



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, 3 de septiembre 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: Aprendizaje automático, cuantificación de incertidumbre, predicción conformal, estimación del perfil biológico, estimación de edad, antropología forense.

Resumen

La estimación del perfil biológico es una de las principales tareas de la Antropología Forense, empleada en procesos de identificación humana y en procesos legales en los que se requiere conocer la edad de personas vivas. Durante la última década, se han logrado avances significativos en la aplicación de técnicas de Machine Learning para este propósito, con el objetivo de reducir la subjetividad y posibilitar un estudio más riguroso y controlado de las capacidades de los métodos de estimación del perfil biológico, gracias a un entorno de variables más estandarizado y medible. Sin embargo, estos modelos suelen proporcionar únicamente una predicción puntual (como una edad concreta o una probabilidad de clase), obviando por completo la cuantificación de la incertidumbre inherente a la predicción. Esta limitación es crítica en un contexto forense, donde la fiabilidad de una estimación puede variar enormemente entre casos debido a factores como la calidad de la muestra, la variabilidad biológica intrínseca del individuo o las propias limitaciones del modelo entrenado.

Este trabajo propone la integración de la predicción conformal en el flujo de trabajo de estimación del perfil biológico. Este marco no se plantea como un sustituto de la predicción puntual, sino como un complemento esencial que la enriquece, generando intervalos de predicción (para regresión) y conjuntos de predicción (para clasificación) con garantías estadísticas sólidas de cobertura, válidas para cualquier distribución de los datos. Así, en lugar de simplemente afirmar que cierto individuo tiene “20 años” de edad esperada (más técnicamente conocida como edad biológica), el modelo podría también indicar una edad en el intervalo [18, 22] años con un 95 % de confianza, o en un problema de clasificación de sexo, devolver el conjunto *{masculino, femenino}* en casos ambiguos, reflejando de manera transparente y rigurosa la incertidumbre del modelo.

Para demostrar su utilidad —y costes— con una aplicación práctica, se emplea un conjunto de imágenes de radiografías maxilofaciales de individuos entre los 14 y los 25 años de 12 países distintos alrededor del globo. Se utiliza un modelo de red neuronal convolucional para resolver distintos problemas que predicen edades (en regresión y clasificación) de los individuos a partir de estas y se aplican técnicas de predicción conformal, generando intervalos de predicción calibrados en regresión y conjuntos de etiquetas en clasificación. Los resultados se evalúan empíricamente para verificar el cumplimiento de las garantías de cobertura marginal (por ejemplo, un 95 %), al tiempo que se analizan propiedades cruciales como el tamaño promedio de los intervalos o conjuntos, y la adaptatividad (la correlación entre el tamaño del conjunto/intervalo y la dificultad de la predicción en esa instancia).

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

David González Durán

Keywords: Machine learning, uncertainty quantification, conformal prediction, biological profile estimation, age estimation, forensic anthropology.

Abstract

Biological profile estimation is one of the main tasks of Forensic Anthropology, used in human identification processes and in legal proceedings where it is necessary to determine the age of living persons. Over the last decade, significant advances have been made in the application of Machine Learning techniques for this purpose, aiming to reduce subjectivity and enable a more rigorous and controlled study of the capabilities of biological profile estimation methods, thanks to a more standarized and measurable environment of variables. However, these models often provide only a point prediction (such as a specific age or a class probability), completely ignoring the quantification of the uncertainty inherent to the prediction. This limitation is critical in a forensic context, where the reliability of an estimate can vary enormously between cases due to factors such as sample quality, the intrinsic biological variability of the individual, or the inherent limitations of the trained model.

This work proposes the integration of conformal prediction into the biological profile estimation workflow. This framework is not intended as a substitute for point prediction, but as an essential complement that enriches it, generating prediction intervals (for regression) and prediction sets (for classification) with strong statistical coverage guarantees, valid for any data distribution. Thus, instead of simply stating that a certain individual has an expected age of “20 years” (more technically known as biological age), the model could also indicate an age in the interval [18, 22] years with 95 % confidence, or in a sex classification problem, return the set $\{\text{male}, \text{female}\}$ in ambiguous cases, transparently and rigorously reflecting the model’s uncertainty.

To demonstrate its usefulness—and costs—with a practical application, a dataset of maxillofacial radiographs from individuals between 14 and 25 years old from 12 different countries around the globe is used. A convolutional neural network model is used to solve different problems predicting age (in regression and classification) of the individuals from these images, and conformal prediction techniques are applied, generating calibrated prediction intervals in regression and label sets in classification. The results are empirically evaluated to verify compliance with marginal coverage guarantees (e.g., 95 %), while analyzing crucial properties such as the average size of the intervals or sets, and adaptivity (the correlation between the size of the set/interval and the difficulty of the prediction for that instance).

Yo, **David González Durán**, alumno de la doble titulación de Ingeniería Informática y Administración y Dirección de Empresas 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 3 de septiembre de 2025.

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

D. Javier Vénema Rodríguez, Esdudiante de Doctorado del programa de 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 3 de septiembre de 2025.

Los directores:

Pablo Mesejo Santiago

Javier Vénema Rodríguez

Agradecimientos

A todo el sistema educativo y universitario público, que ha confiado en mí e invertido en mi futuro. Mi gratitud por proveer la infraestructura, el talento docente y los recursos indispensables que han cimentado mi preparación.

A mis padres, por todo el apoyo incondicional que me han brindado durante toda la vida, pero especialmente en este último tramo, con los esfuerzos que han supuesto mis estudios universitarios, lejos de mi ciudad. Gracias por animarme a perseguir mis metas, por el gran sacrificio económico que han hecho y por socorrerme cada vez que lo he necesitado.

A mi compañera de piso y amiga durante los últimos cinco años, María. Por transformar nuestro apartamento en un verdadero hogar, un refugio lleno de calidez donde siempre me sentí cómodo y seguro. Por tu cuidado incondicional en los días buenos y, especialmente, en los menos buenos. Gracias por ser mi familia lejos de casa y por hacer que estos años hayan sido, sin duda, una de las mejores etapas de mi vida.

A mi novio, Juan, que me ha acompañado en mi último curso universitario. Su apoyo emocional fue crucial en los momentos más duros, y su ayuda fue invaluable: escuchó mis monólogos sobre el TFG con paciencia infinita, hizo preguntas que clarificaron mis ideas y me brindó una confianza inquebrantable. Su capacidad para escuchar y su constante aliento hicieron que todo fuera más llevadero.

A la Delegación de Estudiantes de Ingenierías Informática y de Telecomunicación, por ser el canal que me permitió conectar con la realidad de la escuela y participar activamente en su mejora. Gracias por mostrarme que la universidad no es solo un lugar de formación individual, sino una comunidad donde la colaboración y el apoyo mutuo son posibles. Con este agradecimiento, quiero poner en valor la representación estudiantil: humilde, crítica y ambiciosa, con la convicción de trabajar no en beneficio propio, sino pensando en quienes formarán parte de la universidad en el futuro.

A mis tutores, Pablo y Javier, por la incuestionable ayuda que me han ofrecido en todo momento, por tener tanta paciencia con todos los problemas surgidos, por sus valiosas críticas y consejos que han elevado la calidad de este proyecto, y por transmitirme su pasión por la investigación. Ha sido un privilegio aprender a vuestro lado.

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	3
1.3. Objetivos	6
1.4. Planificación temporal	6
1.5. Planificación económica	9
2. Fundamentos teóricos	12
2.1. Machine Learning	12
2.1.1. Problemas de regresión	13
2.1.2. Problemas de clasificación	14
2.2. Deep Learning	15
2.2.1. El perceptrón multicapa	15
2.2.2. Entrenamiento y validación de la red	17
2.2.3. Redes Neuronales Convolucionales	20
2.2.4. Transfer Learning	25
2.3. Incertidumbre	26
2.3.1. Incertidumbre en <i>machine learning</i>	27
2.3.2. Cuantificación de la incertidumbre en <i>machine learning</i>	28
2.4. Predicción conformal	29
2.4.1. Propiedades de la predicción conformal	30
2.4.2. Algoritmo conformal	32
3. Estado del arte	35
3.1. Estimación de la edad en antropología forense	35
3.2. Estimación de la edad en antropología forense usando <i>machine learning</i> . .	38
3.3. Cuantificación de la incertidumbre para la estimación de la edad	40

4. Materiales y métodos	43
4.1. Conjunto de datos disponibles	43
4.2. Métodos propuestos	45
4.2.1. Arquitectura empleada	45
4.2.2. Regresión cuantílica	45
4.2.3. Métodos de predicción conformal para regresión	47
4.2.4. Calibración de probabilidades en clasificación	49
4.2.5. Métodos de predicción conformal para clasificación	50
5. Experimentación	59
5.1. Problemas propuestos	59
5.2. Protocolo de validación experimental	60
5.3. Preprocesado de los datos	61
5.4. Esquema general de los experimentos realizados	62
5.4.1. Problema de estimación de edad	62
5.4.2. Problema de estimación de mayoría de edad	63
5.4.3. Problema de clasificación de edad	64
5.5. Evaluación del rendimiento	67
5.5.1. Métricas para regresión	67
5.5.2. Métricas para clasificación	69
5.5.3. Tests estadísticos	70
5.6. Experimentación para la estimación de edad	71
5.6.1. Entrenamiento de los modelos	71
5.6.2. Resultados	74
5.7. Experimentación para la estimación de mayoría de edad	83
5.7.1. Entrenamiento de los modelos	83
5.7.2. Resultados	84
5.8. Experimentación para la clasificación de edad	90
5.8.1. Entrenamiento de los modelos	90
5.8.2. Resultados	90
6. Conclusiones y trabajos futuros	98
6.1. Conclusiones	98
6.2. Valoración del trabajo realizado	100
6.3. Trabajos futuros	101
A. Intervalos de valores razonables	115
B. Problema de estimación de sexo	118

Índice de figuras

1.1.	Procedimiento secuencial para la identificación forense basada en el esqueleto humano (<i>skeleton-based forensic identification</i>)	4
1.2.	Evolución de hallazgos/identificación de cadáveres en España (2010-2024).	5
1.3.	Evolución del número de diligencias preprocesales de determinación de edad abiertas en España (2011-2023).	5
1.4.	Diagrama de modelo de regresión que usa predicción conformal, el cual, además de proporcionar una estimación puntual del valor esperado, entrega un intervalo de predicción con un nivel de confianza del 95 %.	7
1.5.	Diagrama de Gantt inicial del proyecto	7
1.6.	Diagrama de Gantt final del proyecto	8
2.1.	Esquema gráfico del funcionamiento de una unidad artificial de un perceptrón multicapa.	16
2.2.	Esquema gráfico de obtención de probabilidades mediante sigmoide y <i>softmax</i> en problemas de clasificación binaria y multiclase.	17
2.3.	Arquitectura simplificada de un MLP.	17
2.4.	Esquema gráfico de la aplicación de un filtro convolucional sobre una región de una imagen.	21
2.5.	Esquema gráfico de aplicación de <i>max pooling</i> con un filtro 2×2 y <i>stride</i> de 1.	22
2.6.	Esquema gráfico de aplicación de un filtro convolucional de 3×3 con <i>zero-padding</i> de 1 y <i>stride</i> de 2.	23
2.7.	Diagrama de la arquitectura de la red neuronal convolucional “AlexNet”.	24
2.8.	Esquema gráfico del funcionamiento de neuronas con <i>dropout</i>	25
2.9.	Diagrama del proceso de <i>fine-tuning</i> de un modelo de red neuronal en una nueva tarea.	26
2.10.	Ejemplo de predicción conformal en problemas de regresión y clasificación.	29
2.11.	Ejemplo adversario mal clasificado por un modelo de <i>machine learning</i> entrenado con datos textuales.	30
2.12.	Esquema gráfico de conjuntos de predicción bajo distintas nociones de cobertura: sin cobertura garantizada, con cobertura marginal y con cobertura condicional.	31

2.13. Determinación del umbral de no conformidad para intervalos simétricos y asimétricos	33
3.1. Evolución de las publicaciones que usan técnicas de antropología para la estimación de edad.	36
3.2. Hallazgos radiológicos en un posible menor con edad disputada: criterio de edad mínima para la determinación de edad.	37
3.3. Evolución de las publicaciones que usan técnicas de antropología y métodos ML para la estimación de edad.	38
3.4. Procedimiento secuencial clásico de <i>machine learning</i> para la extracción de características antropológicas.	39
3.5. Metodología de construcción de un modelo <i>end-to-end</i>	39
3.6. Cronograma de desarrollo de la unión epifisaria.	40
3.7. Distribución del error por edad real para un modelo de estimación de edad. .	41
3.8. Estudio del error en métodos propuestos para la estiamción de edad. . .	42
4.1. Histograma de edades de los individuos del conjunto de datos disponible diferenciado por sexo.	44
4.2. Visualización de la función de pérdida <i>pinball</i> para cada valor de error. .	46
4.3. Ejemplo de selección de clases para el conjunto conformal en LAC. . . .	51
4.4. Diagrama ilustrativo de la división de ejemplos utilizada en MCM. . . .	52
5.1. Esquema visual del modelos de regresión propuesto.	60
5.2. Diagrama de división del <i>dataset</i> en <i>train</i> , <i>validation</i> y <i>test</i>	61
5.3. Diagrama de división del <i>dataset</i> en <i>train</i> , <i>validation</i> , <i>calibration</i> y <i>test</i> . .	61
5.4. Esquema de experimentación para la estimación de edad.	63
5.5. Esquema de experimentación para la estimación de mayoría de edad. .	64
5.6. Esquema de experimentación para la clasificación de edad.	66
5.7. Ejemplo de matriz de confusión para un modelo de estimación de sexo. .	70
5.8. Adaptación de la arquitectura ResNeXt50 para el problema de la estimación de edad.	72
5.9. Problema de estimación de edad: Gráfica de dispersión de la cobertura empírica frente a la amplitud media del intervalo de predicción.	77
5.10. Problema de estimación de edad: Histogramas del amplitud del intervalo de predicción con diferenciación por cobertura, correspondientes a los métodos QR y CQR.	80
5.11. Problema de estimación de edad: Mapa de calor de la cobertura empírica en base a la amplitud del intervalo de predicción por cada método de predicción en media de sus ejecuciones.	81
5.12. Problema de estimación de edad: Gráficos de líneas de la cobertura empírica del intervalo de predicción (%) para cada método en función de la edad cronológica entera de los individuos, diferenciando por sexo.	82

5.13. Problema de estimación de edad: Gráficos de líneas de la amplitud media del intervalo de predicción para cada método en función de la edad cronológica entera de los individuos, diferenciando por sexo.	82
5.14. Problema de estimación de mayoría de edad: Gráfica de dispersión de la cobertura empírica frente al tamaño medio de conjunto de predicción.	86
5.15. Problema de estimación de la mayoría de edad: Mapa de calor de la cobertura empírica en base al tamaño del conjunto por cada método de predicción a lo largo de 10 ejecuciones.	88
5.16. Problema de estimación de mayoría de edad: Diagrama de líneas de la cobertura empírica en base al sexo y la edad cronológica por cada método de predicción a lo largo de 10 ejecuciones.	89
5.17. Problema de estimación de mayoría de edad: Diagrama de líneas del tamaño medio de conjunto de predicción en base al sexo y la edad cronológica por cada método de predicción a lo largo de 10 ejecuciones.	89
5.18. Problema de clasificación de edad: Gráfica de dispersión de la cobertura empírica frente al tamaño medio de conjunto de predicción.	92
5.19. Problema de clasificación de edad: Mapa de calor de cobertura empírica en base al tamaño del conjunto por cada método de predicción a lo largo de las distintas ejecuciones.	95
5.20. Problema de clasificación de edad: Gráficos de líneas de la cobertura empírica del intervalo de predicción (%) para cada método en función de la edad cronológica entera de los individuos, diferenciando por sexo.	96
5.21. Problema de clasificación de edad: Gráficos de líneas de tamaño medio del conjunto de predicción para cada método en función de la edad cronológica entera de los individuos, diferenciando por sexo.	96
A.1. Ejemplo de intervalo de confianza para la media poblacional.	116
A.2. Ejemplo de intervalo de credibilidad para la media poblacional.	117
A.3. Intervalos de predicción (95 % de confianza) construidos con CQR para estimación de edad.	117
B.1. Problema de estimación de sexo: Gráfica de dispersión de la cobertura empírica frente al tamaño medio de conjunto de predicción.	118
B.2. Problema de estimación de sexo: Mapa de calor de la cobertura empírica en base al tamaño del conjunto por cada método de predicción a lo largo de 10 ejecuciones.	120
B.3. Problema de estimación de sexo: Proporción de instancias bien clasificadas, mal clasificadas e indeterminadas de los métodos propuestos.	121
B.4. Problema de estimación de sexo: Diagrama de líneas de la cobertura empírica en base al sexo y la edad cronológica por cada método de predicción a lo largo de 10 ejecuciones.	121
B.5. Problema de estimación de sexo: Diagrama de líneas del tamaño medio de conjunto de predicción en base al sexo y la edad cronológica por cada método de predicción a lo largo de 10 ejecuciones.	121

Índice de tablas

4.1. Lista de instituciones participantes en la recolección de los datos e imágenes dentales utilizados en el trabajo.	44
4.2. Comparativa de métodos propuestos para problemas de regresión.	50
4.3. Comparativa de métodos propuestos para problemas de clasificación.	58
5.1. Problema de estimación de edad: Error absoluto medio y error cuadrático medio obtenidos por cada método de predicción a lo largo de distintas ejecuciones.	74
5.2. Problema de estimación de edad: Resultados de la prueba <i>post-hoc</i> de Tukey HSD para el error absoluto medio entre pares de métodos.	75
5.3. Problema de estimación de edad: Resultados de la prueba <i>post-hoc</i> de Tukey HSD para el error cuadrático medio entre pares de métodos.	75
5.4. Problema de estimación de edad: Cobertura empírica y amplitud media del intervalo de predicción obtenidos por cada método de predicción a lo largo de distintas ejecuciones.	76
5.5. Problema de estimación de edad: Resultados de la prueba <i>post-hoc</i> de Games-Howell para cobertura empírica entre pares de métodos.	77
5.6. Problema de estimación de edad: Resultados de la prueba <i>post-hoc</i> de Games-Howell para la amplitud media del intervalo de predicción entre pares de métodos.	78
5.7. Problema de estimación de edad: <i>Mean Interval Score</i> obtenidos por cada método de predicción a lo largo de distintas ejecuciones.	78
5.8. Problema de estimación de mayoría de edad: Exactitud, sensibilidad y especificidad obtenidos por cada método de predicción a lo largo de distintas ejecuciones.	84
5.9. Problema de estimación de mayoría de edad: Cobertura empírica y tamaño medio del conjunto de predicción obtenidos por cada método de predicción a lo largo de las distintas ejecuciones.	86
5.10. Problema de estimación de mayoría de edad: Resultados de la prueba <i>post-hoc</i> de Tukey HSD para la cobertura empírica entre pares de métodos.	87
5.11. Problema de clasificación de edad: Cobertura empírica obtenida por cada método de predicción a lo largo de las distintas ejecuciones.	91
5.12. Problema de clasificación de edad: Tamaño medio del conjunto de predicción obtenido por cada método a lo largo de las distintas ejecuciones.	92

5.13. Problema de clasificación de edad: Resultados de la prueba <i>post-hoc</i> de Tukey HSD para la cobertura empírica entre pares de métodos.	93
5.14. Problema de clasificación de edad: Resultados de la prueba <i>post-hoc</i> de Tukey HSD para el tamaño medio del conjunto de predicción entre pares de métodos.	93
B.1. Problema de estimación de mayoría de sexo: Valores de exactitud, cobertura empírica y tamaño medio del conjunto obtenidas por cada método de predicción a lo largo de 10 ejecuciones.	119

Lista de Abreviaturas

- AF** Antropología Forense. 1–4, 8, 38, 59, 62
- APS** Adaptive Prediction Sets. 7, 52–57, 64, 65, 94
- CNN** Convolutional Neural Network. 20, 23–25
- CP** Conformal Prediction. 8, 29–33, 48, 50, 52, 59, 61–65, 68, 69
- CQR** Conformalized Quantile Regression. 48, 49, 69, 82, 83
- DL** Deep Learning. 15
- DNN** Deep Neural Network. 15, 16, 20, 48
- FC** Fully Connected. 23, 24, 72
- IA** Inteligencia Artificial. 4, 12, 28
- ICP** Inductive Conformal Prediction. 47, 49, 64
- ID** Identificación Humana. 2, 3
- LAC** Least-Ambiguous set-valued Classifiers. 7, 50–53, 64, 65, 94
- MCM** Mondrian Confidence Machine. 52, 65
- ML** Machine Learning. 4, 6, 12, 14, 15, 26–30, 38, 39, 49, 59
- MLP** MultiLayer Perceptron. 15–17, 20, 23
- PB** Perfil Biológico. 3, 4, 6, 38
- QR** Ruantile Regression. 45, 49, 63, 69, 72
- RAPS** Regularized Adaptive Prediction Sets. 7, 55–57, 64, 65, 94
- SAPS** Sorted Adaptive Prediction Sets. 56, 57, 65, 94
- SSH** Secure Shell. 10
- UQ** Uncetainty Quantification. 28, 29, 32, 45

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.

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.

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

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].

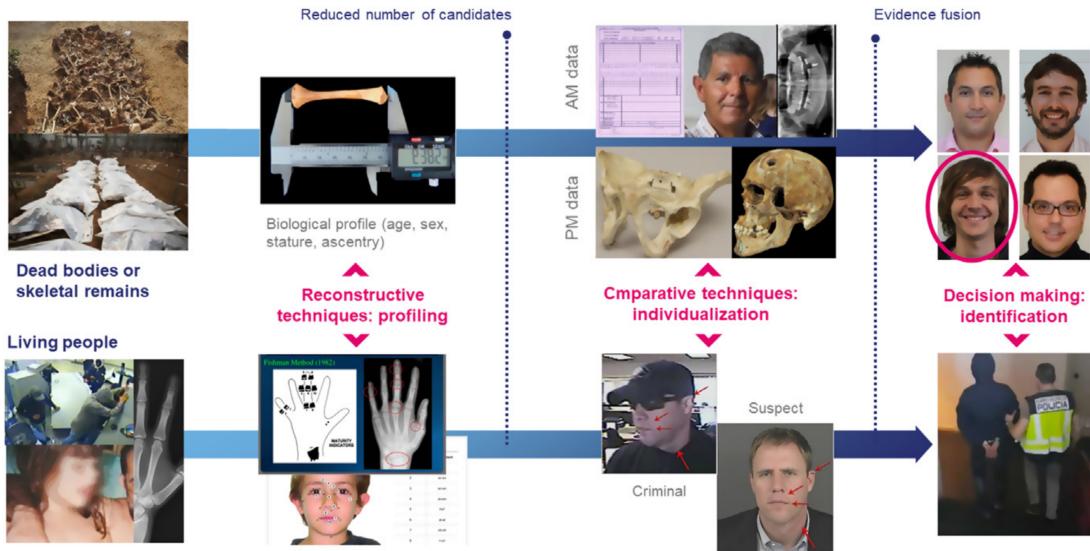


Figura 1.1: Procedimiento secuencial para la identificación forense basada en el esqueleto humano (*skeleton-based forensic identification*) [20].

embargo, su aplicación puede presentar ambigüedades en su formulación que den lugar a interpretaciones 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 [26-30] como de edad [30-35]. Sin embargo, aún mejorando la exactitud de las predicciones, la variabilidad biológica inherente al desarrollo esquelético humano continúa representando un desafío para la interpretación de estas. No todas las predicciones tienen el mismo nivel de confianza o fiabilidad, y los modelos no incluyen información sobre esta incertidumbre en sus predicciones. Ya en [36] se introducía no solo la necesidad de identificar el método adecuado para estimar la edad a partir de los elementos disponibles, sino también de evaluar su confiabilidad.

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 1.2 [37]. En 2024 se ha alcanzado una cifra record, —en gran parte debido a las inundaciones de la DANA Valencia del mismo año—, 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 franquis-

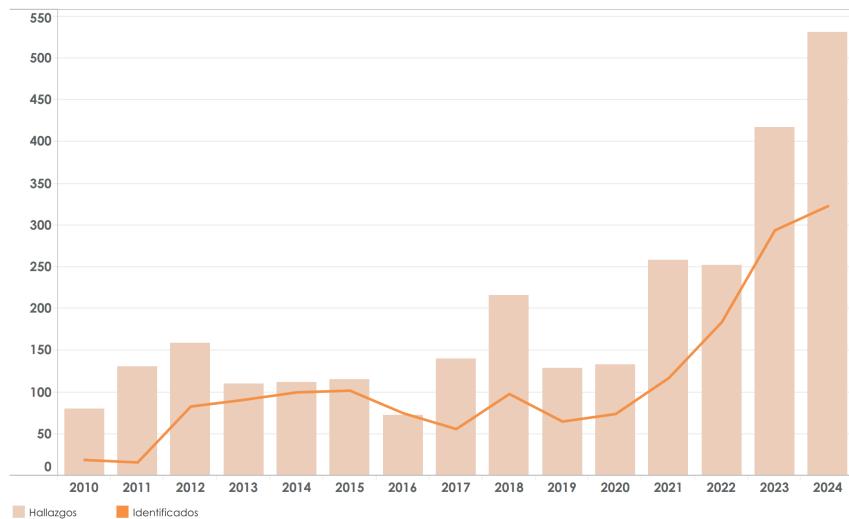


Figura 1.2: Evolución de hallazgos/identificación de cadáveres en España (2010-2024) [37].

mo, 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 [38].

- En España, se ha registrado en la última década (2013-2023) un aumento significativo en la llegada de Menores Extranjeros No Acompañados [39-42], que ha disparado consigo el número de diligencias abiertas para la determinación de su edad, como se ve reflejado en la Figura 1.3.

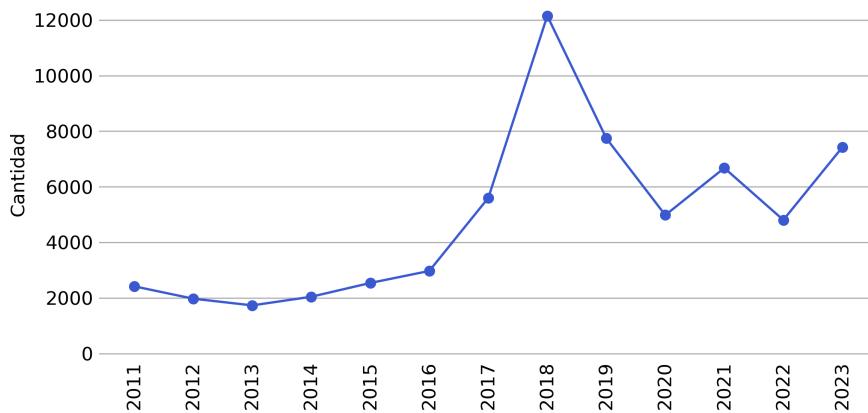


Figura 1.3: Evolución del número de diligencias preprocesales de determinación de edad abiertas en España (2011-2023). Elaboración propia a partir de [39-42].

- La relevancia de la ciencia forense en la identificación de víctimas y la protección de la dignidad humana ha convertido su aplicación en un pilar fundamental de los derechos humanos y la justicia internacional, naciendo así la **acción forense humanitaria** [43]. 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 [44].

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.

Este Trabajo de Fin de Grado tiene un doble objetivo:

- Defender la cuantificación de incertidumbre como herramienta esencial en ML, ofrecer un panorama de métodos destacados, analizando superficialmente sus ventajas y limitaciones, y centrarnos en la predicción conformal y sus técnicas más populares.
- Aplicar la predicción conformal a un contexto práctico como es el problema de estimación del PB, centrándose en la estimación de edad 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.

