



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. Identificación humana y estimación del perfil biológico	2
1.2. Motivación	5
1.3. Objetivos	7
1.4. Planificación temporal del proyecto	9
1.5. Planificación económica del proyecto	9
2. Fundamentos teóricos	11
2.1. Machine Learning	11
2.1.1. Problemas de regresión	13
2.1.2. Problemas de clasificación	13
2.2. Deep Learning	14
2.2.1. El perceptrón multicapa	15
2.2.2. Entrenamiento y validación de la red	17
2.2.3. Redes Neuronales Convolucionales	21
2.2.4. Transfer Learning	27
2.3. Incertidumbre	28
2.3.1. Incertidumbre en <i>machine learning</i>	29
2.3.2. Cuantificación de la incertidumbre en <i>machine learning</i>	30
2.4. Predicción conformal	31
2.4.1. Propiedades de la predicción conformal	32
2.4.2. Algoritmo conformal	34

3. Estado del arte	39
3.1. Estimación de la edad en antropología forense	39
3.2. Estimación de la edad en antropología forense usando <i>machine learning</i>	42
3.3. Cuantificación de la incertidumbre para la estimación de la edad	43
3.4. Estimación del sexo en antropología forense	45
4. Materiales y métodos	49
4.1. Conjunto de datos disponibles	49
4.2. Problemas planteados	52
4.2.1. Problema de estimación de edad	52
4.2.2. Estimación de mayoría de edad	52
4.2.3. Clasificación combinada de mayoría de edad y sexo . .	53
4.3. Métodos propuestos	54
4.3.1. Arquitectura empleada	54
4.3.2. Regresión cuántica	54
4.3.3. Métodos de predicción conformal para regresión . . .	56
4.3.4. Calibración de probabilidades en clasificación	59
4.3.5. Métodos de predicción conformal para clasificación .	59
5. Experimentación	65
5.1. Protocolo de validación experimental	65
5.2. Experimentos propuestos	67
5.2.1. Comparativa de métodos para la estimación de edad .	67
5.2.2. Comparativa de métodos para la estimación de mayoría de edad	68
5.2.3. Comparativas de métodos para la clasificación combinada de mayoría de edad y sexo	68
5.3. Entrenamiento de los modelos	69
5.3.1. Preparación de los datos de entrenamiento	69
5.3.2. Adaptación de la red para la estimación de edad . .	70
5.3.3. Adaptación de la red para la estimación de mayoría de edad	72

5.3.4. Adaptación de la red para la clasificación combinada de mayoría de edad y sexo	72
5.4. Métricas usadas en los experimentos	73
5.4.1. Métricas para regresión	73
5.4.2. Métricas para clasificación	76
5.5. Resultados	77
5.5.1. Resultados para la estimación de edad	77
5.5.2. Resultados para la estimación de mayoría de edad . .	88
5.5.3. Resultados para la clasificación combinada de sexo y mayoría de edad	90
6. Conclusiones y trabajos futuros	93
6.1. Conclusiones	93
6.1.1. Conclusiones sobre mejor método	93
6.2. Trabajos futuros	94

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.	Línea de regresión del modelo de regresión propuesto en [33] que predice la estatura a partir de la longitud de la tibia.	6
1.3.	Evolución de hallazgos/identificación de cadáveres en España (2010-2024) [39].	7
1.4.	Evolución del número de diligencias preprocesales de determinación de edad abiertas en España (2011–2023). Elaboración propia a partir de [41-44].	8
2.1.	Esquema visual del funcionamiento de una unidad artificial. Adaptado de [61].	16
2.2.	Diagrama de obtención de probabilidad en problemas de clasificación. Adaptado de [63].	17
2.3.	Arquitectura simplificada de un MLP. Recuperado de [64]. .	18
2.4.	Esquema gráfico de la aplicación de un filtro convolucional sobre una región de una imagen.	22
2.5.	Esquema gráfico de <i>max pooling</i> con un filtro 2x2 y <i>stride</i> de 1. Recuperado de la Figura 14.12 de [60].	24
2.6.	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 [65].	25
2.7.	Diagrama del funcionamiento de neuronas con <i>dropout</i> . Recuperado de la Figura 5.29 de [65].	26
2.8.	Diagrama de <i>fine-tuning</i> de un modelo en una nueva tarea. Recuperado de la Figura 19.2 de [60].	28

2.9. Ejemplo de predicción conformal en problemas de regresión (arriba) y clasificación (abajo). Recuperado de [92].	32
2.10. Ejemplo adversario mal clasificado por un modelo de ML entrenado con datos textuales. Adaptado de la Figura 2 de [79], original de [93].	33
2.11. Conjuntos de predicción bajo distintas nociones de cobertura: sin cobertura garantizada, con cobertura marginal y con cobertura condicional. Recuperado de [91].	37
2.12. Determinación del umbral de no conformidad para intervalos simétricos y asimétricos.	38
3.1. Hallazgos radiológicos en un posible menor con edad disputada: criterio de edad mínima para la determinación de edad. Recuperado de la Figura 1 de [21].	41
3.2. Procedimiento secuencial clásico de ML para el método propuesto en [127].	42
3.3. Metodología de construcción de un modelo <i>end-to-end</i> . Recuperado de [38].	43
3.4. Cronograma de desarrollo de la unión epifisaria. Recuperado de [3], original de [130].	46
3.5. Distribución del error por edad real para el modelo propuesto en [29].	47
3.6. Estudio del error en los métodos 2D propuestos en [128]. .	47
4.1. Histograma de edad de los individuos del conjunto de datos disponible.	51
4.2. Gráficas de densidad y de caja de edad por sexo de los individuos del conjunto de datos disponible.	51
4.3. Esquema visual del modelos de regresión propuesto.	53
4.4. Visualización de la función de pérdida <i>pinball</i> para cada valor de error.	56
5.1. Diagrama de división del <i>dataset</i> en <i>train</i> , <i>validation</i> y <i>test</i> . .	66
5.2. Diagrama de división del <i>dataset</i> en <i>train</i> , <i>validation</i> , <i>calibration</i> y <i>test</i>	66
5.3. Matriz de confusión para la estimación de sexo según el modelo <i>random forest</i> propuesto en [145].	77

5.4. Gráfica de dispersión <i>Empirical Coverage-Mean Precition Interval Width</i>	80
5.5. Histogramas del amplitud del intervalo de predicción con diferenciación por cobertura, correspondientes a los modelos QR y CQR.	84
5.6. Análisis comparativo de la cobertura empírica y el ancho medio del intervalo de predicción por edad cronológica para los diferentes métodos evaluados.	87
5.7. Gráfica de dispersión <i>Empirical Coverage-Mean Prediction Set Size</i>	89
5.8. Matrices de confusión conformal correspondientes a tres modelo de 'base', LAC y MCM.	90

Índice de tablas

4.1. Instituciones participantes en la recolección de datos e imágenes	50
4.2. Comparativa de métodos propuestos de CP para problemas de regresión.	60
5.1. Error absoluto medio y error cuadrático medio obtenidos por cada método de predicción a lo largo de distintas ejecuciones.	78
5.2. Resultados de la prueba <i>post-hoc</i> de Tukey HSD para MAE entre pares de modelos.	79
5.3. Resultados de la prueba <i>post-hoc</i> de Tukey HSD para MSE entre pares de modelos.	79
5.4. Cobertura empírica y amplitud media del intervalo de predicción obtenidos por cada método de predicción a lo largo de distintas ejecuciones.	80
5.5. Resultados de las predicciones obtenidas por los modelos para el problema de estimación de edad en cada ejecución.	81
5.6. Cobertura empírica del intervalo de predicción obtenida por cada método de predicción para distintas franjas de amplitud de intervalos.	83
5.7. Cobertura empírica y amplitud media del intervalo de predicción obtenidos por cada método de predicción para distintas edades cronológicas.	86
5.8. Exactitud (<i>accuracy</i>) obtenida por cada método de predicción a lo largo de las distintas ejecuciones.	88
5.9. Cobertura empírica y tamaño medio del conjunto de predicción obtenidos por cada método de predicción a lo largo de las distintas ejecuciones.	89

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. Aunque tradicionalmente asociada al estudio de restos humanos esqueletizados o en descomposición, la AF también contribuye a la estimación del perfil biológico en individuos vivos, especialmente en contextos legales.

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), ya sea en restos esqueletizados donde los tejidos blandos se han deteriorado hasta el punto de que estas características no pueden determinarse mediante inspección visual, o en personas vivas mediante técnicas no invasivas como análisis radiográficos o morfológicos.
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. Identificación humana y estimación del perfil biológico

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.

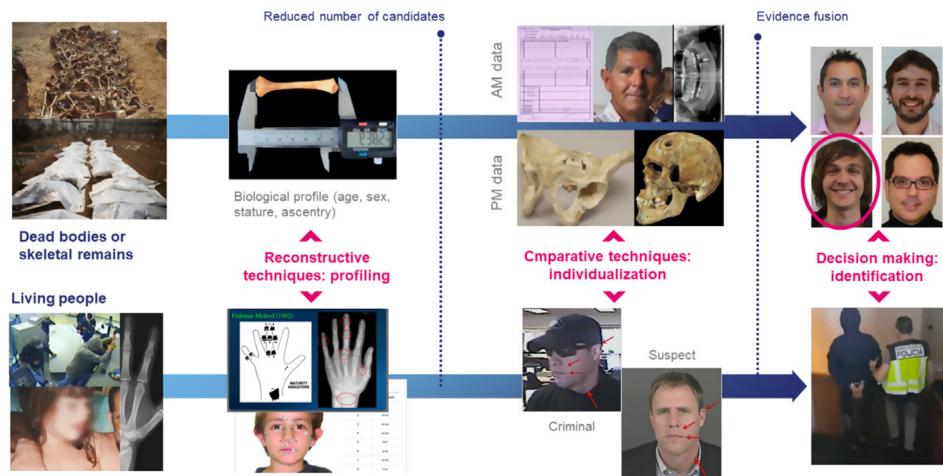


Figura 1.1: Procedimiento secuencial para la identificación forense basada en el esqueleto humano (*skeleton-based forensic identification*) [20].

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 o fisiológica 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 interinterpretaciones 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].

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 [32] 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 frequentista² [33-35]. Un ejemplo de este tipo de análisis se ilustra en la Figura 1.2, donde se examina la distribución probabilística del error residual arrojado por el modelo de regresión propuesto en [33].

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 [34, 36, 37] o etiquetas únicas en clasificación [36, 38], sin cuantificar la incertidumbre asociada a cada predicción.

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

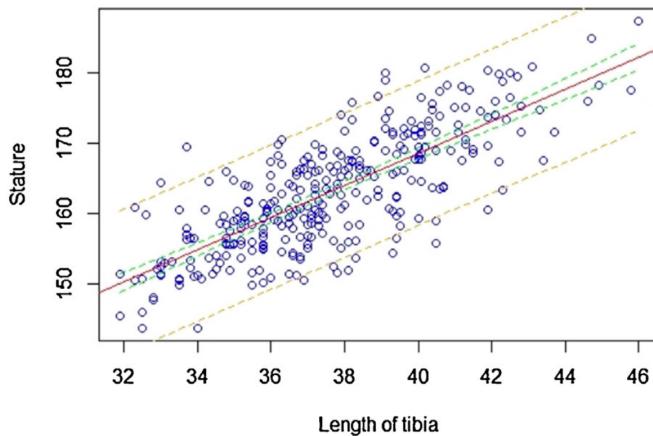


Figura 1.2: Línea de regresión del modelo de regresión propuesto en [33] que predice la estatura a partir de la longitud de la tibia. En rojo, la línea de regresión; en verde, la línea de los intervalos de confianza del 95 %; y en naranja, la línea de los intervalos de predicción al 95 % de confianza.

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 [39]. 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.
- 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 [40].
- En España, se ha registrado en la última década (2013-2023) un aumento significativo en la llegada de Menores Extranjeros No Acompañados [41-44], 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

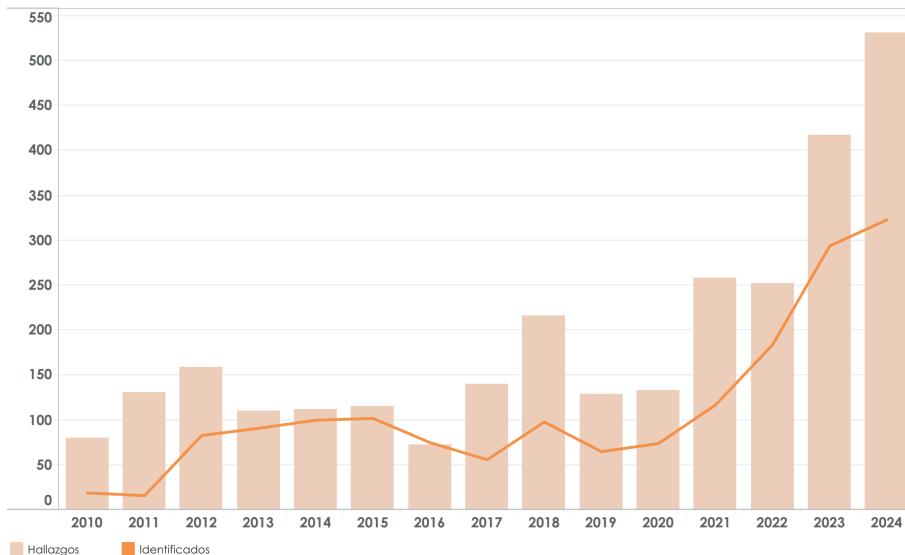


Figura 1.3: Evolución de hallazgos/identificación de cadáveres en España (2010-2024) [39].

pilar fundamental de los derechos humanos y la justicia internacional, naciendo así la **acción forense humanitaria** [45]. 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 [46].

1.3. Objetivos

La **predicción conformal** 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 o conjuntos de predicción que reflejan la confiabilidad de la predicción en cada caso particular.

Este Trabajo de Fin de Grado tiene un doble objetivo:

- desde un prisma teórico, defender la cuantificación de incertidumbre como herramienta esencial en ML, ofrecer un panorama de métodos destacados, analizando sus ventajas y limitaciones, y centrarnos en la predicción conformal y sus variantes más populares.

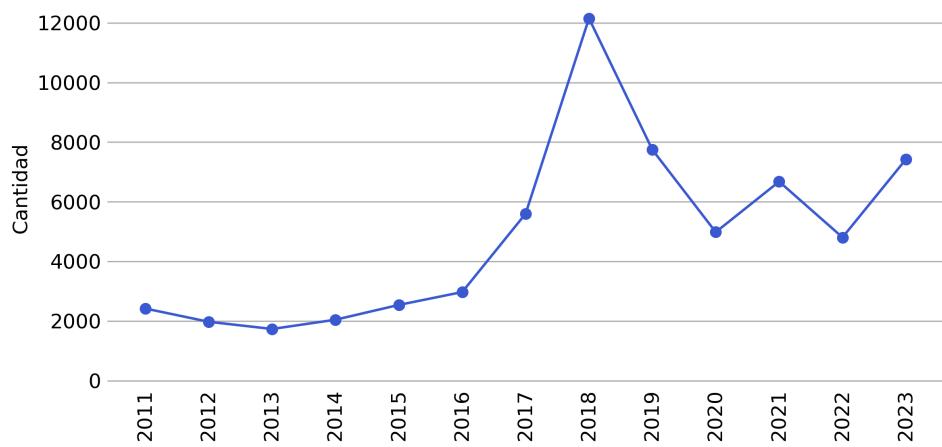


Figura 1.4: Evolución del número de diligencias preprocesales de determinación de edad abiertas en España (2011–2023). Elaboración propia a partir de [41–44].

- aplicar la predicción conformal a un contexto práctico como es el problema de estimación del PB, centrándose en la estimación de edad y de sexo a partir de datos biológicos e imágenes médicas.

De esta forma, podremos incorporar la incertidumbre propia del problema a resolver y del modelo entrenado para él, para, en aquellos casos más confusos, 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 [16,20]) en problemas de regresión, en ambos casos para un nivel de confianza determinado.

Pablo: Esto no creo que quede muy claro. Generalmente, se proponen uno o dos objetivos principales, y luego se presentan una serie de objetivos parciales, cuya consecución asegura el cumplimiento de los objetivos principales. Pero aquí no me queda claro si estos son los objetivos parciales... entiendo que sí. De ser así, no dudes en indicarlo con claridad, diciendo sin ambages que estos son objetivos parciales.

Por tanto, podemos desgranar los objetivos en:

- Estudiar de forma exhaustiva la bibliografía sobre predicción conformal y sus diversas variantes, así como de la estimación de sexo y edad, centrando 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 contener el valor real en un conjunto o intervalo de predicción, útiles para la toma de decisiones fundamentadas en contextos prácticos donde la interpretabilidad y robustez son críticas.

1.4. Planificación temporal del proyecto

1.5. Planificación económica del proyecto

Este trabajo ha sido realizado con dos equipos independientes:

- Un ordenador portátil personal, empleado principalmente para la redacción y compilación de este documento en L^AT_EX.
- Un clúster de computación proporcionado por el Instituto de Ciencia de Datos e Inteligencia Artificial (DaSCI), de la Universidad de Granada, al que se accedió mediante conexión SSH. Para el desarrollo de los distintos modelos y variantes de predicción conformal, se utilizó el entorno de desarrollo Visual Studio Code de Microsoft.

En cuanto al software, todo el empleado es gratuito y de código abierto.

Pablo: Algo que yo creo que podría ser de utilidad al lector es mostrar un esquema genérico visual de lo que piensas hacer a nivel práctico, por ejemplo, a la hora de estimar la edad empleando una red neuronal. Yo incluiría un esquema de aprendizaje automático, como la Fig. 3 de <https://link.springer.com/1022-07981-0>, pero añadiendo elementos visuales que indiquen y subrayen la idea de que se emplean intervalos de predicción (con un conjunto de calibración, por ejemplo) para cuantificar la incertidumbre en las predicciones proporcionadas. Dicho de otro modo: queremos una figura que, de un vistazo, permita entender cómo vas a combinar estimación de la edad (por ejemplo) e intervalos de predicción (y estos últimos de dónde salen). Aunque luego los detalles se presenten más adelante, un diagrama en la introducción creo que ayudaría a aterrizar las ideas principales.

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 [47]: 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 [48], 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 [49, 50] o los árboles de decisión [51], 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 [52].

El ML diferencia tres tipos de aprendizaje en base a tres tipos de retroalimentación [53]:

- **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 [54].
- **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 [55].
- **Aprendizaje no supervisado**, en el que el agente tampoco dispone de valores de salida, solo de entrada [54], 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.

El objetivo en el aprendizaje supervisado es establecer una hipótesis que se ajuste de forma óptima a los ejemplos futuros. Para ello, se presupone 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 [53].

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 [54], 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 residuo¹, 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.

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 variante 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.

¹A pesar de que en la literatura más especializada —que veremos a continuación—, los términos “error” y “residuo” se distinguen.

- 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 [54], que será el tipo de clasificación en el que nos centraremos. 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.

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 [56]. 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.

Las redes neuronales tienen su origen en el intento de modelar las redes de neuronas del cerebro humano [49]. Se requirió de numerosas contribuciones teóricas —como el perceptrón [50] o el algoritmo de *backpropagation* [57, 58], 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 [56], 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)² [53]. 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) [59]. 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.

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* [60]. 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 [49, 50], 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

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 [53].

producir una salida que propaga a las unidades de la siguiente capa (véase la Figura 2.1).

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.

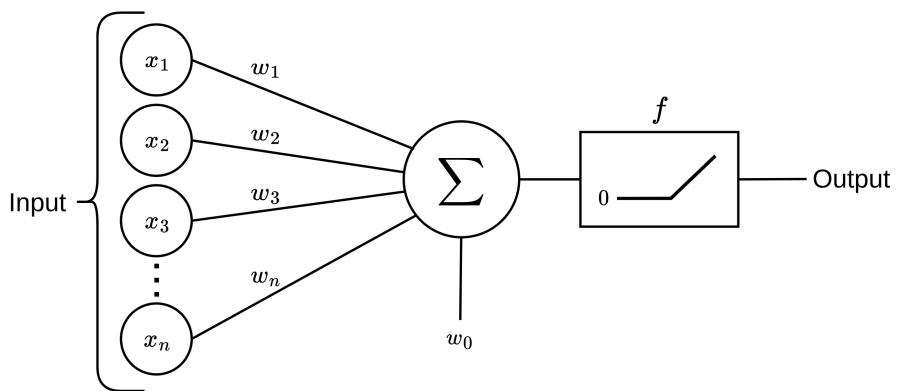


Figura 2.1: Esquema visual del funcionamiento de una unidad artificial. Adaptado de [61].

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 2.3):

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).

⁴Sin ella, el MLP se reduciría a una simple combinación lineal de las entradas, incapaz de representar jerarquías de características [60].

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 [62].

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.

- **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 multiclas (véase la Figura 2.2).

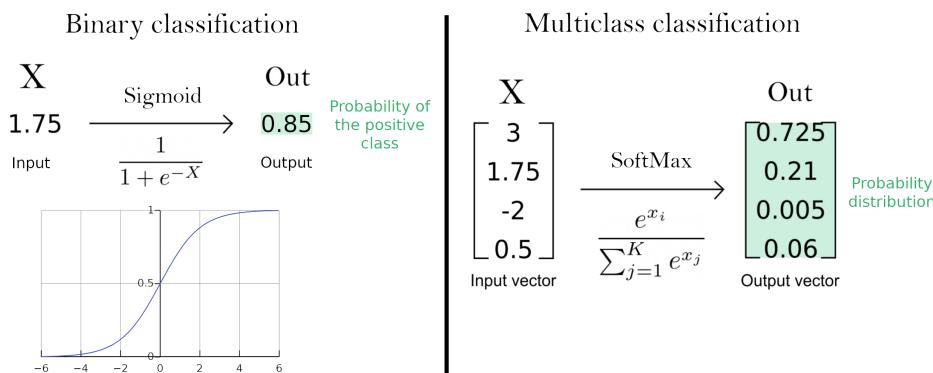


Figura 2.2: Diagrama de obtención de probabilidad en problemas de clasificación. Adaptado de [63].

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

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.

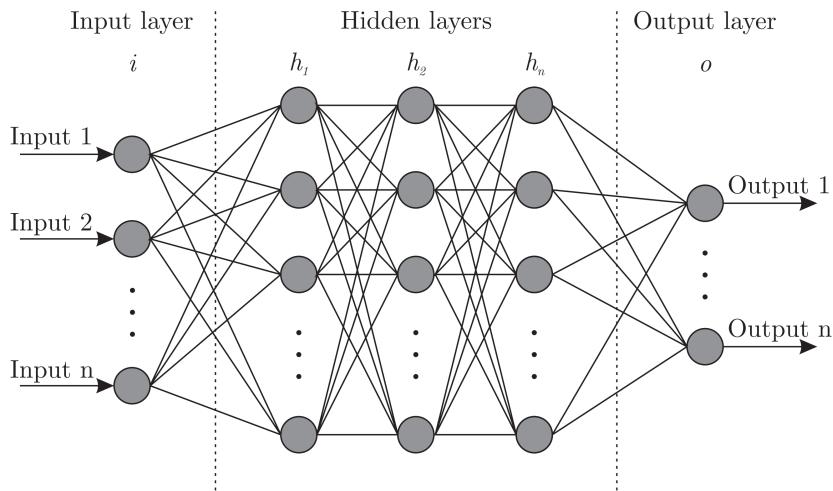


Figura 2.3: Arquitectura simplificada de un MLP. Recuperado de [64].

