2021. [Citado en págs. 12, 21, 23, 28, 35].
- [57] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Heidelberg: Springer-Verlag, 2006. [Citado en págs. 12, 18].
- [58] E. Alpaydin, *Introduction to Machine Learning*, 2nd. The MIT Press, 2010. [Citado en págs. 12, 42].
- [59] M. A. Bidmos, O. I. Olateju, S. Latiff, T. Rahman y M. E. Chowdhury, “Machine learning and discriminant function analysis in the formulation of generic models for sex prediction using patella measurements,” *International Journal of Legal Medicine*, vol. 137, n.º 2, págs. 471-485, 2023. [Citado en pág. 19].

- [60] L. Porto, L. Lima, A. Franco, D. Pianto, C. Machado y F. Vidal, “Estimating sex and age from a face: a forensic approach using machine learning based on photo-anthropometric indexes of the Brazilian population,” *International journal of legal medicine*, vol. 134(6), págs. 2239-2259, 2020. [Citado en pág. 20].
- [61] J. G. Sam Lau y D. Nolan, *Cross Validation*, Consultado el 26/05/2025, 2023. URL: https://learningds.org/ch/16/ms_cv.html. [Citado en pág. 23].
- [62] Y. LeCun, Y. Bengio y G. Hinton, “Deep Learning,” *Nature*, vol. 521, págs. 436-44, 2015. [Citado en págs. 22, 23].
- [63] D. E. Rumelhart, G. E. Hinton y R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, págs. 533-536, 1986. [Citado en pág. 23].
- [64] P. J. Werbos, *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*. USA: Wiley-Interscience, 1994. [Citado en pág. 23].
- [65] Red Hat, *Deep learning*, Consultado el 10/05/2025, 2023. URL: <https://www.redhat.com/es/topics/ai/what-is-deep-learning>. [Citado en pág. 23].
- [66] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. [Citado en págs. 24, 25, 29, 31, 32, 36, 38, 39].
- [67] Code World, *Understanding ML & DL in python*, Consultado el 19/05/2025, 2022. URL: <https://codeworld.tistory.com/2>. [Citado en pág. 24].
- [68] V. M. Vargas, D. Guijo-Rubio, P. A. Gutiérrez y C. Hervás-Martínez, “ReLU-Based Activations: Analysis and Experimental Study for Deep Learning,” en *Advances in Artificial Intelligence*, E. Alba, G. Luque, F. Chicano, C. Cotta, D. Camacho, M. Ojeda-Aciego, S. Montes, A. Troncoso, J. Riquelme y R. Gil-Merino, eds., Cham: Springer International Publishing, 2021, págs. 33-43. [Citado en pág. 25].
- [69] G. Furnieles, *Sigmoid and SoftMax Functions in 5 minutes*, Consultado el 26/05/2025, 2022. URL: <https://towardsdatascience.com/sigmoid-and-softmax-functions-in-5-minutes-f516c80ea1f9/>. [Citado en pág. 26].
- [70] F. Bre, J. Gimenez y V. Fachinotti, “Prediction of wind pressure coefficients on building surfaces using Artificial Neural Networks,” *Energy and Buildings*, vol. 158, 2017. [Citado en pág. 27].
- [71] R. Szeliski, *Computer Vision: Algorithms and Applications*, 1st. Berlin, Heidelberg: Springer-Verlag, 2010. [Citado en págs. 26, 29, 34, 35].

- [72] L. Bottou, “Large-Scale Machine Learning with Stochastic Gradient Descent,” *Proc. of COMPSTAT’2010*, págs. 177-186, 2010. [Citado en pág. 28].
- [73] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy y P. T. P. Tang, *On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima*, 2017. URL: <https://arxiv.org/abs/1609.04836>. [Citado en pág. 28].
- [74] I. Goodfellow, Y. Bengio y A. Courville, *Deep Learning*. MIT Press, 2016. [Citado en pág. 29].
- [75] NVIDIA, *Convolutional Neural Network*, Consultado el 21/05/2025, 2025. URL: <https://www.nvidia.com/en-eu/glossary/convolutional-neural-network/>. [Citado en pág. 30].
- [76] S. Chen, E. Dobriban y J. Lee, “Invariance reduces Variance: Understanding Data Augmentation in Deep Learning and Beyond,” *ArXiv*, 2019. URL: <https://api.semanticscholar.org/CorpusID:198895147>. [Citado en pág. 33].
- [77] A. Zhang, Z. C. Lipton, M. Li y A. J. Smola, *Dive into Deep Learning*, 2021. [Citado en pág. 33].
- [78] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever y R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research*, vol. 15, n.º 56, págs. 1929-1958, 2014. [Citado en pág. 34].
- [79] J. Tompson, R. Goroshin, A. Jain, Y. LeCun y C. Bregler, *Efficient Object Localization Using Convolutional Networks*, 2015. URL: <https://arxiv.org/abs/1411.4280>. [Citado en pág. 34].
- [80] S. Ioffe y C. Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, 2015. URL: <https://arxiv.org/abs/1502.03167>. [Citado en pág. 34].
- [81] S. Santurkar, D. Tsipras, A. Ilyas y A. Madry, *How Does Batch Normalization Help Optimization?* 2019. URL: <https://arxiv.org/abs/1805.11604>. [Citado en pág. 34].
- [82] S. Arora, Z. Li y K. Lyu, *Theoretical Analysis of Auto Rate-Tuning by Batch Normalization*, 2018. URL: <https://arxiv.org/abs/1812.03981>. [Citado en pág. 34].
- [83] Joint Committee for Guides in Metrology (JCGM), *International vocabulary of metrology — Basic and general concepts and associated terms (VIM)*, VIM 2008 version with minor corrections, JCGM 200:2012, Consultado el 30/05/2025, JCGM, Sèvres, France, 2012. URL: https://www.bipm.org/documents/20126/2071204/JCGM_200_2012.pdf. [Citado en pág. 37].