Podemos dividir este objetivo a su vez en los siguientes:

- Estudiar de forma exhaustiva la bibliografía sobre predicción conformal y algunas de sus técnicas, así como de la estimación de edad.
- Implementar, entrenar y validar modelos de regresión y clasificación a los que aplicar la inferencia conformal, y comparar aproximaciones heurísticas de cuantificación de incertidumbre con los métodos conformales.
- Realizar una primera aproximación a un marco interpretable y con garantías estadísticas para la estimación del perfil biológico (véase un ejemplo práctico en la Figura 1.4), 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

Partimos de que el Proyecto de Fin de Grado tiene asignado 12 créditos ECTS, lo que equivale a 360 horas de trabajo². Estas 360 horas se distribuyen a lo largo del segundo cuatrimestre del curso 2024/2025, constando de 66 días lectivos, lo que resulta en una

²De 25 a 30 horas por crédito según [45].

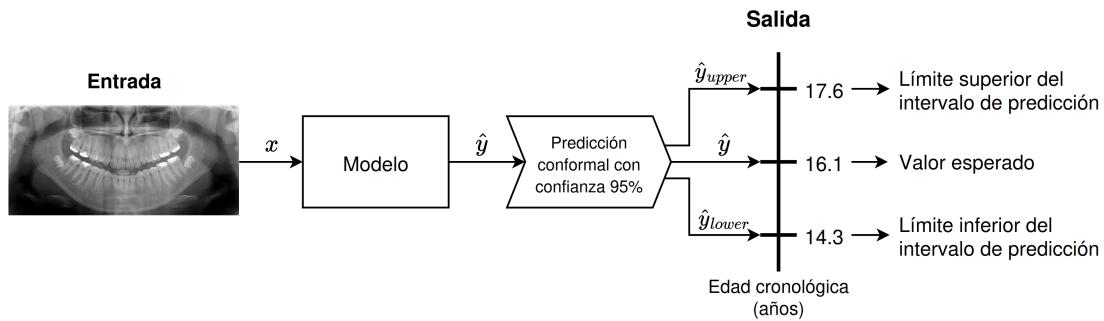


Figura 1.4: Diagrama de modelo de regresión que usa predicción conformal, el cual, además de proporcionar una estimación puntual del valor esperado, entrega un intervalo de predicción con un nivel de confianza del 95 %. Esta salida se lee de la siguiente manera: “la edad esperada del individuo es de 16.1 años y, con una confianza del 95 %, la edad real del individuo está entre los 14.3 y los 17.6 años”.

carga de trabajo de aproximadamente 5.45 horas al día, aproximadamente 5 horas y media. La planificación inicial del proyecto se presenta en el diagrama de Gantt de la Figura 1.5.

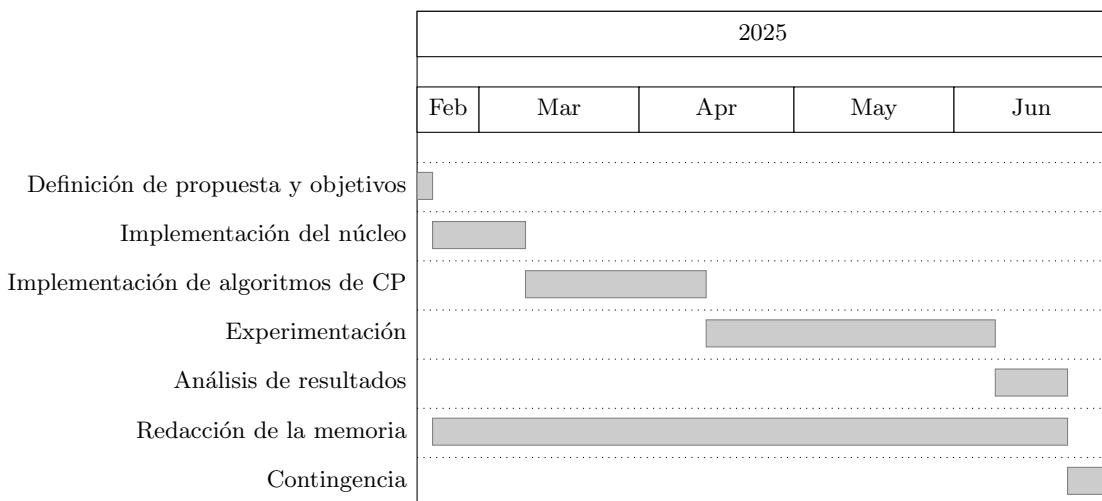


Figura 1.5: Diagrama de Gantt inicial del proyecto

No obstante, debido a ajustes en la orientación conceptual del trabajo, modificaciones en el enfoque metodológico y determinadas circunstancias personales de salud, fue necesario realizar cambios sobre la planificación inicial, posponiéndose finalmente la entrega del trabajo al mes de septiembre.

La organización temporal del trabajo se desarrolló finalmente de la siguiente manera (véase la Figura 1.6):

- Febrero: se redactó la propuesta del proyecto, la cual fue aceptada en el plazo de una semana.
- Marzo: la primera parte del mes se destinó a la implementación de algunas técnicas de predicción conformal en clasificación(LAC, APS y RAPS) sobre el conjunto de datos CIFAR-10. A mediados de mes se concedió acceso al clúster SLURM, lo que permitió comenzar la organización del proyecto, la implementación del núcleo del

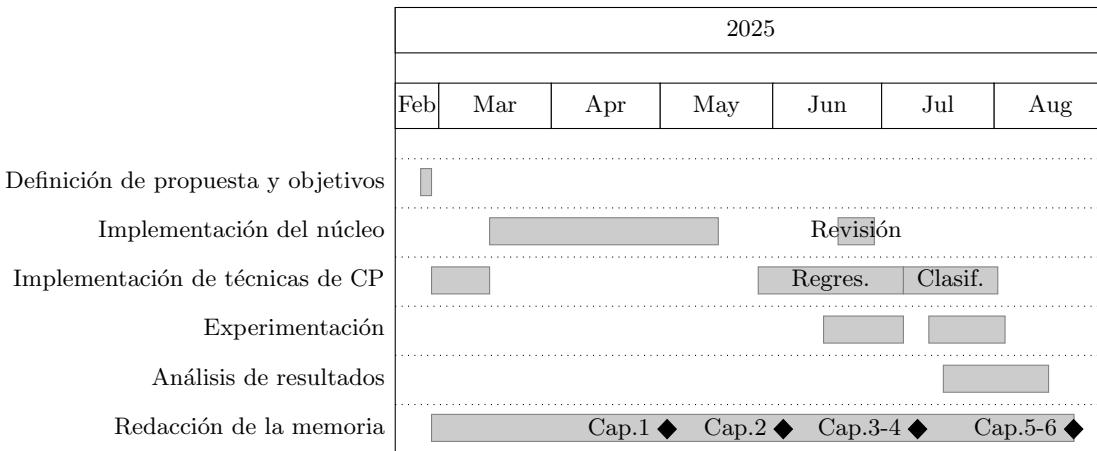


Figura 1.6: Diagrama de Gantt final del proyecto

código y la configuración del entorno de trabajo mediante un proceso iterativo de prueba y error orientado a obtener una solución más cómoda y flexible.

- Abril: se llevó a cabo la búsqueda, revisión, organización y análisis de bibliografía, con especial atención a la literatura en AF. A mediados de mes se obtuvo acceso al conjunto de datos definitivo que sería utilizado en el proyecto. Fue necesario adaptar el núcleo de código previamente desarrollado, dedicándose varias semanas a la experimentación y mejora del modelo de red neuronal convolucional. Se aprovechó para organizar el código y subirlo a un repositorio.
- Mayo: a principios de mes se completó una primera versión del capítulo inicial (*Introducción*). Durante la primera mitad del mes se continuó con la optimización del entrenamiento e inferencia del modelo, en paralelo con la redacción del capítulo de Fundamentos teóricos.
- Junio: a comienzos de mes se obtuvo la primera versión completa del capítulo segundo (*Fundamentos teóricos*). En este momento se decidió centrar los esfuerzos en la parte de regresión, lo que condujo a la implementación de las técnicas de CP aplicadas a este contexto. Durante el proceso se identificó la necesidad de reformar nuevamente el núcleo del código, con el fin de reducir repeticiones y dotarlo de mayor flexibilidad, permitiendo mediante argumentos entrenar toda la red o únicamente la cabeza, seleccionar la técnica de CP, realizar inferencia, entre otras opciones. De forma paralela, se avanzó en la redacción del capítulo tercero (*Estado del arte*) y del capítulo cuarto (*Materiales y métodos*).
- Julio: a principios de mes se dispuso de una versión preliminar de la memoria que incluía la parte de regresión hasta la presentación de resultados, aunque sin discusión. Tras esto, el trabajo se orientó hacia la implementación de técnicas de CP para clasificación, en problemas derivados del enfoque de regresión. Asimismo, se desarrolló un sistema de almacenamiento más organizado de los resultados, de manera que pudieran ser cargados posteriormente en un cuaderno Jupyter para su análisis.
- Agosto: se discutieron los resultados obtenidos tanto en regresión como en clasificación, se completaron apartados pendientes de la memoria y se llevaron a cabo ajustes finales. Finalmente, se redactó la conclusión y el *abstract*.

Finalmente, el trabajo se desarrolló entre el 24 de febrero y el 22 de agosto, considerando únicamente los días laborables, con excepción de los períodos del 4 al 8 de junio y del 3 al 10 de agosto, debido a vacaciones. Durante los días laborables se dedicó una media de 6 horas diarias, mientras que los sábados se trabajó aproximadamente 3 horas diarias³.

Teniendo en cuenta esta distribución, se puede estimar el total de horas trabajadas de la siguiente manera:

- Días laborables efectivos: $122 \text{ días} \times 6\text{h/día} = 732\text{h}$
- Sábados: $23 \text{ días} \times 3\text{h/día} = 69\text{h}$

Por tanto, el total aproximado de horas dedicadas al proyecto asciende a 801 horas, lo que supera por mucho la cifra teórica de 360 horas. Este tiempo de más se puede atribuir a varias causas principales:

- Complejidad del proyecto: Nunca antes me había enfrentado a un trabajo de esta complejidad, por lo que la planificación inicial no podía anticipar todas las dificultades técnicas y conceptuales.
- Adaptación y organización del código: La necesidad de modificar constantemente el núcleo del código para hacerlo más modular y flexible ha supuesto tiempo adicional no contemplado en la estimación inicial.
- Iteraciones y mejoras del modelo: Se llevaron a cabo múltiples iteraciones de entrenamiento y ajuste de hiperparámetros de los modelos CNN, así como experimentos para mejorar la eficiencia y la estabilidad de las predicciones, lo que incrementó considerablemente la carga de trabajo.
- Búsqueda, análisis y revisión bibliográfica: La investigación en literatura especializada, especialmente en antropología forense, materia en la que soy nuevo, requirió un tiempo prolongado de lectura, análisis y síntesis para incorporarlo a la memoria.
- Circunstancias externas y aprendizaje: Ajustes en la planificación por cuestiones de salud, aprendizaje de nuevas herramientas (PyTorch, SLURM, Jupyter), problemas con el software (drivers de CUDA) y adaptaciones metodológicas también contribuyeron a la ampliación del tiempo dedicado.

1.5. Planificación económica

Se ha dispuesto de todos los materiales necesarios para la realización del proyecto de manera gratis. Aún así, en este apartado se hace una estimación del coste económico de desarrollar el trabajo.

Este trabajo ha sido realizado con dos equipos independientes:

- **Ordenador portátil personal:** Asus Zephyrus G14, modelo GA401 de 2021, empleado principalmente para la redacción y compilación de este documento en L^AT_EX, así como el análisis exploratorio de datos y el tratamiento de resultados. El

³Estimación realizada de manera aproximada.

sistema operativo empleado es Linux Mint (versión 22.1), de acceso libre y, por tanto, gratuito. El portátil fue adquirido por 1100€ en marzo de 2022. Para la estimación de la amortización se ha considerado una vida útil de 4 años, criterio habitual en contabilidad empresarial y acorde con la obsolescencia tecnológica de este tipo de dispositivos:

$$\begin{aligned}\text{Coste imputado} &= \frac{\text{Precio de adquisición}}{\text{Vida útil}} \times \frac{\text{Tiempo de uso}}{\text{en el proyecto}} \\ &= \frac{1100 \text{ €}}{48 \text{ meses}} \times 4 \text{ meses} = 91.67 \text{ €}\end{aligned}$$

- **Clúster de computación DaSCI:** proporcionado por el Instituto de Ciencia de Datos e Inteligencia Artificial de la Universidad de Granada, al que se accede mediante conexión SSH.

El clúster cuenta con nodos diversos: con y sin GPU, con entre 40 y 60 CPUs y aproximadamente 122 GB de memoria RAM cada uno. Los nodos con GPU presentan diferentes modelos de gráfica: NVIDIA TITAN XP y NVIDIA Titan RTX, con CUDA 11.7 o superior. Todos los nodos cuentan con sistemas Linux compatibles con Pytorch. La implementación del código será adaptada para ejecutarse en cualquiera de los nodos con GPU.

Se considera un total de 720 horas de uso de GPU, el doble de las horas teóricas estimadas del trabajo, con el objetivo de cubrir tanto la fase de experimentación como posibles reentrenamientos y ajustes de hiperparámetros, garantizando un margen de seguridad frente a imprevistos y evitando interrupciones por falta de recursos durante el desarrollo y la validación del modelo.

Para simplificar la estimación del coste, se considera un único nodo con la GPU más económica, NVIDIA Titan XP, ya que los requisitos computacionales del proyecto no justificaban el uso de hardware más potente.

La estimación del coste de uso de la GPU se basa en precios de referencia de servicios *cloud* equivalentes. La NVIDIA Titan XP no es una GPU de centro de datos, pero se aproxima —en rendimiento para operaciones F32 y memoria— a una NVIDIA Tesla P100, con un precio de referencia entre 1.2-1.5 € por hora. Asumiendo un valor medio de 1.35 €/h, el coste total imputado del uso de la GPU en el proyecto sería:

$$\begin{aligned}\text{Coste GPU} &= \text{Tiempo GPU} \times \text{Coste por tiempo} \\ &= 720 \text{ h} \times 1.35 \text{ €/h} = 972 \text{ €}\end{aligned}$$

En cuanto al software, se ha recurrido exclusivamente a herramientas de código abierto y libre distribución:

- **Linux Mint:** es el sistema operativo del ordenador personal.
- **Visual Studio Code:** empleado como entorno de desarrollo integrado y como interfaz de conexión al clúster mediante SSH.
- **TeX Live:** distribución de L^AT_EX, utilizada para la redacción, compilación y maquetación del documento.

- **Python:** lenguaje de programación utilizado para la implementación de los algoritmos.
- **Jupyter Notebook:** para experimentación interactiva y análisis exploratorio de datos y resultados.
- **PyTorch:** framework de referencia en el desarrollo de modelos de aprendizaje profundo.
- **Matplotlib y Seaborn:** bibliotecas de visualización para la representación gráfica de resultados, métricas y distribuciones.

Finalmente, en cuanto a recursos humanos, el coste asociado a la mano de obra se divide en dos categorías:

- **Investigación y desarrollo:** aunque el proyecto ha sido realizado por el autor, estudiante universitario, se considera el salario de un ingeniero de IA junior en España, aproximadamente 21€/h⁴. Basado en las horas dedicadas al proyecto, el coste hipotético sería:

$$\begin{aligned}\text{Coste Investigador} &= \text{Tiempo de trabajo} \times \text{Coste por tiempo} \\ &= 360 \text{ h} \times 21 \text{ €/h} = 7560 \text{ €}\end{aligned}$$

- **Mantenimiento del clúster:** no se considera un coste adicional, ya que está incluido en el servicio de la nube proporcionado.

El coste total estimado, considerando todos los componentes anteriormente citados, ascendería a **8 623.67€**.

⁴Este valor se obtuvo en el portal de información laboral: [46].

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 importancia 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.

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

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 variantes de clasificación, que pueden diferenciarse según diversos criterios:

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

No obstante, la mayoría de los problemas estudiados en la literatura de ML, y concretamente en antropología forense, corresponden a clasificación binaria o multiclasa, con etiquetas únicas y asignación precisa [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. Estas 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*, DNN) 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 DNN 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 artificial**. 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 producir una salida que propaga a las unidades de la siguiente capa (véase la Figura 2.1).

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

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

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].

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

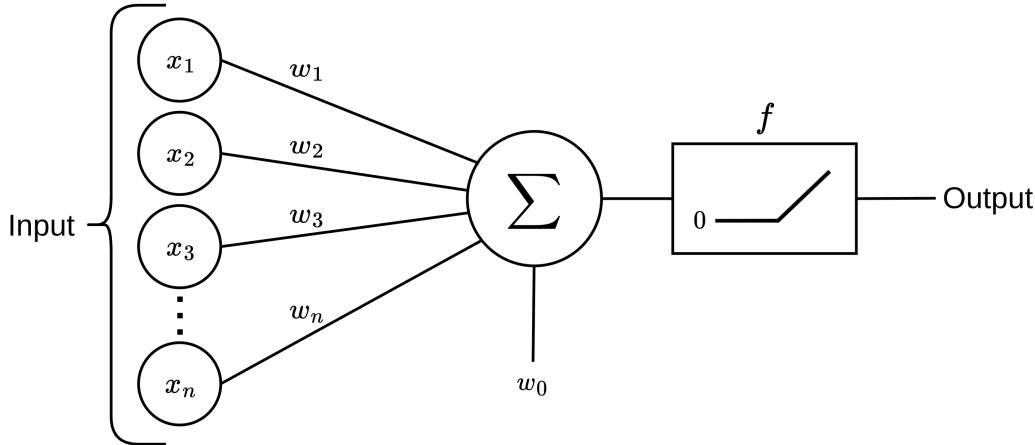


Figura 2.1: Esquema gráfico del funcionamiento de una unidad artificial de un perceptrón multicapa. Adaptado de [61].

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):

- **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—;

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 DNN, 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.

- en problemas de clasificación, esta capa tendrá una sola unidad —generalmente, con activación sigmoid— en clasificación binaria, o múltiples unidades —con activación *softmax*⁷— en clasificación multiclase (véase la Figura 2.2).

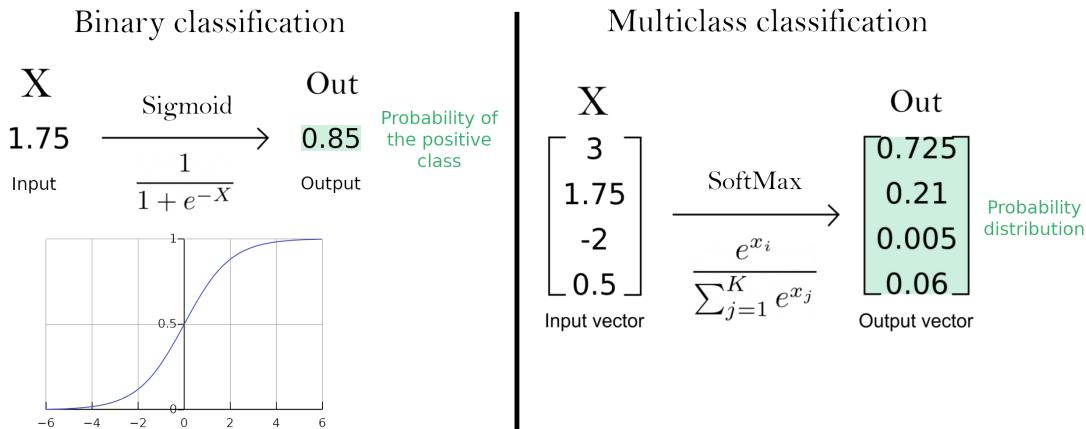


Figura 2.2: Esquema gráfico de obtención de probabilidades mediante sigmoides y *softmax* en problemas de clasificación binaria y multiclase, respectivamente. Adaptado de [63].

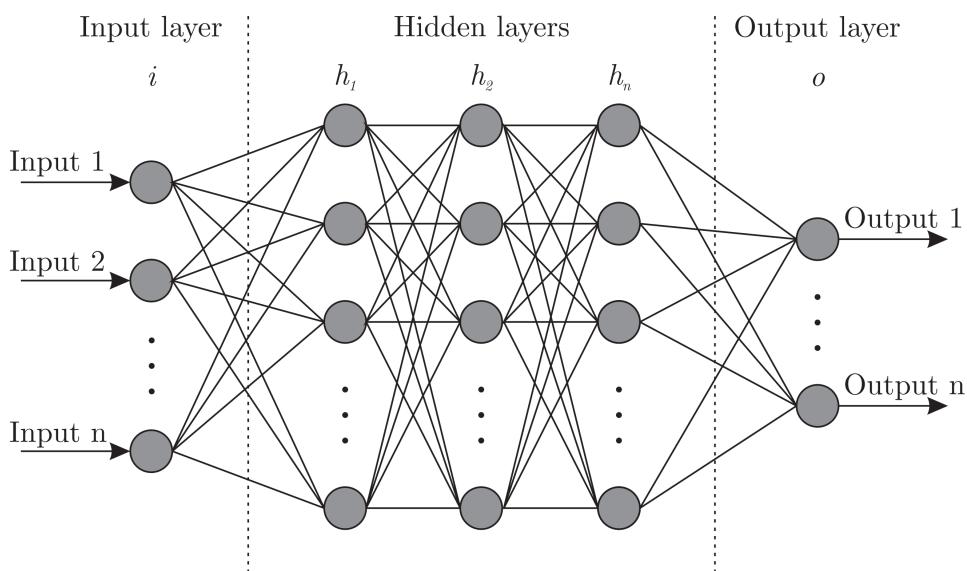


Figura 2.3: Arquitectura simplificada de un MLP. Recuperado de [64].

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

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.

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.

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

En clasificación, las funciones de pérdida más comunes son la entropía cruzada (*cross-entropy loss*) para problemas de clasificación binaria y multiclas, 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 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 parallelización 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, CNN*) 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 CNN. 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 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 CNN 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.

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.

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

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

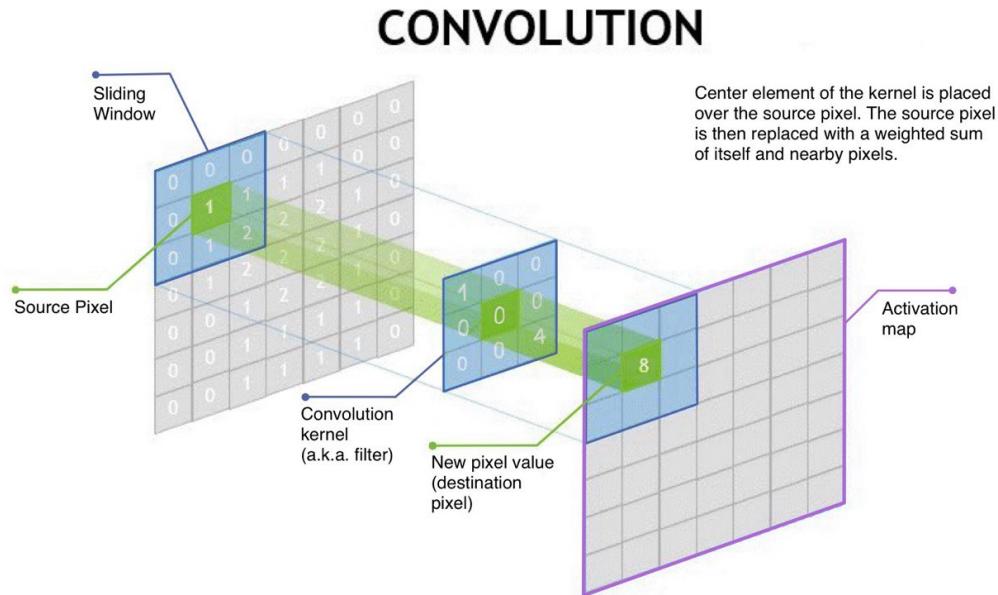


Figura 2.4: Esquema gráfico de la aplicación de un filtro convolucional de 3×3 sobre una región de una imagen. Adaptado de [69].

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 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 emplear usar capas de *pooling* o usar *stride* mayor de 1 en los filtros de las capas convolucionales.

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

Capas de pooling

Las **capas de agrupación (pooling layers)** tienen como objetivo principal compimir 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 diversas operaciones de *pooling*, entre los que destacan:

- **Max pooling**, que calcula el máximo valor de regiones del mapa de características, y lo usa para crear un mapa de características reducido (véase la Figura 2.5).
- **Average pooling**, que reemplaza el valor máximo del *max pooling* por el cálculo de la media entre los valores de la región.

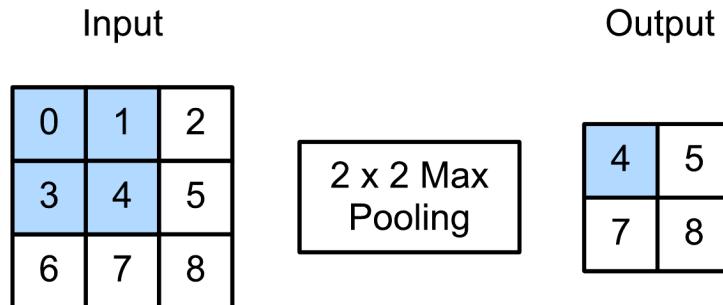


Figura 2.5: Esquema gráfico de aplicación de *max pooling* con un filtro 2×2 y *stride* de 1. Recuperado de la Figura 14.12 de [60].

La región de aplicación del *pooling*, al igual que en la convolución, viene determinada por ciertos parámetros, definidos por el diseñador, como el tamaño de filtro (que suele ser de 2×2), 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*.

Reducción de dimensionalidad mediante *stride* aumentado

En una convolución estándar, un filtro se desliza sobre la imagen con un desplazamiento determinado. El *stride* indica cuántos píxeles se mueve el filtro en cada paso:

- *Stride* de 1: El filtro se mueve un píxel a la vez. Esto mantiene la mayor parte de la información espacial.
- *Stride* de 2: El filtro se mueve dos píxeles a la vez, saltándose algunas posiciones. Esto reduce el tamaño de la salida (*feature map*), lo que se conoce como submuestreo o *downsampling*. Véase la Figura 2.6.

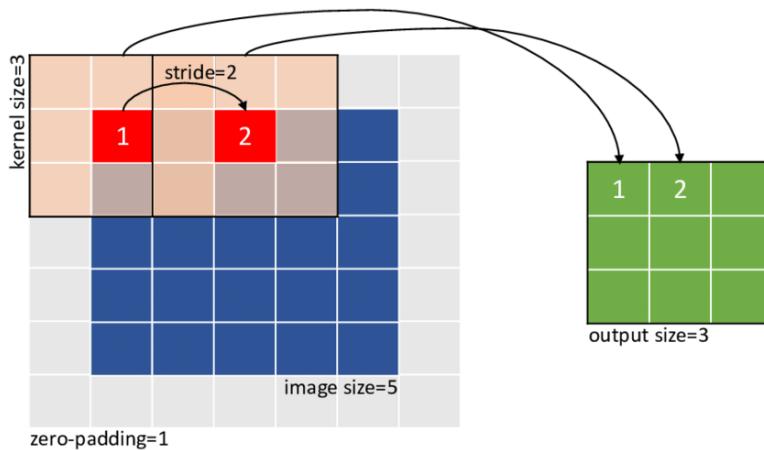


Figura 2.6: Esquema gráfico de aplicación de un filtro convolucional de 3×3 con *zero-padding* de 1 y *stride* de 2. Recuperado de la Figura 3 de [70]. El *zero-padding* consiste en agregar filas y columnas de ceros alrededor de la imagen, de forma que los bordes puedan ser procesados por el filtro sin reducir excesivamente el tamaño del *feature map* resultante. Esto permite que la convolución considere todas las regiones de la imagen, incluyendo los bordes, y facilita la preservación de la información espacial. Con un *stride* de 2, el filtro se desplaza de dos en dos píxeles, realizando simultáneamente la extracción de características y el *subsampling*, lo que genera una reducción controlada del tamaño de salida.