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 [65]:

- **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 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 métrica 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, que mide la diferencia promedio al cuadrado entre las predicciones y los valores reales, o el error absoluto medio, que calcula la diferencia promedio en valor absoluto⁸.

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 el *Stochastic Gradient Descent*, 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

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

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 *Stochastic Gradient Descent* puro [66]—, 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 [67], 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 [53]. 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 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*** [68].

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

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 [60].

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 ineficiencia por el elevado número de parámetros requeridos [65].

Precisamente para estos casos, otras arquitecturas profundas resultan más apropiadas. Las **redes neuronales convolucionales (Convolutional Neural Network, 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 2.4).

Este proceso se repite al desplazar el filtro por toda la imagen mediante una **ventana deslizante**, generando un **mapa de activación**, que permite

¹⁰Aunque, como veremos a continuación, filtro y *kernel* a la hora de hablar de capas convolucionales, no son técnicamente lo mismo.

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

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

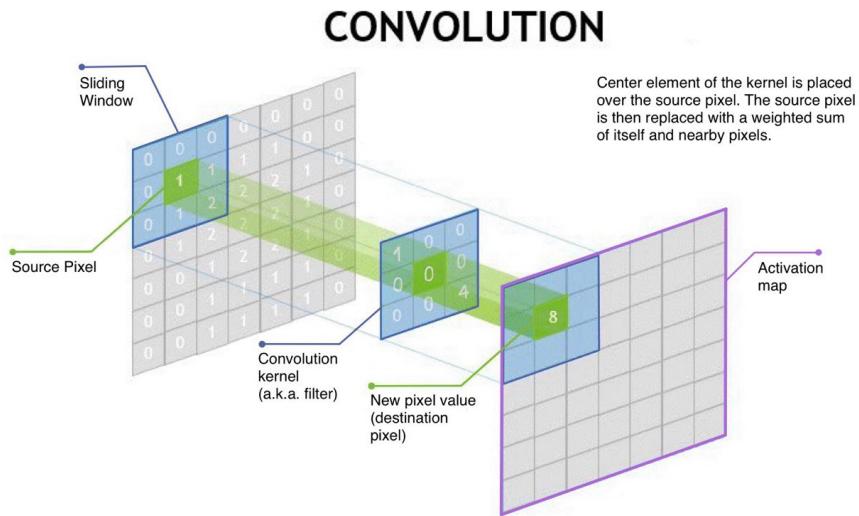


Figura 2.4: Esquema gráfico de la aplicación de un filtro convolucional sobre una región de una imagen. Adaptado de [69].

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 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 entrada¹² 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 píxeles¹³), 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

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.

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.

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 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*) [60], 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. 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:

Me he estado informando después y resulta que ResNeXt no utiliza pooling, sino stride aumentado, por lo que debería eliminar este apartado y explicar cómo funciona el stride.
(AGOSTO)

- **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 2.5).
- **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

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

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*.

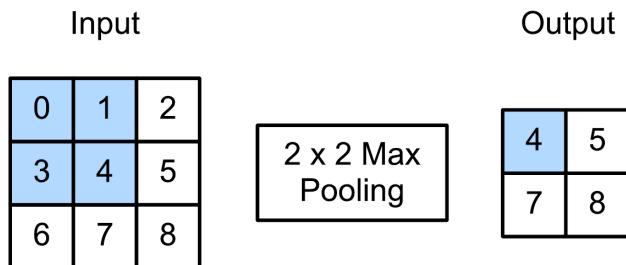


Figura 2.5: Esquema gráfico de *max pooling* con un filtro 2x2 y *stride* de 1. Recuperado de la Figura 14.12 de [60].

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/regresor¹⁵ 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.

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

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 extractor 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 FC, toma estas representaciones para realizar la tarea específica de salida, ya sea clasificación o regresión.

En la Figura 2.6 se puede observar un ejemplo de arquitectura CNN completa.

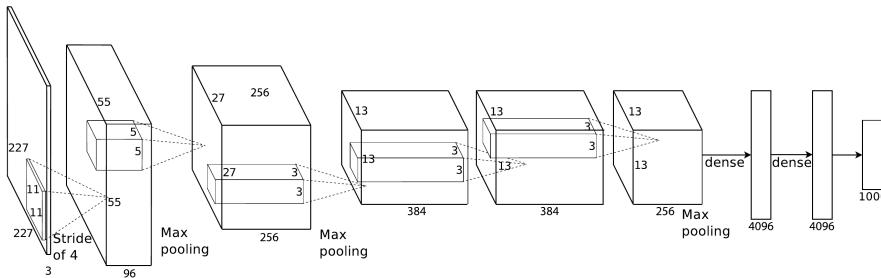


Figura 2.6: Esquema gráfico de la arquitectura conocida como “AlexNet”, diseñada para resolver un problema de clasificación con 1000 clases. Recuperado de la Figura 5.39 de [65]. Esta arquitectura presenta una serie de capas convolucionales con funciones de activación no lineales ReLU y max pooling, que formarían el *backbone* y una serie de capas FC (*classifier*), con una capa final softmax, que aplimentaría una función de pérdida de entropía cruzada multiclasa.

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** [70, 71]: 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.

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** [72]: 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 2.7). En CNNs suele aplicarse a capas *fully-connected*, aunque existen variantes como *Spatial Dropout* [73] que elimina canales completos en capas convolucionales, forzando una distribución más robusta de características.

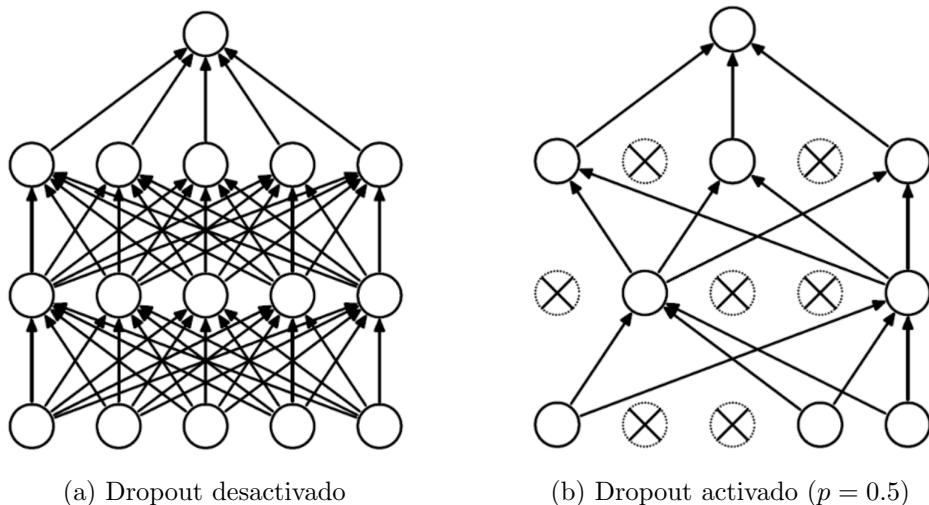


Figura 2.7: Diagrama del funcionamiento de neuronas con *dropout*. Recuperado de la Figura 5.29 de [65]. Cuando se evalúa el modelo, todas las unidades funcionan correctamente (a). Durante el entrenamiento, algunas son “apagadas” (b).

- **Batch normalization** [74]: 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 [75] y reduce la sensibilidad a la tasa de aprendizaje [76], permitiendo establecer valores más altos. En CNNs se aplica típicamente después de las capas convolucionales y antes de la función de activación

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)

Por completar
(AGOSTO)

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 [53]. En redes neuronales, el aprendizaje consiste en ajustar pesos, y en el caso del *transfer learning*, estos pesos se inicializan con valores previamente optimizados para una tarea fuente, en lugar de comenzar con valores aleatorios (véase la Figura 2.8).

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. Por ejemplo, en [38] 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.

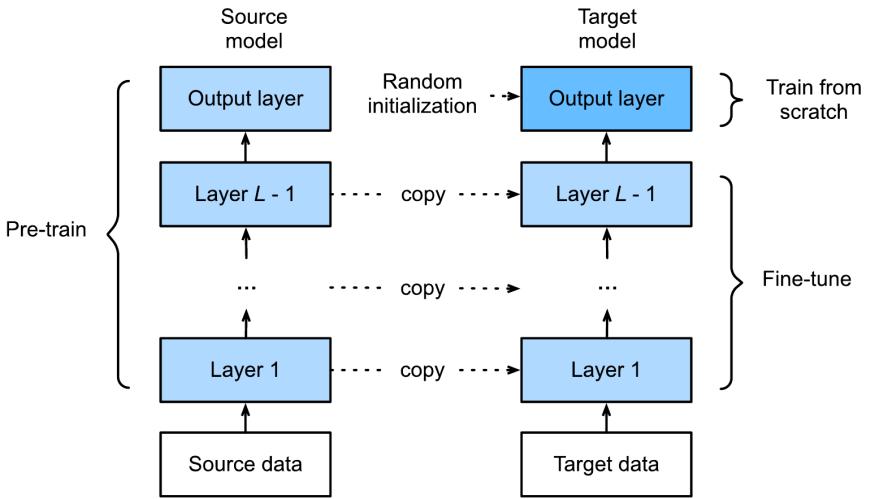


Figura 2.8: Diagrama de *fine-tuning* de un modelo en una nueva tarea. Recuperado de la Figura 19.2 de [60]. La capa final de salida es entrenada desde cero para la nueva tarea. El resto de capas son inicializadas con los pesos previos.

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 caractériza la dispersión de los valores que podrían atribuirse razonablemente al **mensurando**, que es como se denomina a la magnitud a ser medida. [...] El parámetro puede ser, por ejemplo, una desviación estándar, o la amplitud

¹⁶Ciencia de las mediciones y sus aplicaciones [77].

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*) [78] y el Vocabulario Internacional de Metrología (*Vocabulaire international de métrologie*) [77].

de un intervalo con un nivel de confianza establecido” [78].

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 máximo posible a dicho valor, minimizando las distintas componentes del error [78]:

- para el error aleatorio, esto se logra aumentando el número de observaciones, ya que su distribución tiende a 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.

Sin embargo, en la práctica no existen reglas claras para distinguir 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 al enfoque anterior, se presenta el **enfoque basado en la incertidumbre** [78], 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 potencial del instrumento —o modelo en ML—.

2.3.1. Incertidumbre en *machine learning*

Las fuentes de incertidumbre pueden ser muy variadas, y su identificación requiere en muchos casos de conocimiento específico del problema. No obstante, en términos prácticos, suelen considerarse dos tipos de incertidumbre en las predicciones realizadas en ML [79, 80]:

- La **incertidumbre aleatoria o estocástica** procede de la variabilidad aleatoria de un fenómeno. Es irreducible por naturaleza, aunque se disponga de más datos. Un ejemplo típico es el resultado de lanzar una moneda al aire. Incluso el mejor modelo solo será capaz de dar probabilidades para las dos posibles salidas, sin una respuesta definitiva. En el contexto de la estimación de la edad forense, esta incertidumbre

se manifiesta en las diferentes edades biológicas que pueden obtenerse para individuos de la misma edad cronológica. Se sabe que existe una correlación entre edad biológica y la cronológica, pero esta no es perfecta, debido a que existe variabilidad inherente al problema.

- 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, las limitaciones teóricas y prácticas del modelo escogido, etc. A diferencia de la incertidumbre aleatoria, esta sí es reducible por naturaleza; puede reducirse con más datos, mejores modelos o mayor comprensión del problema.

A estos, se les puede añadir un tercer tipo: el *drift* [80, 81], 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—; o la toma de imágenes médicas con una máquina distinta a la que se empleó para obtener las imágenes con las que se ha entrenado el modelo.

2.3.2. Cuantificación de la incertidumbre en *machine learning*

El desarrollo de las técnicas modernas de ML se asocia con un enfoque basado en el error, centrándose en la minimización y cuantificación del error en predicción. Este enfoque ha permitido que el aprendizaje automático despliegue un gran potencial en multitud de aplicaciones. Sin embargo, cuando se trata de aplicaciones críticas —como la medicina, los sistemas financieros o el control de infraestructuras— surge una necesidad esencial: no solo importa cuán precisa es una predicción, sino también cuán confiable es [82]. En respuesta a esta necesidad, durante la última década se ha producido un creciente interés y desarrollo de técnicas orientadas a la explicabilidad e interpretabilidad de la IA [83-86] y la cuantificación de la incertidumbre [80, 87, 88].

Mientras la explicabilidad de la IA busca entender las razones detrás de cada predicción centrándose en el estudio del modelo y arquitectura concretos [89], la **cuantificación de la incertidumbre (*uncertainty quantification*, UQ)** evalúa el grado de confianza en las predicciones realizadas y se centra en caracterizar las fuentes de variabilidad y posible error en los datos, el modelo y el entorno de aplicación [80].

Existe una gran variedad de técnicas de UQ. Estas técnicas pueden clasificarse de distintas formas:

- Algunas son *model-agnostic*, es decir, pueden aplicarse a cualquier tipo de modelo sin requerir acceso a su estructura interna; otras son *model-specific*, diseñadas para aprovechar características particulares del modelo subyacente.
- Algunas técnicas suponen que los datos siguen ciertas distribuciones estadísticas explícitas, mientras que otras operan sin realizar tales suposiciones.
- También existen métodos que asumen intercambiabilidad entre observaciones, frente a aquellos que no lo hacen y requieren estructuras de dependencia más complejas.
- **Procesos de regresión gausiana (*gaussian process regression, GPR*):**
- **Técnicas bayesianas:** Monte Carlo Dropout, redes neuronales bayesianas (*bayesian neural networks, BNNs*)
- **Técnicas *ensemble*:**
- Métodos deterministas:

2.4. Predicción conformal

La **predicción conformal** (*conformal prediction, CP*) [90, 91] es un marco teórico para la UQ en modelos de ML, que proporciona intervalos o conjuntos de predicción con garantías estadísticas de cobertura, esto es, para una entrada dada x , el marco de CP genera un conjunto de posibles salidas $\hat{C}(x) \subseteq Y$ que garantiza, con una probabilidad predefinida $1 - \alpha$, que la verdadera etiqueta o valor y esté contenida en $\hat{C}(x)$ (véanse los ejemplos de la Figura 2.9).

Para construir los conjuntos de predicción conformal, se requiere dividir el conjunto de datos disponible en al menos dos partes: un conjunto de entrenamiento, usado para ajustar el modelo base, y un **conjunto de calibración**, usado para calibrar la predicción conformal, tal y como veremos en los siguientes apartados. Con esto también se busca reducir la variabilidad de las predicciones puntuales, que pueden ser sensibles a pequeños cambios en los datos de entrada, como el ejemplo de la Figura 2.10. Estas garantías son válidas bajo el supuesto mínimo de intercambiabilidad de los datos, sin

Pablo: De hecho, aquí repites cosas... Tal y como dices, se ve que está por completar... Aún así, como consejo para todo lo que falta por completar: no te rayes ni líes mucho. Ya llevas 131 páginas y, bajo ningún concepto, queremos superar por mucho este valor (150 páginas podría estar bien, pero mucho más de eso... me parecería excesivo). Da pinceladas que demuestran que entiendes de lo que hablas, pero sin entrar en demasiado detalle. Si alguien quiere saber más: que se lea las referencias que incluyas y/o que te pregunte durante la



Regression task: age estimation

Model prediction: `24`

MAPIE prediction interval: `[20, 29]`
(with 90% confidence)



Classification task: species identification

Model prediction: `zebra`

MAPIE prediction set: `{zebra, horse}`
(with 90% confidence)

Figura 2.9: Ejemplo de predicción conformal en problemas de regresión (arriba) y clasificación (abajo). Recuperado de [92]. MAPIE es una biblioteca de Python para la cuantificación de incertidumbre, principalmente con técnicas de CP.

requerir hipótesis sobre la distribución subyacente de los mismos. La intercambiabilidad de los datos se refiere a que el orden de las observaciones no aporta información adicional, es decir, la distribución conjunta es invariante ante cualquier permutación de los índices.

2.4.1. Propiedades de la predicción conformal

La CP garantiza que las predicciones contengan el valor verdadero con al menor una probabilidad $1 - \alpha$, donde α es el nivel de significación:

$$P(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})) \geq 1 - \alpha$$

Esta propiedad se denomina **cobertura marginal válida** [94], y se cumple para todas las entrada X , siempre y cuando los datos sean intercambiables (*interexchangeable*). Esta intercambiabilidad en imágenes implicaría que todas fueran tomadas en condiciones similares: mismo dominio,

Original Sentence	Adversarial Example
<p>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</p>	<p>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</p>
Negative sentiment	Positive sentiment

Figura 2.10: Ejemplo adversario mal clasificado por un modelo de ML entrenado con datos textuales. Adaptado de la Figura 2 de [79], original de [93]. 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”. Con la CP se busca que predicciones no solo proporcionen una etiqueta puntual, sino un conjunto de posibles etiquetas que capture de manera robusta la incertidumbre asociada al ejemplo de entrada.

distribución de valores de salida, iluminación, resolución, estilo, etc. Sin embargo, la CP no asegura **cobertura condicional válida** [95]; es decir, no es posible garantizar cobertura para todos los subgrupos de datos sin hacer suposiciones fuertes o sacrificar utilidad práctica, en concordancia con el conocido *No Free Lunch Theorem* [96]. En la Figura 2.11 se presenta una noción de la diferencia entre cobertura marginal y condicional.

Además, el conjunto de calibración debe ser estadísticamente representativo de la distribución completa de los datos. Esto crea un compromiso fundamental: asignar más muestras a la calibración mejora la precisión de los intervalos predictivos, pero a costa de reducir el tamaño del conjunto de entrenamiento, lo que potencialmente puede empeorar el rendimiento del modelo base.

Algunas características deseables en los métodos de CP son:

- **Independencia del modelo (*model-agnostic*)**: que no requiera reentrenar el modelo ni modificar su arquitectura, permitiendo su aplicación *post-hoc* a modelos preentrenados.
- **Independencia del dominio (*domain-agnostic*)**: que pueda manejar entradas de cualquier tipo sin restricciones.
- **Predicción adaptativa (*adaptive prediction*)**: se refiere a que el intervalo o conjunto de valores varía su tamaño (o forma) en función de la incertidumbre asociada a cada predicción individual. En general, cuanto más rica y específica sea la información que el método utiliza sobre las predicciones y su incertidumbre, más adaptativo será.

- **Ser eficiente computacionalmente:** es preferible que tanto la calibración como la inferencia no sean computacionalmente muy costosas.
- **Robustez frente a datos ruidosos y detección de datos *out-of-distribution***¹⁸: que los intervalos reflejen adecuadamente la incertidumbre ante datos corruptos o fuera del dominio de entrenamiento, ya sea detectando de que el dato es anómalo o dando un intervalo/conjunto de predicción muy amplio.

En general, existe un *trade-off* entre flexibilidad y precisión: los métodos dependientes del modelo, del dominio o incluso del propio problema (incluyendo información experta), que asumen ciertas distribuciones de los datos, pueden cuantificar mejor la incertidumbre aleatoria y epistémica, y producir intervalos más ajustados e informativos.

2.4.2. Algoritmo conformal

Existen multitud de métodos de CP. Generalmente, estas dependen del tipo de problema a resolver: regresión [97-99], clasificación [100-102], series temporales [103-105]¹⁹, o detección de anomalías [106], entre otros.

A pesar de su diversidad, todos los algoritmos conformales comparten un elemento clave: la definición de una **función de no conformidad**, una heurística que mide la incertidumbre asociada a cada predicción. Intuitivamente, esta función actúa como una medida de discrepancia entre el valor predicho y el valor observado, y permite determinar cuán “extraña” o “no conforme” es una nueva observación respecto al comportamiento esperado del modelo.

La implementación de la predicción conformal consta de los siguientes pasos:

1. Se divide el conjunto de datos disponible en dos subconjuntos: un conjunto de entrenamiento, utilizado para ajustar el modelo predictivo (es decir, para entrenar el modelo), y un conjunto de calibración, que se reserva exclusivamente para estimar la incertidumbre mediante el cálculo de las puntuaciones de no conformidad. Esta separación permite que la estimación del intervalo de predicción sea independiente del proceso de entrenamiento, lo cual es crucial para garantizar la validez estadística del método.

¹⁸Los datos *out-of-distribution* son datos que no provienen de la misma distribución que los datos con los que se entrenó el modelo.

¹⁹El marco de CP clásico asume que los datos son intercambiables, una propiedad que no se cumple en las series temporales debido a la dependencia secuencial entre observaciones. A pesar de ello, se han desarrollado diversas extensiones del enfoque conformal para adaptarse a estos datos.

2. Se entrena el modelo predictivo utilizando únicamente el conjunto de entrenamiento.
3. **Calibración conformal:** Este proceso se da tras el entrenamiento del modelo. En este, se calculan las **puntuaciones de no conformidad (nonconformity scores)** R^{20} .

$$R = \{R_1, R_2, \dots, R_n\}$$

donde n es el número de ejemplos del conjunto de calibración.

Estas puntuaciones se derivan a partir de una heurística que combina al menos el valor real y el predicho del problema con otras posibles fuentes de información, como las entradas o incluso representaciones internas del modelo²¹. Bajo las garantías estadísticas que ofrece el marco teórico de la CP, esta flexibilidad muestra un gran potencial para ser integrada con otros métodos de UQ, ampliando así sus aplicaciones y mejorando la robustez de las estimaciones de incertidumbre.

Independientemente del diseño específico de la función de no conformidad, esta debe cumplir una condición esencial: las puntuaciones deben ser intercambiables entre el conjunto de calibración y las nuevas instancias. En otras palabras, deben ser idénticamente distribuidos. Esta propiedad es crucial para que CP garantice cobertura marginal válida a un nivel de confianza determinado. Por tanto, aunque existe flexibilidad en el diseño de la función de no conformidad, su elección debe considerar tanto la capacidad para capturar incertidumbre útil como el cumplimiento del supuesto de intercambiabilidad.

A continuación, se calcula el **umbral de no conformidad**. Para un nivel de significancia α , se selecciona el $(1 - \alpha)(1 + 1/n)$ -ésimo cuantil²² de las puntuaciones de no conformidad obtenidas en el conjunto de calibración (véase la Figura 2.12).