- [84] Joint Committee for Guides in Metrology (JCGM), *Evaluation of measurement data — Guide to the expression of Uncertainty in Measurement (GUM), GUM 1995 with minor corrections*, JCGM 100:2008, Consultado el 30/05/2025, JCGM, Sèvres, France, 2008. URL: https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf. [Citado en págs. 37, 38].
- [85] J. R. Berrendero. “Materiales del libro de Estadística,” visitado 2 de jun. de 2025. URL: <https://verso.mat.uam.es/~joser.berrendo/libro-est/>. [Citado en pág. 38].
- [86] Y. Romano, E. Patterson y E. Candes, “Conformalized quantile regression,” *Advances in neural information processing systems*, vol. 32, 2019. [Citado en pág. 39].
- [87] R. Luo y Z. Zhou, “Conformal thresholded intervals for efficient regression,” en *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, págs. 19 216-19 223. [Citado en pág. 39].
- [88] M. Sadinle, J. Lei y L. Wasserman, “Least ambiguous set-valued classifiers with bounded error levels,” *Journal of the American Statistical Association*, vol. 114, n.º 525, págs. 223-234, 2019. [Citado en pág. 39].
- [89] Y. Romano, M. Sesia y E. Candes, “Classification with valid and adaptive coverage,” *Advances in neural information processing systems*, vol. 33, págs. 3581-3591, 2020. [Citado en págs. 39, 43].
- [90] A. Angelopoulos, S. Bates, J. Malik y M. I. Jordan, “Uncertainty sets for image classifiers using conformal prediction,” *arXiv preprint arXiv:2009.14193*, 2020. [Citado en págs. 39, 43].
- [91] E. Hüllermeier y W. Waegeman, “Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods,” *Machine Learning*, vol. 110, págs. 457-506, 2021. [Citado en págs. 40, 43].
- [92] J. Gama, “A survey on learning from data streams: current and future trends,” *Progress in Artificial Intelligence*, vol. 1, págs. 45-55, 2012. [Citado en pág. 40].
- [93] I. Steinwart y A. Christmann, “Estimating conditional quantiles with the help of the pinball loss,” *Bernoulli*, vol. 17, n.º 1, págs. 221-225, 2011. [Citado en pág. 41].
- [94] J. Vermorel. “Quantile Regression,” LOKAD Quantitive Supply Chain, visitado 2 de jun. de 2025. URL: <https://www.lokad.com/quantile-regression-time-series-definition/>. [Citado en pág. 41].
- [95] R. Koenker, *Quantile Regression* (Econometric Society Monographs). Cambridge University Press, 2005. [Citado en pág. 42].

- [96] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya et al., “A review of uncertainty quantification in deep learning: Techniques, applications and challenges,” *Information fusion*, vol. 76, págs. 243-297, 2021. [Citado en pág. 42].
- [97] S. T. Tokdar y J. B. Kadane, “Simultaneous linear quantile regression: a semiparametric Bayesian approach,” *Bayesian Analysis*, vol. 7, n.º 1, págs. 51-72, 2012. [Citado en pág. 42].
- [98] J. Feldman y D. Kowal, “Bayesian Quantile Regression with Subset Selection: A Posterior Summarization Perspective,” *arXiv preprint arXiv:2311.02043*, 2023. [Citado en pág. 42].
- [99] M. Sato, J. Suzuki, H. Shindo e Y. Matsumoto, “Interpretable Adversarial Perturbation in Input Embedding Space for Text,” en *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018)*, Stockholm, Sweden: International Joint Conferences on Artificial Intelligence, 2018, págs. 4323-4330. [Citado en pág. 43].
- [100] S. Xie, R. Girshick, P. Dollár, Z. Tu y K. He, “Aggregated residual transformations for deep neural networks,” en *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, págs. 1492-1500. [Citado en pág. 50].
- [101] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li y L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” en *2009 IEEE conference on computer vision and pattern recognition*, IEEE, 2009, págs. 248-255. [Citado en pág. 50].
- [102] C. Guo, G. Pleiss, Y. Sun y K. Q. Weinberger, “On calibration of modern neural networks,” en *International conference on machine learning*, PMLR, 2017, págs. 1321-1330.