Capas *Fully-Connected*

Como hemos visto hasta ahora, en las CNN, 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/regresor¹⁴ que toma todas las señales procesadas por las capas anteriores y predice la clase a la que pertenece la imagen o el valor objetivo.

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

Diseño de la CNN para problemas de clasificación y regresión

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

¹⁴Si bien, independientemente de la tarea —regresión o clasificación—, a esta parte de la red se le denomina clasificador

- 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 2.7 se puede observar un ejemplo de arquitectura CNN completa.

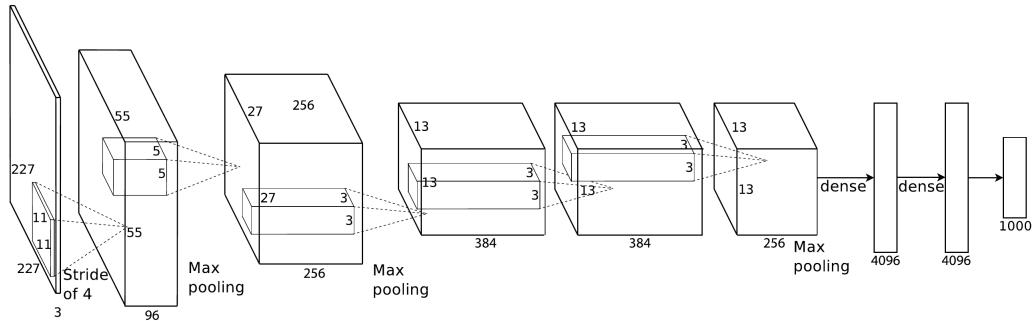


Figura 2.7: Diagrama de la arquitectura de la red neuronal convolucional “AlexNet”, diseñada para resolver un problema de clasificación con 1000 clases. Recuperado de la Figura 5.39 de [65]. Esta arquitectura presenta una serie de capas convolucionales con funciones de activación no lineales ReLU y *max pooling*, que formarían el *backbone* y una serie de capas FC (*classifier*), con una capa final *softmax*, que alimenta 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 CNN:

- Data augmentation** [71, 72]: 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** [73]: 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 2.8). En CNN suele aplicarse a capas FC, aunque existen variantes como *Spatial Dropout* [74] que elimina canales completos en capas convolucionales, forzando una distribución más robusta de características.

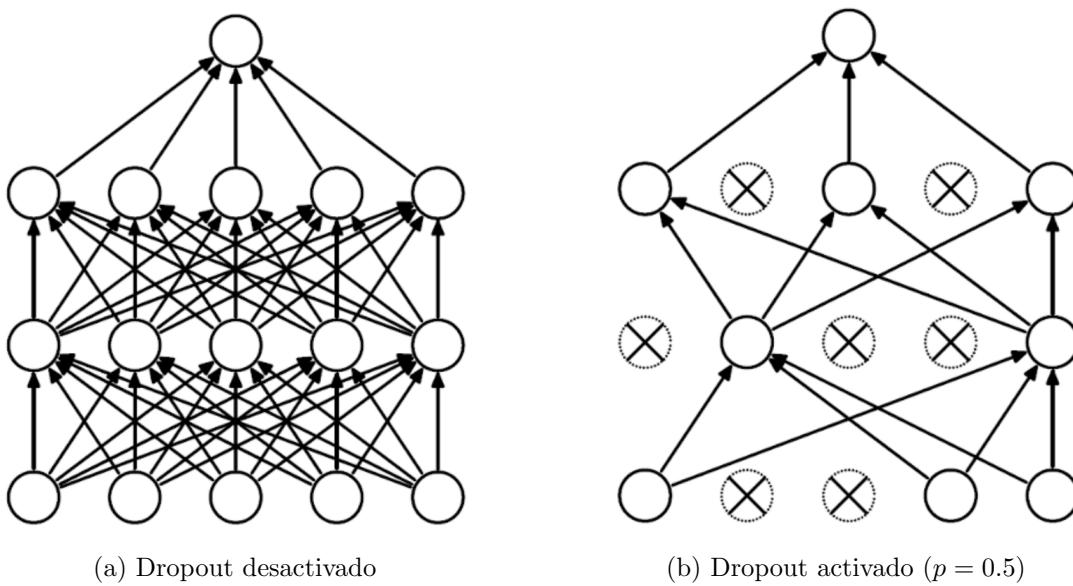


Figura 2.8: Esquema gráfico del funcionamiento de neuronas con *dropout*. Recuperado de la Figura 5.29 de [65]. Cuando se evalúa el modelo, todas las unidades funcionan correctamente (a). Durante el entrenamiento, algunas son “apagadas” (b).

- ***Batch normalization*** [75]: 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 [76] y reduce la sensibilidad a la tasa de aprendizaje [77], permitiendo establecer valores más altos. En CNNs se aplica típicamente después de las capas convolucionales y antes de la función de activación

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 2.9).

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 [29] 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.

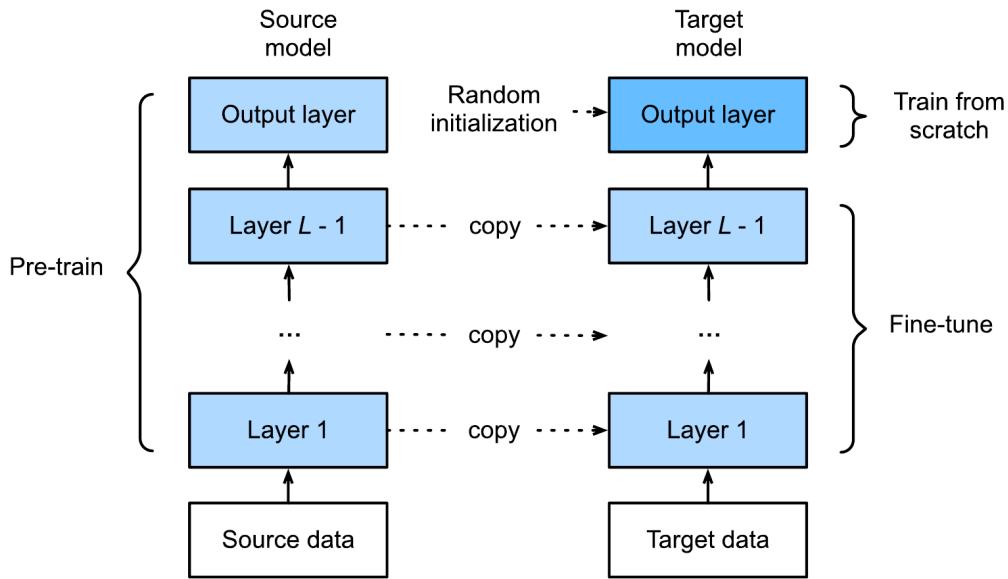


Figura 2.9: Diagrama del proceso de *fine-tuning* de un modelo de red neuronal en una nueva tarea. Recuperado de la Figura 19.2 de [60]. La capa final de salida es entrenada desde cero para la nueva tarea. El resto de capas son inicializadas con los pesos previos.

2.3. Incertidumbre

La metrología¹⁵, 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*)¹⁶ define el **error** como una “medición imperfecta” de la magnitud observada, que puede estar causada por efectos aleatorios (componente aleatoria del error) y por efectos sistemáticos (componente sistemática del error, más conocida como **sesgo**). Por otro lado, define a la **incertidumbre** como “parámetro, asociado con el resultado de una medición, que caracteriza la dispersión de los valores que podrían atribuirse razonablemente al **mensurado**, que es como se denomina a la magnitud a ser medida. [...] El parámetro puede ser, por ejemplo, una desviación estándar, o la amplitud de un intervalo con un nivel de confianza establecido” [79].

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.

¹⁵Ciencia de las mediciones y sus aplicaciones [78].

¹⁶Este 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*) [79] y el Vocabulario Internacional de Metrología (*Vocabulaire international de métrologie*) [78].

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 [79]:

- 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** [79], 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 [80, 81]:

- 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** [81, 82], 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 [83]. 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 [84-87] y la cuantificación de la incertidumbre [81, 88, 89].

Mientras la explicabilidad de la IA busca entender las razones detrás de cada predicción centrándose en el estudio del modelo y arquitectura concretos [90], 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 [81].

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 técnicas que asumen intercambiabilidad entre observaciones, frente a aquellos que no lo hacen y requieren estructuras de dependencia más complejas.

Algunas técnicas específicas de UQ en modelos ML se dividen en cuatro grupos:

- **Procesos de regresión gausiana (*gaussian process regression*)**: Modelos no paramétricos que proporcionan una distribución completa sobre funciones posibles, permitiendo obtener predicciones con intervalos de confianza naturales. Son especialmente útiles en problemas de regresión con pocos datos y cuando la estimación de incertidumbre es crítica [91].
- **Redes neuronales bayesianas**: *Markov Chain Monte Carlo* [92], *Variational Inference* [93] y *Monte Carlo Dropout* [94] incorporan incertidumbre de manera explícita en los parámetros del modelo o en la estructura de la red, generando distribuciones sobre las predicciones en lugar de valores puntuales.
- **Técnicas *ensemble***: Agrupan múltiples modelos independientes o variantes del mismo modelo para mejorar la robustez de las predicciones. La variabilidad entre los miembros del *ensemble* se utiliza como medida de incertidumbre. Ejemplos incluyen *bagging*, *boosting* y *deep ensembles* [95].

- **Métodos deterministas:** Aproximaciones que estiman la incertidumbre sin recurrir a técnicas probabilísticas, generalmente mediante reformulaciones del modelo que generan límites superiores e inferiores sobre las predicciones, como los métodos basados en *predicción conformal* [96-98].

En este trabajo exploraremos la predicción el método determinista de predicción conformal, de los métodos más flexibles que hay: mayormente *model-agnostic* y aplicable sobre cualquier distribución de datos, si bien sí asume intercambiabilidad entre observaciones. Esta es una técnica fácil de implementar y proporciona intervalos de predicción válidos para un nivel de confianza, permitiendo cuantificar la incertidumbre de manera robusta con conjuntos de datos limitados.

2.4. Predicción conformal

La **predicción conformal** (*conformal prediction*, CP) [98, 99] 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 2.10).



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 2.10: Ejemplo de predicción conformal en problemas de regresión (arriba) y clasificación (abajo). Recuperado de [100]. MAPIE es una biblioteca de Python para la cuantificación de incertidumbre, principalmente con técnicas de CP.

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 2.11. Estas garantías son válidas bajo el supuesto mínimo de intercambiabilidad de los datos, sin 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.

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 2.11: Ejemplo adversario mal clasificado por un modelo de ML entrenado con datos textuales. Adaptado de la Figura 2 de [80], original de [101]. 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.

2.4.1. Propiedades de la predicción conformal

La CP garantiza que las predicciones contengan el valor verdadero con al menos 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** [102], y se cumple para todas las entradas X , siempre y cuando los datos sean intercambiables (*interexchangeable*). Esta intercambiabilidad implica que todas fueran tomadas en condiciones similares: mismo dominio, distribución de valores de salida, iluminación, resolución, estilo, etc. Sin embargo, la CP **no asegura cobertura condicional válida** [103]; 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* [104]. En la Figura 2.12 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 ejemplos 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.

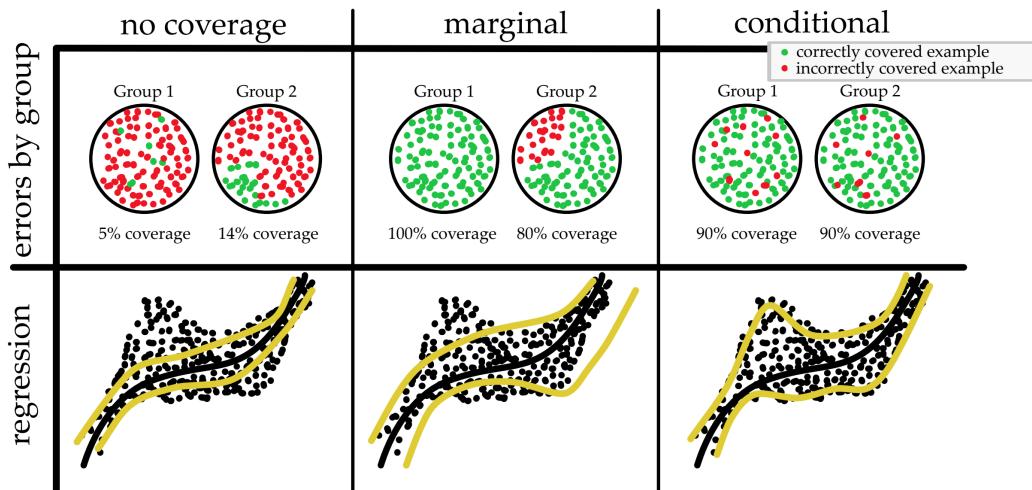


Figura 2.12: Esquema gráfico de conjuntos de predicción bajo distintas nociones de cobertura: sin cobertura garantizada, con cobertura marginal y con cobertura condicional. Recuperado de [98]. En la parte inferior de la figura se muestra la amplitud de los intervalos de predicción (en línea amarilla) 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.

Algunas características deseables en las técnicas 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, tanto datos estructurados como no estructurados. En este trabajo todos los métodos estudiados presentan esta característica.
- **Predictión adaptativa (*adaptive prediction*):** se refiere a que el intervalo o conjunto de predicción varía su tamaño en función de la incertidumbre asociada a cada predicción individual. En general, cuanto más rica y específica sea la información que la técnica utiliza sobre las predicciones y su incertidumbre, más adaptativa será la predicción conformal. Esta adaptatividad aproxima a la cobertura condicional.
- **Ser eficiente computacionalmente:** es deseable que tanto la fase de calibración como la inferencia sean rápidas y no introduzcan una sobrecarga significativa en comparación con la predicción puntual.

Aunque la mayoría de los métodos de predicción conformal presentan una sobrecarga computacional moderada, esta puede variar dependiendo del enfoque utilizado. Por ejemplo, las técnicas basadas en *split conformal* —las únicas que exploraremos en este trabajo— suelen ser más eficientes que aquellas que requieren reentrenamiento múltiple, como en algunos métodos *Cross conformal prediction* o *Jackknife+*¹⁷.

2.4.2. Algoritmo conformal

Existen multitud de técnicas de CP. Generalmente, estas dependen del tipo de problema a resolver: regresión [96, 107, 108], clasificación [97, 109, 110], series temporales [111-113]¹⁸, o detección de anomalías [114], entre otros. A pesar de su diversidad, todos los algoritmos conformales comparten un elemento clave: la definición de una **función de no conformidad** $NC(x_i, y_i)$, 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 entregar el modelo de forma habitual—, 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.
2. Se entrena el modelo predictivo con los ejemplos del conjunto de entrenamiento.
3. **Calibración conformal:** En este, se calculan las **puntuaciones de no conformidad (nonconformity scores)** R^{19} :

$$R = \{NC(x_i, y_i)\}_{i=1}^n$$

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 propias entradas o incluso representaciones internas del modelo²⁰. Bajo las garantías estadísticas que ofrece el marco teórico de la CP, esta flexibilidad muestra un gran potencial para ser integrada con otras técnicas de UQ, ampliando así sus aplicaciones y mejorando la robustez de las estimaciones de incertidumbre.

¹⁷ *Cross conformal prediction* [105] y *Jackknife+* [106] son métodos de CP que mejoran las garantías de cobertura de técnicas de CP mediante el uso más eficiente de los datos disponibles para calibración y entrenamiento, de forma análoga a la validación cruzada (*cross-validation*) o el *leave-one-out*, respectivamente.

¹⁸ El 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.

¹⁹ En la literatura, a este vector de puntuaciones se le suele denominar como R por ‘*residual*’ o E por ‘*error*’.

²⁰ Cabe señalar que una técnica 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.

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 distribuidas. 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 significación α , se selecciona el $(1 - \alpha)(1 + 1/n)$ -ésimo cuantil²¹ de las puntuaciones de no conformidad obtenidas en el conjunto de calibración (véase la Figura 2.13).

$$\delta_\alpha = \text{Quantile}_{\lceil(1-\alpha)(1+1/n)\rceil}(\{NC(x_i, y_i)\}_{i=1}^n)$$

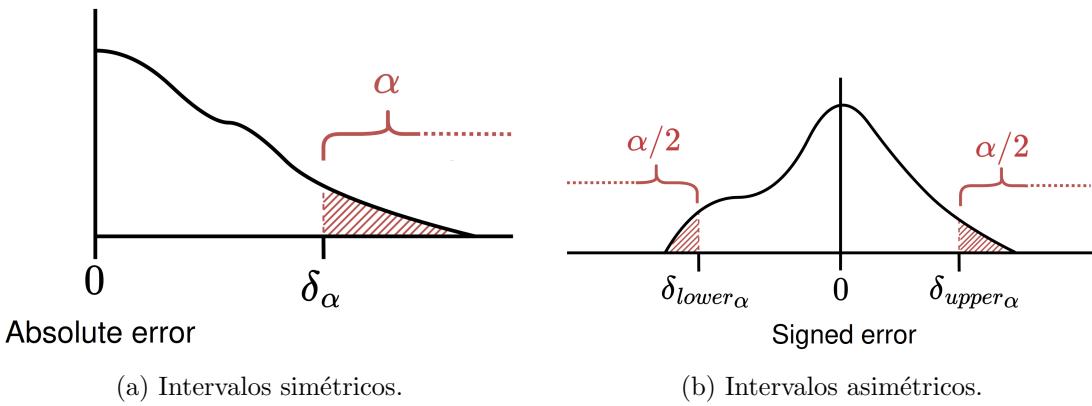


Figura 2.13: Determinación del umbral de no conformidad para intervalos simétricos y asimétricos. En (a), el error es absoluto, y el umbral se calcula como se ha especificado anteriormente. En (b), el error tiene signo, y hay dos umbrales de incertidumbre, uno por cada cola, calculado como el cuantil con significación $\alpha/2$ de los errores negativos y de los errores positivos, respectivamente para el umbral inferior y el umbral superior.

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:

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

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 la técnica *Inductive Conformal Prediction* [96] para problemas de regresión, —que describiremos en mayor profundidad en el Capítulo 4—, se utiliza el error absoluto como función de no conformidad:

$$NC(x_i, y_i) = |y_i - \hat{f}(x_i)|$$

²¹La corrección $(1 + 1/n)$ asegura validez estadística para conjuntos de tamaño finito.

El umbral de no conformidad se obtiene como el $(1 - \alpha)(1 + 1/n)$ -ésimo cuantil de las puntuaciones de no conformidad (véase la Figura 2.13a):

$$\delta_\alpha = Quantile_{\lceil(1-\alpha)(1+1/n)\rceil}(\{NC(x_i, y_i)\}_{i=1}^n)$$

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:

$$NC(x_i, y_i) = |y_i - \hat{f}(x_i)|$$

entonces al imponer la condición $NC(x_{n+1}, y_{n+1}) \leq \delta_\alpha$, se obtiene:

$$|y_{n+1} - \hat{f}(x_{n+1})| \leq \delta_\alpha$$

Despejando y_{n+1} , se obtiene:

$$\hat{f}(x_{n+1}) - \delta_\alpha \leq y_{n+1} \leq \hat{f}(x_{n+1}) + \delta_\alpha$$

Por tanto, el intervalo de predicción conformal está dado por:

$$y_{n+1} \in [\hat{f}(x_{n+1}) - \delta_\alpha, \hat{f}(x_{n+1}) + \delta_\alpha]$$

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

A continuación se presenta un ejemplo numérico para ilustrar la construcción del intervalo. Supongamos que x_i es un vector de unos determinados indicadores osteológicos de un individuo con edad cronológica conocida y_i (en años). Entrenamos un modelo de regresión \hat{f} en un conjunto de entrenamiento y reservamos un conjunto de calibración de n individuos identificados. Se supone un conjunto de calibración de $n = 50$ individuos y $\alpha = 0.10$. Calculamos los residuos absolutos $NC(x_i, y_i)$ en calibración y obtenemos su cuantil, $(1 - \alpha)(1 + 1/n) = 0.90 \times 1.02 = 0.918$ (percentil 91.8). Si ese cuantil es $\delta_{0.10} = 3.3$ años y para un caso nuevo $\hat{f}(x_{n+1}) = 34.7$ años, entonces:

$$y_{n+1} \in [34.7 - 3.3, 34.7 + 3.3] = [31.4, 38.0] \text{ años}$$

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. El número de publicaciones que usan técnicas de antropología para la estimación de edad ha aumentado continuamente en las últimas décadas (véase la Figura 3.1¹), lo que indica un interés creciente en la antropología forense por métodos más precisos y confiables de estimación de edad.

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.

Las técnicas de estimación de edad presentan diferencias significativas en individuos maduros e inmaduros [115]. 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 dental² [116], los tiempo de aparición y cambios en la morfología de centros de osificación³, y los tiempos de fusión de los centros primarios (también denominados diáfisis) y secundarios (epífisis) [117, 118]. 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 [119]. En

¹La query usada ha sido TITLE-ABS-KEY ((age AND (estimation OR assessment) AND (bone OR anthropology OR skeleton))) el día 21 de agosto de 2025.

²La 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.

³La 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 comienza el proceso de formación ósea durante el desarrollo embrionario, fetal, infantil y adolescente.

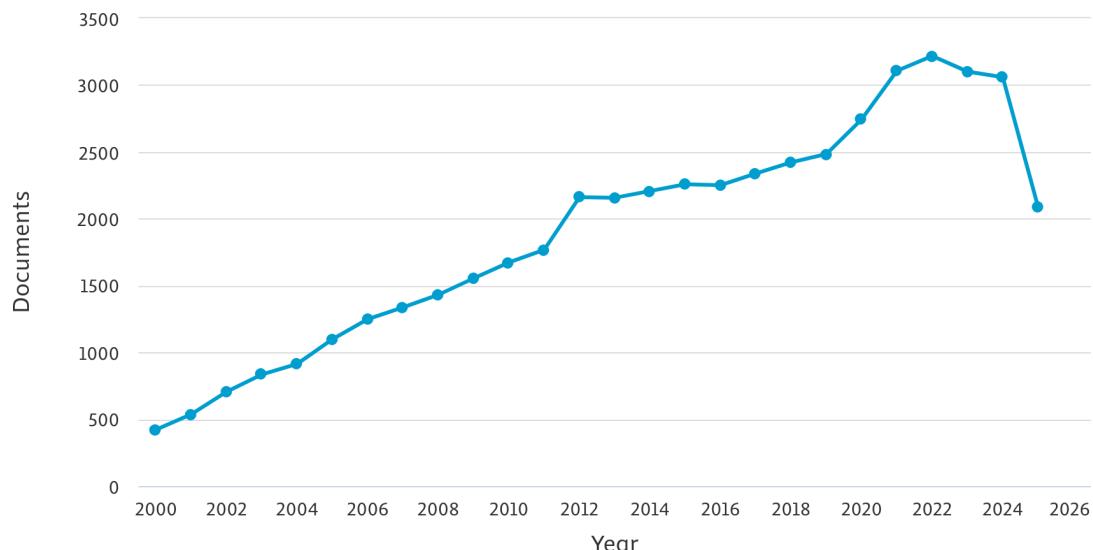


Figura 3.1: Evolución de las publicaciones que usan técnicas de antropología para la estimación de edad. Información recuperada de Scopus.

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 [118].

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 ambientales⁴ [122, 123]. Actualmente, se recomienda un análisis conjunto del proceso degenerativo de síntesis pública (propuesto en [124]) y de las transparencias en las raíces dentales [125]. Cuando este no es posible, pueden emplearse otros métodos [126], como el análisis de la superficie del ilion [127], el examen del extremo esternal de la cuarta costilla [128], o el estudio de los procesos de obliteración de las suturas craneales [129].

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 Legal⁵ 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 3.2).

⁴Por ejemplo, artículos como [120, 121] 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.

⁵La 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.

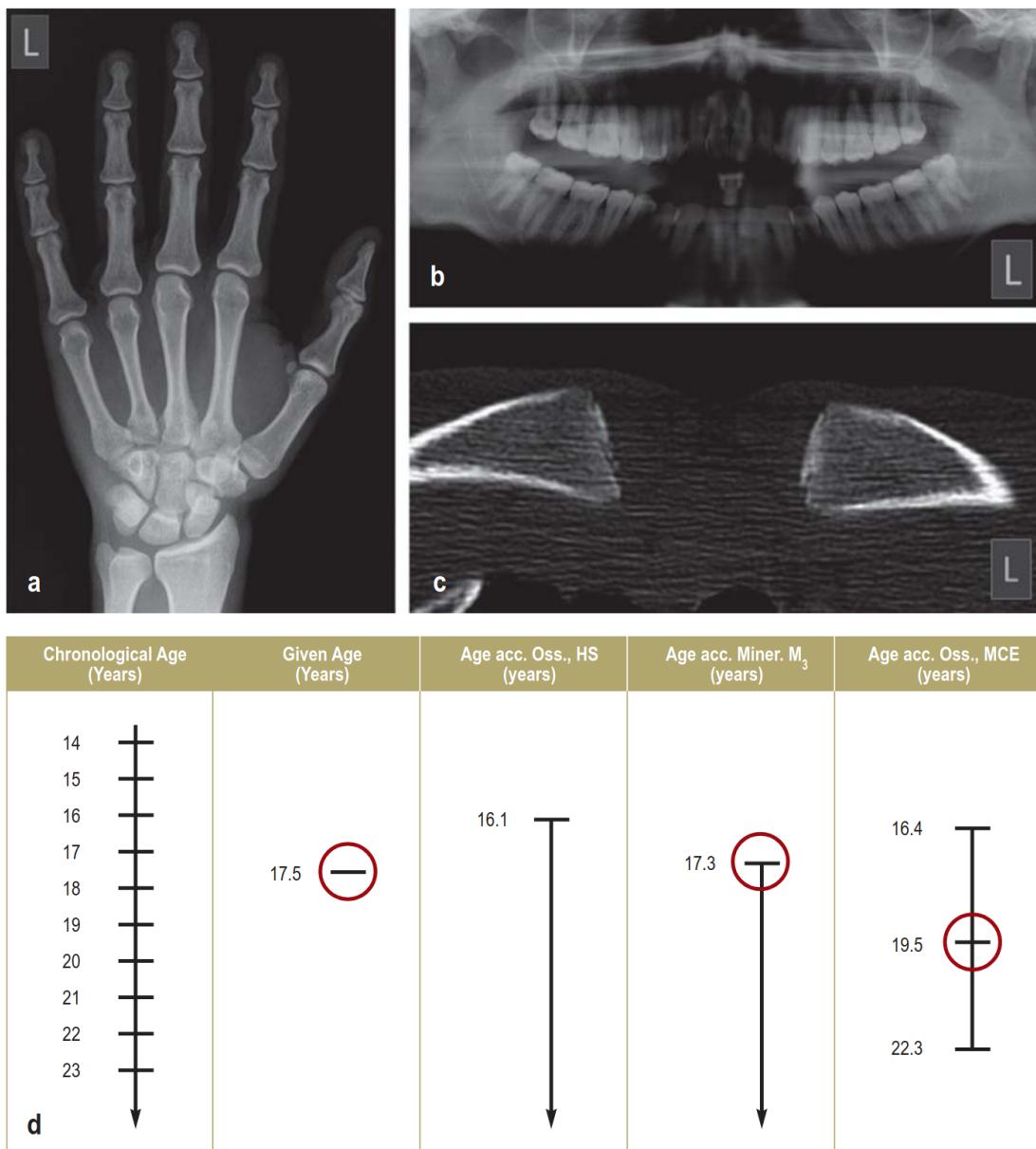


Figura 3.2: Hallazgos radiológicos en un posible menor con edad en disputa. Criterio de edad mínima para la determinación de edad. Recuperado de la Figura 1 de [21]. El sujeto masculino afirma tener 17.5 años. Tanto la historia clínica como el examen físico no revelan signos de alteraciones en el desarrollo. 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*

En la última década, ha aumentado el uso de técnicas de antropología forense combinadas con métodos de ML para la estimación de la edad, como se ilustra en la Figura 3.3⁶.