4. **Inferencia conformal:** Para cada nueva instancia x_{n+1} se genera una predicción puntual y_{n+1} utilizando el modelo entrenado. Luego, se construye un conjunto o intervalo de predicción $\Gamma(x_{n+1})$ a partir de la predicción puntual y el umbral de no conformidad $\hat{q}_{1-\alpha}$, tal que se garantiza con nivel de confianza $1 - \alpha$ que el verdadero valor y_{n+1} pertenezca al conjunto:

²⁰En la literatura, a este vector de puntuaciones se le suele denotar como R por ‘residual’ o E por ‘error’, aunque no tiene por qué corresponderse con estas variables.

²¹Cabe señalar que un método será independiente del modelo y del dominio cuando solo tenga en cuenta las salidas del modelo y los valores reales del problema para realizar la CP.

²²La corrección $(1 + 1/n)$ asegura validez estadística para conjuntos de tamaño finito.

$$y_{n+1} \in \Gamma_{1-\alpha}(x_{n+1})$$

La forma de construir $\Gamma(x_{n+1})$ depende de cómo se haya definido la función de no conformidad durante la fase de calibración. Por ejemplo, en el método *Inductive Conformal Prediction* [97] para problemas de regresión, —que describiremos en mayor profundidad en el Capítulo 4—, se utiliza el error absoluto como función de no conformidad. De esta forma, el umbral de no conformidad $\hat{q}_{1-\alpha}$ se considera el quantil $1 - \alpha$ del error que arroja el modelo. Y, en la inferencia, se toma como intervalo de predicción:

$$\Gamma_\alpha(x_{n+1}) = [\hat{y}_{n+1} - q_{1-\alpha}, \hat{y}_{n+1} + q_{1-\alpha}]$$

La construcción del intervalo de predicción surge directamente de despejar el valor real y_{n+1} en la expresión que iguala la función de no conformidad, evaluada sobre la nueva instancia, con el umbral de no conformidad. En este caso en el que se emplea el error absoluto, es decir:

$$R(y_i, \hat{y}_i) = |y_i - \hat{y}_i|$$

entonces al imponer la condición $R(y_{n+1}, \hat{y}_{n+1}) \leq q_{1-\alpha}$, se obtiene:

$$|y_{n+1} - \hat{y}_{n+1}| \leq q_{1-\alpha}$$

Despejando y_{n+1} , se obtiene el intervalo:

$$y_{n+1} = \hat{y}_{n+1} \pm q_{1-\alpha}$$

Esta expresión determina los límites inferior y superior del intervalo de predicción conformal para dicha instancia.

Pablo: Enriquece esta explicación teórica con un ejemplo aplicado de juguete, en donde tienes un modelo caja negra que realiza la predicción. Muestra cómo calculas $q_{1-\alpha}$ en base a un conjunto de calibración con 10 valores reales. Yo creo que quedaría todo mucho más claro. Y luego genera el intervalo de predicción para una estimación concreta.

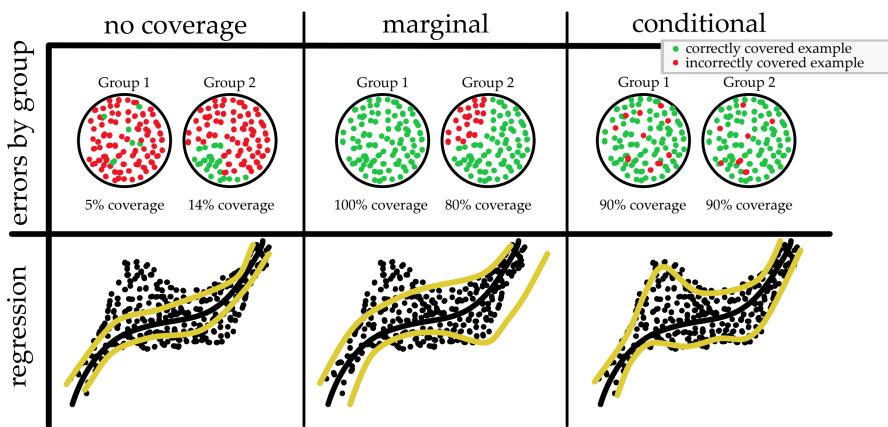
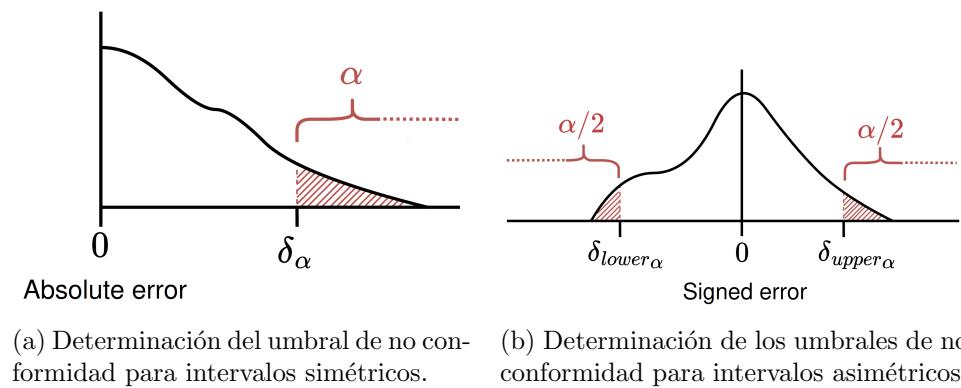


Figura 2.11: Conjuntos de predicción bajo distintas nociones de cobertura: sin cobertura garantizada, con cobertura marginal y con cobertura condicional. Recuperado de [91]. En la parte inferior de la figura se muestra la amplitud de los intervalos de predicción (en línea amarilla) generados en un problema de regresión. En la parte superior, las instancias se dividen en dos grupos; los valores reales contenidos en los intervalos se representan en verde, mientras que los no contenidos aparecen en rojo. La primera columna ilustra un caso con intervalos de predicción demasiado estrechos, lo que resulta en una baja cobertura: la mayoría de los valores reales quedan fuera del intervalo. En la segunda columna, los intervalos son más amplios y permiten capturar una mayor proporción de los valores reales, alcanzando una cobertura marginal del 90 % en el conjunto total. Sin embargo, esta cobertura no se distribuye equitativamente: el error se concentra en una región específica dentro de uno de los grupos, lo que indica ausencia de cobertura condicional. Finalmente, en la tercera columna, los intervalos se ajustan a la distribución de las predicciones, logrando cobertura marginal como condicional del 90 %, con un error repartido de manera uniforme entre regiones y grupos, reflejando una calibración más precisa y equitativa del modelo.



(a) Determinación del umbral de no conformidad para intervalos simétricos.

(b) Determinación de los umbrales de no conformidad para intervalos asimétricos.

Figura 2.12: Determinación del umbral de no conformidad para intervalos simétricos y asimétricos. En (a), el error es absoluto, y el umbral se calcula como se ha especificado anteriormente. En (b), el error tiene signo, y hay dos umbrales de incertidumbre, uno por cada cola, calculado como el cuantil con significación $\alpha/2$ de los errores negativos y de los errores positivos, respectivamente para el umbral inferior y el umbral superior.

Capítulo 3

Estado del arte

3.1. Estimación de la edad en antropología forense

En ausencia de documentación escrita confiable y cuando otros métodos como los genéticos o dactilares no son viables, los métodos más precisos para estimar la edad se basan en el análisis del estado de los huesos del cuerpo humano. Los huesos experimentan cambios continuos a lo largo de la vida, y estas transformaciones progresivas permiten determinar la *edad biológica* de un individuo. Esta edad refleja la etapa de desarrollo en la que se encuentra el esqueleto dentro del proceso de cambios que ocurren desde el nacimiento hasta la vejez [3]. Cabe destacar que la edad biológica no siempre coincide con la edad cronológica —el tiempo transcurrido desde el nacimiento—, pero ambas guardan una correlación significativa, lo que permite aproximaciones razonables en contextos forenses, antropológicos o médicos.

Incluir gráfica con el número de publicaciones para estimación de la edad en antropología forense

Las técnicas de estimación de edad presentan diferencias significativas en individuos maduros e inmaduros [107]. La diferencia radica en el grado de desarrollo esquelético y dental: en inmaduros, el esqueleto y la dentición no están completamente formados, por lo que los métodos se basan en patrones de crecimiento y osificación; en contraste, en maduros (con desarrollo completo), las técnicas se enfocan en cambios degenerativos, como el deterioro articular o la pérdida ósea.

La estimación en cuerpos subadultos (individuos que no han alcanzado la madurez esquelética) se basa en el desarrollo y erupción dental¹ [108], los tiempo de aparición y cambios en la morfología de centros de osificación², y

¹La erupción dental es el proceso natural mediante el cual los dientes se desplazan desde el interior del hueso maxilar o mandibular hasta alcanzar su posición definitiva en la boca, atravesando las encías.

²La osificación es el proceso natural mediante el cual el cartílago o tejido conectivo se convierte en hueso. Los centros de osificación son regiones específicas del esqueleto donde

los tiempos de fusión de los centros primarios (también denominados diáfisis) y secundarios (epífisis) [109, 110]. Los métodos de mayor precisión se basan en el desarrollo dental, dado que estos, para una determinada edad cronológica, muestran menor variabilidad que el esqueleto [111]. En ausencia de estos, se recurre al análisis de la epífisis de diferentes huesos, cuya formación y fusión son clave para la estimación de la edad esquelética [110].

La valoración en adultos es más compleja, dado que el desarrollo de la dentadura se ha completado, así como el crecimiento del esqueleto ha cesado [3], por lo que los indicadores se basan más en características del deterioro óseo; pero la variabilidad de estas aumenta con la edad debido al efecto acumulativo de las influencias ambientales³ [114, 115]. Actualmente, se recomienda un análisis conjunto del proceso degenerativo de síntesis pública (propuesto en [116]) y de las transparencias en las raíces dentales [117]. Cuando este no es posible, pueden emplearse otros métodos [118], como el análisis de la superficie del ilion [119], el examen del extremo esternal de la cuarta costilla [120], o el estudio de los procesos de obliteración de las suturas craneales [121].

Sin embargo, cuando la estimación de edad se realiza en personas vivas, no se tiene acceso a sus huesos de forma directa. En estos casos se recurren a otro tipo de métodos como exámenes físicos o toma de imágenes médicas. El Grupo de Estudio para el Diagnóstico Forense de Edad (AGFAD) de la Sociedad Alemana de Medicina Legal⁴ ha publicado recomendaciones estandarizadas sobre cómo llevar a cabo evaluaciones de edad en personas vivas. En estas incluyen estudios como [21]: historial clínico, examen físico, radiografía de una mano, radiografía panorámica maxilofacial y si está indicado, una tomografía computerizada de cortes finos de la epífisis mediales de las clavículas. Se suelen combinar múltiples métodos para una mayor exactitud en la predicción. Dependiendo de los asuntos legales, se requerirá la estimación de la edad mínima del individuo o su edad más probable [21] (véase un ejemplo en la Figura 3.1).

comienza el proceso de formación ósea durante el desarrollo embrionario, fetal, infantil y adolescente.

³Por ejemplo, artículos como [112, 113] indican que la obesidad puede causar que se sobreestime la edad del cuerpo, mientras que personas con una complexión más ligera o bajo peso corporal tienden a presentar una infraestimación de la edad.

⁴La Arbeitsgemeinschaft für Forensische Altersdiagnostik (AGFAD) es una organización alemana, compuesta por expertos multidisciplinares. Ha publicado protocolos estandarizados para la estimación de edad en personas vivas, logrando reconocimiento y aplicación a nivel internacional.

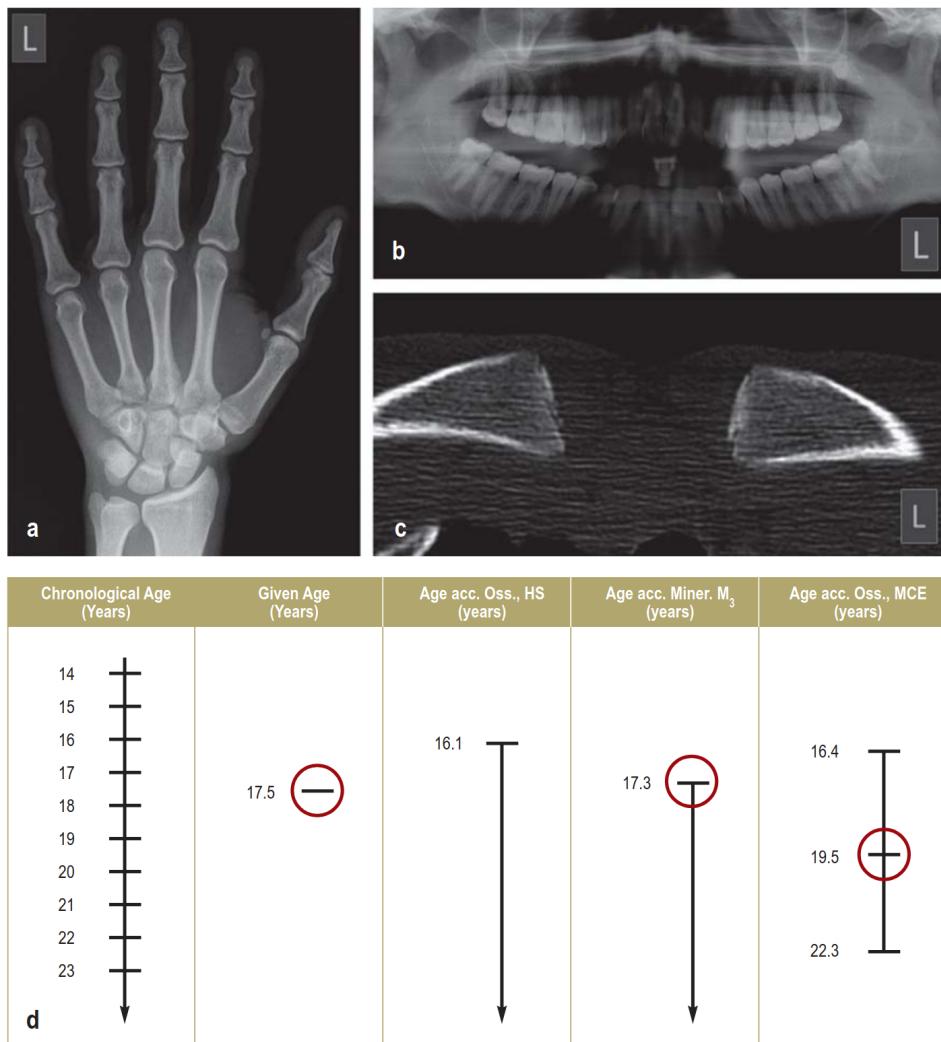


Figura 3.1: Hallazgos radiológicos en un posible menor con edad en disputa. Criterio de edad mínima para la determinación de edad. Recuperado de la Figura 1 de [21]. El sujeto masculino afirma tener 17,5 años. Tanto la historia clínica como el examen físico no revelan signos de alteraciones en el desarrollo. Se presentan las siguientes imágenes: una radiografía de la mano izquierda en (a), una radiografía panorámica maxilofacial en (b) y una tomografía computarizada de las epífisis mediales de la clavícula en (c). En la imagen (d) se muestran los rangos de edad estimados según los diferentes indicadores radiológicos. Las edades mínimas asociadas a las etapas de desarrollo observadas son 16,1 años, 17,3 años y 16,4 años, respectivamente. La edad mínima del individuo queda determinada por la mayor de estas estimaciones, es decir, 17,3 años. Esta edad mínima estimada es consistente con la edad declarada por el examinado.

3.2. Estimación de la edad en antropología forense usando *machine learning*

Incluir gráfica con el número de publicaciones para estimación de la edad en antropología forense usando ML

Los métodos manuales de estimación del PB se basan en la evaluación visual y en el análisis morfométrico de rasgos esqueléticos. Sin embargo, su aplicación demanda conocimiento especializado, pueden presentar ambigüedades en su formulación que den lugar a interpretaciones variables [122], y están sujetos a posibles errores de medición [24], sesgando el proceso y reduciendo su fiabilidad. Estas limitaciones han motivado el desarrollo de métodos automatizados basados en ML, que siguen dos enfoques principales.

El primero consiste en tomar un método clásico de AF y automatizar sus etapas mediante herramientas computacionales. Para ello, el método debe definir:

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 que opere sobre estas características predefinidas.

La Figura 3.2 ilustra un ejemplo de este enfoque. Entre las propuestas destacadas podemos mencionar BoneXpert [123], que empleaba técnicas clásicas de ML y demostró robustez en poblaciones diversas, tanto en origen geográfico como en condiciones clínicas de adquisición de imágenes [124-126].

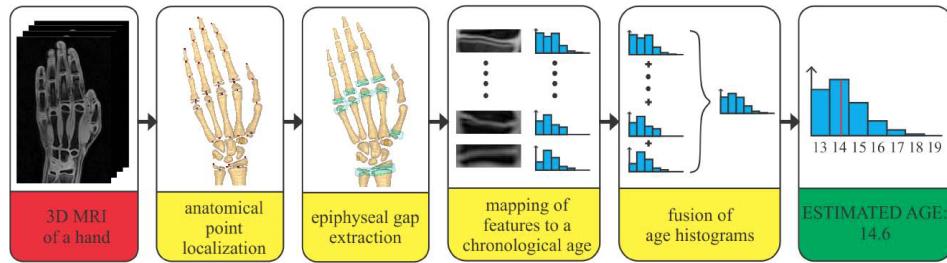


Figura 3.2: Procedimiento secuencial clásico de ML para el método propuesto en [127]. Las características se extraen manualmente o con herramientas independientes del modelo.

En cambio, el auge del *deep learning*, impulsó el aprendizaje extremo a extremo (*end-to-end learning*), donde un único modelo aprende de manera automática tanto la extracción de características como la clasificación/regresión a partir de los datos en bruto. Las redes neuronales convolucionales

consiguen eliminar la dependencia de criterios antropológicos preestablecidos, y permiten al modelo extraer por sí mismo las características más relevantes para la tarea en que se entrenan.

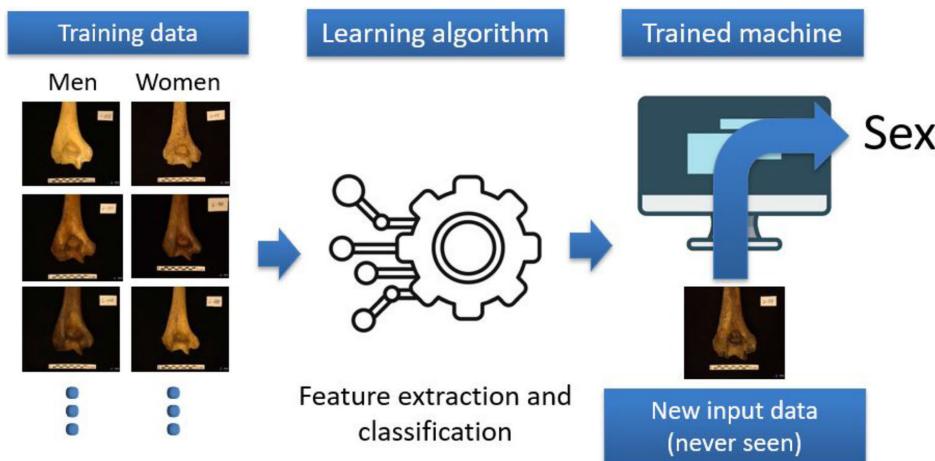


Figura 3.3: Metodología de construcción de un modelo *end-to-end*. Recuperado de [38]. La CNN aprende a extraer las características de la manera más conveniente para resolver el problema en el que se le entrena, en este caso estimación de sexo.

Siguiendo este paradigma, se han propuesto numerosos modelos basados en redes convolucionales. Un ejemplo destacado es el propuesto en [128], el cual, entrenado con resonancias magnéticas 3D de manos, aprende a identificar las características más relevantes para la estimación de edad automáticamente. Este modelo se ha consolidado como estado del arte en su dominio, alcanzando un error absoluto medio de 0.37 ± 0.51 años⁵. Además, los autores demostraron su adaptabilidad al procesar imágenes 2D de radiografías, logrando también un rendimiento líder en el ámbito de rayos X.

3.3. Cuantificación de la incertidumbre para la estimación de la edad

La mayoría de trabajos académicos de AF no presentan un enfoque explícito en la cuantificación de incertidumbre de las predicciones, pero sí evalúan la confiabilidad de los métodos propuestos, a través del análisis del error.

Incluir gráfica con el número de publicaciones para cuantificación de la incertidumbre en estimación de la edad

⁵Esta notación, como veremos a continuación, representa el error absoluto medio y su desviación estándar.

3.3. Cuantificación de la incertidumbre para la estimación de la edad

El enfoque principal en la evaluación de estos métodos consiste en comparar la edad cronológica (la *ground truth*) con la edad biológica (la estimada), las cuales no siempre presentan una correlación directa [129].

Es por ello que los métodos manuales suelen estimar intervalos de edad o, en casos específicos —especialmente en contextos legales—, valores de edad mínima probable. Esto se debe a la variabilidad biológica entre individuos, influenciada por factores genéticos, ambientales, nutricionales y de salud, que impide establecer una edad cronológica exacta a partir de los indicadores empleados. En general, los intervalos son determinados en base a una población de referencia, y se suele escoger un intervalo que cubra un 95 % de los casos esperados [118] (véase un ejemplo en la Figura 3.4).

Por otro lado, los modelos de ML suelen generar predicciones puntuales (valores únicos) sin proporcionar intervalos de confianza o distribuciones probabilísticas asociadas. De esta forma, la edad biológica se trata como una construcción artificial —cuyos valores son los predichos por el modelo—, que, idealmente, representan las edades cronológicas más probables en un continuo de cambios observado en un dominio concreto, que son generalmente imágenes médicas.

La métrica más empleada para esta evaluación es el error absoluto medio \pm la desviación estándar. Esta cuantifica el error absoluto promedio entre la edad cronológica y la edad biológica predicha, proporcionando una medida de la precisión del modelo; y la desviación estándar indica la dispersión de estos errores, reflejando la consistencia del modelo en sus predicciones. Otras métricas como el coeficiente de correlación de Pearson (r) o el coeficiente de determinación (R^2) —aunque menos frecuentes— aportan información sobre la relación lineal entre predicciones y valores reales.

Sin embargo, estas métricas pueden esconder sesgos en ciertos grupos etarios⁶, y un modelo con mal desempeño en la población general, puede arrojar buenos resultados en algunos grupos específicos, o viceversa. Es por ello que el análisis se puede completar empleando las métricas en subpoblaciones específicas, apoyándose en representaciones gráficas, permiten visualizar las relaciones no lineales entre variables, así como identificar patrones, tendencias o valores atípicos. Entre ellas destacan:

- Gráficas de dispersión: comparando edad cronológica vs. edad biológica, o mostrando errores en función de las edades cronológica o biológica (véase la Figura 3.5).
- Histogramas, gráficos de densidad o diagramas de cajas para representar la distribución de errores. En la Figura 3.6 vemos un ejemplo en el que plasman un histograma escrito como texto.