Figura 3.3: Evolución de las publicaciones que usan técnicas de antropología y métodos ML para la estimación de edad. Información recuperada de Scopus.

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 [130], 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 3.4 ilustra un ejemplo de este enfoque. Entre las propuestas destacadas podemos mencionar BoneXpert [131], que empleaba técnicas clásicas de ML y demostró robustez en poblaciones diversas, tanto en origen geográfico como en condiciones clínicas de adquisición de imágenes [132-134].

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

⁶La query usada ha sido TITLE-ABS-KEY ((age AND estimation AND (bone OR anthropology OR skeleton)) AND ((deep AND learning) OR (artificial AND intelligence))) el día 21 de agosto de 2025.

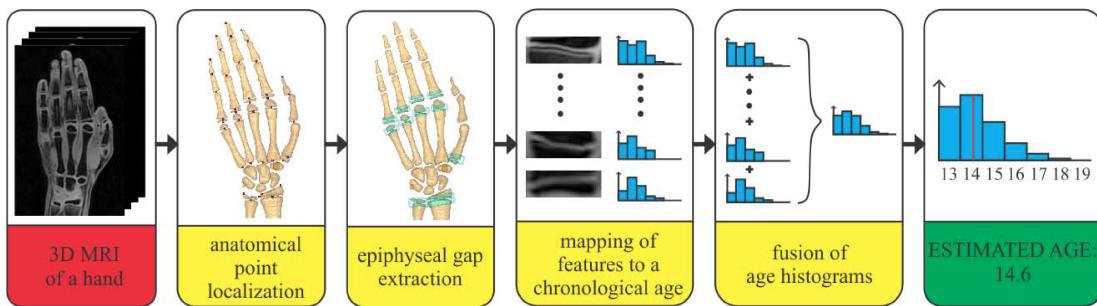


Figura 3.4: Procedimiento secuencial clásico de ML para la extracción de características antropológicas, propuesto en [135]. Las características se extraen manualmente o con herramientas independientes del modelo.

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 entrena.

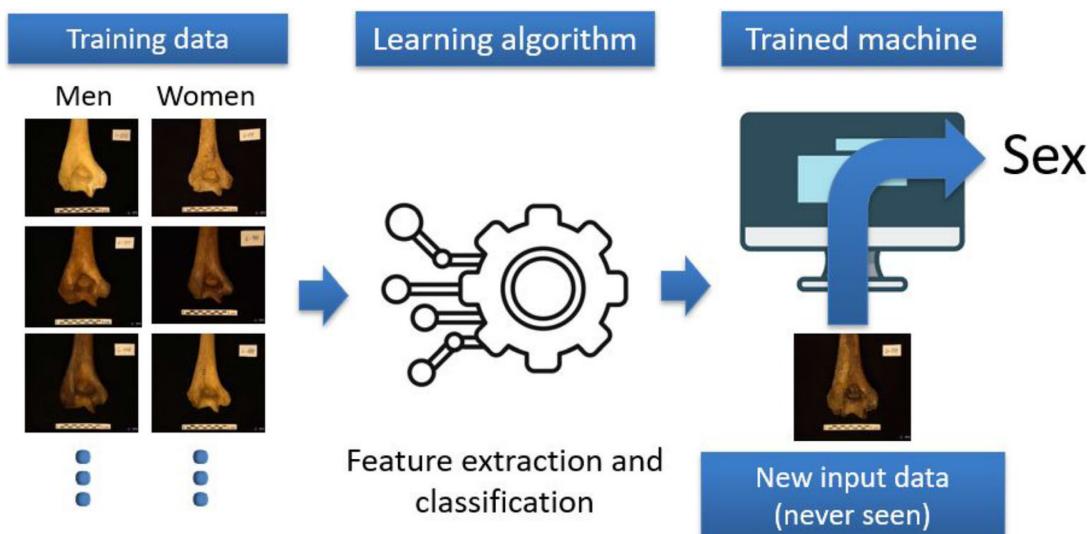


Figura 3.5: Metodología de construcción de un modelo *end-to-end*. Recuperado de [29]. La CNN aprende a extraer las características de la manera más conveniente para resolver el problema en el que se le entrena, en este caso estimación de sexo.

Siguiendo este paradigma, se han propuesto numerosos modelos basados en redes convolucionales. Un ejemplo destacado es el propuesto en [136], 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 años⁷. Además, los autores demostraron su adaptabilidad al procesar imágenes 2D de radiografías, logrando también un rendimiento líder en el ámbito de rayos X.

⁷Esta 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

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.

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 [137].

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 [126] (véase un ejemplo en la Figura 3.6).

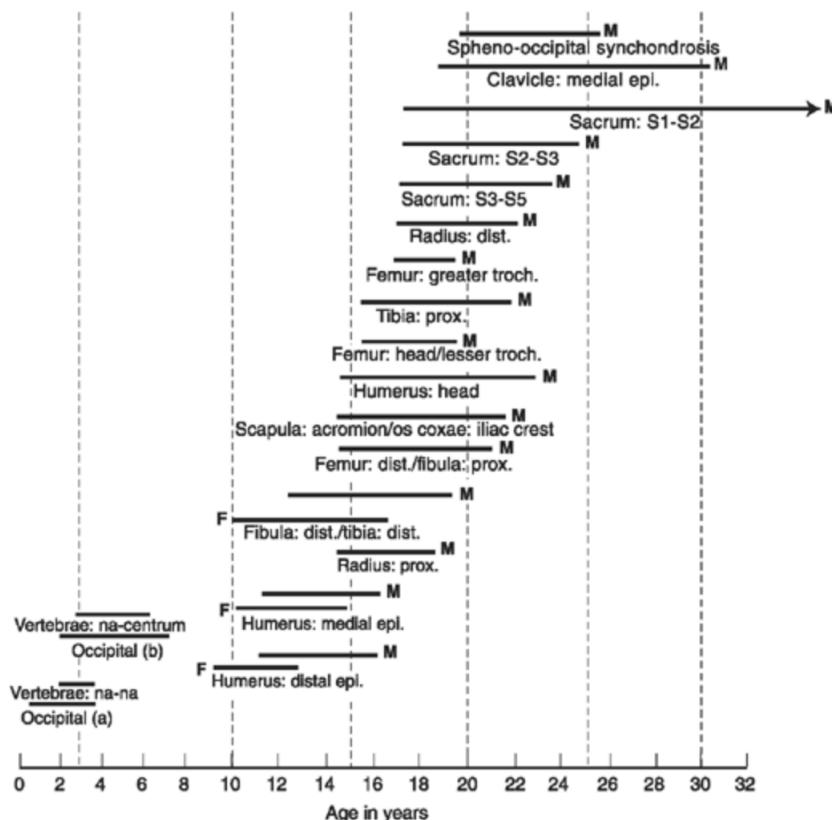


Figura 3.6: Cronograma de desarrollo de la unión epifisaria. Recuperado de [3], original de [138]. Este gráfico combina información de diferentes fuentes sobre los rangos de edades en los que ocurre la fusión de diversas epífisis del esqueleto humano, representado con una línea que indica la variabilidad 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.

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 etarios⁸, 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 3.7).
- Histogramas, gráficos de densidad o diagramas de cajas para representar la distribución de errores. En la Figura 3.8 vemos un ejemplo en el que plasman un histograma escrito como texto.

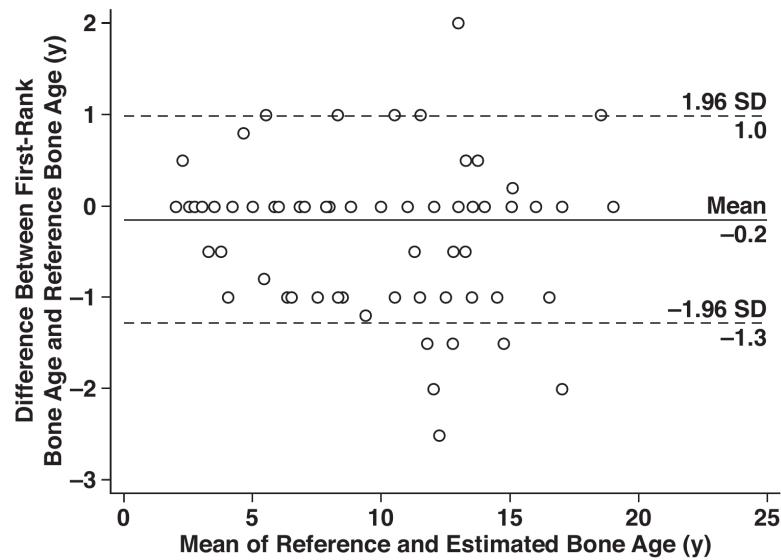


Figura 3.7: Distribución del error por edad real para el modelo propuesto en [31]. Recuperado de la Figura 3. Esta visualización permite observar el error en cada instancia predicha para diferentes edades reales.

⁸Los 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.

		BAR1 → BAR1 (mean \pm std [years]) error histogram [years] (-3 -2 -1 0 +1 +2 +3)	BAR2 → BAR2 (mean \pm std [years]) error histogram [years] (-3 -2 -1 0 +1 +2 +3)	CA → CA (mean \pm 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	0.54 ± 0.60 2 9 102 214 82 9 0	0.57 \pm 0.60 0 11 89 197 68 4 0	0.86 ± 0.63 6 28 95 140 93 21 0
	cnn-2d-bones-carpal	0.49 \pm 0.60 1 13 66 238 92 8 0	0.57 \pm 0.60 1 3 57 198 100 11 0	0.77 ± 0.60 0 23 89 157 93 20 0
	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	0.70 ± 0.61 1 13 105 157 124 17 0	0.68 ± 0.65 3 5 73 173 137 25 1	1.00 ± 0.73 5 20 80 124 107 33 7
	cnn-2d-bones-carpal	0.66 \pm 0.61 0 7 80 174 132 24 0	0.60 \pm 0.62 3 8 102 194 97 13 0	0.90 ± 0.70 2 29 67 153 92 29 5
	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	0.62 ± 0.61 3 22 207 371 206 26 0	0.62 ± 0.63 3 16 162 370 205 29 1	0.93 ± 0.69 11 48 175 264 200 54 7
	cnn-2d-bones-carpal	0.57 \pm 0.61 1 20 146 412 224 32 0	0.58 \pm 0.61 4 11 159 392 197 24 0	0.83 ± 0.66 2 52 156 310 185 49 5
	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 3.8: Estudio del error en los métodos 2D propuestos en [136]. Recuperado de la Tabla 2 de [136]. 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. El *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.

Hay un total de 10 739 ejemplos, de los que 5756 son de individuos de sexo femenino y 4983 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 esta figura también podemos comprobar la distribución de ejemplos en cada edad por sexo, y observamos que hay ligeramente más ejemplos del sexo femenino que del masculino para prácticamente todas las edades.

¹Los modelos empleados fueron: *Planmeca Promax Digital Panoramic*; *Sirona ORTHOPHOS-XG*, *ORTHOPHOS-DS*, y *SIDEXIS*. Las constantes radiológicas usadas fueron de 66 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.

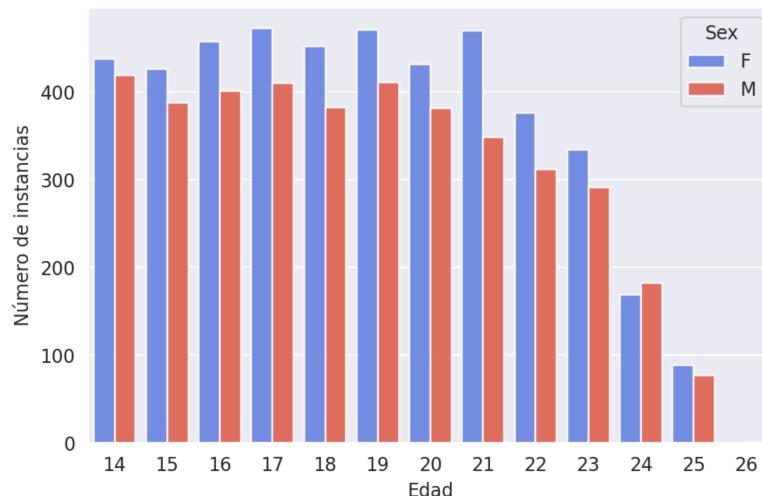


Figura 4.1: Histograma de edades de los individuos del conjunto de datos disponible diferenciado por sexo.

En conclusión, el dataset no presenta desbalances extremos, lo que permite un análisis representativo de la población incluida. No obstante, será necesario examinar con mayor detalle la infrarrepresentación de la población 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. Métodos propuestos

La implementación de todos los métodos a continuación expuestos se encuentra disponible en el siguiente repositorio: [esdavide2910/tfg-bioprofile-uncertainty](https://github.com/esdavide2910/tfg-bioprofile-uncertainty).

4.2.1. Arquitectura empleada

Se avanza que los problemas planteados en el siguiente capítulo parten de imágenes bidimensionales de las radiografías panorámicas maxilofaciales como entrada, y estimación de edad o sexo a la salida.

Como modelo, se ha decidido emplear una CNN, dado su buen desempeño en tareas de visión por computador. Específicamente, se ha optado por la arquitectura ResNeXt50 [139] preentrenada en Imagenet [140] 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.

4.2.2. 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 solo en arquitectura 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* [141]. 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 4.2 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.

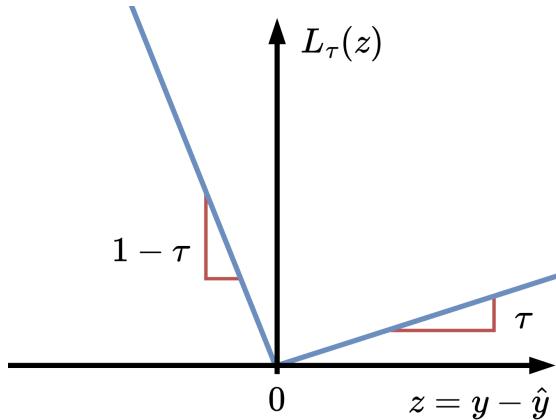


Figura 4.2: Visualización de la función de pérdida *pinball* para cada valor de error. Adaptado de la Figura 1 de [107]. Esta concretamente muestra la función de pérdida para un cuantil cercano a cero, ya que es más permisivo con los errores positivos que con los negativos, lo cual empujará sus predicciones hacia la parte inferior de la distribución objetivo.

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),

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:

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

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 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 [107]. 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.

4.2.3. 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 [106] 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 (también conocida como *Split Conformal Prediction*) [96] 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:

$$NC(x_i, y_i) = |y_i - \hat{f}(x_i)|$$

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

$$\delta_\alpha = Quantile_{\lceil(1-\alpha)(1+1/n)\rceil}(\{NC(x_i, y_i)\}_{i=1}^n)$$

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

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

Este método de CP presenta varias ventajas:

- **Model-agnostic:** Es completamente independiente del modelo y su arquitectura, ya que no utiliza representaciones internas del modelo y solo emplea una salida.
- **Bajo coste computacional:** El método introduce una fase adicional tras el entrenamiento, denominada calibración, en la que se calculan las puntuaciones de no conformidad sobre el conjunto de calibración ($O(n_{calib})$) y se determina el umbral correspondiente mediante el cálculo de cuantiles ($O(n_{calib} \log n_{calib})$). En consecuencia, el coste global es $O(n_{calib} \log n_{calib})$. Aun cuando este orden de complejidad es superior al lineal, en la práctica el tiempo requerido resulta despreciable en comparación con el entrenamiento de modelos complejos, como *ensembles* o DNN. La inferencia conformal no introduce ningún cambio de orden ($O(1)$), ya que tan solo calcula la puntuación de no conformidad de la nueva instancia para compararla con el umbral.

Sin embargo, también presenta una importante limitación: los **intervalos generados son simétricos y no adaptativos**, es decir, todos tienen el mismo ancho ($2q_{1-\alpha}$), no permitiéndose adaptarse a la incertidumbre específica de cada predicción.

Conformalized Quantile Regression (CQR)

Como su nombre indica, este método se realiza sobre la regresión cuantílica. La CQR [107] 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 [107], que incluye la calibración de ambas colas para obtener intervalos asimétricos [142]. 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 las puntuaciones de no conformidad sobre los datos del conjunto de calibración como las diferencias entre los valores observados y los límites del intervalo predictivo:

$$\begin{aligned} NC_{lower}(x_i, y_i) &= \hat{q}_{lower}(x_i) - y_i \\ NC_{upper}(x_i, y_i) &= y_i - \hat{q}_{upper}(x_i) \end{aligned}$$

donde $\hat{q}_{upper}(x_i)$ y $\hat{q}_{lower}(x_i)$ 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 :

$$\delta_{lower_\alpha} = Quantile_{\lceil(1-\alpha)(1+1/n)\rceil}(\{NC_{lower}(x_i, y_i)\}_{i=1}^n)$$

$$\delta_{upper_\alpha} = Quantile_{\lceil(1-\alpha)(1+1/n)\rceil}(\{NC_{upper}(x_i, y_i)\}_{i=1}^n)$$

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

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

En comparación con ICP:

- CQR no es **independiente del modelo**, ya que al implementar QR, requiere de que la arquitectura sea una red neuronal o un modelo *ensemble*. Por tanto, además, requiere de reentrenamiento (a no ser que se partiera de un modelo ya con QR implementada).
- CQR comparte un **similar coste 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.
- 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.

Para cada método se han entrenado 10 modelos independientes desde cero, con el objetivo de capturar la variabilidad inherente al proceso de entrenamiento.

En la Tabla 4.2 observamos un cuadro comparativo de los distintos métodos presentados para la estimación interválica.

4.2.4. Calibración de probabilidades en clasificación

Los modelos de clasificación producen puntuaciones discriminativas, como *logits* o distancias a un hiperplano, pero no siempre reflejan probabilidades reales para la predicción. La **calibración de probabilidad** en problemas de clasificación en ML se refiere al ajuste de estas puntuaciones de salida para que reflejen más fielmente probabilidades de predicción.

La calibración de probabilidades busca que cuando un modelo diga “X % de probabilidad”, eso corresponda de verdad a la frecuencia observada en los datos. Por ejemplo, si un modelo asigna una probabilidad del 70 % a un conjunto de instancias, aproximadamente 7 de cada 10 deberían pertenecer a la clase positiva si el modelo está bien calibrado.

Existen distintas técnicas de post-procesamiento para mejorar la calibración. En este trabajo se empleará **Temperature Scaling** [143], algoritmo usado principalmente en redes neuronales profundas. Este ajusta los *logits* dividiéndolos por un parámetro de temperatura $T > 0$, de manera que:

Característica	ICP	QR	CQR
Model-agnostic	Sí	Sí	Sí
Cobertura marginal global garantizada	Sí	No	Sí
Cobertura condicional global garantizada	No	No	No
Intervalos adaptativos	No	Sí	Sí
Coste calibración	$O(n \log(n))$	–	$O(n \log(n))$
Coste inferencia (por predicción)	$O(1)$	$O(1)$	$O(1)$

Tabla 4.2: Comparativa de métodos propuestos para problemas de regresión. Las técnicas de CP son las únicas que garantizan cobertura marginal. Se recuerda que no existe ninguna técnica que garantice la cobertura global condicional.

- Si $T > 1$, las puntuaciones se suavizan, disminuyendo la confianza excesiva del modelo.
- Si $T < 1$, las puntuaciones se vuelven más extremas, aumentando la confianza.

El parámetro T se aprende sobre un conjunto de validación independiente, optimizando la entropía cruzada de las predicciones ajustadas con respecto a las etiquetas reales. De esta manera, se mejora la calibración sin afectar la discriminación del modelo, es decir, las clases más probables siguen siendo las mismas.

Entre sus ventajas destacan:

- Muy simple: un solo parámetro.
- Mantiene la discriminación de las clases.
- Efectivo para corregir sobreconfianza típica en redes profundas.

Se ha escogido este algoritmo ya que es el utilizado en trabajos recientes sobre técnicas de Conformal Prediction para clasificación [110, 144].

4.2.5. Métodos de predicción conformal para clasificación

Least-Ambiguous set-valued Classifiers (LAC)

LAC [97] 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** [145], 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:

$$NC(x_i, y_i) = 1 - \hat{\pi}_{y_i}(x_i)$$

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 las puntuaciones de no conformidad:

$$\delta_\alpha = Quantile_{\lceil(1-\alpha)(1+1/n)\rceil}(\{NC(x_i, y_i)\}_{i=1}^n)$$

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 (véase la Figura 4.3):

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

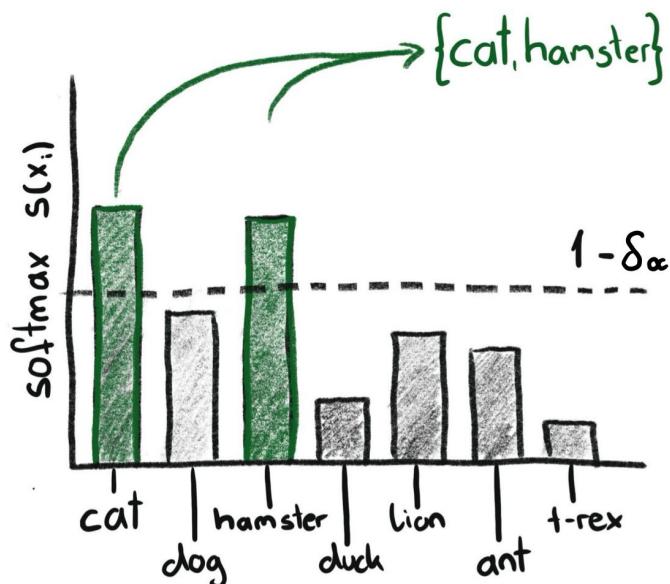


Figura 4.3: Ejemplo de selección de clases para el conjunto conformal en LAC. Se escogen aquellas clases que presentan una puntuación softmax mayor a $1 - \delta_\alpha$. Adaptado de [146].

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

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.

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

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. Este enfoque no ajusta dinámicamente el tamaño del conjunto de predicción en función de la incertidumbre del modelo para cada instancia.
- **Bajo coste computacional:** Al igual que pasaba en regresión, solo añade coste computacional en la calibración, con el cálculo de puntuaciones de no conformidad ($O(n_{calib})$) y la obtención del umbral de no conformidad ($O(n_{calib} \log n_{calib})$). No añade coste a la inferencia ($O(1)$).

Mondrian Confidence Machine (MCM)

MCM [147] 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 y los umbrales de no conformidad de forma independiente para cada una (véase la Figura 4.4).

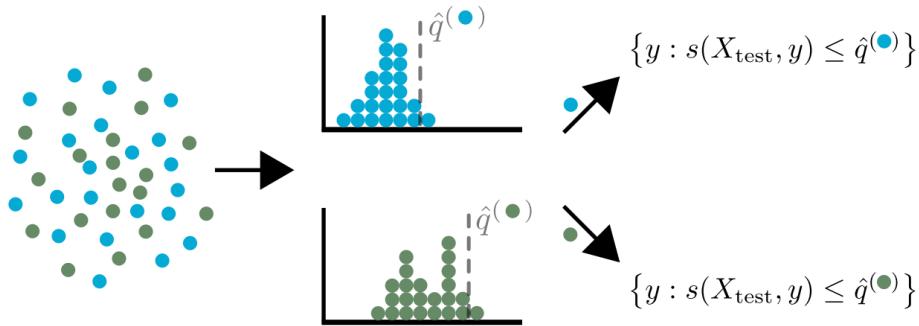


Figura 4.4: Diagrama ilustrativo de la división de ejemplos utilizada en MCM. Para cada clase se calcula un umbral de no conformidad a partir de las puntuaciones de no conformidad de los ejemplos pertenecientes a esa clase.

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 APS, 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.

- **Coste computacional comparable a LAC:** Al igual que LAC, MCM calcula las puntuaciones de no conformidad para todas las instancias del conjunto de calibración. Posteriormente, para cada clase se ordenan las puntuaciones de su subconjunto para determinar el umbral correspondiente. Si hay K clases y n_k instancias por clase, el coste total de esta fase es $\sum_{k=1}^K O(n_k \log n_k)$, que en la práctica se aproxima a $O(n_{calib} \log n_{calib})$, manteniendo la complejidad algorítmica de LAC. En la fase de inferencia, la predicción para una nueva instancia requiere únicamente calcular la puntuación de no conformidad y compararla con el umbral de su clase, lo que supone un coste lineal respecto al número de clases, $O(K)$, que se sigue traduciendo en tiempos prácticamente despreciables para la inmensa mayoría de problemas. Por tanto, MCM sigue siendo eficiente.

Adaptive Prediction Sets (APS)

APS [109], 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:

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

donde (k) es el índice de la clase con la k mayor probabilidad, la función de no conformidad se define como:

$$NC(x_i, y_i) = \sum_{j=1}^k \hat{\pi}_{(j)}(x_i) \text{ donde } (k) = y_i$$

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 multiclas, 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:

$$\delta_\alpha = Quantile_{\lceil (1-\alpha)(1+1/n) \rceil}(\{NC(x_i, y_i)\}_{i=1}^n)$$

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:

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

Componente aleatoria Una extensión útil de APS consiste en introducir una componente aleatoria durante la fase de calibración para compensar la tendencia del método a la *sobrecolección*. En este ajuste, cuando se calcula la puntuación de no conformidad para una instancia de calibración, se decide excluir la contribución de la última clase incluida (aquella correspondiente a la etiqueta verdadera y_i) con una probabilidad proporcional al exceso de cobertura —respecto al nivel de confianza $1 - \alpha$ — que dicha clase genera.

Formalmente, sea:

$$V_i = \frac{NC(x_i, y_i) - (1 - \alpha)}{\hat{\pi}_{(k)}(x_i)} \quad \text{con} \quad (k) = y_i$$

donde A es la puntuación de no conformidad estándar de APS y $\hat{\pi}_{(k)}$ la probabilidad de la última clase incluida. Se genera un valor aleatorio $u_i \sim \mathcal{U}(0, 1)$ y, si $u_i > V_i$ y la clase verdadera no está en primera posición ($k \geq 2$), se elimina completamente la contribución de dicha clase de la suma acumulada. En caso contrario, se mantiene la puntuación original.

De este modo, las instancias en las que la última clase aporta un exceso notable de cobertura tienen mayor probabilidad de ser recortadas, reduciendo así el umbral de no conformidad δ_α y produciendo conjuntos de predicción más pequeños, sin comprometer significativamente la cobertura global. En la fase de inferencia, el procedimiento de construcción de $\Gamma_\alpha(x_{n+1})$ se mantiene idéntico, pero el umbral reducido provoca que se necesiten menos clases para alcanzar el criterio, generando así **conjuntos de predicción más pequeños** en promedio, y que siguen garantizando estadísticamente la cobertura marginal.

Las principales características de APS son las siguientes:

- Este algoritmo 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 que los de LAC** [109, 110]. Este fenómeno es un *trade-off* inherente al intentar aproximar la cobertura condicional sin asumir distribuciones subyacentes.

- El **coste computacional resulta más elevado que los dos algoritmos previos**. Durante la fase de calibración, cada ejemplo requiere ordenar las probabilidades de las clases para calcular su puntuación de no conformidad, lo que implica un coste total de $O(n_{calib} \cdot K \log K)$ donde K es el número total de clases. En la fase de inferencia, el procedimiento se repite para cada nueva instancia: se ordenan las clases por probabilidad y se van acumulando hasta superar el cuantil de

calibración, alcanzando un coste de $O(K \log K)$ por predicción. Aunque este orden es superior al del método clásico, en la práctica sigue siendo asumible para valores moderados de K .

Regularized Adaptive Prediction Sets (RAPS)

RAPS [110] es una variante del método APS, que introduce modificaciones clave para reducir el tamaño de los conjuntos de predicción, especialmente en escenarios con muchas clases, donde APS tiende a generar conjuntos excesivamente grandes. El objetivo principal de RAPS es mantener la propiedad de cobertura marginal deseada, al tiempo que se obtienen conjuntos de predicción más compactos y útiles en la práctica.

RAPS extiende la función de no conformidad utilizada en APS mediante la incorporación de un término de regularización que penaliza explícitamente la inclusión de clases con baja probabilidad en conjuntos de predicciones de tamaño ya elevado.

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

- k_{reg} representa el tamaño mínimo del conjunto de predicción a partir del cual se comenzará a aplicar penalización. Es decir, los conjuntos de predicción de tamaño menor o igual a k_{reg} no serán penalizados, ya que se asume que, si todos los conjuntos tuvieran como máximo ese tamaño, la cobertura marginal aún se mantendría.

El valor de este hiperparámetro se determina empíricamente observando, en el conjunto de calibración, cuál es el menor tamaño de conjunto que cumple con la cobertura deseada en una fracción suficientemente alta de las instancias.

- λ , un parámetro de regularización que penalizará más a aquellos conjuntos que superen k_{reg} etiquetas predichas cuanto mayor número de etiquetas tengan.

Este hiperparámetro se determina típicamente a través de validación en un conjunto de datos independiente al de calibración, mediante búsqueda de hiperparámetros que minimicen el tamaño medio del conjunto de predicción sin comprometer significativamente la cobertura marginal. En la práctica, se suele probar con varios valores posibles para λ y seleccionar el que logre el mejor equilibrio entre concisión y cobertura en el conjunto de validación.

Una vez determinados los valores de estos hiperparámetros, se calculan las puntuaciones de no conformidad de la siguiente manera:

$$NC(x_i, y_i) = \sum_{j=1}^k \hat{\pi}_{(j)}(x_i) + \lambda(k - k_{reg})^+ \text{ donde } (k) = y_i$$

El procedimiento de calibración y predicción en RAPS sigue la misma estructura general que APS, pero utiliza la función de no conformidad regularizada en lugar de la acumulación pura de probabilidades. Así, el umbral calibrado se calcula como:

$$\delta_\alpha = Quantile_{\lceil(1-\alpha)(1+1/n)\rceil} (\{NC(x_i, y_i)\}_{i=1}^n)$$

Y el conjunto de predicción para una nueva instancia x_{n+1} se construye como

$$\Gamma_\alpha(x_{n+1}) = \{(1), \dots, (k)\} \text{ donde}$$

$$k = \min \left\{ j : \sum_{i=1}^j \pi_{(i)}(x_{n+1}) + \lambda(k - k_{reg})^+ \geq \delta_\alpha \right\}$$

Gracias a la regularización, RAPS tiende a generar **conjuntos de predicción más pequeños que APS**, especialmente cuando la clase verdadera se encuentra entre las más probables (y por tanto el hiperparámetro k_{reg} tiene un valor bajo).

Al igual que APS, también admite una **componente aleatoria**, que funciona de igual manera: durante la fase de calibración, se calcula para cada instancia la probabilidad de excluir la contribución de la última clase incluida —aquella correspondiente a la etiqueta verdadera— considerando también el término de regularización, en función del exceso de cobertura que produce respecto al nivel nominal $1 - \alpha$.

Al tratarse de una extensión de APS, RAPS tiene un **orden algorítmico idéntico a APS**, aunque en la práctica RAPS puede ser ligeramente más costoso debido al paso extra de regularización.

Sorted Adaptive Prediction Sets (SAPS)

SAPS [144] propone un enfoque distinto a métodos previos como APS y RAPS. Los autores identifican una limitación importante en estos algoritmos: las probabilidades producidas por la capa *softmax* suelen seguir una distribución con cola larga, lo que facilita la inclusión de clases poco probables en los conjuntos de predicción. Esto lleva a la generación de conjuntos innecesariamente grandes, que reducen la utilidad práctica del método.

SAPS argumenta que muchas de estas probabilidades de baja magnitud representan información redundante o poco útil para la tarea de predicción conforme. En lugar de utilizar todo el vector de probabilidades, propone construir los conjuntos de predicción únicamente a partir de dos elementos clave:

- la probabilidad más alta (asociada a la clase predicha como más probable), y
- el orden de clasificación de las clases según el modelo.

A partir de esta representación reducida, las clases se ordenan por probabilidad decreciente y se define una función de no conformidad que combina la probabilidad máxima con la posición k de la clase verdadera en el ranking. El objetivo es que la inclusión de una clase dependa más de su relevancia relativa que de su probabilidad absoluta, evitando el efecto adverso de las colas largas.

A diferencia de los anteriores métodos adaptativos, este método incluye la **componente aleatoria** desde su planteamiento inicial. La función de no conformidad propuesta es:

$$NC(x_i, y_i) = \hat{\pi}_{max}(x_i) + \lambda(k - 2 + u_i)$$

donde:

- $\hat{\pi}_{\max}(x_i)$ corresponde a la probabilidad más alta predicha por el modelo,

- λ es un parámetro de regularización que incrementa la penalización a medida que la clase verdadera se aleja de las primeras posiciones del ranking,
- k indica la posición de y_i en el ranking de salida (con $k = 1$ para la clase más probable, $k = 2$ para la segunda, y así sucesivamente),
- $u_i \sim \mathcal{U}(0, 1)$ introduce un componente aleatorio continuo que suaviza la penalización y evita empates exactos.

Observe que la penalización efectiva se aplica sobre el término $k - 2 + u_i$: no se penaliza la clase en primera posición ($k = 1$), mientras que para las siguientes posiciones la penalización crece de forma lineal. La inclusión de u_i hace que la penalización de la última clase considerada se distribuya de manera uniforme en el intervalo $(k - 1, k)$, lo que garantiza aleatoriedad controlada y una calibración más estable.

El umbral de no conformidad δ_α se calcula a partir de las puntuaciones $NC(x_i, y_i)$ en el conjunto de calibración, de forma análoga a APS y RAPS:

$$\delta_\alpha = Quantile_{\lceil(1-\alpha)(1+1/n)\rceil} (\{NC(x_i, y_i)\}_{i=1}^n)$$

En la fase de inferencia, el conjunto de predicción para una nueva instancia x_{n+1} se construye añadiendo las clases en orden decreciente de probabilidad, calculando para cada una la puntuación de no conformidad y deteniéndose cuando esta supera el umbral δ_α :

$$\begin{aligned} \Gamma_\alpha(x_{n+1}) &= \{(1), \dots, (k)\} \text{ donde} \\ k &= \min \{j : \hat{\pi}_{max}(x_{n+1}) + \lambda(j - 1) \geq \delta_\alpha\} \end{aligned}$$

Gracias a esta estrategia basada en orden y probabilidad máxima, SAPS intenta generar conjuntos de predicción más compactos que APS y RAPS en escenarios con distribuciones de salida muy sesgadas, manteniendo la cobertura marginal deseada y reduciendo la inclusión de clases irrelevantes.

El **coste computacional es idéntico al de APS** tanto en orden algorítmico como en las operaciones realizadas con las implementaciones realizadas para este trabajo.

En la Tabla 4.3 observamos un cuadro comparativo de los distintos métodos presentados para la estimación de conjuntos predictivos.

Característica	LAC	MCM	(R/S)APS
Model-agnostic	Sí	Sí	Sí
Cobertura marginal global garantizada	Sí	Sí	Sí
Cobertura condicional por clase garantizada	No	Sí	No
Conjuntos de tamaño variable	Sí	Sí	Sí
Conjuntos adaptativos	No	No	Sí
Coste calibración	$O(n \log(n))$	$O(n \log(n))$	$O(nK \log(K))$
Coste inferencia (por predicción)	$O(1)$	$O(1)$	$O(K \log(K))$

Tabla 4.3: Comparativa de métodos propuestos para problemas de clasificación.

Capítulo 5

Experimentación

5.1. Problemas propuestos

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: El problema de **estimación de edad (*age estimation*)** 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 5.1). 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 simplificaba la arquitectura.

2. Estimación de la mayoría de edad: Un problema inmediatamente derivado del anterior es la **estimación de mayoría de edad (*age majority estimation*)**, útil en contextos legales donde es necesario determinar si una persona ha alcanzado 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”.
3. **Estimación de la edad legal resuelta como un problema de clasificación multiclasa:** Se propone un problema de estimación de edad, pero planteado como problema de clasificación multiclasa, donde cada edad —como valor entero— es una clase independiente. El potencial para aunar el planteamiento de un problema de regresión con uno de clasificación viene de la mano de la CP, que, aplicada al problema de clasificación, permite generar conjuntos de etiquetas que toleran la cercanía entre clases, de forma que errores pequeños en el valor predicho (por ejemplo, predecir 19 en lugar de 20) no se consideren fallos completos.

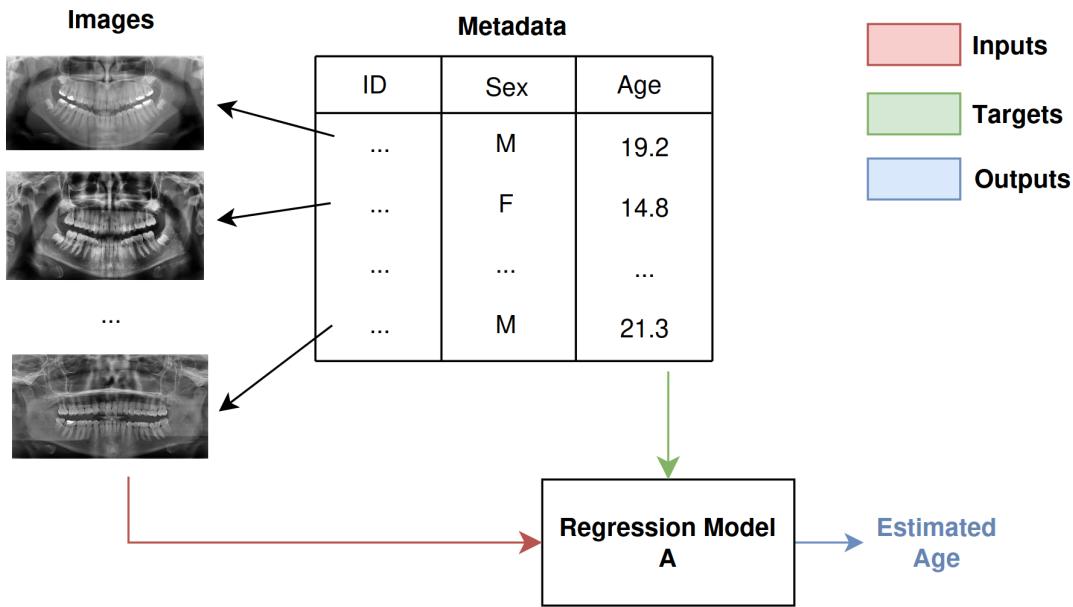


Figura 5.1: Esquema visual del modelos de regresión propuesto. El modelo solo tiene radiografías maxilofaciales como entrada.

5.2. Protocolo de validación experimental

Como se ha descrito en el capítulo previo, 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 5.2 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 mismas instancias permanezcan en los mismos subconjuntos. Esto permite garantizar que ningún modelo preentrenado reutilice datos previamente utilizados en etapas de validación o calibración,

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”, o una de edad 18.2 y sexo femenino como “18.0_F”.

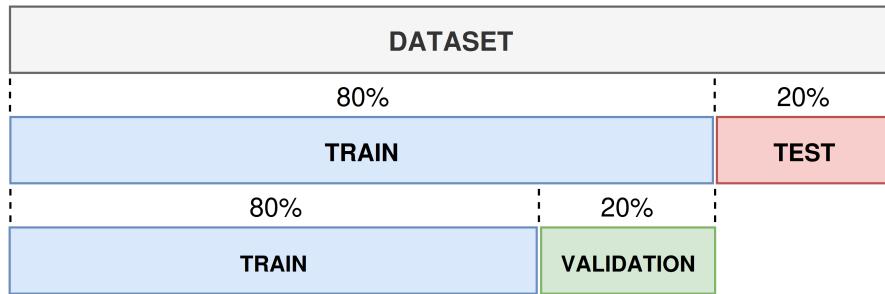


Figura 5.2: Diagrama de división del *dataset* en *train*, *validation* y *test*.

algo especialmente relevante dado que los problemas abordados están jerárquicamente relacionados (la clasificación de sexo y mayoría de edad se deriva directamente de la estimación 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 [148]. 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 [149] (que recomienda dedicar entre un 10 % y 30 % de datos de entrenamiento a calibración), tal y como se muestra en la Figura 5.3.

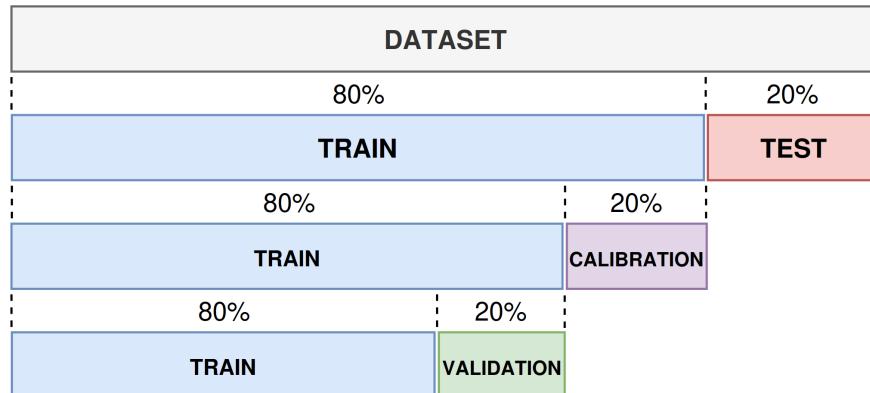


Figura 5.3: Diagrama de división del *dataset* en *train*, *validation*, *calibration* y *test*.

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, validación 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.

5.3. Preprocesado de los datos

Dado que las imágenes del conjunto de datos disponible son significativamente más anchas que altas, se han normalizado todas las dimensiones a 448×224 píxeles para

homogenizar las entradas del modelo³. 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 o iluminación ligeramente variable:

- 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 %.

Se ha establecido un tamaño de *batch* de 32, tras encontrar preliminarmente un equilibrio entre regularización y buen ritmo de aprendizaje.

5.4. Esquema general de los experimentos realizados

Para cada problema planteado, se propone realizar una comparativa entre distintos métodos, incluyendo tanto predicciones puntuales como interválicas en los casos de regresión, y predicciones de una sola etiqueta o de un conjunto de etiquetas en los casos de clasificación, utilizando tanto heurísticas como métodos de CP. De esta forma queremos evaluar tanto la utilidad tradicional para estimar el valor esperado como la capacidad para proporcionar intervalos de confianza fiables que capturen la incertidumbre predictiva. Todas las métricas se calculan sobre el conjunto de test.

Se requerirá el 95 % de confianza en las predicciones interválicas o de conjunto de etiquetas, que es la cifra de confianza generalmente empleada en AF.

5.4.1. Problema de estimación de edad

Para el problema de estimación de edad se han propuesto los siguientes cuatro métodos (véase la Figura 5.4):

- **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 la desviación típica obtenido en el conjunto de validación, que es una aproximación heurística común para construir intervalos de predicción que asumen normalidad en los errores. Bajo esta suposición, el intervalo debería cubrir aproximadamente el 95 % de los casos, aunque en la práctica esta cobertura puede verse afectada si los residuos no siguen una distribución normal o presentan heterocedasticidad. Este método sirve como *baseline* para comparar la mejora que aportan técnicas más sofisticadas.
- **Método ‘ICP’:** Implementa el método *Inductive Conformal Prediction* para la CP.

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

- **Método ‘QR’:** Este método implementa *Quantile Regression*. Utiliza tres cuantiles

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

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

- **Método ‘CQR’:** Este método implementa *Conformalized Quantile Regression*, con los mismos cuantiles que QR.

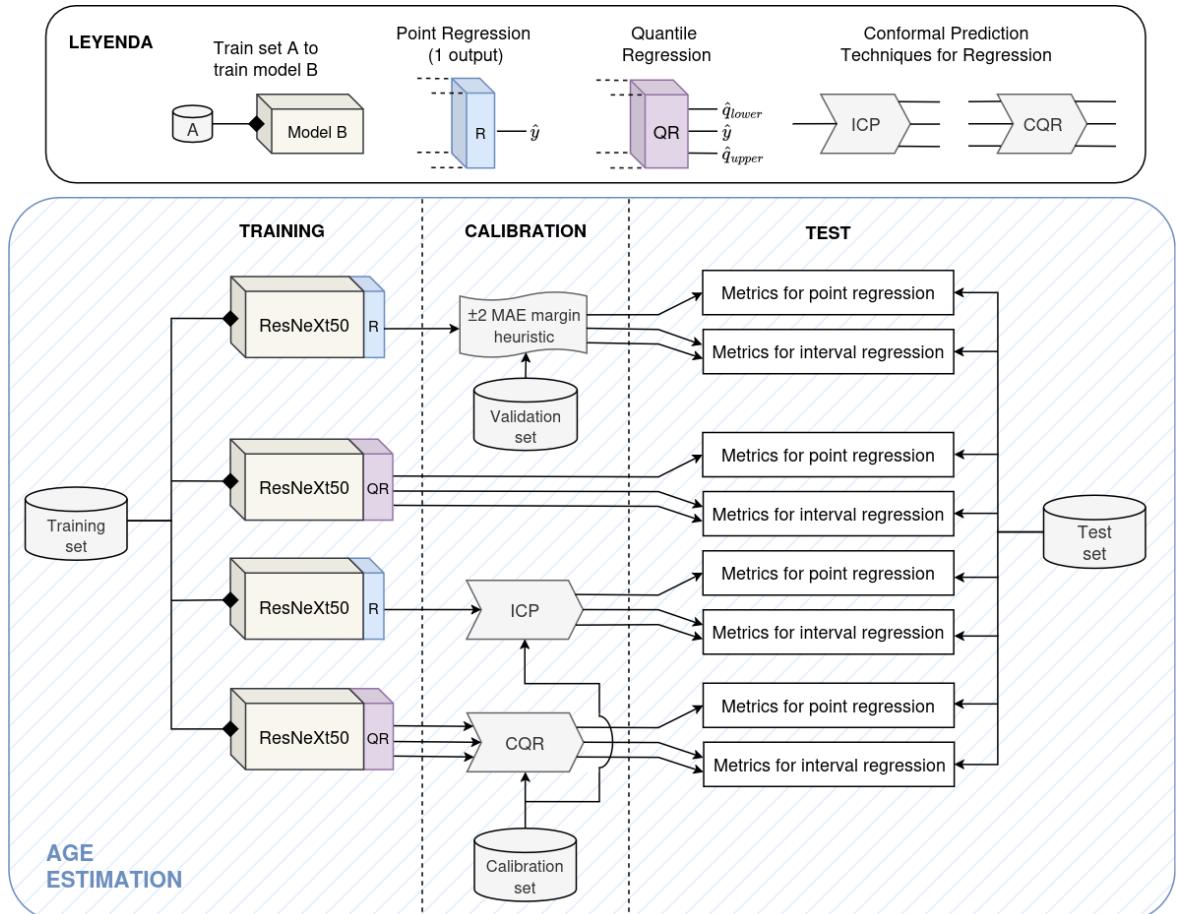


Figura 5.4: Esquema de experimentación para la estimación de edad. Cada modelo se entrena por separado.

R: Regresión puntual

QR: Quantile Regression

ICP: Inductive Conformal Prediction

CQR: Conformalized Quantile Regression

5.4.2. Problema de estimación de mayoría de edad

Respecto al problema de estimación de mayoría de edad, se han propuesto los siguientes tres métodos (véase la Figura 5.5):

- **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 estimación de edad, 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 las técnicas APS y RAPS de CP para clasificación, ya que APS es teóricamente equivalente a LAC en problemas de clasificación binaria, y RAPS no resulta aplicable en dicho contexto.

En este caso, también se han obtenido 10 modelos independientes para cada método.

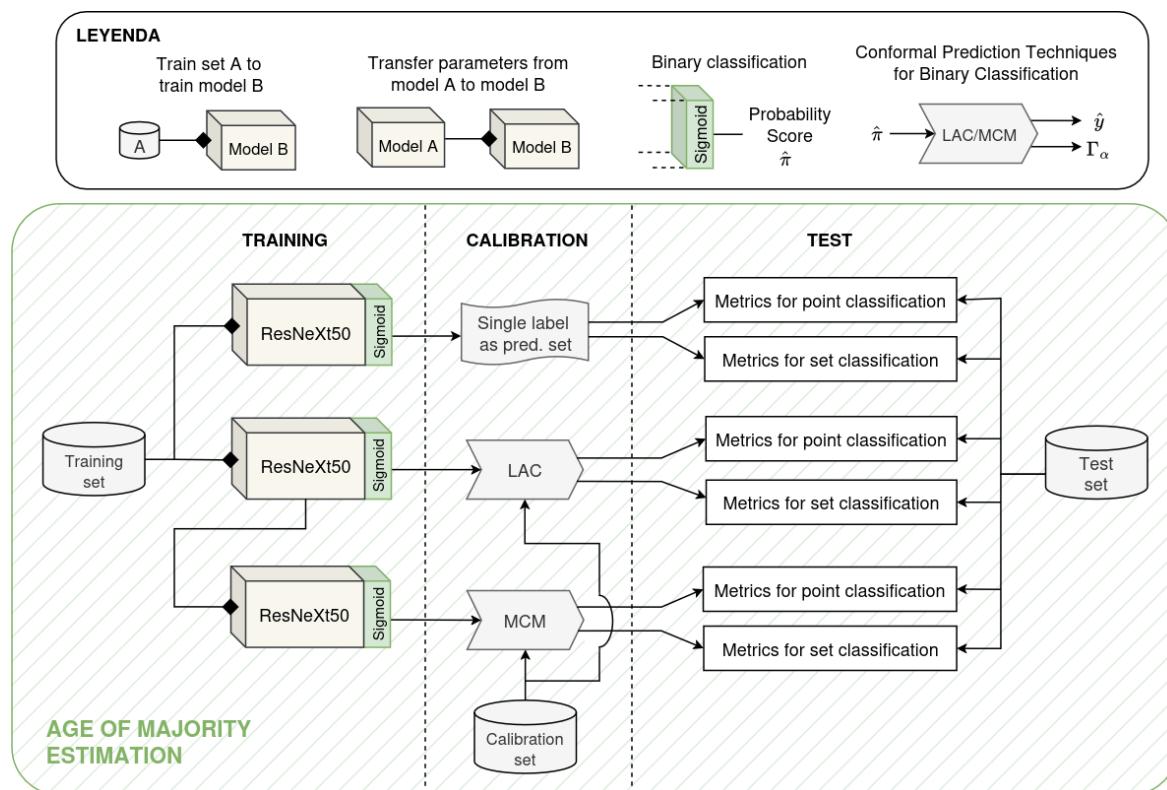


Figura 5.5: Esquema de experimentación para la estimación de mayoría de edad.

LAC: Least-Ambiguous set-valued Classifiers

MCM: Mondrian Confidence Machine

5.4.3. Problema de clasificación de edad

Para el problema de clasificación de edad, se ha empleado la técnica de **calibración de probabilidades Temperature 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 del *softmax*. 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 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 parámetros, dado que el algoritmo es determinista y produce resultados consistentes para un mismo conjunto de datos y parámetros.

Los métodos propuestos para este problema son (véase la Figura 5.6):

- **Método ‘base’:** Dado que los valores de edad se discretizaron en clases y considerando la cercanía entre estas —como se observó en el análisis de regresión—, para evitar evitar que el conjunto de predicción sea demasiado estrecho y no capture la incertidumbre inherente entre clases adyacentes, se propone un enfoque alternativo: se construyen conjuntos de predicción agregando clases hasta que la suma de sus puntuaciones de softmax supere el 95 %. Este método sirve como *baseline* y no utiliza ningún método de CP. El modelo se entrena a partir de un modelo ‘base’ previamente entrenado para el problema de estimación de mayoría de edad.
- **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 estimación de mayoría de edad.
- **Método ‘MCM’:** Implementa la técnica MCM para CP. El modelo será exactamente el mismo que el de LAC para este mismo problema.
- **Método ‘APS’:** Implementa la técnica APS para CP, con componente aleatoria para tamaños de conjunto de predicción más ajustados. El modelo será exactamente el mismo que el de LAC para este mismo problema.
- **Método ‘RAPS’:** Implementa la técnica RAPS para CP, también con componente aleatoria. El modelo será exactamente el mismo que el de LAC para este mismo problema.
- **Método ‘SAPS’:** Implementa la técnica SAPS para CP. Usará el mismo modelo que LAC.

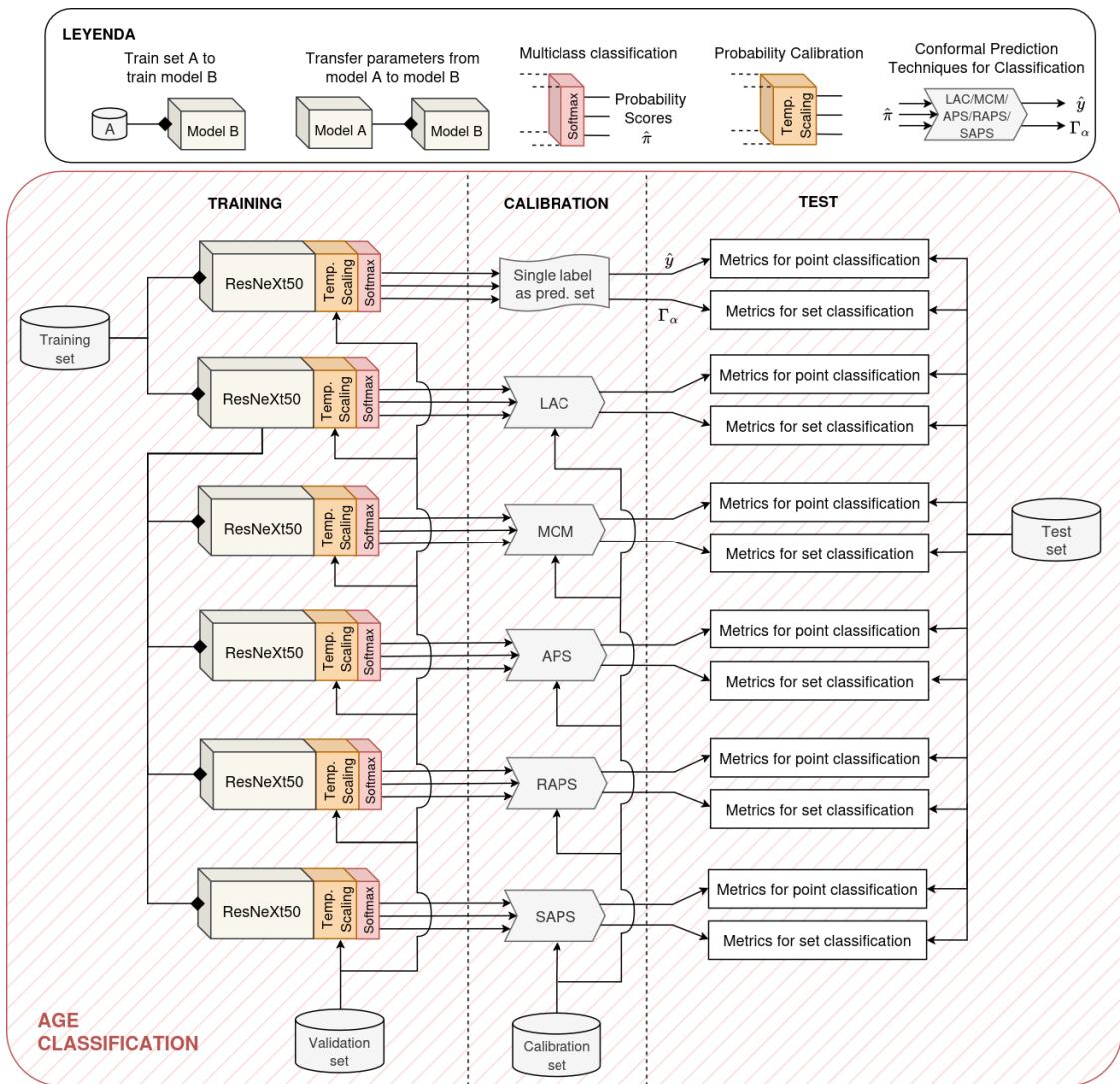


Figura 5.6: Esquema de experimentación para la clasificación de edad.