⁶Los grupos etarios son intervalos de edad utilizados para clasificar a la población o a los sujetos de estudio en categorías específicas según su edad cronológica.

3.4. Estimación del sexo en antropología forense

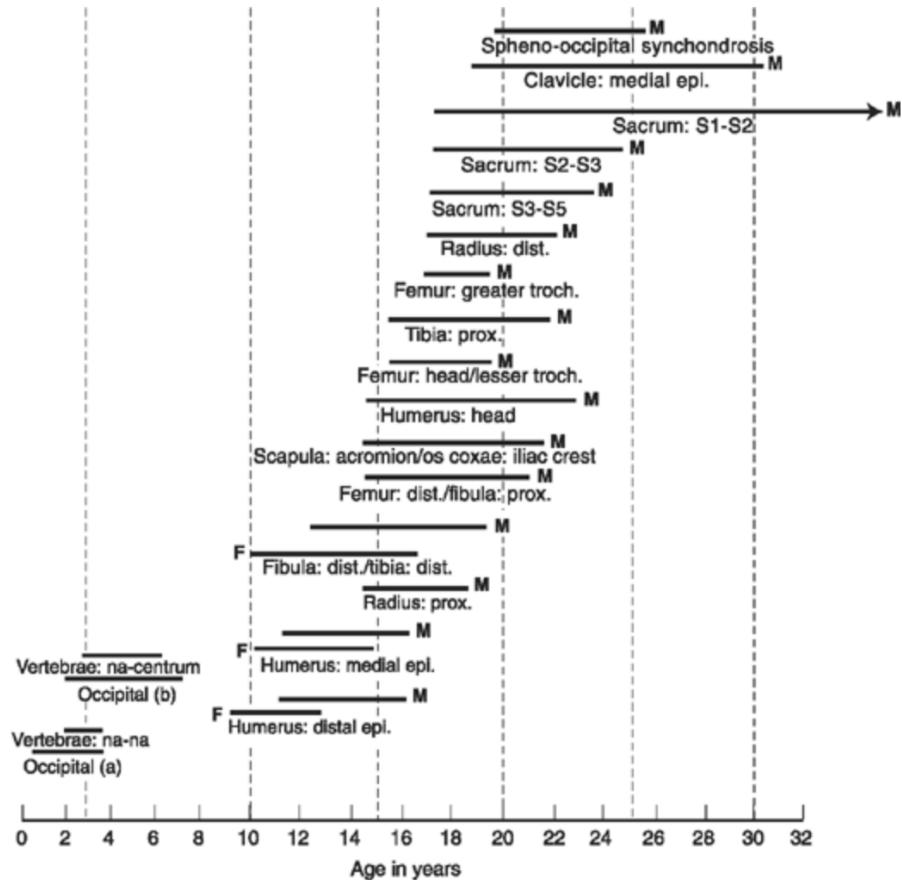


Figura 3.4: Cronograma de desarrollo de la unión epifisaria. Recuperado de [3], original de [130]. Este gráfico combina información de diferentes fuentes sobre los rangos de edades en los que ocurre la fusión de diversas epífisis del esqueleto humano, representado con una línea que indica la variabilidad del momento en que puede producirse dicha fusión con un 95 % de confianza, permitiendo estimar la edad del individuo a partir del grado de unión observado.

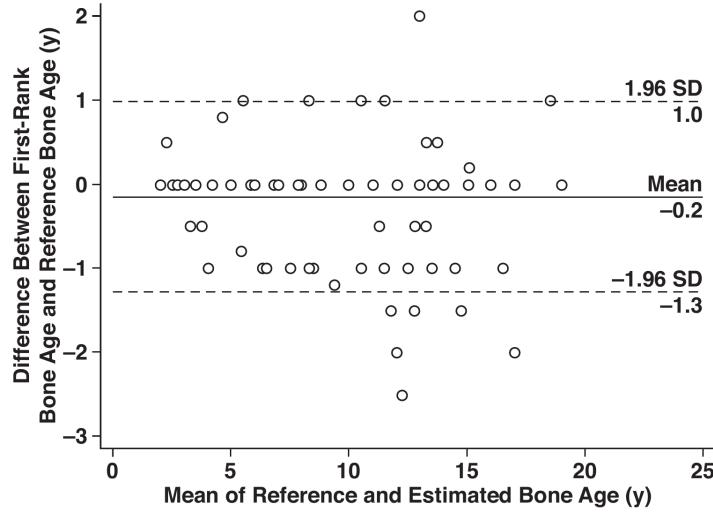


Figura 3.5: Distribución del error por edad real para el modelo propuesto en [29]. Recuperado de la Figura 3. Esta visualización permite observar el error en cada instancia predicha para diferentes edades reales.

		BAR1 → BAR1 (mean ± std [years]) error histogram [years] (-3 -2 -1 0 +1 +2 +3)	BAR2 → BAR2 (mean ± std [years]) error histogram [years] (-3 -2 -1 0 +1 +2 +3)	CA → CA (mean ± std [years]) error histogram [years] (-3 -2 -1 0 +1 +2 +3)
M	BoNet (Spampinato et al., 2017)	0.68	0.74	—
	cnn-2d-bones	—	—	—
	cnn-2d-bones-carpal	0.54 ± 0.60 2 9 102 214 82 9 0 0.49 ± 0.60	0.57 ± 0.60 0 11 89 197 68 4 0 0.57 ± 0.60	0.86 ± 0.63 6 28 95 140 93 21 0 0.77 ± 0.60
	cnn-2d-hand	0.58 ± 0.64 2 14 97 207 86 10 2	0.66 ± 0.65 1 12 99 175 69 13 0	0.89 ± 0.67 8 40 92 137 90 15 1
F	BoNet (Spampinato et al., 2017)	0.79	0.75	—
	cnn-2d-bones	—	—	—
	cnn-2d-bones-carpal	0.70 ± 0.61 1 13 105 157 124 17 0 0.66 ± 0.61	0.68 ± 0.65 3 5 73 173 137 25 1 0.60 ± 0.62	1.00 ± 0.73 5 20 80 124 107 33 7 0.90 ± 0.70
	cnn-2d-hand	0.89 ± 0.75 2 41 108 135 95 29 7	0.77 ± 0.70 2 31 107 156 96 23 2	1.20 ± 0.96 16 37 63 89 47 21 3
ALL	BoNet (Spampinato et al., 2017)	0.73	0.74	—
	cnn-2d-bones	—	—	—
	cnn-2d-bones-carpal	0.62 ± 0.61 3 22 207 371 206 26 0 0.57 ± 0.61	0.62 ± 0.63 3 16 162 370 205 29 1 0.58 ± 0.61	0.93 ± 0.69 11 48 175 264 200 54 7 0.83 ± 0.66
	cnn-2d-hand	0.73 ± 0.72 4 55 205 342 181 39 9	0.72 ± 0.68 3 43 206 331 165 36 2	1.03 ± 0.82 24 77 155 226 137 36 4

Figura 3.6: Estudio del error en los métodos 2D propuestos en [128]. Recuperado de la Tabla 2 de [128]. Se observa que se muestra tanto el error absoluto medio \pm desviación estándar como un histograma de errores, que permite ver la distribución general de los errores.

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 12 países distintos (véase en la tabla 4.1), obtenidas con distintos modelos de máquinas de rayos X¹. Este conjunto de datos ha sido proporcionado 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 “sample” (clasificación según el origen geográfico de la radiografía).
- imágenes bidimensionales de radiografías panorámicas maxilofaciales, con una imagen asociada a cada individuo 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 ha ignorado el campo “sample”, 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 análisis ni en el entrenamiento de los modelos, centrándose exclusivamente en las variables de edad, sexo e imagen.

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 del Sur	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.

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.

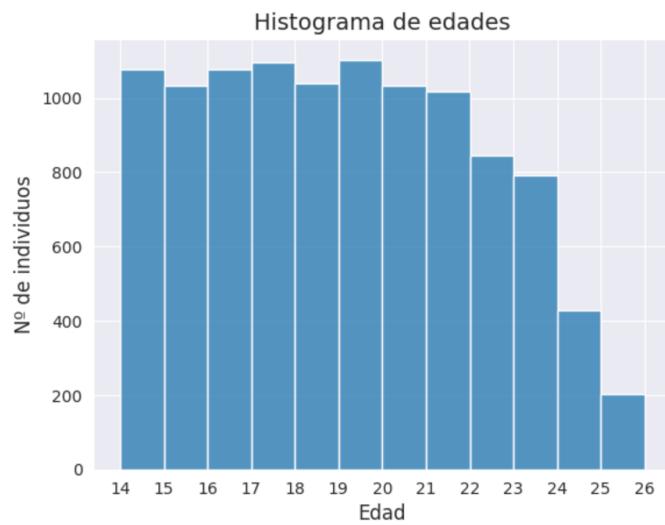


Figura 4.1: Histograma de edad de los individuos del conjunto de datos disponible.

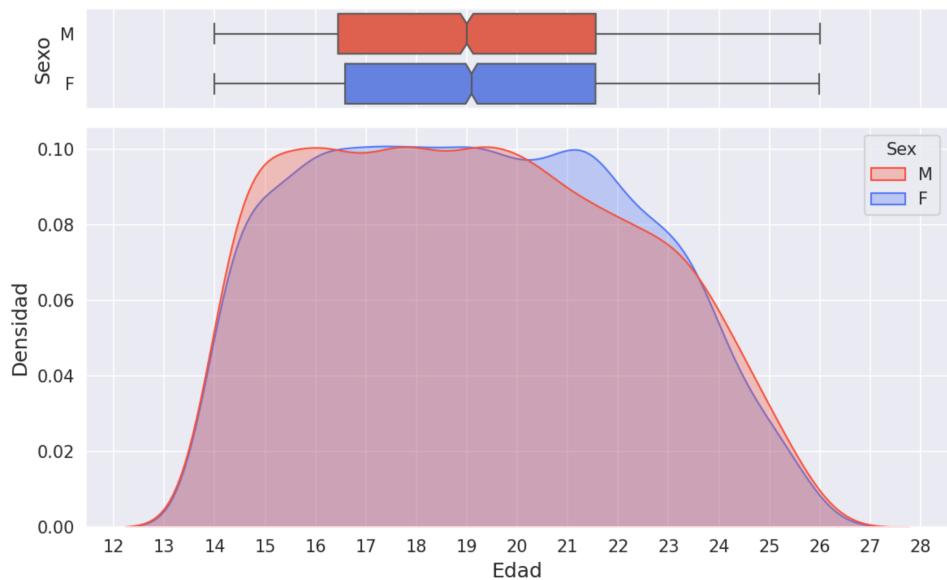


Figura 4.2: Gráficas de densidad y de caja de edad por sexo de los individuos del conjunto de datos disponible.

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. La división de ambos conjuntos se hizo de forma estratificada, de lo que se asume que la distribución será igual en ambos datasets.

4.2. Problemas planteados

Como se ha mencionado anteriormente, y con el objetivo de validar los métodos de predicción conformal en diferentes tipos de problemas, este trabajo se centra en tres casuísticas que, si bien están relacionadas en el ámbito de la AF, se tratan de diferente forma en el campo del ML:

1. estimación de la edad legal resuelta como un problema de regresión;
2. estimación de la edad legal planteada como un problema de clasificación binaria (mayor o menor de 18 años); y
3. estimación simultánea de la edad legal y el sexo planteada como un problema de clasificación multiclasa, combinando en cuatro clases sexo (masculino o femenino) y la edad (mayor o menor a 18 años).

4.2.1. Problema de estimación de edad

El problema de **estimación de edad** (*age estimation*, AE) consiste en predecir la edad cronológica de un individuo en una escala continua, lo que lo define como un problema de regresión.

Para ello, se ha escogido usar las imágenes de radiografías maxilofaciales como entrada del algoritmo (véase la Figura 4.3). Inicialmente se consideró incluir el sexo como metadato adicional en el modelo; sin embargo, se descartó tras observar de manera preliminar que no tenía un impacto significativo en el rendimiento del modelo, además de que su exclusión simplifica la arquitectura.

Para agosto: Un anexo que demuestre esto, yo ya lo he comprobado empíricamente

4.2.2. Estimación de mayoría de edad

Un problema inmediatamente derivado del anterior es la **estimación de mayoría de edad** (*assessment of the age of majority*, AAM), útil en contextos legales donde es necesario determinar si una persona ha alcanzado

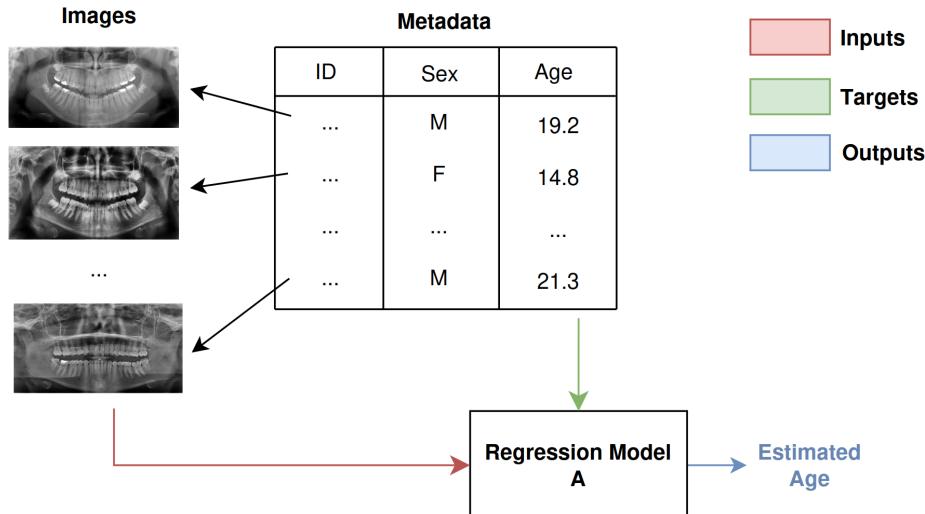


Figura 4.3: Esquema visual del modelos de regresión propuesto. El modelo solo tiene radiografías maxilofaciales como entrada.

la mayoría de edad. Este se trata de un problema de clasificación binaria, en el que el objetivo es asignar a cada individuo una de dos clases: “menor de edad” o “mayor de edad”.

4.2.3. Clasificación combinada de mayoría de edad y sexo

Finalmente, se propone ampliar el anterior problema a una **clasificación combinada de mayoría de edad y sexo (*age of majority and sex classification, AMSC*)**. Este se trata de un problema de clasificación multiclasa, en el que se asigna a cada individuo a una de las posibles combinaciones entre el estatus de mayoría de edad (mayor o menor) y el sexo (masculino o femenino). En total, se consideran cuatro clases: “menor masculino”, “menor femenino”, “mayor masculino” y “mayor femenino”.

Este problema busca evaluar si es posible estimar el sexo de una persona a partir de radiografías maxilofaciales. Para ello, se exploran características morfológicas relevantes, como la forma y tamaño de los caninos [131], así como los contornos de la rama mandibular y el mentón [132], cuya morfología presenta una fuerte correlación con el sexo biológico.

4.3. Métodos propuestos

4.3.1. 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 —y sexo, opcionalmente— como entrada, y estimación de edad a la salida.

Como modelo, empleamos una CNN, dado su buen desempeño en tareas de visión por computador. Específicamente, utilizamos la arquitectura ResNeXt50 [133] preentrenada en Imagenet [134] como punto de partida. Aunque ResNeXt50 fue entrenado originalmente para una tarea de clasificación, se puede adaptar fácilmente a tareas de regresión —como la estimación de edad— reemplazando su capa final por una capa de salida adecuada. Por otro lado, a pesar de haber sido entrenado en un dominio diferente al de nuestro problema, el uso de pesos preentrenados ofrece una ventaja significativa: permite una inicialización más robusta que comenzar desde cero, ya que la arquitectura ya ha aprendido a extraer patrones visuales básicos, como bordes y texturas, mediante filtros genéricos.

4.3.2. Regresión cuantílica

La **regresión cuantílica** (*quantile regression, QR*) es un tipo de regresión que, a diferencia de la regresión puntual, predice intervalos o cuantiles específicos de la distribución de la variable respuesta, en lugar de solo su media. Esta técnica parte de la noción de que la inferencia estadística no se limita a un valor único, sino que puede representarse mediante una distribución de valores probables, de la cual es posible estimar ciertos cuantiles para describir la variabilidad del comportamiento de la variable objetivo.

En este sentido, la regresión cuantílica permite modelar límites inferiores y superiores (por ejemplo, el percentil 10 % y 90 %) para capturar la incertidumbre o heterocedasticidad en los datos. No debe confundirse con una técnica de UQ, ya que no modela explícitamente la incertidumbre epistémica ni proporciona garantías estadísticas de cobertura como lo hacen los métodos de predicción conformal. Sin embargo, puede utilizarse como parte de un enfoque para cuantificar la incertidumbre aleatoria o condicional al estimar intervalos de predicción directamente a partir de los datos.

Esta técnica de regresión puede implementarse en modelos de redes neuronales y modelos tipo *ensemble*, aunque su implementación difiere significativamente.

En redes neuronales, esta regresión requiere de:

- Definir una capa de salida con múltiples neuronas, una por cada cuantil deseado (\hat{q}_τ). Por ejemplo, para obtener una región del 90 % con predicción puntual, tendríamos que inferir los cuantiles 0.05 y 0.95 para los límites inferior y superior, respectivamente, junto con el cuantil 0.5 para la predicción central.
- Cambiar la función de pérdida para la estimación de cuantiles. En general, se suele utilizar la pérdida *pinball* [135]. La **función de pérdida pinball** es una generalización de la función de pérdida $L1^2$, que penaliza las predicciones de manera asimétrica según el error es positivo o negativo. Para un cuantil $\tau \in (0, 1)$, se define como:

$$L_\tau(y, \hat{q}_\tau) = \begin{cases} \tau \cdot (y - \hat{q}_\tau) & \text{si } y \geq \hat{q}_\tau \\ (1 - \tau) \cdot (\hat{q}_\tau - y) & \text{si } y < \hat{q}_\tau \end{cases}$$

La Figura 4.4 ilustra cómo la pérdida penaliza de forma desigual los errores positivos y negativos. Mientras que la pérdida $L1$ se centra en ajustar la mediana (cuantil 0.5), la pérdida pinball permite dirigir una salida del modelo en cualquier cuantil deseado. Esto es especialmente útil cuando se desea modelar distribuciones asimétricas y capturar diferentes percentiles de la variable de salida, en lugar de asumir una distribución de errores simétrica, como la normal.

A diferencia de con la función de pérdida $L1$, que trata todos los errores como absolutos y busca ajustar la mediana (cuantil 0.5) de la distribución, la *pinball loss* permite enfocar la salida del modelo en cualquier cuantil específico. Esto es especialmente útil para capturar diferentes percentiles de la variable de salida, y modelar la variabilidad en las predicciones de forma más detallada.

Esta función de pérdida, aplicada a múltiples salidas (cada una asociada a un cuantil específico), busca que las predicciones del modelo cubran la proporción deseada de los datos dentro del intervalo definido por parejas de cuantiles (τ_1, τ_2), tratando de cumplir así con un criterio de cobertura probabilística. Por ejemplo: con dos salidas $\tau_1 = 0.05$ y $\tau_2 = 0.95$, se busca que el 90 % las observaciones reales (y) estén entre los límites predichos de los dos cuantiles ($\hat{q}_{0.05}$ y $\hat{q}_{0.95}$).

Además, como ya se comentó al inicio, se puede incluir una tercera salida para el cuantil $\tau_3 = 0.5$, correspondiente a la mediana de la

²También conocida como error absoluto medio, cuantifica la diferencia entre los valores predichos por un modelo y los valores reales como la diferencia absoluta entre cada par:

$$L1 \text{ loss} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

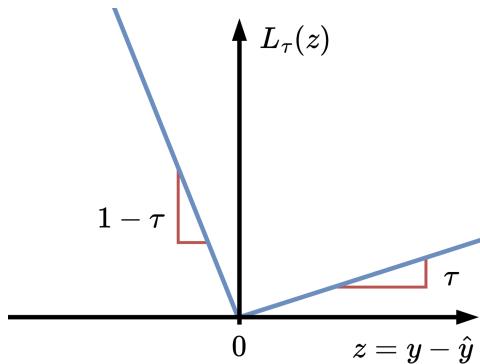


Figura 4.4: Visualización de la función de pérdida *pinball* para cada valor de error. Adaptado de la Figura 1 de [98]. Esta concretamente muestra la función de pérdida para un cuantil cercano a cero, ya que es más permisivo con los errores positivos que con los negativos, lo cual empujará sus predicciones hacia la parte inferior de la distribución objetivo.

distribución condicional, que actúa como una predicción puntual y es equivalente a minimizar la pérdida $L1$.

Finalmente, el valor arrojado por la función de pérdida conjunta de los cuantiles se suele expresar como la media de las pérdidas para cada cuantil:

$$\mathcal{L}_{total} = \frac{1}{Q} \sum_{i=1}^Q L_{\tau_i}(y, \hat{q}_{\tau_i})$$

donde Q es el número de cuantiles empleados.

Por tanto, este tipo de regresión da una estimación puntual \hat{y} (correspondiente a $\hat{q}_{0.5}$) y una estimación interválica formada por límites inferior y superior $[\hat{q}_{lower}, \hat{q}_{upper}]$. Este enfoque es ampliamente aplicable y obtiene intervalos adaptativos a la heterocedasticidad de los datos [98]. Sin embargo, no tiene garantías estadísticas de cobertura bajo distribuciones arbitrarias de errores. Es por ello que se requiere de herramientas adicionales para garantizar la cobertura.

4.3.3. Métodos de predicción conformal para regresión

Todos los métodos propuestos en este trabajo son *split calibration*, es decir, los datos de entrenamiento se dividen en dos subconjuntos: entrenamiento y calibración. No hemos implementado métodos *cross-calibration* como [136] dado que requieren un mayor coste computacional. Además, en los experimentos preliminares, *split calibration* demostró ser suficiente para

obtener valores razonablemente buenos de cobertura marginal y una eficiencia adecuada en los intervalos de predicción.

Inductive Conformal Prediction (ICP)

La ICP [97] fue el primer método de predicción conformal desarrollada para problemas de regresión. Su planteamiento es muy simple: consiste en añadir un margen a las predicciones puntuales, calculado a partir de un cuantil del error absoluto observado en un conjunto de calibración independiente. Este margen permite construir intervalos de predicción que contienen el valor real con una probabilidad determinada previamente (por ejemplo, 90 % o 95 %).