LAC: Least-Ambiguous set-valued Classifiers

MCM: Mondrian Confidence Machine

APS: Adaptive Prediction Sets

RAPS: Regularized Adaptive Prediction Sets

SAPS: Sorted Adaptive Prediction Sets

Temp. Scaling: Temperature Scaling

5.5. Evaluación del rendimiento

5.5.1. 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 valor esperado 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.

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

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 compromiso 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 mantener 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. Sin embargo, valores excesivamente pequeños conducen inevitablemente a coberturas pueden indicar ..., especialmente cuando el problema muestra una variabilidad inherente A diferencia de la cobertura empírica, no sabemos cuál es el valor óptimo,

- La **mean interval score** [150] 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 del límite inferior al valor real ($l_i - y_i$).
- $\frac{2}{\alpha} (y_i - u_i) \mathbb{I}[y_i > u_i]$ penaliza los casos en que el valor verdadero y_i está por encima del límite superior u_i , proporcionalmente a la distancia del límite superior al valor real ($y_i - u_i$).

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 las predicciones interválicas:

- **Gráfica de dispersión de Cobertura Empírica - Amplitud Media del Intervalo de Predicción:** 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.
- **Histograma 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.

Solo tiene sentido analizar el histograma para aquellos métodos que dan intervalos de predicción de tamaño variable, como es en nuestro caso QR y CQR.

5.5.2. 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.

Para la clasificación de etiqueta única usaremos:

- La **matriz de confusión** es una herramienta fundamental que permite visualizar el rendimiento de modelos de clasificación, tanto binarios como multiclas. Esta muestra una tabla con tantas columnas y filas como clases haya. En un eje, se representan las clases reales (etiquetas verdaderas), y en el otro eje, las clases predichas por el modelo. Cada celda de la matriz indica la cantidad de ejemplos que pertenecen a una clase real específica y que han sido clasificados como una clase predicha específica (véase la Figura 5.7). 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 **exactitud (accuracy)** es la proporción de instancias totales bien clasificadas.

Por otro lado, para la clasificación multietiqueta emplearemos:

- 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.

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

Esta variable se puede obtener o bien en todos los ejemplos del conjunto, o en subpoblaciones específicas de este.

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

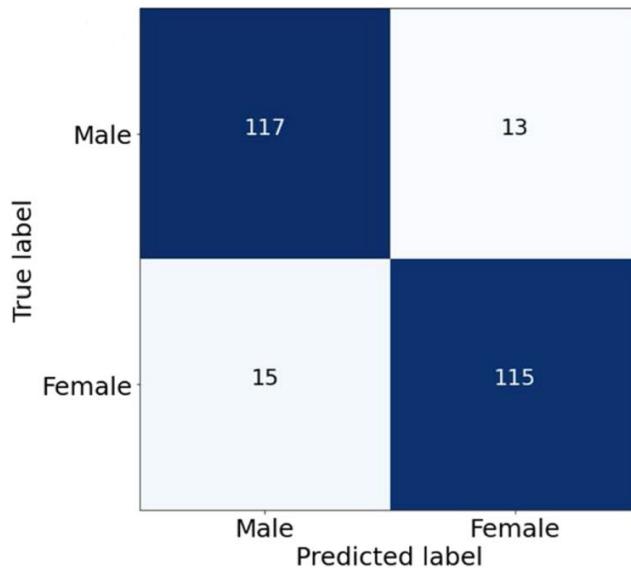


Figura 5.7: Ejemplo de matriz de confusión para el modelo de estimación de sexo propuesto en [151].

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

Y, finalmente, también usaremos elementos visuales para valorar el desempeño de las predicciones interválicas, como pueden ser:

- **Gráfica de dispersión de Cobertura Empírica - Tamaño Medio del Conjunto de Predicción:** 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.
- **Histograma de tamaños de conjuntos de predicción:** De similar forma a como se plantean los histogramas de amplitudes de intervalos para regresión, se puede plantear histograma donde cada tamaño es el número de clases del conjunto predicho. Este nos permite ver la distribución de tamaños de conjuntos de predicción. Aquellos con pocas clases serán más informativos, mientras que tamaños más grandes pueden revelar una incertidumbre elevada.

5.5.3. Tests estadísticos

En los casos en los que las diferencias en una métrica entre métodos presenten valores intercalados o solapamientos aparentes, se aplican tests estadísticos para determinar si las diferencias observadas son significativas, evitando basarnos únicamente de la comparación visual de medias o medianas.

En el análisis de comparación de métodos de predicción, se seleccionaron diferentes pruebas estadísticas según el cumplimiento de los supuestos de normalidad y homocedasticidad de los datos [152]:

1. **ANOVA clásico + Tukey HSD:** Esta combinación se utiliza cuando los residuos del modelo cumplen los supuestos de normalidad (Shapiro-Wilk) y homocedasticidad (Levene). La ANOVA permite evaluar si existen diferencias significativas en la media de la métrica entre los grupos, mientras que Tukey HSD realiza comparaciones por pares controlando el error tipo I, proporcionando intervalos de confianza para la diferencia de medias. Este enfoque es apropiado cuando las varianzas son similares y los datos siguen una distribución aproximadamente normal.
2. **Welch ANOVA + Games-Howell:** Cuando se cumple la normalidad pero no se cumple la homocedasticidad, se recurre a Welch ANOVA, que ajusta los grados de libertad para compensar la desigualdad de varianzas. Para las comparaciones *post-hoc* se utiliza Games-Howell, que es robusto frente a varianzas desiguales y tamaños de grupo distintos. Esta combinación permite detectar diferencias entre grupos sin asumir igualdad de varianzas, manteniendo el control del error tipo I.

En todos las pruebas globales, las hipótesis son:

- **Hipótesis nula (H_0):** No existen diferencias en la métrica analizada entre los métodos comparados, asumiendo que las medias (o medianas, en el caso de pruebas no paramétricas) son iguales.
- **Hipótesis alternativa (H_1):** Al menos un método difiere significativamente de los demás.

En las pruebas *post-hoc* por pares, las hipótesis son:

- **Hipótesis nula (H_0):** Cada par de métodos comparados no presenta diferencias significativas en la métrica.
- **Hipótesis alternativa (H_1):** La métrica de un método difiere significativamente de la de otro método.

Estas comparaciones permiten identificar específicamente qué grupos presentan diferencias significativas, controlando el error tipo I⁴ mediante correcciones apropiadas según la prueba utilizada (Tukey HSD o Games-Howell).

5.6. Experimentación para la estimación de edad

5.6.1. Entrenamiento de los modelos

Como se venía anticipando en el anterior capítulo, adaptaremos la arquitectura del modelo ResNeXt50 para el problema de regresión:

- El extracto de características no necesita ser modificado, ya que mantiene la proporcionalidad de las dimensiones a lo largo de sus bloques convolucionales, independientemente del tamaño de las imágenes de entrada.

⁴El error tipo I ocurre cuando se rechaza la hipótesis nula (H_0) siendo esta en realidad verdad.

- Se sustituye la cabecera predeterminada de *average pooling* con capa FC por un *adaptive average pooling*, seguido de una capa *flatten*, y dos bloques densos consecutivos, cada uno compuesto por una capa *batch normalization*, una capa de *dropout* y una capa FC, con una activación ReLU entre ambos bloques. La primera capa FC contiene 4096 neuronas, la segunda 512, y finalmente se incluye una capa de salida de una sola neurona. Véase la Figura 5.8. Esta configuración ha sido seleccionada siguiendo la recomendación de los tutores, quienes cuentan con experiencia previa en el trabajo con este conjunto de datos.

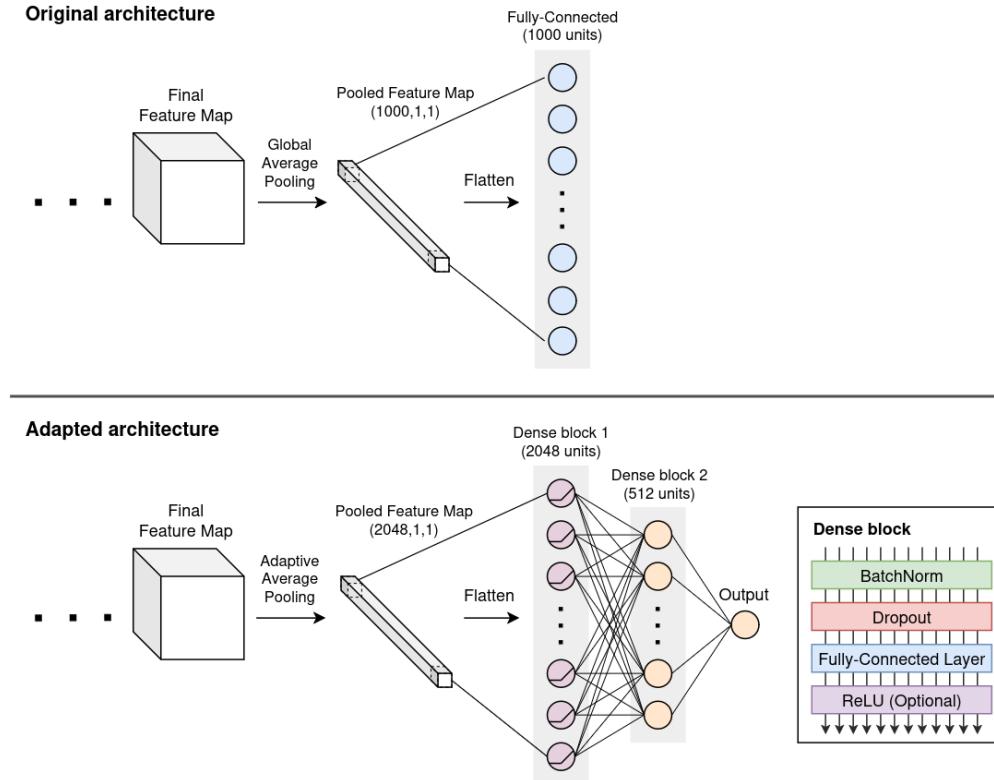


Figura 5.8: Adaptación de la arquitectura ResNeXt50 para el problema de la estimación de edad. En la parte superior de la imagen se muestra la cabecera original de ResNeXt50 diseñada para CIFAR1000. En la parte inferior, se presenta la nueva cabecera adaptada al problema de regresión, con una única salida. Se representa un esquema de la función ReLU con fondo violeta sobre los nodos que emplean esta función de activación. Cabe destacar que, en el caso de la *Quantile Regression*, el modelo genera tres salidas en lugar de una.

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.

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 [153]. 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* [154]. 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 entranan 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 menor⁵. 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. 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.

El tiempo de entrenamiento medio —incluyendo en este tanto entrenamiento como validación de la red hasta la última época— ha sido de 1 hora y 24 minutos, mientras que el tiempo de calibración ha supuesto 4 minutos y 45 segundos de media. Por tanto, el entrenamiento supone un 94 % del tiempo total de cómputo para la puesta en marcha del modelo, y la calibración el restante, considerándose residual en términos prácticos, más todavía considerándose que se está incluyendo en este el tiempo de infancia puntual del conjunto de calibración. Esta distribución temporal confirma que **la sobrecarga asociada a la inferencia conformal es mínima comparado con el coste inicial de entrenamiento del modelo**, lo que refuerza su viabilidad para aplicaciones prácticas donde la eficiencia operativa es crítica.

⁵Se 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.

5.6.2. Resultados

Análisis de métricas para la estimación puntual de edad

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 tan solo unas centésimas en las métricas evaluadas. No obstante, un análisis estadístico riguroso entre los valores obtenidos (véase el Análisis Estadístico 5.1) reveló diferencias significativas entre métodos tanto en el MAE como el MSE. Los resultados identificaron los siguientes patrones:

- No existen diferencias significativas entre los modelos QR y base en ninguna métrica, al igual que tampoco entre los modelo 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.
- Los modelos conformales (ICP y CQR) mostraron errores significativamente mayores ($p < 0.01$) que los modelos no conformales (base y QR). Esto era esperable, pues los métodos conformales tienen menos ejemplos para entrenarse y, por tanto, generalizan peor.

Ejecución	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	1.17	1.20	1.16	1.18	2.40	2.48	2.34	2.46
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	1.18	1.19	1.17	1.17	2.40	2.44	2.41	2.40
Ejecución 10	1.15	1.20	1.15	1.19	2.29	2.48	2.34	2.51
Media	1.16	1.20	1.16	1.18	2.36	2.48	2.39	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. Se marca en negrita la media con mejor valor para cada métrica.

Análisis estadístico 5.1: MAE y MSE en el problema de estimación de edad

El test ANOVA indicó diferencias significativas tanto en MAE ($F(3, 36) = 27.754$, $p < 0.001$) como en MSE ($F(3, 36) = 17.284$, $p < 0.001$), cumpliéndose los supuestos de normalidad (Shapiro-Wilk, $p > 0.5$ para ambas métricas) y homocedasticidad (Levene, $p > 0.7$). Para identificar qué pares de modelos presentaban diferencias

significativas, se aplicó la prueba *post-hoc* de comparaciones múltiples Tukey HSD (véanse las Tablas 5.2 y 5.3).

Modelo 1	Modelo 2	Dif. media	Valor <i>p</i>	IC 95 %	Signif.
CQR	ICP	0.0128	0.0299	[0.001, 0.0246]	Sí
CQR	QR	-0.0199	0.0003	[-0.0317, -0.0081]	Sí
CQR	base	-0.0209	0.0002	[-0.0327, -0.0091]	Sí
ICP	QR	-0.0327	<0.0001	[-0.0445, -0.0209]	Sí
ICP	base	-0.0337	<0.0001	[-0.0455, -0.0219]	Sí
QR	base	-0.001	0.9959	[-0.0128, 0.0108]	No

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

Método 1	Método 2	Dif. media	Valor <i>p</i>	IC 95 %	Signif.
CQR	ICP	0.04	0.1397	[-0.0087, 0.0887]	No
CQR	QR	-0.0542	0.0243	[-0.103, -0.0055]	Sí
CQR	base	-0.0779	0.0007	[-0.1267, -0.0292]	Sí
ICP	QR	-0.0942	<0.0001	[-0.143, -0.0455]	Sí
ICP	base	-0.1179	<0.0001	[-0.1667, -0.0692]	Sí
QR	base	-0.0237	0.5625	[-0.0724, 0.025]	No

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

Análisis de métricas para la estimación interválica de edad

A continuación, la Tabla 5.4 presenta las métricas sobre las predicciones interválicas de los métodos.

- El método ‘QR’ es el que muestra menor cobertura, con apenas un 88.8 % de media, muy por debajo del valor nominal. Esto se da porque las regiones generadas por la regresión cuantílica son las más estrechas de entre los métodos.
- El resto de métodos sí alcanzan coberturas más próximas al 95 %, tal como cabía esperar dada su fundamentación estadística (aunque, recordemos, ello implica asumir normalidad en los residuos para que el método ‘base’ logre cubrir el 95 %). Las diferencias de cobertura entre métodos se explican principalmente por la amplitud de los intervalos; se constata la relación de compromiso o *trade-off* entre la cobertura y amplitud de los intervalos: a mayor amplitud del intervalo, mayor cobertura, y viceversa. Esta relación se visualiza claramente en la Figura 5.9, donde se podría trazar una curva de tendencia ascendente que refleja dicha correlación positiva entre ambas métricas.

Dicho esto, los métodos conformales (ICP y CQR) logran aproximarse mejor a la cobertura nominal: ICP sobrecubriendo en un 0.18 % de media, y CQR infracubriendo en un 0.11 % de media, frente a la infracobertura del 0.65 % de media en el método ‘base’. De hecho, el Análisis Estadístico 5.2 revela diferencias significativas en cobertura entre ‘base’ y los dos métodos conformales (si bien no entre ICP y CQR). Esta menor cobertura del método ‘base’ se explica por sus suposiciones más restrictivas, ya que asume normalidad en los residuos, lo que solo garantiza una cobertura del 95 % bajo dicha hipótesis. En cambio, los métodos conformales no requieren este supuesto de normalidad y ajustan empíricamente la distribución de los errores, logrando así una calibración más precisa de la incertidumbre e intervalos predictivos más fiables.

Ejecución	Cobertura Empírica (%)				Amplitud Media del Intervalo			
	base	ICP	QR	CQR	base	ICP	QR	CQR
	93.77	94.54	88.10	94.89	5.79	6.40	4.62	6.04
Ejecución 1	94.10	95.40	89.54	95.49	5.81	6.29	4.67	6.22
Ejecución 2	94.38	95.59	88.94	94.70	5.87	6.30	4.67	5.75
Ejecución 3	94.52	94.56	89.22	95.03	5.84	6.20	4.67	6.16
Ejecución 4	94.80	95.35	87.78	94.47	5.95	6.27	4.59	6.04
Ejecución 5	94.28	95.03	87.73	94.66	5.92	6.15	4.62	5.88
Ejecución 6	94.66	94.66	88.15	94.89	5.94	6.24	4.60	5.98
Ejecución 7	94.52	94.47	89.64	94.42	5.88	6.18	4.68	5.85
Ejecución 8	94.42	95.49	89.27	95.21	5.89	6.42	4.67	6.09
Ejecución 9	94.10	95.68	89.59	95.12	5.94	6.33	4.65	5.99
Ejecución 10	94.35	95.18	88.80	94.89	5.88	6.28	4.64	6.00
Media	94.35	95.18	88.80	94.89	5.88	6.28	4.64	6.00

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. Se marcan en negrita las métricas de aquellos métodos que logran una cobertura próxima o superior al 95 %.

Análisis estadístico 5.2: Cobertura empírica y Amplitud Media del Intervalo de predicción en el problema de estimación de edad

En cuanto a cobertura empírica, se cumple normalidad en la distribución (Shapiro-Wilk, $p > 0.3$), pero se descarta homocedasticidad (Levene, $p < 0.05$), por lo que se puede aplicar el test Welch ANOVA, que revela diferencias globales significativas entre métodos: $F(3, 36) = 178.81$, $p < 0.0001$. Para identificar qué pares de métodos presentaban diferencias significativas, se aplicó la prueba *post-hoc* de Games-Howell (véase la Tabla 5.5).

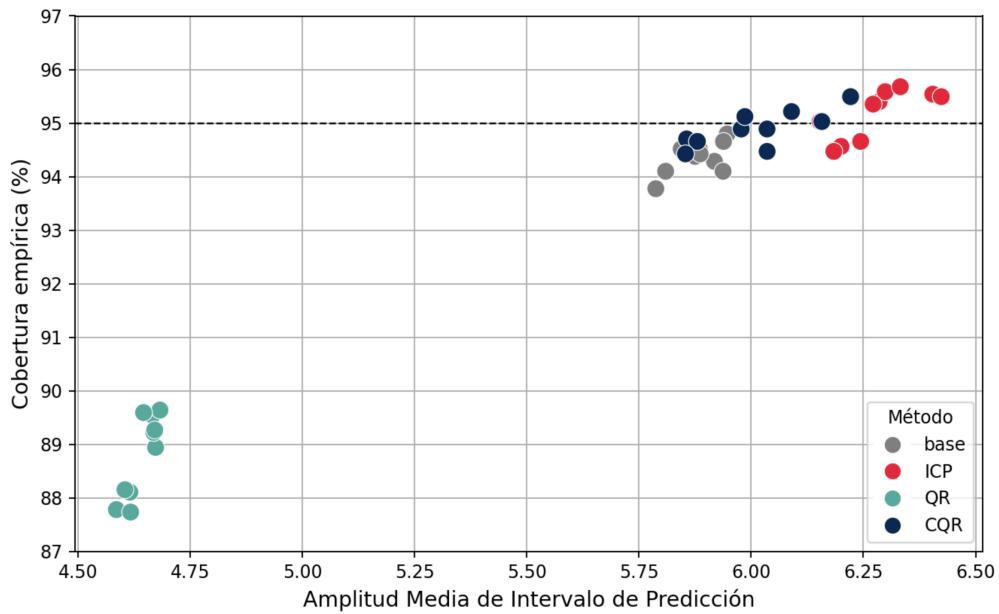


Figura 5.9: Gráfica de dispersión de la Cobertura empírica frente a la Amplitud media del intervalo de predicción. Existe una relación de compromiso entre la cobertura y la amplitud de los intervalos: al aumentar una, generalmente también lo hace la otra, y viceversa. Los métodos más eficaces son aquellos que alcanzan una cobertura empírica cercana o superior al valor nominal (0.95), manteniendo al mismo tiempo una amplitud media lo más baja posible. Estos métodos se sitúan idealmente en la esquina superior izquierda del gráfico.

Modelo 1	Modelo 2	Dif. media	Valor <i>p</i>	Signif.
base	ICP	-0.0082	0.0012	Sí
base	QR	0.0556	<0.0001	Sí
base	CQR	-0.0053	0.0076	Sí
ICP	QR	0.0638	<0.0001	Sí
ICP	CQR	0.0029	0.4	No
QR	CQR	-0.0609	<0.0001	Sí

Tabla 5.5: Resultados de la prueba *post-hoc* de Games-Howell para cobertura empírica entre pares de métodos. Se muestran la diferencia media entre grupos, el valor *p* ajustado, el intervalo de confianza al 95 % y si la diferencia es estadísticamente significativa ($\alpha = 0.05$).

Respecto a las amplitudes medias de los intervalos de predicción, se evaluaron las diferencias de los métodos utilizando un test Welch ANOVA, dado que los residuos de los datos seguían una distribución normal (Shapiro-Wilk, $p > 0.4$), pero no cumplían con la homocedasticidad (Levene, $p < 0.05$). El Welch ANOVA reveló diferencias globales significativas entre los métodos: $F(3, 18.27) = 1829.13$, $p < 0.0001$. Tras esto, se aplicó la prueba *post-hoc* de Games-Howell para comparar las marcas por pares de métodos (véase la Tabla 5.6).

Modelo 1	Modelo 2	Dif. media	Valor <i>p</i>	Signif.
base	ICP	-0.39718	<0.0001	Sí
base	QR	1.2399	<0.0001	Sí
base	CQR	-0.1161	0.0523	No
ICP	QR	1.6372	<0.0001	Sí
ICP	CQR	0.2811	0.0002	Sí
QR	CQR	-1.3561	<0.0001	Sí

Tabla 5.6: Resultados de la prueba *post-hoc* de Games-Howell para la amplitud media del intervalo de predicción entre pares de métodos. Se muestran la diferencia media entre grupos, el valor *p* ajustado, el intervalo de confianza al 95% y si la diferencia es estadísticamente significativa ($\alpha = 0.05$).

En la Tabla 5.7 apreciamos cómo los métodos ‘base’ y CQR logran significativamente menores valores de *interval score* que ICP, indicando que tienen un mejor equilibrio entre cobertura y tamaño del intervalo. En consecuencia, estos métodos se perfilan como los que logran mejor equilibrio cobertura/amplitud del intervalo, si bien CQR logra aproximarse más a la cobertura requerida, perfilándose como la opción *a priori* más ventajosa, con garantías de cobertura para el nominal requerido e intervalos de predicción ajustados.

Ejecución	Mean Interval Score			
	base	ICP	QR	CQR
Ejecución 1	8.14	8.23	8.69	7.95
Ejecución 2	7.99	8.11	8.30	7.94
Ejecución 3	7.90	8.24	8.57	8.02
Ejecución 4	7.98	8.26	8.86	8.01
Ejecución 5	7.96	8.14	9.04	8.28
Ejecución 6	8.15	8.21	9.01	7.91
Ejecución 7	8.02	8.25	9.00	7.96
Ejecución 8	8.01	8.14	8.61	8.15
Ejecución 9	8.01	7.16	8.54	7.83
Ejecución 10	8.14	8.00	8.51	7.97
Media	8.02	8.17	8.71	8.01

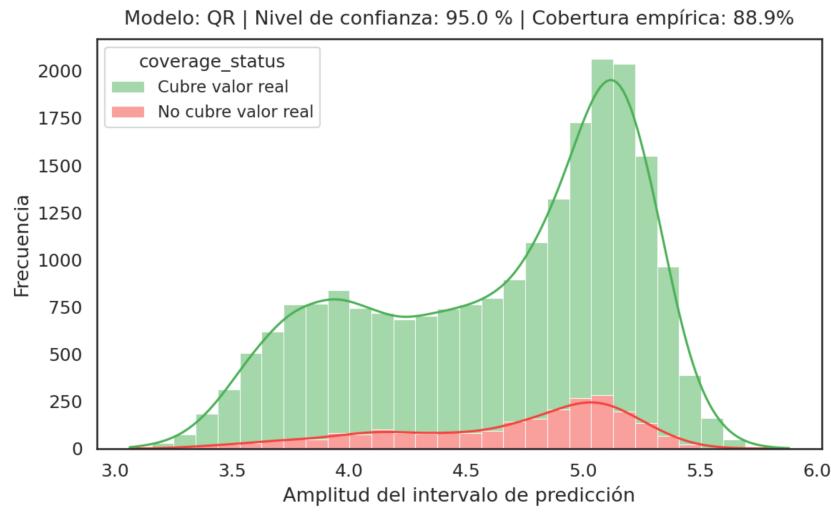
Tabla 5.7: *Mean Interval Score* obtenidos por cada método de predicción a lo largo de distintas ejecuciones. Se presentan los valores para cada ejecución individual, así como la media final de cada métrica. Se marca en negrita la mejor marca en la métrica media.