Por ello, la función de no conformidad es el error absoluto de la predicción respecto al valor real. En el proceso de calibración, se calculan los errores absolutos para cada ejemplo en el conjunto de calibración:

$$R = \left\{ |y_i - \hat{f}(x_i)| \right\}_{i=1,\dots,n}$$

Luego, el umbral de no conformidad para un nivel de confianza $1 - \alpha$ se calcula como el cuantil $(1 - \alpha)(1 + 1/n)$ de R :

$$\delta_\alpha = \text{Quantile}_{\lceil(1-\alpha)(1+1/n)\rceil}(R)$$

Finalmente, para una instancia x_{n+1} , el intervalo de predicción $C(x_{n+1})$ se construye como:

$$\hat{C}_\alpha(x_{n+1}) = [\hat{f}(x_{n+1}) - \delta_\alpha, \hat{f}(x_{n+1}) + \delta_\alpha]$$

Este método de CP presenta varias ventajas:

- **Model-agnostic y domain-agnostic:** Es independiente tanto del modelo como del dominio, ya que no utiliza representaciones internas del modelo ni de las entradas.
- **Bajo coste computacional:** Solo añade coste computacional en la calibración, con el cálculo de puntuaciones de no conformidad en calibración ($\mathcal{O}(n_{calib})$) y cálculo del umbral de no conformidad ($\mathcal{O}(n_{calib} \log n_{calib})$). La inferencia conformal mantiene el mismo orden de coste que el modelo base ($\mathcal{O}(1)$ por predicción).

Sin embargo, también presenta importantes limitaciones:

- **Intervalo simétrico y no adaptativo:** El intervalo es simétrico, además de tener siempre el mismo ancho ($2q_{1-\alpha}$), no permitiendo adaptarse a la incertidumbre específica de cada predicción.
- **Sensibilidad a datos ruidosos o OOD:** Si el conjunto de calibración contiene *outliers* o viola el supuesto de intercambiabilidad, el umbral $q_{1-\alpha}$ puede inflarse, generando intervalos excesivamente conservadores. Tampoco detecta heterocedasticidad automáticamente.

Conformalized Quantile Regression (CQR)

Como su nombre indica, este método se realiza sobre la regresión cuantílica. La CQR [98] combina la flexibilidad de la regresión cuantílica para estimar directamente los cuantiles condicionales con la garantía de validez estadística proporcionada por la conformalización. Esto permite obtener intervalos de predicción que son asimétricos y adaptativos, ajustándose localmente a la variabilidad y distribución de los datos.

Se ha optado por implementar la segunda definición del intervalo de predicción, presentada en el segundo teorema de [98], que incluye la calibración de ambas colas para obtener intervalos asimétricos [137]. Según el artículo, esta opción mejora las garantías de cobertura, aunque puede implicar un aumento en el ancho del intervalo.

El proceso de calibración de este método se lleva a cabo de la siguiente manera:

- Se calculan dos arrays de puntuaciones de no conformidad sobre los datos del conjunto de calibración como las diferencias entre los valores observados y los límites del intervalo predictivo:

$$\begin{aligned} R_{lower} &= \{\hat{q}_{lower}(x_i) - y_i\}_{i=1,\dots,n} \\ R_{upper} &= \{y_i - \hat{q}_{upper}(x_i)\}_{i=1,\dots,n} \end{aligned}$$

donde $\hat{f}_{upper}(x_i)$ y $\hat{f}_{lower}(x_i)$ representan los límites superior e inferior del intervalo predictivo para la observación x_i , respectivamente, e y_i es el valor observado real.

- Se calcula un umbral de no conformidad para un nivel de confianza dado $1 - \alpha$ como el cuantil $(1 - \alpha)(1 + 1/n)$ de R :

$$\begin{aligned} \delta_{lower\alpha} &= Quantile_{\lceil(1-\alpha)(1+1/n)\rceil}(R_{lower}) \\ \delta_{upper\alpha} &= Quantile_{\lceil(1-\alpha)(1+1/n)\rceil}(R_{upper}) \end{aligned}$$

Tras haber calibrado el modelo, para una instancia x_{n+1} , el intervalo de predicción $C(x_{n+1})$ se construye como:

$$\hat{C}_\alpha(x_{n+1}) = [\hat{q}_{lower}(x_{n+1}) - \delta_{lower_\alpha}, \hat{q}_{upper}(x_{n+1}) + \delta_{upper_\alpha}]$$

CQR, al igual que ICP, es independiente del modelo y del dominio, ya que solo emplea las salidas y valores reales para realizar la calibración. También tiene el mismo orden de eficiencia computacional, puesto que realiza prácticamente las mismas operaciones que ICP, pero para cada límite del intervalo predicho, calibrando los cuantiles inferior y superior de manera independiente para mantener la cobertura deseada.

Sin embargo, CQR logra intervalos asimétricos y adaptativos, dado que la regresión cuantílica estima directamente los cuantiles condicionales de la distribución de la variable objetivo, permitiendo que los límites del intervalo se ajusten según la heterocedasticidad y la forma local de la distribución de los datos, en lugar de asumir una distribución simétrica o constante del error.

En la Tabla 4.2 observamos un cuadro comparativo de los distintos métodos propuestos de CP.

4.3.4. Calibración de probabilidades en clasificación

4.3.5. Métodos de predicción conformal para clasificación

Least-Ambiguous set-valued Classifiers (LAC)

Hacer este apartado. No debería ser muy largo. Presentar solo el método de Platt Scaling.

LAC [100] es el primer método propuesto de predicción conformal para problemas de clasificación. Propone un enfoque de clasificación de conjuntos de valores (*set-valued classification*) en el que, en lugar de asignar una única etiqueta a cada instancia, se selecciona un conjunto de etiquetas que garanticen un nivel de confianza predeterminada por el usuario.

La función de no conformidad es conocida como **probabilidad inversa** o **hinge loss** [138], y se calcula como la unidad menos la probabilidad de la clase verdadera³ o, lo que es lo mismo, la suma de valores de probabilidad de todas las clases salvo la correspondiente a la etiqueta verdadera:

$$R = \{1 - \hat{\pi}_{y_i}(x_i)\}_{i=1,\dots,n}$$

³Se le denomina probabilidad a un valor de certeza que realmente no tiene garantías estadísticas, ya que proviene directamente de la salida *softmax* o sigmoide del modelo. Estas salidas no están necesariamente bien calibradas ni corresponden a verdaderas probabilidades, si bien el término se utiliza frecuentemente por motivos de simplicidad y comunicación.

Característica	base	ICP	QR	CQR
Cobertura Marginal	No garantizada	Garantizada	No garantizada	Garantizada
Cobertura Condicionada	No garantizada	No garantizada	No garantizada	No garantizada, pero approxima
Model-agnostic	Sí	Sí	Sí	Sí
Domain-agnostic	Sí	Sí	Sí	Sí
Intervalos simétricos/asimétricos	Simétricos	Simétricos	Asimétricos	Asimétricos
Intervalos adaptativos	No	No	Sí	Sí
Coste calibración	No existe calibración	$O(n \log(n))$	No existe calibración	$O(n \log(n))$
Coste inferencia (por predicción)	$O(1)$	$O(1)$	$O(1)$	$O(1)$

Tabla 4.2: Comparativa de métodos propuestos de CP para problemas de regresión.

donde $\hat{\pi}_{y_i}(x_i)$ es la probabilidad para la clase de la etiqueta verdadera⁴.

El umbral de no conformidad para un nivel de confianza $1 - \alpha$ se calcula como el cuantil $(1 - \alpha)(1 + 1/n)$ de R :

$$\delta_\alpha = Quantile_{\lceil(1-\alpha)(1+1/n)\rceil}(R)$$

El conjunto de predicción conformal de una nueva instancia x_{n+1} se construye como las clases cuyas probabilidades superan la unidad menos el umbral de no conformidad:

$$\Gamma_\alpha(x_{n+1}) = \{k | \hat{\pi}_k(x_{n+1}) \geq 1 - \delta_\alpha\}$$

Así, se seleccionan aquellas clases cuya probabilidad es lo suficientemente alta como para superar el umbral de no conformidad previamente calculado. No obstante, puede ocurrir que, para ciertas instancias, ninguna clase alcance dicho umbral, lo que resultaría en un conjunto de predicción vacío. Para evitar esta situación, se ha optado por incluir en estos casos todas las clases posibles dentro del conjunto de predicción. Esta elección responde a una estrategia conservadora: ante la falta de evidencia suficiente para respaldar alguna clase en particular con el nivel de confianza requerido, lo más prudente es no excluir ninguna posibilidad, y así reflejar una alta incertidumbre.

Algunas propiedades de este método son:

- **Model agnostic:** Es independiente del modelo, ya que solo necesita el vector de puntuaciones predictivas $\hat{\pi}(x_i)$ y la etiqueta verdadera para cada instancia y_i .
- **Conjuntos de predicción no adaptativos:** A pesar de poder presentar conjuntos con distinto número de clases predichas, emplea un único umbral calibrado globalmente sobre todas las muestras y clases por igual.
- **Bajo coste computacional:** Solo añade coste computacional en la calibración, con el cálculo de puntuaciones de no conformidad ($\mathcal{O}(n_{calib})$) y la obtención del umbral de no conformidad ($\mathcal{O}(n_{calib})log n_{calib}$). No añade coste a la inferencia ($\mathcal{O}(1)$).

⁴ $\hat{\pi}(x_i)$ es el vector de probabilidades de las clases para la instancia i .

Falta añadir una imagen que refleje la intuición detrás de esta técnica

Mondrian Confidence Machine (MCM)

(MCM) [139] es un método estrechamente relacionado con LAC, ya que emplea el mismo esquema general de CP. Sin embargo, introduce una diferencia clave: en lugar de aplicar un único umbral global para todas las clases, MCM segmenta el conjunto de calibración por clase y calcula las puntuaciones de no conformidad y los umbrales de decisión de forma independiente para cada una.

A continuación, se detallan sus principales características diferenciadas de LAC:

- **Garantiza cobertura condicional por clase**, lo cual es muy útil en conjuntos desbalanceados. A diferencia de LAC, que ofrece cobertura marginal sobre el conjunto total, MCM busca asegurar que cada clase individual cumpla el nivel de cobertura deseado, lo que favorece una distribución más equitativa del error.
- **Conjuntos de predicción parcialmente adaptativos**: Estos son adaptativos respecto a cada clase, aunque no por muestra, ya que emplea un umbral de no conformidad por cada clase, pero los aplica igual a todas las muestras.
- **Coste computacional ligeramente superior a LAC**: En la calibración, se requiere calcular las puntuaciones de no conformidad y el umbral de no conformidad para cada clase, lo cual puede aumentar los tiempos linealmente en base al número de clases. La inferencia sigue manteniendo la eficiencia. No obstante, sigue siendo un método eficiente y apto para entornos de predicción en tiempo real, siempre que el número de clases no sea excesivo.

Adaptive Prediction Sets (APS)

APS [101], como sugiere su nombre, tiene como objetivo generar conjuntos de predicción adaptativos, cuyo tamaño se ajusta dinámicamente en función de la incertidumbre del modelo para cada muestra. De este modo, se busca que las predicciones sean más informativas y reflejen con mayor precisión la confianza del modelo.

La función de no conformidad utilizada en APS evalúa, para cada instancia, la probabilidad total acumulada en aquellas clases que el modelo considera al menos tan probables como la clase verdadera. En otras palabras, se calcula como la suma de las probabilidades predichas para todas las clases cuya probabilidad es mayor o igual a la asignada a la etiqueta correcta.

Sea el vector $\hat{\pi}$ ordenado en orden decreciente:

$$\hat{\pi}_{(1)}(x_i) \geq \hat{\pi}_{(2)}(x_i) \geq \dots \geq \hat{\pi}_{(K)}(x)$$

donde (k) es el índice de la clase con la k mayor probabilidad.

Entonces, el conjunto de puntuaciones de no conformidad se define como:

$$R = \left\{ \sum_{j=1}^k \hat{\pi}_{(j)}(x_i) \text{ donde } (k) = y_i \right\}_{i=1,\dots,n}$$

Cabe destacar que, en el caso particular de clasificación binaria, esta medida de no conformidad coincide exactamente con la utilizada en el método LAC, ya que la acumulación se limita a una o dos clases. Por tanto, ambos métodos resultan equivalentes en este escenario. Sin embargo, divergen en problemas multiclase, donde las puntuaciones de no conformidad de APS son más permisivas que las de LAC, ya que reconocen que un modelo puede identificar características comunes entre varias clases y generar valores probabilísticos repartidos. No existe incertidumbre cuando la puntuación probabilística más alta corresponde a la clase verdadera. Por tanto, APS penaliza menos los casos en que la clase correcta está entre las más probables, aunque no necesariamente en primer lugar.

A partir de las puntuaciones de no conformidad en el conjunto de calibración, se calcula el umbral de no conformidad de la manera habitual:

$$\delta_\alpha = Quantile_{\lceil(1-\alpha)(1+1/n)\rceil}(R)$$

Tras la calibración, para una nueva instancia x_{n+1} , se calcula la distribución de probabilidad ordenada en orden decreciente, y se suman de forma acumulada las probabilidades desde la clase más probable hasta que dicha suma sea mayor o igual que el umbral calibrado. El conjunto de predicción $\Gamma_\alpha(x_{n+1})$ se forma entonces incluye todas las clases correspondientes a ese conjunto acumulado:

$$\Gamma_\alpha(x_{n+1}) = \{(1), \dots, (k)\} \text{ donde } k = \min \left\{ j : \sum_{i=1}^j \hat{\pi}_{(i)}(x_{n+1}) \geq \delta_\alpha \right\}$$

Este algoritmo, al igual que LAC, solo garantiza cobertura marginal, pero genera **conjuntos de predicción más adaptativos** respecto a la incertidumbre inherente a la predicción de cada instancia. A diferencia de métodos con umbrales fijos, ajusta dinámicamente el tamaño de los conjuntos según la confianza del modelo en regiones específicas del espacio de características.

Sin embargo, en la práctica se ha observado que esta adaptabilidad conlleva **conjuntos de predicción más grandes en promedio** [101, 102]. Este fenómeno es un *trade-off* inherente al intentar **aproximar la cobertura condicional** sin asumir distribuciones subyacente, que analizaremos en profundidad con nuestros datos en la experimentación.

Regularized Adaptive Prediction Sets (RAPS)

RAPS [102] es una variante del método APS, que introduce modificaciones clave para mejorar su eficiencia práctica, como es la penalización a conjuntos de predicción demasiado grandes, además de un factor aleatorio de ajuste.

Para ello, se introducen dos parámetros en la función de no conformidad:

- k_{reg} , que es el tamaño óptimo del conjunto de predicción (en el sentido de que, si todas los conjuntos de predicción tuvieran ese tamaño, se alcanzaría la cobertura deseada)
- λ , un parámetro de regularización que penalizará más a aquellos conjuntos que superen k_{reg} etiquetas predichas cuanto mayor valor tenga.

...

La calibración de este método se realiza de la siguiente manera:

$$R = \left\{ \sum_{j=1}^k \hat{\pi}_{(j)}(x_i) + \lambda(k - k_{reg})^+ \text{ donde } (k) = y_i \right\}_{i=1,\dots,n}$$

Por completar (JU-LIO)

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*¹. Al proporcionar las particiones predefinidas, se garantiza que no haya contaminación entre los datos de entrenamiento y test, manteniendo así la validez de las métricas obtenidas en el test.

Sin embargo, si se optimizan los parámetros del modelo durante el entrenamiento sin disponer de un conjunto independiente para evaluar su rendimiento, se corre el riesgo de sobreajustarse a los datos de entrenamiento. Es por ello que, además del conjunto de entrenamiento y test, es esencial tener un **conjunto de validación** independiente que permita evaluar el modelo durante su desarrollo, ajustar hiperparámetros y comparar diferentes configuraciones sin contaminar la evaluación final en el conjunto de test. Se consideró realizar validación cruzada (*cross-validation*), pero debido al elevado coste computacional que implica, los resultados satisfactorios obtenidos mediante una simple partición de los datos (*train/validation split*), se decidió prescindir de su aplicación.

En la Figura 5.1 podemos ver la división del *dataset* planteada. Cabe comentar que la división se ha realizado de forma estratificada en base a la edad y el sexo².

Es importante destacar que esta división se mantiene constante en todos los experimentos y para todos los problemas planteados, asegurando que las

¹El *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 no generalizan adecuadamente ante datos nuevos.

²La estratificación se realizó en intervalos de medio año de edad y por sexo; por ejemplo, una instancia con edad 17.7 y sexo masculino se etiquetó como “17.5_M”.

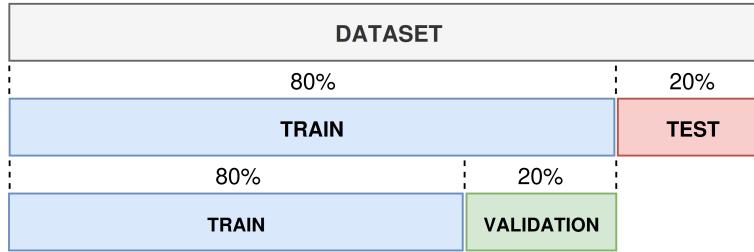


Figura 5.1: Diagrama de división del *dataset* en *train*, *validation* y *test*.

mismas instancias permanezcan en los mismos subconjuntos. Esto permite garantizar que ningún modelo preentrenado reutilice datos previamente utilizados en etapas de validación o calibración, algo especialmente relevante dado que los problemas abordados están jerárquicamente relacionados (la clasificación de sexo y mayoría de edad se deriva directamente de la clasificación de mayoría de edad, que a su vez se deriva de la estimación de edad).

Sin embargo, al emplear métodos de calibración o predicción conformal, si usamos los mismos datos de entrenamiento para la calibración, las probabilidades o intervalos de predicción tenderán a ser optimistas, pues el modelo ha sido entrenado con esos datos [140]. Por tanto, para evitar el sobreajuste y garantizar validez estadística se requiere de un subconjunto de datos adicional: el **conjunto de calibración**. Se ha escogido destinar el 20 % de los ejemplos de entrenamiento para calibración, basándose en los resultados empíricos de [141] (que recomienda dedicar entre un 10 % y 30 % de datos de entrenamiento a calibración), tal y como se muestra en la Figura 5.2.

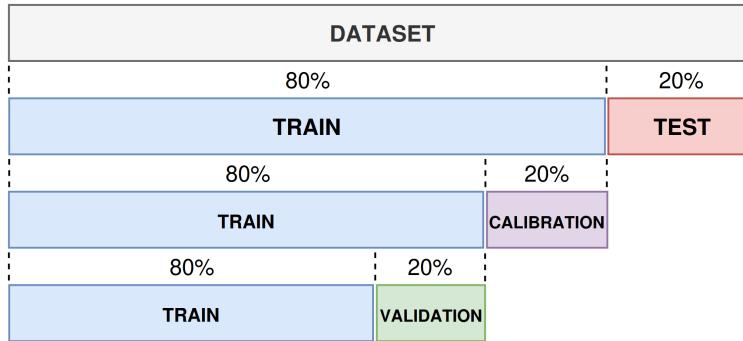


Figura 5.2: Diagrama de división del *dataset* en *train*, *validation*, *calibration* y *test*.

Para una comparativa más justa entre los métodos que usan CP y los que no, se utilizará la siguiente estrategia: los métodos que no emplean CP seguirán el esquema tradicional de división de datos (en entrenamiento, vali-

dación y test), mientras que los métodos basados en CP incorporarán además un conjunto de calibración independiente. Esta diferencia en el diseño experimental nos permitirá cuantificar cómo afecta a la capacidad predictiva de los modelos el hecho de reservar parte de los datos para el proceso de calibración.

5.2. Experimentos propuestos

5.2.1. Comparativa de métodos para la estimación de edad

Se plantea una comparativa entre diversos métodos de predicción para el problema de AE. Todos los métodos presentan tanto predicción puntual como interválica. De esta forma queremos evaluar tanto su utilidad tradicional para estimar el valor esperado como su capacidad para proporcionar intervalos de confianza fiables que capturen la incertidumbre predictiva y sean computacionalmente eficientes. El objetivo es alcanzar el 95 % de confianza en las predicciones interválicas, que es la cifra de confianza generalmente empleada en AF. Los métodos propuestos son los siguientes:

- **Método ‘base’:** Se trata de un modelo de regresión puntual sin técnicas de CP. La predicción interválica se construirá con la predicción puntual ± 2 veces el error absoluto medio obtenido en el conjunto de validación, que es una aproximación heurística común para construir intervalos de predicción que no asumen una distribución de errores específica. Este método sirve como *baseline* para comparar la mejora que aportan las técnicas más sofisticadas.
- **Método ‘ICP’:** Implementa el método ICP, mediante el cual se ...
- **Método ‘QR’:** Este modelo implementa QR. Utiliza tres cuantiles

$$[0.5, \alpha/2, 1 - \alpha/2]$$

para predecir la predicción puntual, límite inferior y límite superior, respectivamente.

- **Método ‘CQR’:** Este modelo implementa

Para cada método se ha entrenado 10 modelos independientes desde cero, con el objetivo de capturar la variabilidad inherente al proceso de entrenamiento. Todas las métricas se calculan sobre el conjunto de test, evaluando tanto predicciones puntuales como interválicas, para garantizar que la evaluación sea objetiva y no esté influenciada por ninguna etapa del entrenamiento o calibración.

5.2.2. Comparativa de métodos para la estimación de mayoría de edad

Todos los métodos propuestos para el problema de AAM presentan una predicción puntual (de una sola etiqueta), además de un conjunto de predicción, formado por una o más etiquetas.

Los métodos propuestos son:

- **Método ‘base’:** Se trata del modelo de clasificación de una sola etiqueta sin uso de técnicas de CP. El conjunto de predicción se considerará aquel formado exclusivamente por la clase más probable. El entrenamiento de este modelo partirá de un modelo ‘base’ ya entrenado para el problema de AE, al cual se realizará un *fine-tuning* de la cabecera. Este método sirve de *baseline* para comparar con el resto.
- **Método ‘LAC’:** Este método implementa la técnica LAC para CP. El entrenamiento del modelo partirá de un modelo ICP ya entrenado para regresión.
- **Método ‘MCM’:** Este método implementa la técnica MCM para CP. El modelo será exactamente el mismo que el de LAC. Solo cambiará la calibración e inferencia conformal.

No se han implementado los otros métodos de clasificación APS y RAPS, puesto que no son aplicables directamente al caso de clasificación binaria.

En este caso, también se han obtenido 10 modelos independientes para cada método, y las métricas se han calculado sobre el conjunto de test.

5.2.3. Comparativas de métodos para la clasificación combinada de mayoría de edad y sexo

Al igual que en el problema de AAM, para el problema de AMSC se ha seguido la misma lógica de evaluación, aplicando tanto predicción puntual como técnicas de CP para obtener conjuntos de predicción.

En este caso, se ha empleado la técnica de calibración de probabilidades *Platt Scaling* para ajustar las salidas del modelo de clasificación multiclase, con el objetivo de mejorar la calidad de las probabilidades utilizadas durante la fase de inferencia conformal. Esta calibración probabilística se realiza antes de aplicar los métodos de CP. Se ha optado por utilizar el conjunto de validación para llevar a cabo dicha calibración de probabilidades, dado que, aunque no es el enfoque más riguroso —ya que lo ideal sería dividir el conjunto de calibración en dos subconjuntos independientes, uno para

No me gusta mucho usar estas siglas en el texto, no sé si debería directamente eliminarlas del trabajo o solo dejarlas para usar en los resultados (para tablas y gráficos, donde no cabe mucho texto)

la calibración de probabilidades y otro para la calibración conformal—esta estrategia mostró buenos resultados en la práctica. Esto se debe a que el conjunto de validación empleado era suficientemente representativo y permitió obtener probabilidades calibradas de manera adecuada. Esta calibración probabilística no afecta a la variabilidad entre modelos con los mismos pesos, dado que el algoritmo es determinista y produce resultados consistentes para un mismo conjunto de datos y parámetros.

Partiendo de esto, los métodos propuestos son:

- **Método ‘base’:** Al igual que en AMM, funciona como un clasificador normal sin métodos de CP, y se usa de *baseline* para comparar con el resto. El entrenamiento de este modelo partirá de un modelo ‘base’ ya entrenado para el problema de AMM.
- **Método ‘LAC’:** Este método implementa la técnica LAC para CP. El entrenamiento de este modelo partirá del modelo ‘LAC’ ya entrenado para el problema de AMM.
- **Método ‘MCM’:** Este método implementa la técnica MCM para CP. El modelo será exactamente el mismo que el de LAC para este mismo problema.
- **Método ‘APS’:** Este método implementa la técnica APS para CP. El modelo será exactamente el mismo que el de LAC para este mismo problema.
- **Método ‘RAPS’:** Este método implementa la técnica RAPS para CP. El modelo será exactamente el mismo que el de LAC para este mismo problema.

Como en los anteriores problemas, se han obtenido 10 modelos independientes para cada método, a partir de los métodos propuestos para el problema de AMM como se ha especificado anteriormente, y las métricas se han calculado sobre el conjunto de test.

5.3. Entrenamiento de los modelos

5.3.1. Preparación de los datos de entrenamiento

Dado que las imágenes del conjunto de datos disponible son significativamente más anchas que altas, se normalizaron todas las dimensiones a

448×224 píxeles para homogenizar las entradas del modelo³. Se ha establecido un tamaño de *batch* de 32, tras encontrar preliminarmente un equilibrio entre regularización y buen ritmo de aprendizaje. Y también se ha realizado *data augmentation* en el conjunto de entrenamiento, introduciendo transformaciones aleatorias en cada época para simular condiciones de posicionamiento del paciente y de la máquina e iluminación ligeramente variables:

- volteo horizontal en la mitad de las imágenes,
- rotación entre -3 y 3 grados,
- traslaciones de hasta el 2 %,
- escalado entre el 95 y 105 %, y
- cambios de brillo y contraste entre 80 y 120 %.

5.3.2. Adaptación de la red para la estimación de edad

Como se venía anticipando en el anterior capítulo, adaptaremos la arquitectura del modelo ResNeXt50 para el problema de regresión. El tamaño de las imágenes de entrada no modifica la arquitectura del modelo, pues el extracto de características conserva la dimensionalidad relativa a través de sus bloques convolucionales. Sustituiremos la última capa del modelo por un *adaptive average pooling*, que permite reducir la dimensionalidad espacial de forma flexible independientemente del tamaño exacto de entrada. A continuación, este tensor de características se aplana en la capa *flatten*.

La salida aplanada pasa por dos bloques densos consecutivos, cada uno compuesto por una capa *batch normalization*, una capa de *dropout* y una capa completamente conectada (FC), con una activación ReLU entre ambos bloques. La primera capa FC contiene 4.096 neuronas, la segunda 512, y finalmente se incluye una capa de salida de una sola neurona. Esta configuración ha sido seleccionada siguiendo la recomendación de los tutores, quienes cuentan con experiencia previa en el trabajo con este conjunto de datos.

Añadir un dibujo con el cambio de cabecera (AGOSTO)

Los componentes clave del *pipeline* de entrenamiento son:

- Error cuadrático medio como función de pérdida en modelos de predicción puntual y *pinball loss* para modelos QR.

³El redimensionado se aplicó de forma consistente a todo el conjunto (entrenamiento, validación, calibración y test), utilizando interpolación bilineal.

El error cuadrático medio es la función de pérdida por defecto para problemas de regresión: los errores siguen una distribución normal, lo que hace que minimizar el MSE equivalga a maximizar la verosimilitud de los datos; penaliza los errores grandes más que los pequeños, lo que ayuda a evitar predicciones extremadamente alejadas de los valores reales; y es derivable en todo su dominio, —además de que su derivada es lineal, lo que facilita el cálculo en la retropropagación— y convexa, lo que garantiza la existencia de un único mínimo global, facilitando la convergencia en problemas lineales.

- Optimizador AdamW [142]. Se ha escogido este optimizador dado que, por lo general, no requiere un ajuste exhaustivo de hiperparámetros para lograr buenos resultados.

Para el entrenamiento de la nueva cabecera, se han congelado todas las capas de la arquitectura salvo las nuevas capas densas, de las cuales se han entrenado los pesos con *learning rate* de 3e-2 y *weight decay* 2e-4 durante dos épocas.

Tras esto, se ha entrenado la red completa. Para ello, se han descongelado todas las capas y se ha aplicado una estrategia de optimización basada en **learning rates discriminativos** combinada con la política de ajuste de *learning rate OneCycle* [143].

En concreto, se han definido diferentes tasas de aprendizaje para cada grupo de capas del modelo, asignadas según su profundidad. Los bloques convolucionales iniciales —más generales y preentrenados— reciben *learning rates* más bajos, mientras que las capas más profundas —específicas de la tarea y recientemente añadidas— se entrena con tasas más altas. Esta asignación se ha realizado mediante una progresión exponencial, que varía desde 1.5e-4 en los bloques más profundos hasta 1.5e-2 en los más superficiales. Este enfoque busca preservar el conocimiento útil de las capas inferiores y permitir una adaptación más rápida en las superiores.

La política OneCycle se ha aplicado individualmente a cada grupo de capas, haciendo que cada uno siga un ciclo de una sola fase: el *learning rate* comienza en un valor inicial bajo, aumenta progresivamente durante las primeras épocas (*warm-up*), y desciende de forma suave hasta un valor final aún menor⁴. Esta estrategia permite acelerar la convergencia en las fases iniciales del entrenamiento y afinar los pesos En las etapas finales, mejorando tanto la estabilidad como el rendimiento del modelo.

⁴Se han mantenido los parámetros por defecto del método OneCycle en PyTorch. Con esta configuración, cada grupo de capas comienza con una tasa de aprendizaje equivalente al 4 % del valor máximo asignado. Durante aproximadamente el 30 % inicial de las épocas, esta tasa crece de forma progresiva, y posteriormente decrece hasta alcanzar el 0,01 % del learning rate máximo.

Esta combinación entre *learning rates* discriminativos y la política de un solo ciclo permite acelerar la convergencia en las primeras etapas del entrenamiento, al tiempo que se mejora la capacidad de generalización mediante un afinado progresivo de los pesos en las fases finales.

El entrenamiento se ha llevado a cabo durante un total de 30 épocas. Para mitigar el riesgo de sobreajuste, se ha implementado una estrategia de *checkpointing*, guardando los pesos del modelo correspondientes a la época en la que se obtuvo la mejor puntuación en el conjunto de validación (menor pérdida). Al finalizar el entrenamiento, se restauran estos pesos, asegurando así que se conserve la versión del modelo con mayor capacidad de generalización.

Tengo que hablar aquí de la adaptación de esta arquitectura y modelo para la Quantile Regression? Ya la expliqué en el capítulo 4, pero no sé si debería ir más bien aquí. Siento que si lo pongo aquí La información estará más ordenada, pero costará más entender la Quantile Regression.

5.3.3. Adaptación de la red para la estimación de mayoría de edad

Dado que la tarea de estimación de mayoría de edad guarda una estrecha relación con la estimación de edad continua, se ha optado por reutilizar el extracto de características previamente entrenado para esta última. Al tratarse de una clasificación binaria cuya frontera de decisión es el umbral de los 18 años, se considera que las representaciones latentes aprendidas por el modelo son igualmente útiles para resolver esta nueva tarea.

En consecuencia, únicamente se ha ajustado la cabecera del modelo, manteniendo congelados los pesos del extracto de características. Se ha empleado el mismo optimizador AdamW que en la tarea de regresión y se ha seguido el mismo procedimiento de entrenamiento descrito para la cabecera: dos épocas con un *learning rate* de 3e-2 y un *weight decay* de 2e-4.

La función de pérdida utilizada en este caso ha sido la ***Binary Cross-Entropy Loss***, adecuada para tareas de clasificación binaria. Esta función combina de forma eficiente una activación sigmoide y la entropía cruzada, lo que permite interpretar la salida del modelo como una probabilidad. Su formulación penaliza de forma asimétrica las predicciones incorrectas, lo que resulta especialmente útil cuando se requiere una buena calibración de las probabilidades de salida.

5.3.4. Adaptación de la red para la clasificación combinada de mayoría de edad y sexo

La clasificación combinada de mayoría de edad y sexo introduce una segunda variable objetivo. Por ello, se ha partido de un modelo preentrenado para la clasificación de mayoría de edad, y se ha procedido a entrenar tanto la cabecera como el conjunto completo de la red.

La última capa del modelo ha sido ajustada para producir cuatro salidas, correspondientes a las clases del problema. La activación *softmax* se aplica durante la inferencia para obtener probabilidades normalizadas.

A diferencia del caso anterior, aquí se ha entrenado tanto la cabecera como la red completa. En la primera fase, se ha entrenado únicamente la cabecera durante dos épocas con los mismos hiperparámetros que en los casos anteriores. Posteriormente, se ha llevado a cabo un *fine-tuning* o de toda la red, aplicando de nuevo la estrategia de *learning rates* discriminativos junto con la política OneCycle, pero reduciendo a la mitad el número de épocas (15) al observarse una convergencia más rápida. Se ha mantenido el uso del optimizador AdamW en todo el proceso.

La función de pérdida utilizada ha sido la *Cross-Entropy Loss*, adecuada para clasificación multiclas mutuamente excluyente. Esta función compara la distribución de probabilidad predicha por el modelo con la distribución real codificada como etiqueta única, y penaliza fuertemente las asignaciones erróneas. Su formulación es robusta, ampliamente utilizada y permite una interpretación probabilística directa de la salida del modelo cuando se combina con una capa de activación *softmax* al final.

5.4. Métricas usadas en los experimentos

5.4.1. Métricas para regresión

En nuestro problema de regresión emplearemos dos tipos de métricas con el objetivo de evaluar aspectos distintos del desempeño del modelo.

Por una parte, las métricas destinadas a las predicciones puntuales se basan fundamentalmente en medir el error entre el valor real (y_i) y el predicho (\hat{y}_i). Estas métricas nos permiten cuantificar directamente la discrepancia entre las estimaciones del modelo (estimación central en modelos de predicción interválica) y la *ground truth*. Las métricas que empleamos para estas predicciones 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.

Javier dijo que le cuadraba más que este apartado etuviera en materiales y métodos. Hay que discutirlo.

También podría reformular este apartado y llamarlo ‘Evaluación de los experimentos’, e incluir tanto métricas como las tests estadísticos que empleo en experimentación.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \in [0, \infty)$$

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.

- 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.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \in [0, \infty)$$

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.

Por otra parte, las métricas aplicadas a las predicciones interválicas examinan tanto la capacidad del modelo para abarcar el valor real dentro del intervalo predicho —conocida como **cobertura** (*coverage*)— como la **amplitud** del mismo, que es el ancho del rango de valores del intervalo de predicción. Generalmente, existe un equilibrio entre ambos aspectos: al aumentar la amplitud, es más probable que el intervalo contenga el valor real, pero esto disminuye la precisión y utilidad práctica de la predicción. Veamos las métricas para este tipo de predicciones:

- La **cobertura empírica** (*empirical coverage*) cuantifica la proporción de valores reales dentro de los intervalos de predicción obtenidos.

$$EC = \frac{1}{n} \sum_{i=1}^n \mathbb{I}[l_i \leq y_i \leq u_i] \in [0, 1]$$

donde l_i y u_i son los límites inferior y superior, respectivamente, de los intervalos de predicción obtenidos mediante inferencia conformal.

Cuanto mayor sea el valor, mejor cobertura ofrece el modelo, si bien coberturas altas suelen conllevar intervalos excesivamente amplios, lo que reduce su utilidad práctica. Es por ello que, empleando métodos de CP, tiene más sentido que el objetivo sea acercarse lo máximo posible a la cobertura marginal nominal ($1 - \alpha$), garantizando así intervalos de predicción que equilibren precisión y fiabilidad sin ser innecesariamente conservadores.

- El **tamaño medio de intervalo de predicción** (*mean prediction interval width*) mide qué tan amplios son en promedio los intervalos predichos.

$$MPIW = \frac{1}{n} \sum_{i=1}^n (u_i - l_i) \in (0, +\infty)$$

Se desea matener este valor lo más pequeño posible, dado un nivel de cobertura adecuado. Valores altos indican intervalos anchos y, por tanto, poco útiles para la toma de decisiones.

- La ***mean interval score*** [144] trata de unificar en una sola métrica el *trade-off* cobertura vs. amplitud del intervalo. Su expresión es la siguiente:

$$\begin{aligned} MIS = \frac{1}{n} \sum_{i=1}^n & \left((u_i - l_i) + \frac{2}{\alpha} (l_i - y_i) \mathbb{I}[y_i < l_i] \right. \\ & \left. + \frac{2}{\alpha} (y_i - u_i) \mathbb{I}[y_i > u_i] \right) \in (0, +\infty) \end{aligned}$$

Al igual que con el *mean interval width*, una puntuación más baja en el *mean interval score* indica un mejor rendimiento del modelo. El primer término $(u_i - l_i)$ representa directamente la amplitud de cada intervalo, mientras que el segundo y tercer términos:

- $\frac{2}{\alpha} (l_i - y_i) \mathbb{I}[y_i < l_i]$ penaliza los casos en que el valor verdadero y_i está por debajo del límite inferior l_i , proporcionalmente a la distancia $(l_i - y_i)$.
- $\frac{2}{\alpha} (y_i - u_i) \mathbb{I}[y_i > u_i]$ penaliza los casos en que el valor verdadero y_i está por encima del límite superior u_i , proporcionalmente a la distancia $(y_i - u_i)$.

Estos dos últimos términos aplican una penalización crecientemente severa cuando las predicciones no cubren el valor verdadero —y lo hacen multiplicando por $2/\alpha$, lo que enfatiza aún más los errores externos a medida que disminuye α , es decir, cuando se busca mayor confianza.

Y, finalmente, también añadiremos elementos visuales para valorar el desempeño de la CP:

- **Gráfica de dispersión *Empirical Coverage-Mean Prediction Interval Width***: Este gráfico permite visualizar el compromiso entre cobertura lograda y tamaño del intervalo. Un buen modelo debería situarse cerca del nivel de confianza objetivo con intervalos lo más cortos posible.

- **Gráficas de densidad de tamaños de intervalos:** Esto nos permitirá analizar la distribución de las longitudes de los intervalos predichos. Una concentración alrededor de valores bajos indica intervalos más informativos, mientras que una distribución amplia o con colas largas puede revelar incertidumbre elevada en ciertos casos. Esta visualización nos será útil para aquellas técnicas que ofrecen intervalos predictivos adaptativos.

5.4.2. Métricas para clasificación

Como con la regresión, diferenciaremos entre las métricas de clasificación de etiqueta única y las de múltiples etiquetas para valorar los conjuntos de predicciones obtenidos con las técnicas de CP.

- La **matriz de confusión** es una herramienta fundamental que permite visualizar el rendimiento de modelos de clasificación, tanto binarios como multiclase. 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 5.3). Idealmente, los valores se concentrarían en la diagonal principal, lo que indicaría que las predicciones coinciden con los valores reales. Prácticamente todas las métricas y visualizaciones parten de la información ofrecida en esta matriz.
- La **cobertura empírica (*empirical coverage*)**, de forma análoga a la regresión, mide la proporción de veces que la etiqueta verdadera está contenida dentro del conjunto predicho.

$$EC = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(y_i \in \Gamma_\alpha(x_i))$$

- El **tamaño medio del conjunto de predicción (*mean prediction set size*)** mide qué cuántas etiquetas, en promedio, incluyen los conjuntos de predicción $\Gamma_\alpha(x)$.

$$MSS = \frac{1}{n} \sum_{i=1}^n |\Gamma_\alpha(x_i)|$$

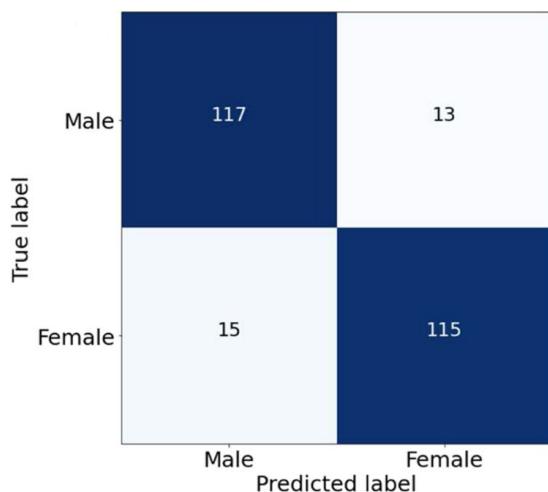


Figura 5.3: Matriz de confusión para la estimación de sexo según el modelo *random forest* propuesto en [145].

5.5. Resultados

5.5.1. Resultados para la estimación de edad

Análisis de métricas para la estimación de edad puntual

La Tabla 5.1 presenta las métricas que evalúan el rendimiento del modelo de regresión en sus estimaciones del valor esperado de edad. En general, se observa **poca variabilidad entre modelos y ejecuciones**, con diferencias de solo unas centésimas en las métricas evaluadas. Sin embargo, el análisis de varianza (ANOVA)⁵ de un factor reveló diferencias estadísticamente significativas entre los modelos para ambas métricas: MAE ($F(3, 28) = 20.38, p < 0.001$) y MSE ($F(3, 28) = 15.09, p < 0.001$). Para identificar qué pares de modelos presentaban diferencias significativas, se aplicó la prueba *post-hoc* de comparaciones múltiples de Tukey (véanse las Tablas 5.2 y 5.3). Los resultados indicaron lo siguiente:

- No se encontraron diferencias significativas entre los modelos ‘QR’ y ‘base’ en ninguna métrica, al igual que tampoco entre los modelos ‘CQR’ e ‘ICP’, lo que sugiere rendimientos similares entre estos pares de modelos. Esto indica que los modelos de regresión cuantílica obtiene resultados equivalentes a los modelos de regresión central.

¿Debería indicar los supuestos a la hora de usar ANOVA y Tukey? Por ejemplo: se asume que cada método presenta distribución normal en las métricas

⁵La aplicación del ANOVA se basó en las suposiciones de independencia entre errores medios por modelo y ejecución, normalidad aproximada de los residuos y homogeneidad de varianzas entre grupos.

Método	Error Absoluto Medio				Error Cuadrático Medio			
	base	ICP	QR	CQR	base	ICP	QR	CQR
Ejecución 1	1.17	1.20	1.17	1.18	2.39	2.50	2.38	2.46
Ejecución 2	1.15	1.18	1.17	1.20	2.33	2.45	2.40	2.49
Ejecución 3	1.17	1.21	1.17	1.17	2.38	2.55	2.42	2.36
Ejecución 4	1.16	1.20	1.14	1.17	2.34	2.47	2.32	2.41
Ejecución 5	1.16	1.21	1.16	1.18	2.37	2.52	2.39	2.42
Ejecución 6	1.15	1.20	1.17	1.17	2.33	2.51	2.40	2.40
Ejecución 7	1.16	1.20	1.18	1.19	2.34	2.48	2.46	2.43
Ejecución 8	1.18	1.20	1.17	1.20	2.39	2.43	2.40	2.47
Ejecución 9								
Ejecución 10								
Media	1.16	1.20	1.16	1.19	2.36	2.49	2.38	2.44

Tabla 5.1: Error absoluto medio y error cuadrático medio obtenidos por cada método de predicción a lo largo de distintas ejecuciones. Se presentan los valores para cada ejecución individual, así como la media final de cada métrica. Punto como separador decimal.

- Los modelos conformales ('ICP' y 'CQR') mostraron errores significativamente mayores ($p < 0.01$) que los modeloss no conformales ('base' y 'QR'). Esto era esperable, pues los métodos conformales tienen menos ejemplos para entrenarse y, por tanto, generalizan peor.

Análisis de métricas para la estimación de edad interválica

A continuación, la Tabla 5.4 presenta las métricas sobre las predicciones interválicas de los métodos. A primera vista, se observan diferencias marcadas entre los métodos conformales y no conformales en las métricas de cobertura empírica y amplitud del intervalo. En particular, los métodos no conformales ('base' y QR) muestran coberturas inferiores al nivel deseado (alrededor del 88-89 % frente al 95 % nominal), lo que indica una infracobertura sistemática. Esto ocurre porque ni la heurística del método 'base' ni las regiones generadas por la regresión cuantílica en QR cuentan con garantías teóricas de cobertura estadística.

En contraste, los métodos conformales (ICP y CQR) sí logran coberturas próximas al valor nominal, tal como se espera dada su fundamentación estadística. Esta mayor cobertura, sin embargo, tiene un costo en cuanto a la amplitud del intervalo, que tiende a ser mayor que en los métodos conforma-

Modelo 1	Modelo 2	Dif. media	Valor <i>p</i>	IC 95 %	Signif.
CQR	ICP	0.0122	0.176	[−0.0036, 0.0279]	No
CQR	QR	-0.0244	0.0013	[−0.0401, −0.0086]	Sí
CQR	base	-0.0250	0.0010	[−0.0407, −0.0092]	Sí
ICP	QR	-0.0365	<0.0001	[−0.0523, −0.0208]	Sí
ICP	base	-0.0371	<0.0001	[−0.0529, −0.0214]	Sí
QR	base	-0.0006	0.9996	[−0.0164, 0.0152]	No

Tabla 5.2: Resultados de la prueba *post-hoc* de Tukey HSD para MAE entre pares de modelos. Se muestran la diferencia media entre grupos, el valor *p* ajustado, el intervalo de confianza al 95 % y si la diferencia es estadísticamente significativa ($\alpha = 0.05$).