Análisis de la cobertura en base al tamaño del intervalo

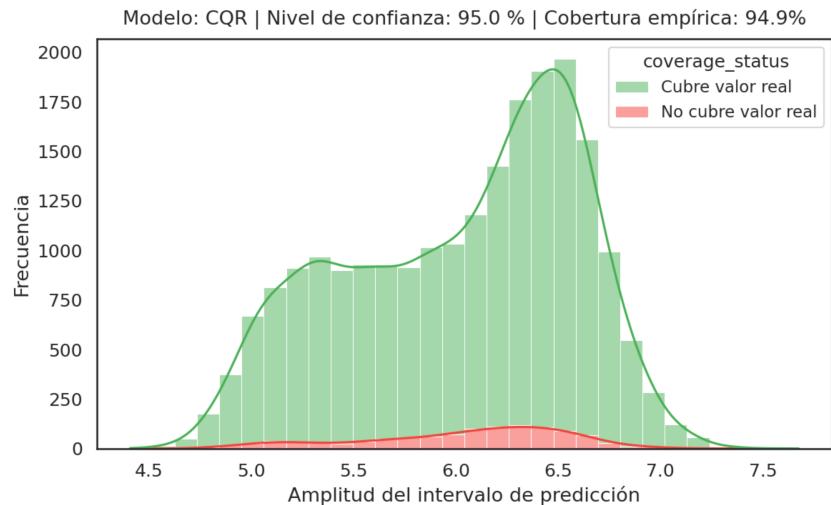
En los métodos donde los intervalos de predicción son adaptativos (QR y CQR), resulta relevante analizar cómo se comporta la cobertura empírica en función de las distintas amplitudes de intervalos de predicción obtenidos de las instancias. La hipótesis subyacente es que intervalos más amplios reflejan una mayor incertidumbre asociada a la predicción, mientras que intervalos más estrechos denotan mayor confianza, de forma

que todos los intervalos lograrían cubrir al nivel de confianza deseado los valores reales. Frente a esta situación, en el peor de los escenarios, los intervalos más estrechos tenderían a infracubrir (es decir, no contener el valor real con la frecuencia esperada) y los intervalos más amplios tenderían a sobrecubrir (conteniendo el valor real más allá del nivel objetivo de confianza). Este escenario sería especialmente negativo dado que implicaría una distribución ineficiente de la incertidumbre, donde solo alcanzaría la cobertura nominal en aquellas predicciones menos informativas o más conservadoras.

La Figura 5.10 muestra los histogramas de la amplitud de los intervalos de predicción para los métodos adaptativos en todas sus ejecuciones, diferenciando el número de instancias que cubre el el valor real de las que no. 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 método 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.



(a) Método QR.



(b) Método CQR.

Figura 5.10: Histogramas de amplitud del intervalo de predicción con diferenciación por cobertura, correspondientes a los métodos QR y CQR. La comparación permite visualizar cómo varía la capacidad de cobertura en función del tamaño 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 Figura 5.11. En ella 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. Podemos observar los siguientes patrones:

- Como era de esperar, los métodos adaptativos (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 del resto de métodos, con intervalos de predicción de tamaño constante.
- CQR es el único método adaptativo que aproxima a la cobertura nominal. Este 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, donde se concentra más de la mitad de instancias. Su peor marca es de un 93.25 %, para instancias con intervalo de predicción de 6 a 6.5 años de longitud, 2 puntos porcentuales por debajo del nominal.

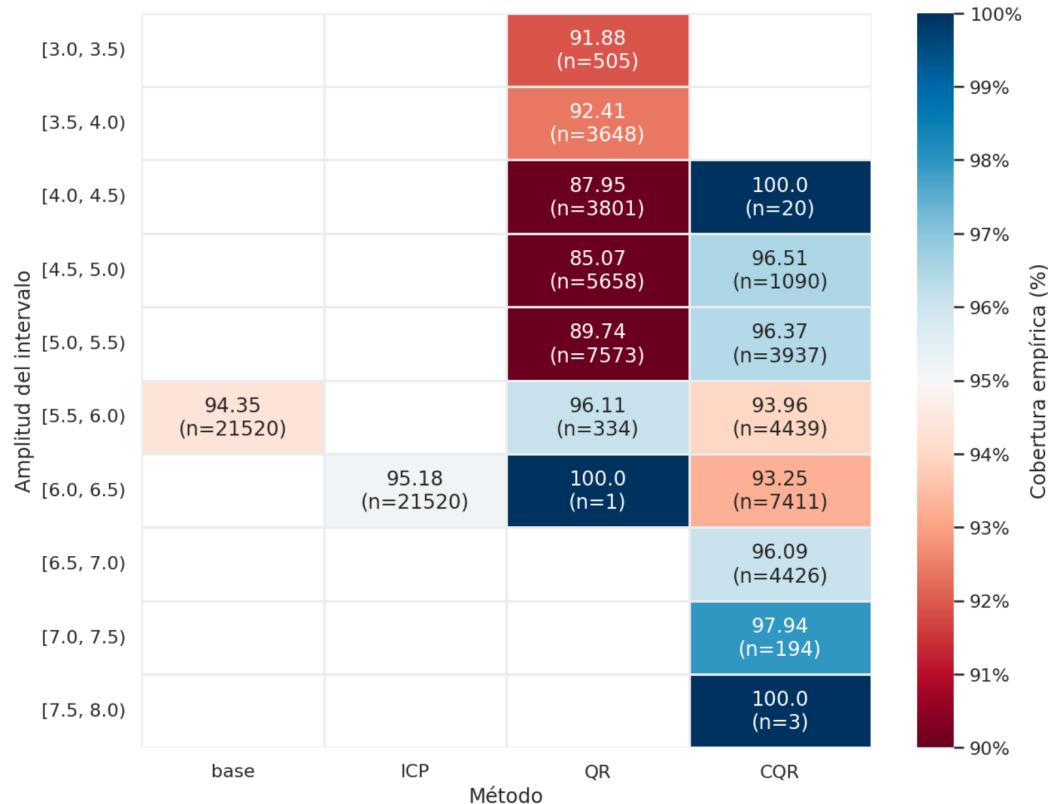


Figura 5.11: Mapa de calor de la cobertura empírica en base a la amplitud del intervalo de predicción por cada método de predicción en media de sus ejecuciones. Se especifica entre paréntesis el número de instancias clasificadas en cada franja de amplitud de intervalo. La escala de colores está centrada en la cobertura nominal (0.95): los valores por debajo de este umbral se representan en tonos rojos, los superiores en tonos azules, y el blanco indica una cobertura empírica equivalente a la nominal.

Análisis de la cobertura en base a la edad cronológica y sexo

Por último, se ha analizado la cobertura en base a la edad real de los individuos y su sexo, ya que resulta crucial identificar posibles sesgos en el desempeño del modelo a lo largo de estas variables.

Las Figuras 5.12 y 5.13 muestran la evolución de la cobertura empírica y el ancho medio de los intervalos de predicción en función de la edad cronológica⁶ y el sexo.

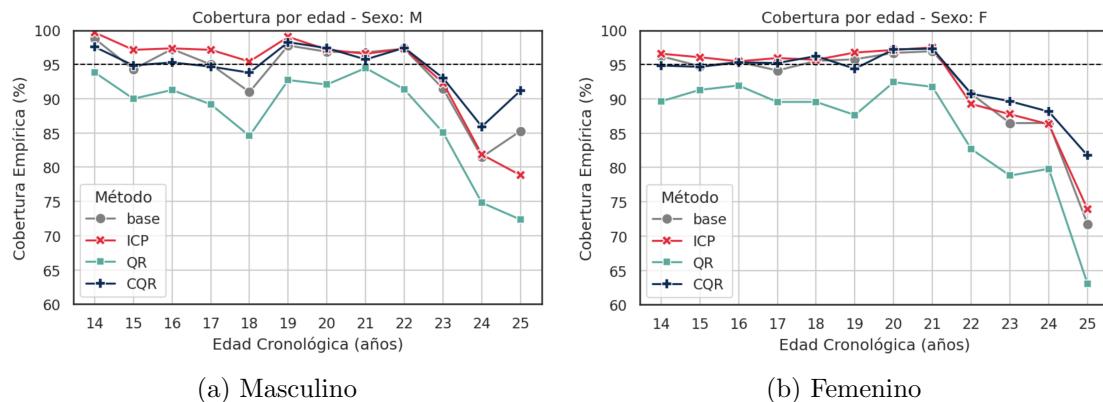


Figura 5.12: Gráficos de líneas de la cobertura empírica del intervalo de predicción (%) para cada método en función de la edad cronológica entera de los individuos, diferenciando por sexo.

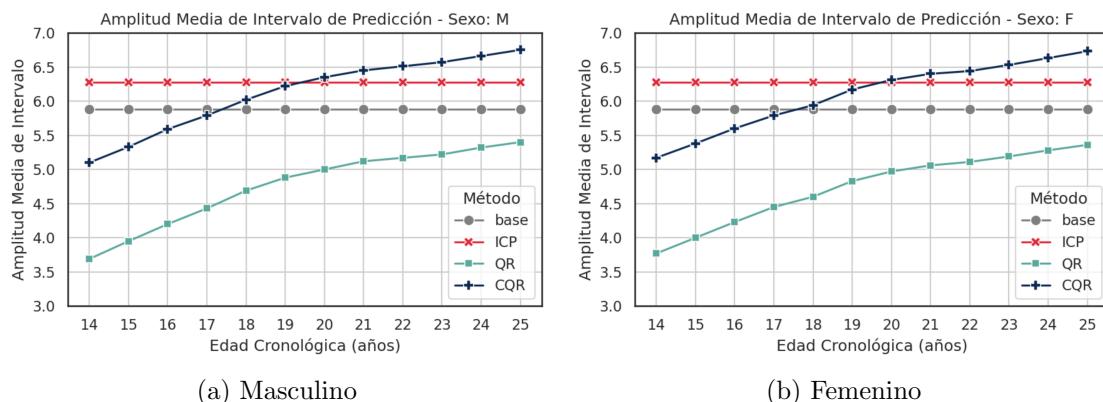


Figura 5.13: Gráficos de líneas de la amplitud media del intervalo de predicción para cada método en función de la edad cronológica entera de los individuos, diferenciando por sexo.

En general, todos los métodos tienden a perder cobertura conforme aumenta la edad cronológica de los individuos, independientemente del sexo. Esta disminución se vuelve especialmente marcada a partir de los 22 años, afectando incluso al método CQR, que hasta entonces presentaba la cobertura más robusta. En particular, CQR mantiene valores cercanos al 95 % en edades tempranas, pero a partir de los 22 años en varones y de los 23 en hembras comienza a descender, alcanzando aproximadamente un 91.5 % y un 82 % de cobertura a los 25 años, en sexo masculino y femenino respectivamente. 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

⁶Parte entera o suelo de la edad real.

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 atribuible a:

- Escasez de ejemplos en edades avanzadas: El conjunto de datos presenta una disminución en el número de muestras a partir de los 23 años, lo que coincide con la reducción en la cobertura predictiva. Esto causaría incertidumbre epistémica.
- Mayor variabilidad fisiológica en adultos jóvenes: A medida que aumenta la edad, los individuos suelen presentar una mayor diversidad en sus características biológicas debido a la acumulación de factores ambientales y estilos de vida [122, 123]. En este caso, esta incertidumbre sería estocástica, ya que es inherente al sistema.

Además, como hemos descrito antes, el **fenómeno es más acusado en el sexo femenino que en el masculino**, a pesar de que había —ligeramente— más ejemplos del sexo femenino que del masculino para los 21 y 22 años. Esto podría deberse a que el sexo femenino suele completar antes la maduración dental que los varones [123], lo que puede hacer que las diferencias interindividuales en hembras puedan deberse más a factores ambientales.

Cabe destacar que este análisis de la cobertura en función de la edad cronológica y el sexo pone de manifiesto una incertidumbre heterogénea a lo largo de las distintas subpopulationes, lo que sugiere la presencia de heterocedasticidad en los errores de predicción. Este comportamiento contradice el supuesto de homocedasticidad asumido en el método ‘base’, que —junto con el método CQR— era el que presentaba el mejor *interval score*.

Discusión de resultados

El método **CQR** se posiciona como el claro ganador en todos los apartados analizados. Este resultado era previsible, ya que se trata del único método conformal y adaptativo considerado en el estudio.

Destaca por presentar la menor amplitud media de los intervalos, manteniendo al mismo tiempo una cobertura muy próxima a la nominal. Además, al ser el único método conformal adaptativo de la lista, ofrece una ventaja estructural frente a los demás. Sus tasas de cobertura empírica son consistentes para diferentes amplitudes de intervalo, y sobresale especialmente en los casos con pocas instancias de edades cronológicas avanzadas, donde logra adaptarse a la mayor incertidumbre ampliando de forma adecuada el intervalo de predicción.

5.7. Experimentación para la estimación de mayoría de edad

5.7.1. Entrenamiento de los modelos

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 extractor 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.

El tiempo de entrenamiento medio de la cabecera ha sido de 12 minutos y 45 segundos, mientras que el tiempo de calibración ha supuesto 4 minutos y 41 segundos de media.

5.7.2. Resultados

Análisis de métricas para la clasificación puntual de mayoría de edad

En la Tabla 5.8 se presentan las métricas que evalúan el rendimiento del modelo de clasificación en sus predicciones de una sola etiqueta.

Método	Exactitud (%)		Sensibilidad (%)		Especificidad (%)	
	base	CP	base	CP	base	CP
Ejecución 1	87.87	86.99	89.07	89.83	86.05	82.65
Ejecución 2	87.87	87.36	89.92	90.99	84.76	81.83
Ejecución 3	87.59	86.52	88.61	88.91	86.05	82.88
Ejecución 4	87.59	87.5	89.07	88.99	85.35	85.23
Ejecución 5	87.64	87.13	90.45	88.22	83.35	85.46
Ejecución 6	87.36	86.76	90.53	90.61	82.53	80.89
Ejecución 7	88.06	87.13	89.07	90.15	86.52	82.53
Ejecución 8	87.41	86.2	87.53	88.45	87.22	82.77
Ejecución 9	87.13	86.99	91.15	89.83	81.01	82.65
Ejecución 10	87.78	87.41	89.30	88.76	85.46	85.35
Media	87.63	87.00	89.47	89.47	84.83	83.22

Tabla 5.8: Exactitud, sensibilidad y especificidad 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. ‘CP’ se refiere a los métodos conformales empleados: LAC y MCM (se recuerda que es el mismo modelo para todos los métodos conformales y, por ello, presentan las mismas predicciones puntuales). Se marca en negrita la media con mejor valor para cada métrica.

El método ‘base’ obtiene una exactitud (*accuracy*) significativamente superior a los métodos conformales (véase el Análisis Estadístico 5.3), principalmente debido a una mayor especificidad, ya que la sensibilidad se mantiene prácticamente igual. Esto sugiere

que los errores del modelo se concentran en la predicción de individuos menores de 18 años. Una posible explicación es que los métodos conformales, al entrenarse con un conjunto de datos más reducido, se ven aún más afectados por el desequilibrio de clases. Como resultado, tienden a favorecer la clase mayoritaria (≥ 18), lo que incrementa los falsos positivos y reduce los verdaderos negativos.

Análisis estadístico 5.3: Exactitud en el problema de estimación de mayoría de edad

La exactitud (*accuracy*) ha mostrado diferencias significativas en los distintos métodos, comprobado mediante test ANOVA: $F(2, 27) = 9.6850$, $p < 0.001$, una vez comprobado el cumplimiento de normalidad (Shapiro-Wilk, $p > 0.05$) y homocedasticidad (Levene, $p > 0.5$). En esta ocasión no se ha aplicado test *post-hoc* por pares, dado que solo hay dos grupos con valores diferentes.

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

La Tabla 5.9 presenta las métricas sobre los conjuntos de predicción de los métodos. Para complementar esta información, estos valores también se representan de manera visual en la Figura 5.14. A partir de ellos, pueden identificarse los siguientes patrones:

- Los métodos conformales logran una cobertura significativamente superior al método ‘base’, como es obvio, dado que este último no está diseñado para garantizar cobertura estadística, sino únicamente para realizar predicciones puntuales.
- Aunque los métodos LAC y MCM muestran tamaños medios del conjunto de predicción muy similares, LAC alcanza una cobertura significativamente superior en prácticamente todas las ejecuciones (véase el Análisis Estadístico 5.4). Esto podría deberse a que MCM calcula un umbral de no conformidad por clase utilizando únicamente las puntuaciones de no conformidad correspondientes a las instancias de esa clase, lo que reduce el tamaño de la muestra utilizada y, en consecuencia, disminuye su representatividad.

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	87.36	94.98	94.14	1	1.20	1.19
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	87.13	94.52	93.87	1	1.19	1.19
Ejecución 10	87.78	94.47	94.47	1	1.19	1.20
Media	87.63	94.60	94.17	1	1.20	1.20

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. Se marcan en negrita las marcas de los métodos conformales.

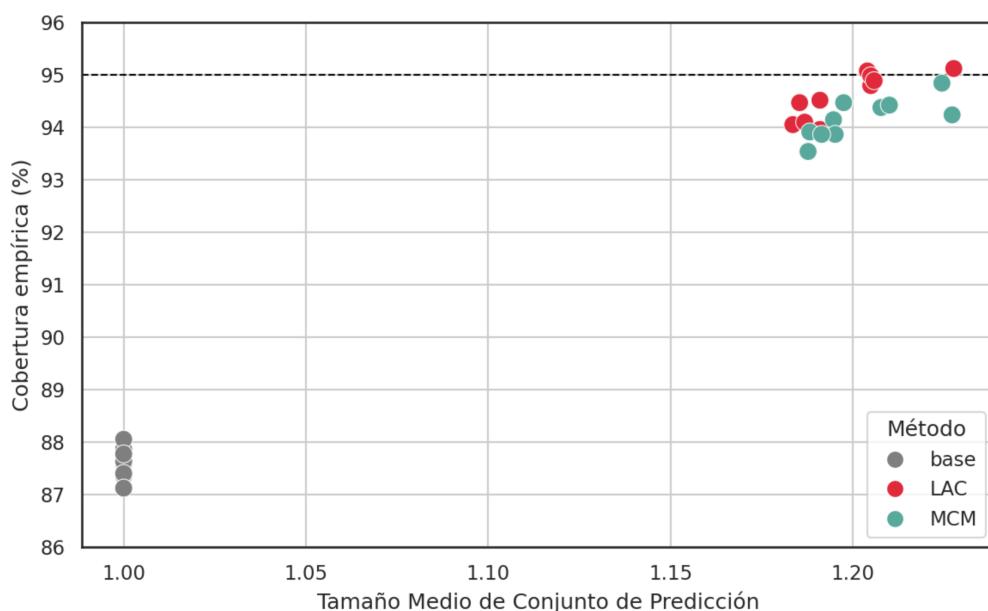


Figura 5.14: Gráfica de dispersión de la cobertura empírica frente al tamaño medio de conjunto de predicción.

Análisis estadístico 5.4: Cobertura empírica en el problema de estimación de mayoría de edad

La cobertura empírica de los métodos cumple normalidad (Shapiro-Wilk, $p > 0.5$) y homocedasticidad (Levene $p > 0.18$). Dado esto, se confirma mediante test ANOVA, que hay diferencias significativas entre las marcas de los distintos métodos ($F(2, 27) = 1097.68, p < 0.001$).

Se aplicó posteriormente la prueba *post-hoc* de comparaciones múltiples Tukey HSD para ver qué métodos presentaban diferencias significativas en la métrica (véase la Tabla 5.10).

Modelo 1	Modelo 2	Dif. media	Valor p	IC 95 %	Signif.
LAC	MCM	-0.0043	0.0415	[-0.0084, -0.0001]	Sí
LAC	base	-0.0697	0.0000	[-0.0738, -0.0655]	Sí
MCM	base	-0.0654	0.0000	[-0.0695, -0.0612]	Sí

Tabla 5.10: Resultados de la prueba *post-hoc* de Tukey HSD para la cobertura empírica entre pares de métodos. Se muestran la diferencia media entre grupos, el valor p ajustado, el intervalo de confianza al 95 % y si la diferencia es estadísticamente significativa ($\alpha = 0.05$).

Análisis de la cobertura en base al tamaño del conjunto

La Figura 5.15 muestra un mapa de calor de la cobertura empírica en base al tamaño del conjunto para los tres métodos planteados. Se observa que el método ‘base’, al no presentar conjuntos de más de un elemento, presenta infracobertura, tal y como hemos podido analizar previamente. En cuanto a los métodos conformales, estos alcanzan, como es de esperar, una cobertura del 100 % en los conjuntos indeterminados. No obstante, el análisis resulta más interesante en los conjuntos de una sola etiqueta, donde se obtienen coberturas superiores al 93 % en ambos métodos conformales, siendo ligeramente mayor en LAC, que alcanza un 93.4 %.

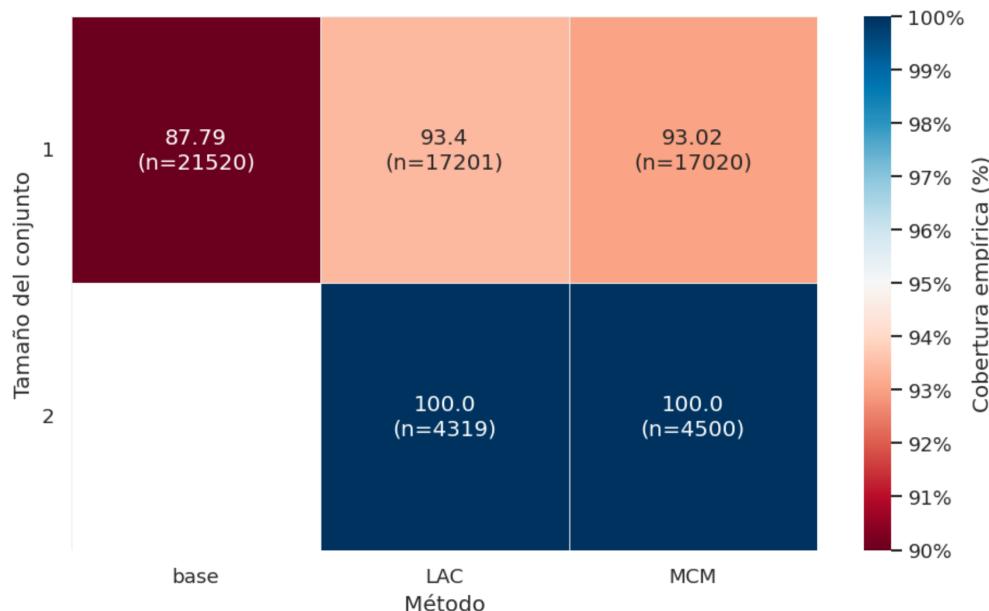


Figura 5.15: Mapa de calor de la cobertura empírica en base al tamaño del conjunto por cada método de predicción a lo largo de 10 ejecuciones. Se especifica entre paréntesis el número de instancias con el número de etiquetas en el conjunto de predicción. La escala de colores está centrada en la cobertura nominal (0.95): los valores por debajo de este umbral se representan en tonos rojos, los superiores en tonos azules, y el blanco indica una cobertura empírica equivalente a la nominal.

Análisis de la cobertura en base al sexo y edad cronológica

Por último, al igual que hicimos en el anterior problema, analizamos la cobertura en base al sexo y a la edad cronológica. Para ello, nos apoyaremos en las Figuras 5.16 y 5.17, donde se muestra la evolución de la cobertura empírica y el tamaño medio del conjunto de predicción en función de la edad cronológica y el sexo de los individuos.

En este caso, se aprecia una tendencia clara y consistente en todos los métodos: la cobertura es elevada tanto en las edades más jóvenes como en las más avanzadas, pero disminuye progresivamente conforme se aproxima la edad cronológica a los 18 años. Este comportamiento resulta coherente, dado que el problema consiste en la estimación de la mayoría de edad, y es precisamente en torno a ese umbral donde se concentra la mayor incertidumbre. En este intervalo, las características morfológicas de individuos menores y mayores de edad tienden a solaparse, lo que dificulta la clasificación y reduce la confianza del modelo en sus predicciones.

Los peores valores de cobertura se registran en los individuos de 17 y 18 años, donde el método ‘base’ presenta una infracobertura en torno al 60-65 %. Los métodos conformales aumentan el ratio de conjuntos indeterminados (de tamaño 2) a medida que aumenta la incertidumbre del modelo, es decir, en aquellos rangos de edad o sexo donde la distinción entre clases resulta más ambigua. De esta forma, aunque sigue existiendo cierta infracobertura, esta es notablemente menor, alcanzando valores superiores al 80 % para ambos sexos en las peores edades, lo que demuestra una mejor calibración y mayor capacidad de los métodos conformales para manejar la incertidumbre en la clasificación. El coste de esto es prácticamente un 50 % de las instancias de estas edades cronológicas indeterminadas, de lo que podemos deducir que la mayoría de los individuos correctamente clasificadas son aquellos casos extremos, es decir, a sujetos cuya morfología maxilofacial

presenta rasgos claramente asociados a la minoría o mayoría de edad, mientras que los casos cercanos al umbral de los 18 años tienden a generar mayor ambigüedad y, por tanto, conjuntos indeterminados.

Esta infracobertura observada en las edades próximas a la mayoría de edad, incluso al aplicar métodos conformales, se ve compensada por las sobrecoberturas registradas en las edades más triviales. Esto plantea la cuestión de si una calibración focalizada en un conjunto de individuos pertenecientes a la franja etaria más conflictiva —aproximadamente entre los 15 y 20 años— podría mejorar la cobertura en esa zona crítica, aunque a costa de un incremento adicional en el tamaño medio de los conjuntos de predicción.

Por último, en relación con el sexo, se observa una cobertura ligeramente inferior en los individuos de sexo femenino en comparación con los masculinos, en consonancia con lo observado en el problema de estimación de edad. Este comportamiento podría deberse a que, como se mencionó anteriormente, la maduración dental suele producirse de forma más temprana en las hembras, lo que genera una mayor variabilidad y, por tanto, una mayor incertidumbre en la clasificación.

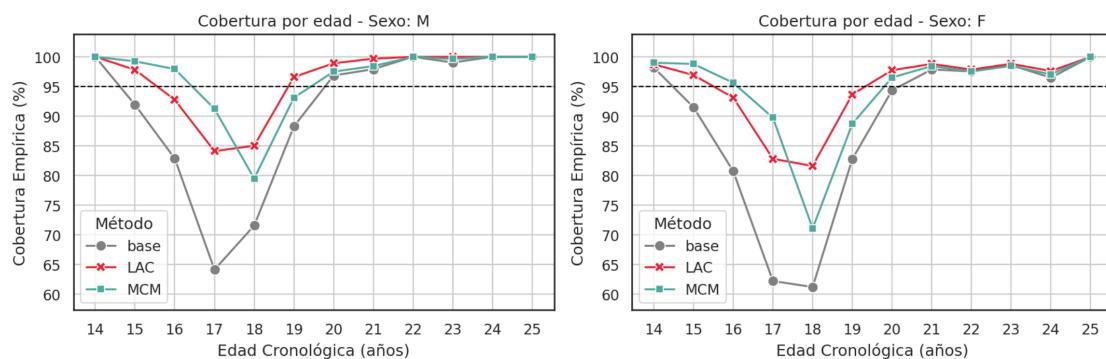


Figura 5.16: Diagrama de líneas de la cobertura empírica en base al sexo y la edad cronológica por cada método de predicción a lo largo de 10 ejecuciones.

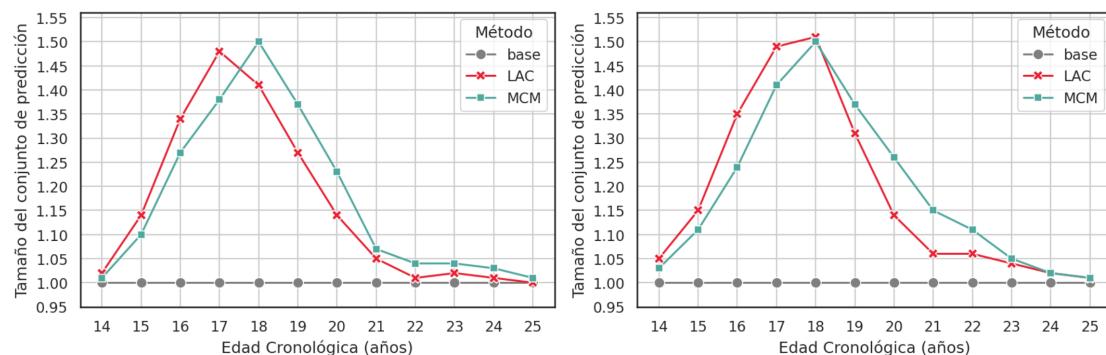


Figura 5.17: Diagrama de líneas del tamaño medio de conjunto de predicción en base al sexo y la edad cronológica por cada método de predicción a lo largo de 10 ejecuciones.