Modelo 1	Modelo 2	Dif. media	Valor <i>p</i>	IC 95 %	Signif.
CQR	ICP	0.0484	0.1353	[−0.0104, 0.1072]	No
CQR	QR	-0.0595	0.0464	[−0.1183, −0.0007]	Sí
CQR	base	-0.0825	0.0035	[−0.1413, −0.0237]	Sí
ICP	QR	-0.1079	0.0002	[−0.1667, −0.0491]	Sí
ICP	base	-0.1309	<0.0001	[−0.1896, −0.0721]	Sí
QR	base	-0.0229	0.7128	[−0.0817, 0.0358]	No

Tabla 5.3: Resultados de la prueba *post-hoc* de Tukey HSD para MSE entre pares de modelos. Se muestran la diferencia media entre grupos, el valor *p* ajustado, el intervalo de confianza al 95 % y si la diferencia es estadísticamente significativa ($\alpha = 0.05$).

les. Esta relación de compromiso o *trade-off* entre cobertura y amplitud de los intervalos —típico en la predicción interválica— se visualiza claramente en la Figura 5.4, donde se observa una alta correlación entre la cobertura empírica y el tamaño del intervalo de predicción.

Sin embargo, CQR presenta unas amplitudes promedias de intervalo significativamente más reducidas que ICP, logrando ambos métodos coberturas muy similares. De hecho, en la Tabla 5.5 apreciamos cómo CQR logra significativamente menores valores de *interval score* que ICP, indicando que CQR tiene un mejor equilibrio entre cobertura y tamaño del intervalo.

En consecuencia, CQR se perfila como una opción más ventajosa, con garantías de cobertura e intervalos de predicción ajustados.

Aplicué ANOVA y Tukey aquí, pero no resultaron muy útiles ya que no tienen en cuenta la relación de correlación entre cobertura empírica y tamaño medio del intervalo, y resultaba en que ICP y CQR no mostraban coberturas significativas.

Método	Cobertura Empírica (%)				Amplitud Media del Intervalo			
	base	ICP	QR	CQR	base	ICP	QR	CQR
Ejecución 1	87.41	94.47	89.03	95.31	4.53	6.17	4.71	6.23
Ejecución 2	87.96	94.84	89.27	94.8	4.57	6.27	4.67	6.11
Ejecución 3	87.73	95.03	88.38	95.45	4.60	6.34	4.65	6.02
Ejecución 4	88.06	94.19	89.5	94.61	4.58	6.04	4.63	5.90
Ejecución 5	87.87	95.03	89.13	94.93	4.63	6.28	4.59	5.92
Ejecución 6	88.62	95.91	89.73	95.17	4.71	6.52	4.66	5.98
Ejecución 7	88.24	95.21	88.8	95.26	4.61	6.33	4.63	6.00
Ejecución 8	87.55	94.7	88.01	95.12	4.64	6.12	4.67	6.08
Ejecución 9								
Ejecución 10								
Media	87.93	94.92	88.98	95.08	4.61	6.26	4.65	6.03

Tabla 5.4: Cobertura empírica y amplitud media del intervalo de predicción obtenidos por cada método de predicción a lo largo de distintas ejecuciones. Se presentan los valores para cada ejecución individual, así como la media final de cada métrica. Punto como separador decimal.

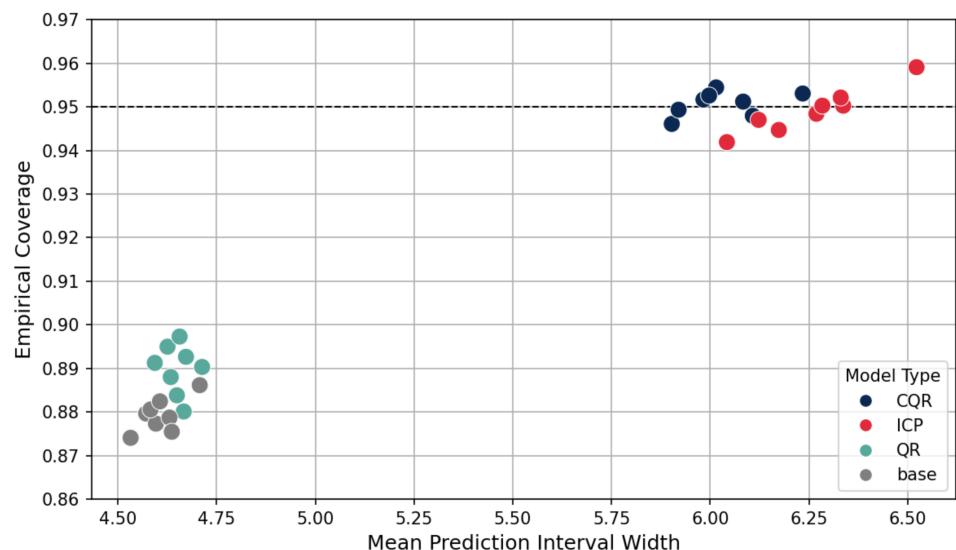


Figura 5.4: Gráfica de dispersión *Empirical Coverage-Mean Precition Interval Width*.

Método	Mean Interval Score			
	base	ICP	QR	CQR
Ejecución 1	9.16	8.17	8.48	8.02
Ejecución 2	8.93	8.21	8.72	8.04
Ejecución 3	8.90	8.24	8.86	7.85
Ejecución 4	8.69	8.00	8.59	7.98
Ejecución 5	8.88	8.27	8.82	7.89
Ejecución 6	8.75	8.23	8.40	8.06
Ejecución 7	8.81	8.19	8.96	7.85
Ejecución 8	8.88	8.03	8.8	7.91
Ejecución 9				
Ejecución 10				
Media	8.88	8.17	8.71	7.95

Tabla 5.5: Resultados de las predicciones obtenidas por los modelos para el problema de estimación de edad en cada ejecución. Punto como separador decimal.

Análisis de la cobertura en base al tamaño del intervalo

En los métodos donde los intervalos de predicción varían en amplitud entre instancias (QR y CQR), resulta relevante analizar cómo se comporta la cobertura empírica en función de dicho tamaño. La hipótesis subyacente es que intervalos más amplios reflejan una mayor incertidumbre asociada a la predicción, mientras que intervalos más estrechos denotan mayor confianza.

Particularmente, se busca determinar si los intervalos más estrechos tienden a infracubrir (es decir, no contienen el valor real con la frecuencia esperada), y si los intervalos más amplios tienden a sobrecubrir (conteniendo el valor real más allá del nivel objetivo de confianza).

En la Figura 5.5 se presentan los histogramas de la amplitud de los intervalos de predicción para dos modelos representativos, uno QR y otro CQR. En cada caso, se diferencia visualmente la proporción de instancias cuya predicción cubre el valor real de aquellas en las que no lo hace. Es notable en ambas figuras la presencia de dos grupos principales de instancias: uno más reducido, asociado a intervalos más estrechos, y otro más numeroso, correspondiente a intervalos de mayor amplitud. Respecto a la cobertura, el modelo QR presenta valores inferiores, lo cual es consistente con su cobertura marginal, que ya se encontraba por debajo del 89 %. En cuanto al ratio entre cobertura e incobertura, este parece mantenerse relativamente estable a lo largo de los distintos rangos de amplitud del intervalo. Sin embargo, para un análisis más detallado y específico sobre cómo varía la cobertura en función del tamaño del intervalo, observemos la información desglosada en la Tabla 5.6.

Esto ocurre, pero no sé por qué. Voy a investigar, pero me temo que va a ser difícil hallar la razón, ya que muy probablemente sea fruto del funcionamiento de la red en regresión cuantílica, y al ser un modelo de caja negra, no pueda hacer nada.

En la Tabla 5.6 se ofrece información detallada sobre la cobertura empírica alcanzada por cada método de predicción (en todas sus ejecuciones) en función de diferentes rangos de amplitud del intervalo de predicción. Esta desagregación permite analizar si existe una relación entre el tamaño del intervalo y la capacidad del modelo para cubrir el valor real.

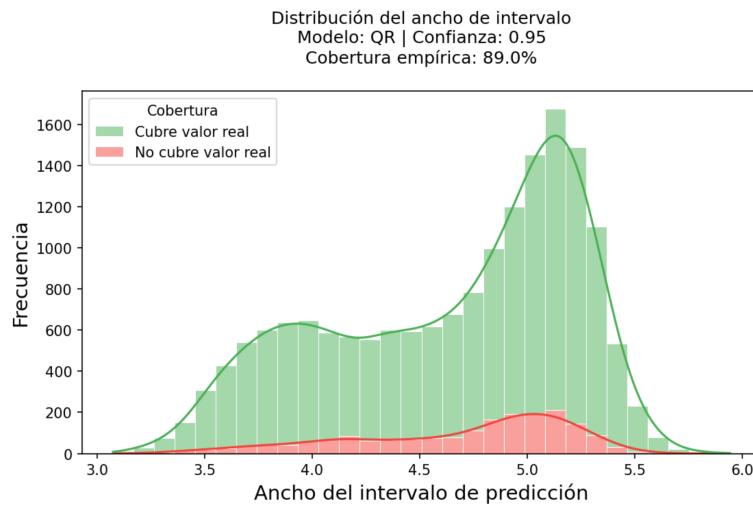
Como era de esperar, los modelos basados en regresión cuantílica (QR y CQR) presentan una mayor diversidad en la amplitud de sus intervalos, dado que generan límites adaptativos y específicos para cada instancia, a diferencia de los métodos conformales de tamaño más constante.

Llama la atención que se logra sobrecobertura tanto en los intervalos más estrechos como en los más amplios, a costa de una infracobertura en los intervalos de amplitud intermedia, concretamente entre 5.5 y 6.5 años, siendo especialmente más bajas en el último medio tramo, donde la cobertura alcanza un 93.4 %.

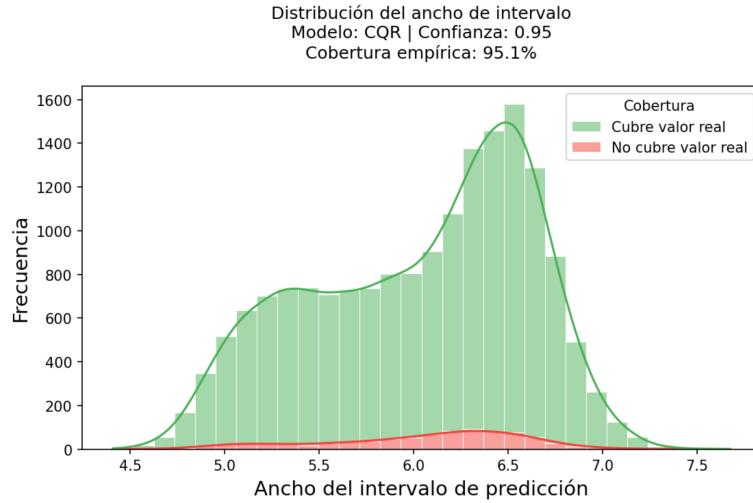
No estoy entrando a hacer valoraciones de lo grave o leve que sea que la cobertura se reduzca de un 95 a un 93.4, porque entiendo que aquí entraría mi subjetividad, y esa debe ir más en las conclusiones que aquí, ¿correcto?

Método	Cobertura Empírica (%)			
	base	ICP	QR	CQR
[4.0, 4.5)	—	—	91.58	100
[4.5, 5.0)	87.93	—	92.85	96.32
[5.0, 5.5)	—	—	88.45	96.47
Amplitud del intervalo	[5.5, 6.0)	—	85.62	94.82
	[6.0, 6.5)	—	94.78	89.43
	[6.5, 7.0)	—	95.91	97.14
	[7.0, 7.5)	—	—	97.37
	[7.5, 8.0)	—	—	100

Tabla 5.6: Cobertura empírica del intervalo de predicción obtenida por cada método de predicción para distintas franjas de amplitud de intervalos. Nota: Los métodos de intervalos de tamaño fijo (como ICP, en este caso) pueden mostrar varias franjas debido a que los tamaños de intervalo pueden variar ligeramente entre entrenamientos para un mismo método.



(a) Histograma de amplitud del intervalo de predicción con diferenciación por cobertura (modelo QR).



(b) Histograma de amplitud del intervalo de predicción con diferenciación por cobertura (modelo CQR).

Figura 5.5: Histogramas del amplitud del intervalo de predicción con diferenciación por cobertura, correspondientes a los modelos QR y CQR. Para cada tipo de método se seleccionó el modelo con el mejor *interval score*. La comparación permite visualizar cómo varía la capacidad de cobertura en función del tamaño del intervalo.

Análisis de la cobertura en base a la edad cronológica

Por último, se ha analizado la cobertura en base a la edad real de los individuos. En la Tabla 5.7 se presentan las métricas interválicas para las instancias de cada edad cronológica⁶. La Figura 5.6 muestra la evolución de la cobertura empírica y el ancho medio de los intervalos de predicción en función de la edad.

Se observa que todos los métodos tienden a reducir su cobertura conforme aumenta la edad cronológica de los individuos. Esta disminución es especialmente notable a partir de los 22 años, afectando incluso al método CQR, que es el método con la cobertura más robusta.

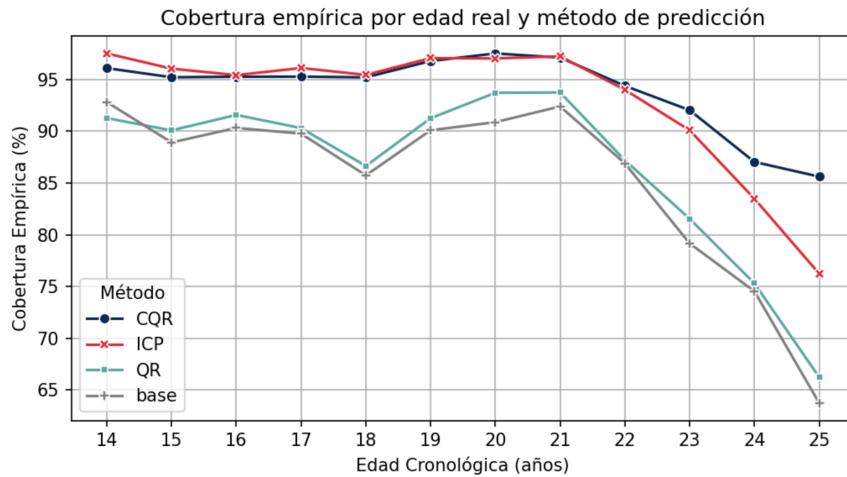
En particular, CQR logra mantener una cobertura cercana al 95 % para individuos de hasta 22 años, pero a partir de los 23 comienza a descender, alcanzando aproximadamente un 85 % en los individuos de 25 años. Este descenso ocurre a pesar de que el tamaño de los intervalos de predicción aumenta de forma sostenida con la edad, lo que indica que, aunque el modelo expresa mayor incertidumbre, no consigue cubrir adecuadamente el valor real. Este patrón refleja que la estimación de la edad biológica se vuelve más incierta conforme avanza la edad cronológica, posiblemente debido a una mayor heterogeneidad fisiológica, ya que este grupo estaba igualmente representado que el resto de edades en el conjunto de entrenamiento.

¿Esto último debería ir en las conclusiones?

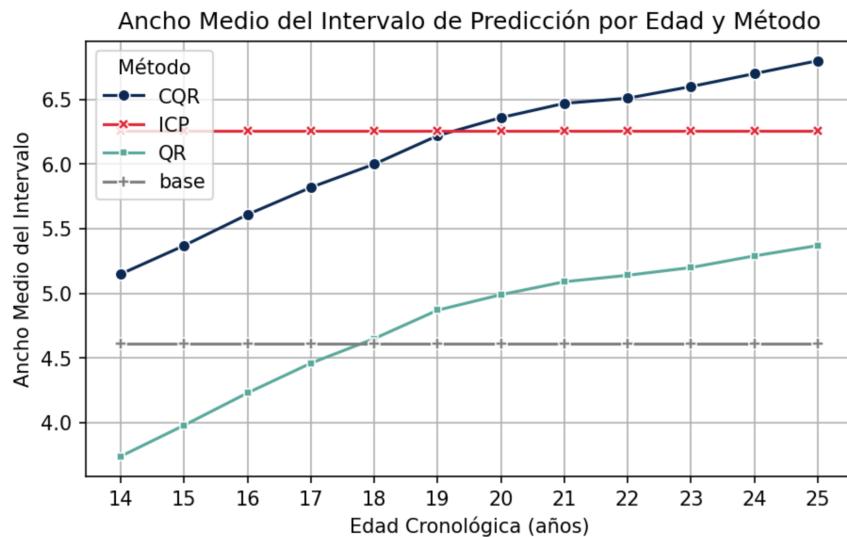
⁶Parte entera (suelo) de la edad real.

Método	Cobertura Empírica (%)				Amplitud Media del Intervalo			
	base	ICP	QR	CQR	base	ICP	QR	CQR
Edad 14	92.82	97.52	91.27	96.1	4.61	6.26	3.74	5.15
Edad 15	88.92	96.05	90.09	95.22	4.61	6.26	3.98	5.37
Edad 16	90.33	95.43	91.58	95.27	4.61	6.26	4.23	5.61
Edad 17	89.77	96.11	90.31	95.28	4.61	6.26	4.46	5.82
Edad 18	85.74	95.45	86.65	95.21	4.61	6.26	4.65	6
Edad 19	90.1	97.07	91.26	96.79	4.61	6.26	4.87	6.22
Edad 20	90.88	97.04	93.72	97.52	4.61	6.26	4.99	6.36
Edad 21	92.4	97.24	93.75	97.12	4.61	6.26	5.09	6.47
Edad 22	86.9	94.01	87.2	94.39	4.61	6.26	5.14	6.51
Edad 23	79.17	90.1	81.55	92.04	4.61	6.26	5.2	6.6
Edad 24	74.54	83.49	75.31	87.04	4.61	6.26	5.29	6.7
Edad 25	63.75	76.25	66.25	85.62	4.61	6.26	5.37	6.8

Tabla 5.7: Cobertura empírica y amplitud media del intervalo de predicción obtenidos por cada método de predicción para distintas edades cronológicas.



(a) Cobertura empírica del intervalo de predicción (%) para cada método en función de la edad cronológica entera de los individuos. Se observa cómo varía la capacidad de cobertura según la edad y el método empleado.



(b) Ancho medio del intervalo de predicción para cada método en función de la edad cronológica entera. Esta gráfica muestra cómo cambia la incertidumbre del modelo con la edad.

Figura 5.6: Análisis comparativo de la cobertura empírica y el ancho medio del intervalo de predicción por edad cronológica para los diferentes métodos evaluados.

5.5.2. Resultados para la estimación de mayoría de edad

Análisis de métricas para la estimación de mayoría de edad puntual

Método	Exactitud		
	(%)		
	base	LAC	MCM
Ejecución 1	87.87	86.99	86.99
Ejecución 2	87.87	87.36	87.36
Ejecución 3	87.59	86.52	86.52
Ejecución 4	87.59	87.5	87.5
Ejecución 5	87.64	87.13	87.13
Ejecución 6	87.27	86.71	86.71
Ejecución 7	88.06	87.13	87.13
Ejecución 8	87.41	86.2	86.2
Ejecución 9			
Ejecución 10			
Media	0.88	0.87	0.87

Tabla 5.8: Exactitud (*accuracy*) obtenida por cada método de predicción a lo largo de las distintas ejecuciones. Se presenta el valor para cada ejecución individual, así como la media final de la métrica. Punto como separador decimal.

Análisis de métricas para la estimación de mayoría de edad en conjunto de predicción

Método	Cobertura Empírica (%)			Tamaño Medio del Conjunto		
	base	LAC	MCM	base	LAC	MCM
Ejecución 1	87.87	94.80	93.91	1	1.20	1.19
Ejecución 2	87.87	95.07	94.38	1	1.20	1.21
Ejecución 3	87.59	95.12	94.24	1	1.23	1.23
Ejecución 4	87.59	93.96	94.42	1	1.19	1.21
Ejecución 5	87.64	94.05	93.54	1	1.18	1.19
Ejecución 6	87.27	93.96	93.63	1	1.18	1.20
Ejecución 7	88.06	94.10	93.87	1	1.19	1.20
Ejecución 8	87.41	94.89	94.84	1	1.21	1.22
Ejecución 9						
Ejecución 10						
Media	87.66	94.49	94.1	1	1.2	1.2

Tabla 5.9: Cobertura empírica y tamaño medio del conjunto de predicción obtenidos por cada método de predicción a lo largo de las distintas ejecuciones. Se presentan los valores para cada ejecución individual, así como la media final de cada métrica. Punto como separador decimal.

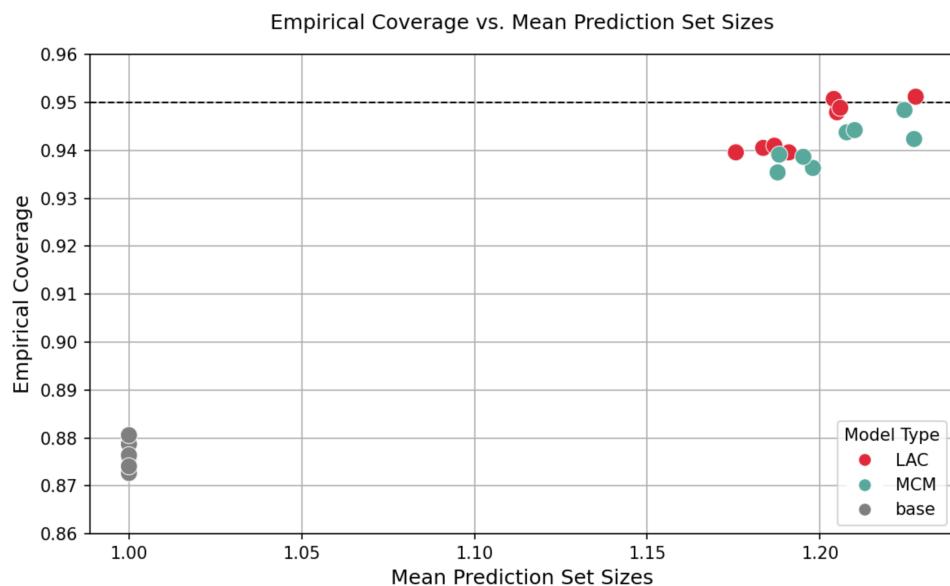


Figura 5.7: Gráfica de dispersión *Empirical Coverage-Mean Prediction Set Size*.

En este apartado se analizará la cobertura mediante matrices de confusión conformal

Análisis de la cobertura en base a la clase

		Conjunto predicho		
		{<18}	{≥18}	{<18,≥18}
Valor real	<18	738	115	0
	≥18	142	1157	0

(a) base

		Conjunto predicho		
		{<18}	{≥18}	{<18,≥18}
Valor real	<18	567	58	228
	≥18	48	1040	211

(b) LAC

		Conjunto predicho		
		{<18}	{≥18}	{<18,≥18}
Valor real	<18	646	28	179
	≥18	83	912	304

(c) MCM

Figura 5.8: Matrices de confusión conformal correspondientes a tres modelo de ‘base’, LAC y MCM.

5.5.3. Resultados para la clasificación combinada de sexo y mayoría de edad

Análisis de métricas para la clasificación se sexo y mayoría de edad

Probablemente MCM empeore a LAC por la menor cantidad de datos a emplear para la calibración (ya que los datos se dividen en cuatro subconjunto dependiendo de la clase a calibrar)

He pensado en utilizar un diagrama de Venn de 4 conjuntos (1 por cada clase) a modo de matriz de confusión conformal

Análisis de la cobertura en base a la clase

Método	Cobertura empírica (%)					Tamaño Medio del Conjunto				
	base	LAC	MCM	APS	RAPS	base	LAC	MCM	APS	RAPS
Ejecución 1	77.79	94.47	95.26	95.72	97.58	1	1.79	1.9	2.32	2.49
Ejecución 2	76.49	95.07	94.8	95.45	97.68	1	1.86	1.89	2.28	2.45
Ejecución 3	76.35	93.77	93.4	94.56	96.61	1	1.91	1.9	2.35	2.52
Ejecución 4	75.23	94.75	94.8	94.7	96.84	1	1.89	1.9	2.34	2.48
Ejecución 5	74.95	93.68	93.82	95.59	97.44	1	1.71	1.76	2.33	2.55
Ejecución 6	77.04	93.4	93.49	95.72	97.17	1	1.83	1.99	2.38	2.52
Ejecución 7	76.07	93.49	93.73	95.49	97.26	1	1.78	1.84	2.4	2.57
Ejecución 8	74.44	94.01	94.42	95.54	97.35	1	1.79	1.82	2.3	2.51
Ejecución 9										
Ejecución 10										
Media	76.05	94.08	94.21	95.35	97.24	1	1.82	1.87	2.34	2.51

Tabla 5.10: Cobertura empírica y tamaño medio del conjunto de predicción obtenidos por cada método de predicción a lo largo de las distintas ejecuciones. Se presentan los valores para cada ejecución individual, así como la media final de cada métrica. Punto como separador decimal.

Capítulo 6

Conclusiones y trabajos futuros

6.1. Conclusiones

6.1.1. Conclusiones sobre mejor método

Tengo que sintetizar y redactar:

CQR es el mejor método empleado para estimación de edad, pero hay mucha incertidumbre en las edades más avanzadas, por lo que es mejor emplearlo cuando hay indicios de que el individuo tiene una edad joven.

Los métodos de CP apenas reducen el rendimiento en sus predicciones puntuales respecto a método sin CP.

Los resultados de los métodos de CP mejorarán a medida que se mejore el desempeño general del modelo.

Teorema de No Free Lunch, pero en la práctica sí hay métodos mejores para dejar de sobrecubrir tanto en algunos grupos para cubrir más en grupos infracubiertos. El objetivo es la cobertura condicional.

Al igual que con las predicciones puntuales, se sigue requiriendo un estudio previo con las métricas en base a las variables específicas de cada problema, para conocer las debilidades y fortalezas del modelo (donde estima mejor y dónde peor el modelo, donde infracubre y dónde sobrecubre).

Dentro del procedimiento de estimación del perfil biológico, una predicción con gran incertidumbre puede indicar que hay que revisar la imagen para repetir la prueba o hacer otra prueba distinta.

El potencial de las herramientas de CP está en la flexibilidad para integrarse con otros métodos de estimación de incertidumbre e incluir informa-

ción del problema específico para reducir la incertidumbre en las predicciones conformales.

¿Es posible recalibrar el modelo para mejorar su cobertura condicional? Por ejemplo, una vez calibrado el modelo con CQR se podría calibrar de nuevo pero con el cálculo de más umbrales sobre los mismos datos de calibración en base a: el decil de tamaño de intervalo, o la edad predicha (parte entera), si bien hay riesgo de poblaciones reducidas no representativas para calibrar varios umbrales; aunque esto también se podría mitigar con número de cuantiles o grupos de edad adaptativos que se ajusten con un conjunto de datos adicional, que podría ser validación. A explorar en un anexo.

Para clasificación, especialmente en aquellos problemas en los que las clases se pisan en el espacio de entrada (p.ej., una misma edad biológica puede corresponderse a una edad entera y la inmediatamente posterior), los métodos tradicionales de clasificación pueden verse forzados a elegir una única clase, incluso cuando existe ambigüedad o solapamiento entre ellas. La CP permite devolver conjuntos de clases en lugar de una sola etiqueta, lo que es particularmente útil en estos escenarios de ambigüedad inherente.

Hay dos maneras de mejorar la cobertura condicional: mejorando la cobertura sobre los grupos infracubiertos concretos con un método de conformal prediction o mejorando las predicciones del modelo sobre los grupos infracubiertos, de manera que el error se asemeje más al de los grupos cubiertos.

6.2. Trabajos futuros

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, 39, 40, 46].
- [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. 3].
- [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. Geserick, “Forensic Age Estimation,” *Deutsches Arzteblatt international*, vol. 113, n.º 4, págs. 44-50, 2016. [Citado en págs. 5, 40, 41].
- [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 págs. 5, 42].
- [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 págs. 5, 47].
- [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] 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. 5].
- [33] 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. 5, 6].
- [34] 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 pág. 5].
- [35] 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 pág. 5].
- [36] 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 pág. 5].

- [37] 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. 200477, 2021. [Citado en pág. 5].
- [38] 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. 5, 27, 43].
- [39] Ministerio del Interior de España, “Informe anual sobre personas desaparecidas 2025,” Ministerio del Interior, inf. téc., 2025. [Citado en págs. 6, 7].
- [40] 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 pág. 6].
- [41] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2024,” Fiscalía General del Estado, Madrid, España, inf. téc., 2024. [Citado en págs. 6, 8].
- [42] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2019,” Fiscalía General del Estado, Madrid, España, inf. téc., 2019. [Citado en págs. 6, 8].
- [43] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2016,” Fiscalía General del Estado, Madrid, España, inf. téc., 2016. [Citado en págs. 6, 8].
- [44] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2013,” Fiscalía General del Estado, Madrid, España, inf. téc., 2013. [Citado en págs. 6, 8].
- [45] 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].
- [46] 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].
- [47] A. Turing, “I.—COMPUTING MACHINERY and INTELLIGENCE,” *Mind*, vol. LIX, n.º 236, págs. 433-460, 1950. [Citado en pág. 11].
- [48] 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].

- [49] 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, 15].
- [50] 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, 15].
- [51] J. R. Quinlan, “Induction of decision trees,” *Machine learning*, vol. 1, págs. 81-106, 1986. [Citado en pág. 11].
- [52] 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].
- [53] S. Russell y P. Norvig, *Artificial Intelligence: A Modern Approach*, 4rd. Prentice Hall Press, 2021. [Citado en págs. 12, 15, 20, 27].
- [54] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Heidelberg: Springer-Verlag, 2006. [Citado en págs. 12-14].
- [55] E. Alpaydin, *Introduction to Machine Learning*, 2nd. The MIT Press, 2010. [Citado en pág. 12].
- [56] Y. LeCun, Y. Bengio y G. Hinton, “Deep Learning,” *Nature*, vol. 521, págs. 436-44, 2015. [Citado en págs. 14, 15].
- [57] 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. 15].
- [58] P. J. Werbos, *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*. USA: Wiley-Interscience, 1994. [Citado en pág. 15].
- [59] 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. 15].
- [60] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. [Citado en págs. 15, 16, 21, 23, 24, 28].
- [61] Code World, *Understanding ML & DL in python*, Consultado el 19/05/2025, 2022. URL: <https://codeworld.tistory.com/2>. [Citado en pág. 16].
- [62] 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. 16].

- [63] 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. 17].
- [64] 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. 18].
- [65] R. Szeliski, *Computer Vision: Algorithms and Applications*, 1st. Berlin, Heidelberg: Springer-Verlag, 2010. [Citado en págs. 18, 21, 25, 26].
- [66] L. Bottou, “Large-Scale Machine Learning with Stochastic Gradient Descent,” *Proc. of COMPSTAT’2010*, págs. 177-186, 2010. [Citado en pág. 20].
- [67] 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. 20].
- [68] I. Goodfellow, Y. Bengio y A. Courville, *Deep Learning*. MIT Press, 2016. [Citado en pág. 20].
- [69] NVIDIA, *Convolutional Neural Network*, Consultado el 21/05/2025, 2025. URL: <https://www.nvidia.com/en-eu/glossary/convolutional-neural-network/>. [Citado en pág. 22].
- [70] 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. 25].
- [71] A. Zhang, Z. C. Lipton, M. Li y A. J. Smola, *Dive into Deep Learning*, 2021. [Citado en pág. 25].
- [72] 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. 26].
- [73] 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. 26].
- [74] 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. 26].

- [75] 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. 26].
- [76] 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. 26].
- [77] 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. 28].
- [78] 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. 28, 29].
- [79] 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. 29, 33].
- [80] V. Nemaní, L. Biggio, X. Huan, Z. Hu, O. Fink, A. Tran, Y. Wang, X. Zhang y C. Hu, “Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial,” *Mechanical Systems and Signal Processing*, vol. 205, pág. 110 796, 2023. [Citado en págs. 29, 30].
- [81] 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. 30].
- [82] E. Begoli, T. Bhattacharya y D. Kusnezov, “The need for uncertainty quantification in machine-assisted medical decision making,” *Nature Machine Intelligence*, vol. 1, n.º 1, págs. 20-23, 2019. [Citado en pág. 30].
- [83] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold y P. M. Atkinson, “Explainable artificial intelligence: an analytical review,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, n.º 5, e1424, 2021. [Citado en pág. 30].

- [84] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez y F. Herrera, “Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence,” *Information fusion*, vol. 99, pág. 101 805, 2023. [Citado en pág. 30].
- [85] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial intelligence*, vol. 267, págs. 1-38, 2019. [Citado en pág. 30].
- [86] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari y U. R. Acharya, “Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022),” *Computer methods and programs in biomedicine*, vol. 226, pág. 107 161, 2022. [Citado en pág. 30].
- [87] 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. 30].
- [88] A. F. Psaros, X. Meng, Z. Zou, L. Guo y G. E. Karniadakis, “Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons,” *Journal of Computational Physics*, vol. 477, pág. 111 902, 2023. [Citado en pág. 30].
- [89] M. Salvi, S. Seoni, A. Campagner, A. Gertych, U. R. Acharya, F. Molinari y F. Cabitza, “Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare,” *International Journal of Medical Informatics*, vol. 197, pág. 105 846, 2025. [Citado en pág. 30].
- [90] V. Vovk, A. Gammerman y G. Shafer, *Algorithmic learning in a random world*. Springer, 2005, vol. 29. [Citado en pág. 31].
- [91] A. N. Angelopoulos y S. Bates, “A gentle introduction to conformal prediction and distribution-free uncertainty quantification,” *arXiv preprint arXiv:2107.07511*, 2021. [Citado en págs. 31, 37].
- [92] Scikit-learn-contrib MAPIE developers. “MAPIE: Model-Agnostic Prediction Interval Estimator.” Accessed: 2025-07-06. URL: <https://mapie.readthedocs.io/en/stable/>. [Citado en pág. 32].
- [93] 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. 33].

- [94] D. Prinster, S. Stanton, A. Liu y S. Saria, “Conformal validity guarantees exist for any data distribution (and how to find them),” *arXiv preprint arXiv:2405.06627*, 2024. [Citado en pág. 32].
- [95] R. Foygel Barber, E. J. Candes, A. Ramdas y R. J. Tibshirani, “The limits of distribution-free conditional predictive inference,” *Information and Inference: A Journal of the IMA*, vol. 10, n.º 2, págs. 455-482, 2021. [Citado en pág. 33].
- [96] D. H. Wolpert y W. G. Macready, “No free lunch theorems for optimization,” *IEEE transactions on evolutionary computation*, vol. 1, n.º 1, págs. 67-82, 1997. [Citado en pág. 33].
- [97] H. Papadopoulos, K. Proedrou, V. Vovk y A. Gammerman, “Inductive confidence machines for regression,” en *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, Springer, 2002, págs. 345-356. [Citado en págs. 34, 36, 57].
- [98] Y. Romano, E. Patterson y E. Candès, “Conformalized quantile regression,” *Advances in neural information processing systems*, vol. 32, 2019. [Citado en págs. 34, 56, 58].
- [99] D. Bethell, S. Gerasimou y R. Calinescu, “Robust uncertainty quantification using conformalised Monte Carlo prediction,” en *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, págs. 20 939-20 948. [Citado en pág. 34].
- [100] 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 págs. 34, 59].
- [101] 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. 34, 62, 64].
- [102] 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. 34, 64].
- [103] C. Xu e Y. Xie, “Conformal prediction interval for dynamic time-series,” en *International Conference on Machine Learning*, PMLR, 2021, págs. 11 559-11 569. [Citado en pág. 34].
- [104] M. Zaffran, O. Féron, Y. Goude, J. Josse y A. Dieuleveut, “Adaptive conformal predictions for time series,” en *International Conference on Machine Learning*, PMLR, 2022, págs. 25 834-25 866. [Citado en pág. 34].

- [105] K. Stankeviciute, A. M Alaa y M. van der Schaar, “Conformal time-series forecasting,” *Advances in neural information processing systems*, vol. 34, págs. 6216-6228, 2021. [Citado en pág. 34].
- [106] R. Laxhammar y G. Falkman, “Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories,” *Annals of Mathematics and Artificial Intelligence*, vol. 74, págs. 67-94, 2015. [Citado en pág. 34].
- [107] D. H. Ubelaker y H. Khosrowshahi, “Estimation of age in forensic anthropology: historical perspective and recent methodological advances,” *Forensic Sciences Research*, vol. 4, n.º 1, págs. 1-9, 2019. [Citado en pág. 39].
- [108] R. Cameriere, L. Ferrante y M. Cingolani, “Age estimation in children by measurement of open apices in teeth,” *International journal of legal medicine*, vol. 120, págs. 49-52, 2006. [Citado en pág. 39].
- [109] L. Scheuer y S. Black, *Developmental Juvenile Osteology*, 1.^a ed. Academic Press, 2000. [Citado en pág. 40].
- [110] J. Adserias-Garriga, *Age estimation: a multidisciplinary approach*. Academic Press, 2019. [Citado en pág. 40].
- [111] S. MacLaughlin, J. Bowman y L. Scheuer, “The relationship between biological and chronological age in the juvenile remains from St Bride’s Church, Fleet Street,” *Annals of Human Biology*, vol. 19, n.º 2, págs. 211-216, 1992. [Citado en pág. 40].
- [112] C. E. Merritt, “The influence of body size on adult skeletal age estimation methods,” *American Journal of Physical Anthropology*, vol. 156, n.º 1, págs. 35-57, 2015. [Citado en pág. 40].
- [113] D. J. Wescott y J. L. Drew, “Effect of obesity on the reliability of age-at-death indicators of the pelvis,” *American Journal of Physical Anthropology*, vol. 156, n.º 4, págs. 595-605, 2015. [Citado en pág. 40].
- [114] D. H. Ubelaker, “Forensic Anthropology: Methodology and Diversity of Applications,” en *Biological Anthropology of the Human Skeleton*. John Wiley & Sons, Ltd, 2018, cap. 2, págs. 43-71. [Citado en pág. 40].
- [115] L. Scheuer y S. Black, *The juvenile skeleton*, 1.^a ed. Elsevier, 2004. [Citado en pág. 40].
- [116] S. Brooks y J. M. Suchey, “Skeletal age determination based on the os pubis: a comparison of the Acsádi-Nemeskéri and Suchey-Brooks methods,” *Human evolution*, vol. 5, págs. 227-238, 1990. [Citado en pág. 40].

- [117] E. Baccino, L. Sinfield, S. Colomb, T. P. Baum y L. Martrille, “The two step procedure (TSP) for the determination of age at death of adult human remains in forensic cases,” *Forensic science international*, vol. 244, págs. 247-251, 2014. [Citado en pág. 40].
- [118] H. Garvin y N. Passalacqua, “Current Practices by Forensic Anthropologists in Adult Skeletal Age Estimation,” *Journal of forensic sciences*, vol. 57, págs. 427-433, 2011. [Citado en págs. 40, 44].
- [119] C. O. Lovejoy, R. S. Meindl, T. R. Pryzbeck y R. P. Mensforth, “Chronological metamorphosis of the auricular surface of the ilium: A new method for the determination of adult skeletal age at death,” *American journal of physical anthropology*, vol. 68, págs. 15-28, 1985. [Citado en pág. 40].
- [120] M. Y. İşcan, S. R. Loth y R. K. Wright, “Metamorphosis at the sternal rib end: A new method to estimate age at death in white males,” *American Journal of Physical Anthropology*, vol. 65, n.º 2, págs. 147-156, 1984. [Citado en pág. 40].
- [121] R. S. Meindl y C. O. Lovejoy, “Ectocranial suture closure: A revised method for the determination of skeletal age at death based on the lateral-anterior sutures,” *American Journal of Physical Anthropology*, vol. 68, n.º 1, págs. 57-66, 1985. [Citado en pág. 40].
- [122] M. J. Berst, L. Dolan, M. M. Bogdanowicz, M. A. Stevens, S. Chow y E. A. Brandser, “Effect of knowledge of chronologic age on the variability of pediatric bone age determined using the Greulich and Pyle standards,” *American Journal of Roentgenology*, vol. 176, n.º 2, págs. 507-510, 2001. [Citado en pág. 42].
- [123] H. H. Thodberg, S. Kreiborg, A. Juul y K. D. Pedersen, “The BoneXpert method for automated determination of skeletal maturity,” *IEEE transactions on medical imaging*, vol. 28, n.º 1, págs. 52-66, 2008. [Citado en pág. 42].
- [124] R. R. van Rijn, M. H. Lequin y H. H. Thodberg, “Automatic determination of Greulich and Pyle bone age in healthy Dutch children,” *Pediatric radiology*, vol. 39, págs. 591-597, 2009. [Citado en pág. 42].
- [125] D. D. Martin, K. Sato, M. Sato, H. H. Thodberg y T. Tanaka, “Validation of a new method for automated determination of bone age in Japanese children,” *Hormone research in paediatrics*, vol. 73, n.º 5, págs. 398-404, 2010. [Citado en pág. 42].
- [126] H. H. Thodberg y L. Sävendahl, “Validation and reference values of automated bone age determination for four ethnicities,” *Academic radiology*, vol. 17, n.º 11, págs. 1425-1432, 2010. [Citado en pág. 42].

- [127] 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 pág. 42].
- [128] D. Štern, C. Payer y M. Urschler, “Automated age estimation from MRI volumes of the hand,” *Medical Image Analysis*, vol. 58, pág. 101538, 2019. [Citado en págs. 43, 47].
- [129] N. Marquez-Grant, “An overview of age estimation in forensic anthropology: perspectives and practical considerations,” *Annals of human biology*, vol. 42, n.º 4, págs. 308-322, 2015. [Citado en pág. 44].
- [130] J. E. Buikstra, “Standards for data collection from human skeletal remains,” *Arkansas archaeological survey research series*, vol. 44, pág. 44, 1994. [Citado en pág. 46].
- [131] N. G. Rao, N. N. Rao, M. Pai y M. Shashidhar Kotian, “Mandibular canine index — A clue for establishing sex identity,” *Forensic Science International*, vol. 42, n.º 3, págs. 249-254, 1989. [Citado en pág. 53].
- [132] A. P. Indira, A. Markande y M. P. David, “Mandibular ramus: An indicator for sex determination-A digital radiographic study,” *Journal of forensic dental sciences*, vol. 4, n.º 2, págs. 58-62, 2012. [Citado en pág. 53].
- [133] 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. 54].
- [134] 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. 54].
- [135] 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. 55].
- [136] R. F. Barber, E. J. Candès, A. Ramdas y R. J. Tibshirani, “Predictive inference with the jackknife+,” *The Annals of Statistics*, vol. 49, n.º 1, págs. 486-507, 2021. [Citado en pág. 56].
- [137] H. Linusson, U. Johansson y T. Löfström, “Signed-error conformal regression,” en *Advances in Knowledge Discovery and Data Mining: 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13-16, 2014. Proceedings, Part I* 18, Springer, 2014, págs. 224-236. [Citado en pág. 58].

- [138] U. Johansson, H. Linusson, T. Löfström y H. Boström, “Model-agnostic nonconformity functions for conformal classification,” en *2017 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2017, págs. 2072-2079. [Citado en pág. 59].
- [139] V. Vovk, D. Lindsay, I. Nouretdinov y A. Gammerman, “Mondrian confidence machine,” *Technical Report*, 2003. [Citado en pág. 62].
- [140] A. Niculescu-Mizil y R. Caruana, “Predicting good probabilities with supervised learning,” en *Proceedings of the 22nd international conference on Machine learning*, 2005, págs. 625-632. [Citado en pág. 66].
- [141] M. Sesia y E. J. Candès, “A comparison of some conformal quantile regression methods,” *Stat*, vol. 9, n.º 1, e261, 2020. [Citado en pág. 66].
- [142] I. Loshchilov y F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017. [Citado en pág. 71].
- [143] L. N. Smith, “A disciplined approach to neural network hyper-parameters: Part 1-learning rate, batch size, momentum, and weight decay,” *arXiv preprint arXiv:1803.09820*, 2018. [Citado en pág. 71].
- [144] T. Gneiting y A. E. Raftery, “Strictly proper scoring rules, prediction, and estimation,” *Journal of the American statistical Association*, vol. 102, n.º 477, págs. 359-378, 2007. [Citado en pág. 75].
- [145] 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. 77].
- [146] 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.
- [147] 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.
- [148] D. D. Martin, D. Deusched, R. Schweizer, G. Binder, H. H. Thodberg y M. B. Ranke, “Clinical application of automated Greulich-Pyle bone age determination in children with short stature,” *Pediatric radiology*, vol. 39, págs. 598-607, 2009.
- [149] D. D. Martin, K. Meister, R. Schweizer, M. B. Ranke, H. H. Thodberg y G. Binder, “Validation of automatic bone age rating in children with precocious and early puberty,” 2011.

- [150] J. G. Sam Lau y D. Nolan, *Cross Validation*, Consultado el 26/05/2025, 2023. URL: https://learningds.org/ch/16/ms_cv.html.
- [151] J. R. Berrendero. “Materiales del libro de Estadística,” visitado 2 de jun. de 2025. URL: <https://verso.mat.uam.es/~joser.berrendo/libro-est/>.
- [152] J. Vermorel. “Quantile Regression,” LOKAD Quantitive Supply Chain, visitado 2 de jun. de 2025. URL: <https://www.lokad.com/quantile-regression-time-series-definition/>.
- [153] R. Koenker, *Quantile Regression* (Econometric Society Monographs). Cambridge University Press, 2005.
- [154] 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.
- [155] J. Feldman y D. Kowal, “Bayesian Quantile Regression with Subset Selection: A Posterior Summarization Perspective,” *arXiv preprint arXiv:2311.02043*, 2023.
- [156] 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.
- [157] 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.