Discusión de resultados

Basándonos únicamente en el criterio de cobertura/tamaño medio del conjunto, el método **LAC** presenta clara ventaja, ya que **ofrece una mayor cobertura al mismo tamaño medio del conjunto que MCM**.

Sin embargo, **si la prioridad en la predicción conformal fuera maximizar la cobertura en los casos de menores, para proteger sus derechos** y minimizar el riesgo de exclusión o clasificación errónea en decisiones sensibles, entonces **el método MCM sería el más adecuado**, ya que ofrece una mayor proporción de aciertos en este grupo etario, incluso a costa de una ligera infracobertura en el resto de la población.

5.8. Experimentación para la clasificación de edad

5.8.1. Entrenamiento de los modelos

Dado que este es un problema directamente derivado del de estimación de edad, se ha optado de nuevo por reutilizar el extractor de características de este. La última capa del modelo ha sido ajustada para producir 12 salidas, correspondientes a las edades enteras del problema (de los 14 a 26 años, ambos inclusive), que son las clases de este. La activación *softmax* se aplica durante la inferencia para obtener probabilidades normalizadas.

Al igual que con la estimación de mayoría de edad, se realizará un ajuste de la nueva cabecera durante 2 épocas, con *learning rate* de 3e-2 y *weight decay* de 2e-4. La función de pérdida utilizada ha sido la ***Cross-Entropy Loss***, adecuada para clasificación multiclase 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.

El tiempo de entrenamiento medio de la cabecera ha sido de 4 minutos y 32 segundos, mientras que el tiempo de calibración ha supuesto 4 minutos y 47 segundos de media. Cabe mencionar que en este problema sí que hay diferencia muy leve de tiempo en la inferencia de métodos. Mientras los métodos ‘base’, LAC y MCM tardan unos 14 segundos en realizar la inferencia con el conjunto de datos de test, los métodos APS, RAPS y SAPS tardan 20-22 segundos, lo que sigue sin ser tiempos muy preocupantes en ningún caso.

5.8.2. Resultados

Análisis de métricas para la clasificación puntual de edad

En este caso no se han analizado las métricas de clasificación de una sola etiqueta, pues no tenía mucho sentido plantearlas tal cual: métricas como la exactitud (*accuracy*) presentan valores muy bajos, ya que existe una gran proximidad entre clases adyacentes y, por tanto, errores que en términos de regresión serían pequeños (por ejemplo, predecir 19 en lugar de 20) se contabilizan como fallos completos en clasificación. También se consideró usar métricas propias de regresión, pero estas obtenían valores artificialmente elevados debido a la discretización previa de la variable objetivo: al forzar las predicciones a valores enteros, se reduce la variabilidad y se exagera la coincidencia con los valores reales.

Análisis de métricas para la clasificación de edad en conjuntos de predicción

Las Tablas 5.11 y 5.12 presenta las métricas sobre los conjuntos de predicción obtenidos con los métodos. Para completar esta información de manera visual, la Figura 5.18 muestra en un gráfico de dispersión la relación de las métricas obtenidas en los diferentes métodos. Llama la atención los resultados extraordinarios de dos métodos:

- El método ‘base’ presenta cobertura del 100 % con tamaño medio del conjunto 13 (el máximo). Esto indica que el conjunto de predicción siempre contiene todas las clases posibles, comportándose de manera no informativa. Al requerir que la suma acumulada de las probabilidades softmax alcance el 95 %, y dada la distribución probabilística del modelo sobre las 13 clases, la estrategia termina incluyendo sistemáticamente la totalidad de las categorías para cumplir con el umbral establecido. Es por ello, que este método será descartado de ahora en adelante, al no tener ningún valor práctico para la toma de decisiones.
- El método MCM presenta de media la mayor cobertura empírica de entre los métodos, pero con tamaños medio de conjuntos también muy superiores, como se evidencia en la Figura 5.18. Esto probablemente se deba a que, en MCM, se calcula el umbral de no conformidad de manera independiente para cada clase utilizando únicamente las instancias pertenecientes a esta. Dado el gran número de clases, cada estimación se realiza con menos datos, lo que incrementa la variabilidad de los umbrales y conduce a intervalos más amplios para garantizar la cobertura deseada. En consecuencia, este método está en clara desventaja respecto al resto.

Método	Cobertura empírica (%)					
	base	LAC	MCM	APS	RAPS	SAPS
Ejecución 1	100.00	94.89	95.96	94.66	95.12	95.26
Ejecución 2	100.00	94.98	95.21	94.01	94.10	95.35
Ejecución 3	100.00	95.40	95.35	94.24	94.28	95.31
Ejecución 4	100.00	95.03	95.91	94.75	94.14	95.26
Ejecución 5	100.00	94.84	95.35	94.24	94.24	95.54
Ejecución 6	100.00	94.24	95.07	94.52	94.89	94.52
Ejecución 7	100.00	94.14	94.80	93.54	93.68	94.66
Ejecución 8	100.00	94.28	94.84	93.31	93.40	95.21
Ejecución 9	100.00	94.75	95.68	94.75	94.89	95.91
Ejecución 10	100.00	95.86	96.00	94.89	95.77	95.96
Media	100.00	94.84	95.42	94.29	94.45	95.30

Tabla 5.11: Cobertura empírica obtenida 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.

Método	Tamaño Medio del Conjunto					
	base	LAC	MCM	APS	RAPS	SAPS
Ejecución 1	13.00	5.85	7.58	6.08	6.03	6.29
Ejecución 2	13.00	6.00	7.63	6.18	5.96	6.31
Ejecución 3	13.00	6.12	7.73	6.07	5.99	6.37
Ejecución 4	13.00	5.99	7.73	6.28	6.02	6.33
Ejecución 5	13.00	5.93	7.48	6.10	5.90	6.29
Ejecución 6	13.00	5.74	7.68	5.97	5.99	6.12
Ejecución 7	13.00	5.75	7.26	5.81	5.73	6.05
Ejecución 8	13.00	5.82	7.46	5.94	5.74	6.23
Ejecución 9	13.00	5.96	7.88	6.30	6.03	6.45
Ejecución 10	13.00	6.16	7.70	6.14	6.08	6.37
Media	13.00	5.93	7.61	6.09	5.95	6.28

Tabla 5.12: Tamaño medio del conjunto de predicción obtenido por cada método a lo largo de las distintas ejecuciones. Se presentan los valores para cada ejecución individual, así como la media final de cada métrica.

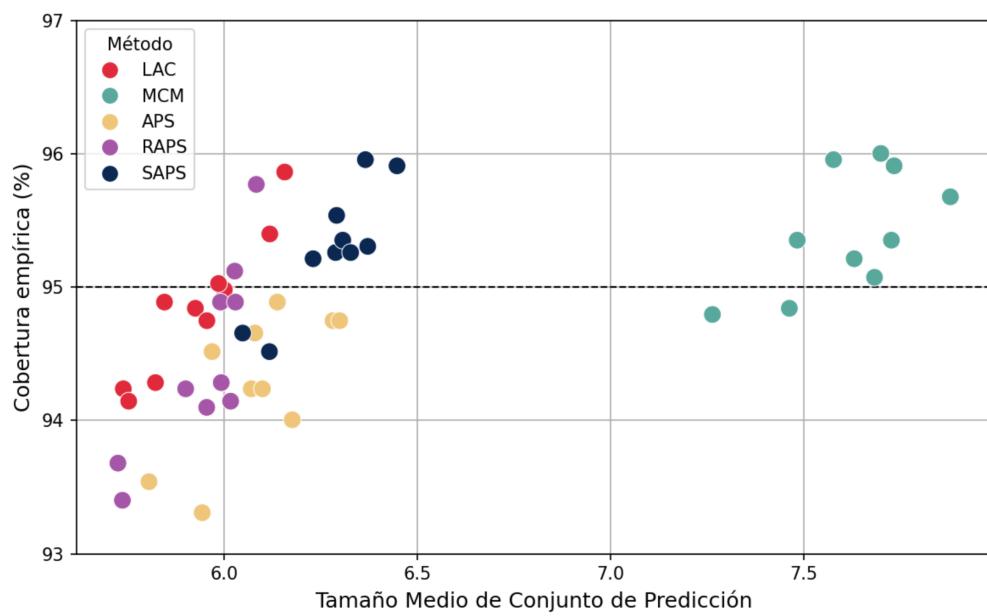


Figura 5.18: Gráfica de dispersión de la Cobertura empírica frente al Tamaño medio de conjunto de predicción. No se incluyen los puntos del método ‘base’ dado que estos están muy alejados y concentrado en cobertura del 100 % y tamaño medio de conjunto 13, dificultando la visualización de la gráfica.

Análisis estadístico 5.5: Cobertura empírica y tamaño medio del conjunto de predicción en el problema de estimación de mayoría de edad

Se llevó a cabo una comparación estadística entre los métodos, —descartando el ‘base’—. La comparación estadística entre los métodos de la primera nube se llevó a cabo mediante un test ANOVA, tanto para la cobertura empírica ($F(5, 36) > 10^5$, $p < 0.001$) como para el tamaño medio del conjunto de predicción ($F(5, 36) > 10^5$, $p < 0.001$). El análisis asume normalidad (Shapiro-Wilk: $p = 0.6174$ para la cobertura empírica y $p = 0.2465$ para el tamaño medio) y homocedasticidad (Levene: $p > 0.65$ en ambas métricas). Los resultados de la prueba post-hoc de Tukey para la comparación por pares de métodos en ambas métricas se presentan en las Tablas 5.13 y 5.14.

Modelo 1	Modelo 2	Dif. media	Valor p	IC 95 %	Signif.
APS	LAC	0.0055	0.1766	[-0.0014, 0.0125]	No
APS	MCM	0.0113	0.0003	[0.0043, 0.0182]	Sí
APS	RAPS	0.0016	0.9628	[-0.0053, 0.0086]	No
APS	SAPS	0.0101	0.0014	[0.0031, 0.017]	Sí
LAC	MCM	0.0058	0.1465	[-0.0012, 0.0127]	No
LAC	RAPS	-0.0039	0.5075	[-0.0109, 0.003]	No
LAC	SAPS	0.0046	0.3521	[-0.0024, 0.0115]	No
MCM	RAPS	-0.0097	0.0024	[-0.0166, -0.0027]	Sí
MCM	SAPS	-0.0012	0.9875	[-0.0082, 0.0057]	No
RAPS	SAPS	0.0085	0.01	[0.0015, 0.0154]	Sí

Tabla 5.13: Resultados de la prueba *post-hoc* de Tukey HSD para la cobertura empírica entre pares de métodos. 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 p	IC 95 %	Signif.
APS	LAC	-0.1558	0.1271	[-0.3383, 0.0267]	No
APS	MCM	1.5265	<0.001	[1.344, 1.709]	Sí
APS	RAPS	-0.1403	0.2043	[-0.3228, 0.0422]	No
APS	SAPS	0.1931	0.0333	[0.0106, 0.3756]	Sí
LAC	MCM	1.6823	<0.001	[0.14998, 1.8648]	Sí
LAC	RAPS	0.0155	0.9992	[-0.167, 0.198]	No
LAC	SAPS	0.3489	<0.001	[0.1664, 0.5314]	Sí
MCM	RAPS	-1.6668	<0.001	[-1.8493, -1.4843]	Sí
MCM	SAPS	-1.3335	<0.001	[-1.516, -1.151]	Sí
RAPS	SAPS	0.3334	<0.001	[0.1509, 0.5159]	Sí

Tabla 5.14: Resultados de la prueba *post-hoc* de Tukey HSD para el tamaño medio del conjunto de predicción entre pares de métodos. Se muestran la diferencia media entre grupos, el valor p ajustado, el intervalo de confianza al 95 % y si la diferencia es estadísticamente significativa ($\alpha = 0.05$).

El resto de los métodos muestra resultados más comparables, formando una nube de

puntos claramente visible en la Figura 5.18. El Análisis Estadístico 5.5 permite establecer las siguientes conclusiones:

- SAPS presenta una cobertura significativamente superior a la de APS y RAPS, aunque con un tamaño medio de conjunto mayor, lo que refleja el compromiso inherente entre fiabilidad y precisión en los métodos de predicción conformal.
- LAC alcanza una cobertura empírica estadísticamente equivalente a SAPS, pero con un tamaño medio de conjunto significativamente menor. Esta combinación de alta cobertura y conjuntos más compactos posiciona a LAC como el método con la mejor relación cobertura-tamaño entre todas las alternativas evaluadas.

Análisis de la cobertura en base al tamaño del conjunto de predicción

De igual manera a como hicimos con el problema de regresión, aquí también analizaremos la cobertura en base al tamaño del conjunto de predicción conformal. La Figura 5.19 presenta un mapa de calor que resume, para cada método, la cobertura empírica obtenida según el número de etiquetas incluidas en el conjunto de predicción.

En términos generales, se observan dos tendencias clave:

- **Cobertura en aumento con el tamaño de los conjuntos:** todos los métodos tienden a mejorar su cobertura a mayor tamaño de conjuntos de predicción devuelven. Esto es esperable, ya que, cuanto más etiquetas tiene el conjunto, más probable es que incluya la de la clase verdadera.
- **Sobre cobertura como síntoma de desequilibrio:** la presencia de sobre cobertura en determinados tamaños implica, inevitablemente, infracobertura en otros. Cuando este patrón se repite y la sobre cobertura se concentra en conjuntos de gran tamaño, suele indicar que el método está “compensando” un mal ajuste en los conjuntos pequeños, lo cual resulta indeseable. En contextos prácticos, esto significa sacrificar precisión en situaciones de alta confianza para inflar artificialmente los resultados en escenarios menos exigentes.

Y ahora, centrándonos en los métodos:

- **MCM:** Genera **conjuntos de predicción excesivamente conservadores**, con un gran número de etiquetas. La mayoría de los conjuntos (un 53 %) tiene más de 7 etiquetas. Además, presenta infracobertura para conjuntos de 7 etiquetas o menos y sobre cobertura sistemática en conjuntos más grandes, lo que evidencia una adaptabilidad limitada al no ajustar adecuadamente el tamaño del conjunto según el nivel de incertidumbre inherente a cada instancia.
- **LAC:** Genera conjuntos de tamaño muy variable, que oscilan entre 1 y 13 etiquetas, si bien más del 90 % tienen entre 3 y 8 etiquetas. Presenta valores de cobertura cercanos al nominal para prácticamente todos los tamaños de conjunto de predicción, siendo la mayor desviación para conjuntos de más de tamaño 3 de 2 puntos porcentuales.
- **APS:** Presenta una **muy alta variabilidad en tamaños de conjuntos de predicción**, con un 96 % de los conjuntos de entre 2 y 9 etiquetas. Al igual que LAC, presenta infracobertura en conjuntos de menos de 7 etiquetas en conjuntos más grandes, pero de manera más pronunciada, evidenciando un mayor desequilibrio en las coberturas en base al tamaño.

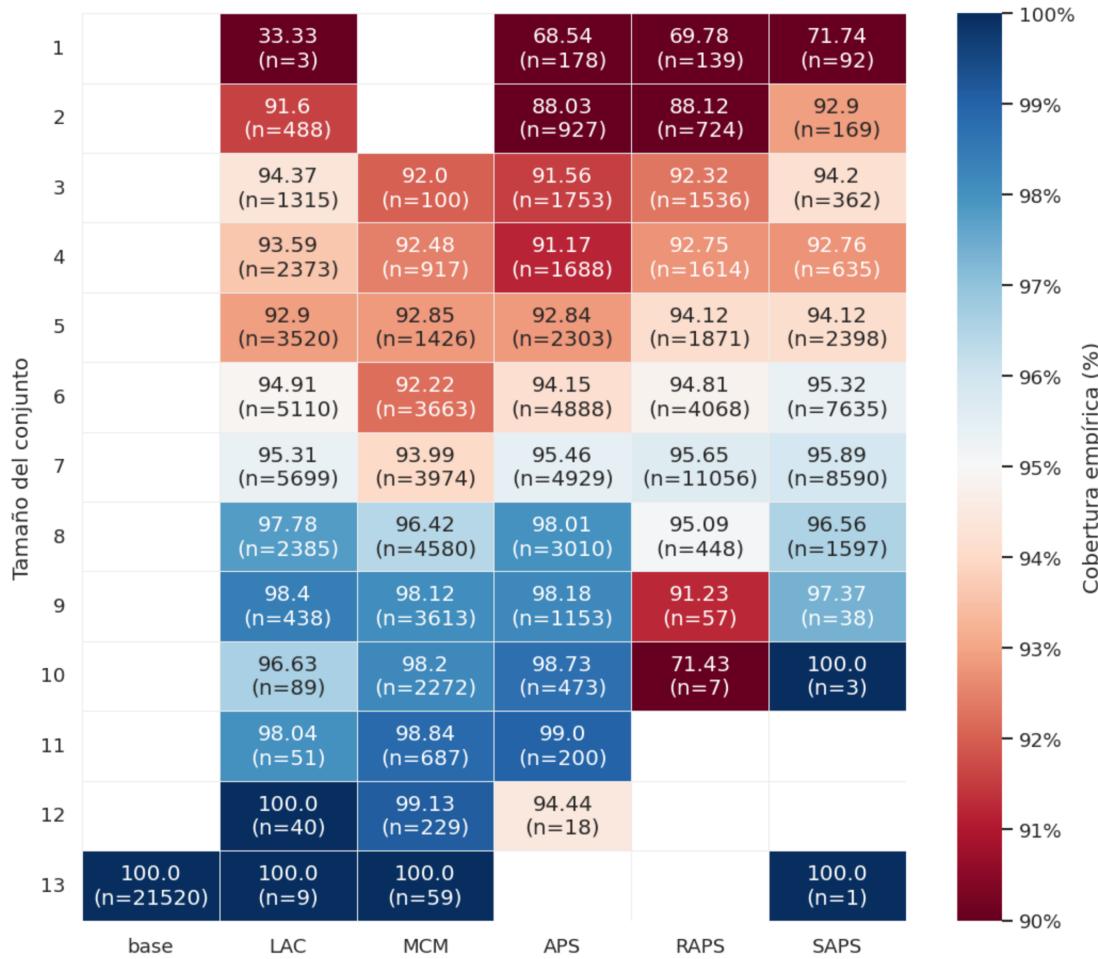


Figura 5.19: Mapa de calor de cobertura empírica en base al tamaño del conjunto por cada método de predicción a lo largo de las distintas ejecuciones. Se especifica entre paréntesis el número de instancias clasificadas en cada franja de amplitud de intervalo. La escala de colores está centrada en la cobertura nominal (0.95): los valores por debajo de este umbral se representan en tonos rojos, los superiores en tonos azules, y el blanco indica una cobertura empírica equivalente a la nominal.

- **RAPS:** La extensión de APS consigue reducir la variabilidad de tamaños del conjunto, con más del 96 % de los conjuntos de entre 2 y 7 etiquetas. Cabe comentar que logra su objetivo de reducir el tamaño de los conjuntos de predicción de APS manteniendo la cobertura marginal global. De hecho, hasta mejora su cobertura: aumenta aquellas marcas en las que APS presenta infracobertura. Sin embargo, obtiene tasas de cobertura bajas para conjuntos de mayor tamaño. Esto puede indicar una **penalización de tamaño excesiva en instancias con alto nivel de incertidumbre**, en las que la regularización de RAPS ha penalizado en exceso la inclusión de clases adicionales.
- **SAPS:** Presenta **conjuntos de predicción muy equilibrados**, con un número de etiquetas ni muy conservador ni excesivamente arriesgado. Un 93 % de los conjuntos de predicción tiene entre 5 y 8 etiquetas. También presenta coberturas equilibradas, con valores más cercanos al nominal que los métodos hermanos (APS y RAPS) para prácticamente todos los tamaños de conjunto.

LAC y SAPS son los dos métodos más equilibrados en este apartado. Presentan los

valores de cobertura más estables para los diferentes tamaños del conjunto de predicción, evitando extremos de infracobertura o sobrecobertura excesiva. La principal diferencia entre ambos es que SAPS presenta tamaños de conjuntos con menor variabilidad que LAC.

Análisis de la cobertura en base a la edad cronológica

Y, en este último apartado, tal y como se hizo con el problema de regresión, se ha analizado la cobertura en base a la edad cronológica de cada individuo, que en este caso es la etiqueta real de cada instancia. Las Figuras 5.20 y 5.21 muestran la relación de la cobertura empírica y el tamaño medio de los conjuntos de predicción para las distintas edades cronológicas, en cada sexo.

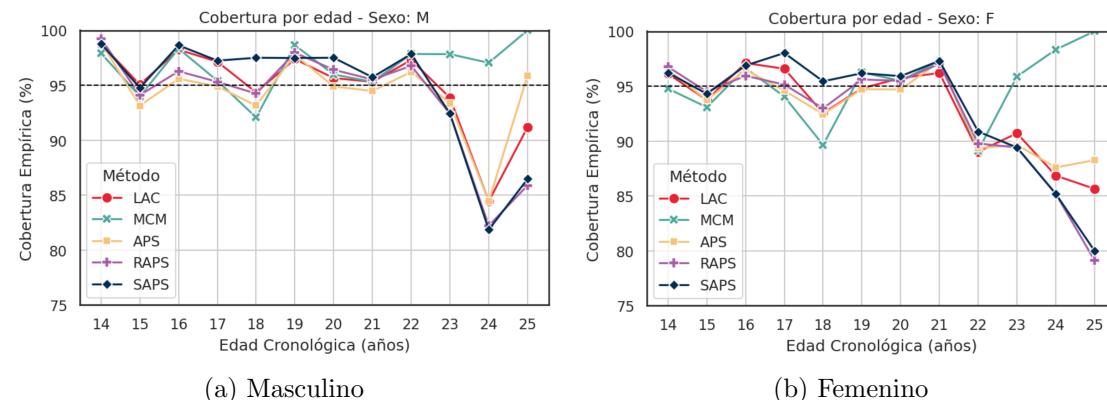


Figura 5.20: Gráficos de líneas de la cobertura empírica del intervalo de predicción (%) para cada método en función de la edad cronológica entera de los individuos, diferenciando por sexo.

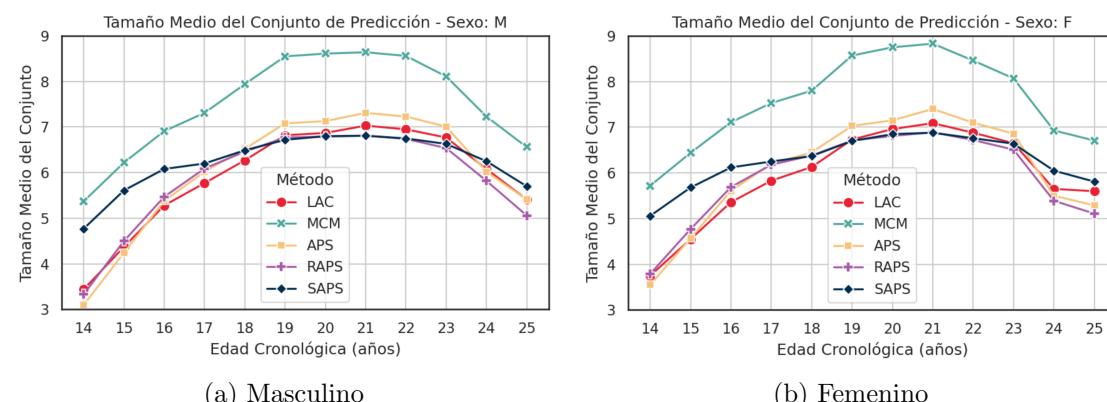


Figura 5.21: Gráficos de líneas del tamaño medio del conjunto de predicción para cada método en función de la edad cronológica entera de los individuos, diferenciando por sexo.

Se observa un patrón general común en casi todos los métodos —salvo MCM—: la cobertura empírica disminuye notablemente para edades avanzadas, especialmente a partir de los 22 años, probablemente debido a la escasez de ejemplos en este rango etario. Sin embargo, a diferencia de con el problema de regresión, donde los intervalos de predicción aumentaban continuamente con la edad, aquí el tamaño medio de los conjuntos de predicción crece hasta un máximo alrededor de los 20-21 años, y posteriormente

disminuye en las edades más avanzadas, lo que refleja una falta de diversidad en las predicciones y una subestimación de la incertidumbre epistémica, resultando en conjuntos potencialmente demasiado optimistas y con cobertura insuficiente para estos grupos subrepresentados

Entre los métodos, se identifican algunos patrones destacables:

- **MCM presenta una alta variabilidad**, con infracobertura y sobrecobertura distribuidas de manera irregular a lo largo de las edades, probablemente debido a la limitada representatividad de las puntuaciones de no conformidad en cada clase.
- **SAPS**, de manera consistente con lo observado en el apartado anterior, **mantiene una mayor estabilidad en el tamaño medio de los conjuntos**. Además, es el método que mejor cobertura logra para edades jóvenes menores de 23, alcanzando coberturas muy cercanas al 95 %, y en la mayoría de casos superándolo.
- El resto de métodos adaptativos y el método LAC presentan tamaño medio de conjunto muy variables para las distintas edades cronológicas. Su cobertura fluctúa alrededor de SAPS, si bien para la mayoría de edades ligeramente por debajo.

Respecto a la variable sexo, se observa una divergencia consistente con los hallazgos del problema de regresión: el sexo femenino presenta un deterioro anticipado de la cobertura, detectable desde los 22 años, mientras que el sexo masculino mantiene coberturas adecuadas hasta los 23 años, donde registra su primera caída significativa.

Discusión de resultados

En definitiva, para este problema, LAC y SAPS se perfilan como los métodos más equilibrados, ya que ambos se aproximan a la cobertura nominal y mantienen una adecuada relación entre cobertura y tamaño medio de los conjuntos.

- **LAC sobresale por** su sencillez de implementación y por **generar conjuntos de tamaño moderado, que alcanzan una cobertura muy próxima a la requerida**, muy eficientes en términos prácticos.
- **SAPS**, por su parte, se caracteriza por **producir conjuntos de predicción con tamaños ligeramente superior a LAC, pero menos variables, ofreciendo una mayor consistencia**. Además, **presenta una mayor adaptatividad respecto al tamaño, con tasas de cobertura más estables** a lo largo de los diferentes tamaños del conjunto.

En cualquier caso, todos los métodos analizados muestran tasas de cobertura muy bajas en instancias correspondientes a edades avanzadas, lo cual puede atribuirse a la escasez de ejemplos en este rango o a la variabilidad fisiológica inherente en edades avanzadas. Sería, por tanto, recomendable disponer de más datos en estas edades para mejorar la capacidad predictiva y la robustez de los modelos, lo que llevaría a una mejor cobertura.

Capítulo 6

Conclusiones y trabajos futuros

6.1. Conclusiones

A la luz de los resultados obtenidos, se puede concluir que el empleo de métodos de predicción conformal constituye una herramienta de gran utilidad, ya que ofrece beneficios significativos en términos de cuantificación de la incertidumbre a un coste computacional muy reducido. Esto resulta especialmente relevante en contextos sensibles, donde la toma de decisiones derivada de estas estimaciones (p. ej., en procedimientos de asilo o investigaciones forenses) incide directamente sobre los derechos fundamentales de las personas.

En primer lugar, se observa que tanto las predicciones puntuales como las conformales mejoran en sus métricas a medida que aumenta el desempeño del modelo subyacente. Es decir, cuanto más preciso resulta el modelo al estimar el valor esperado o clase verdadera de cada instancia, mayor es también la calidad de los intervalos o conjuntos conformales asociados. Por tanto, **el objetivo de mejorar la precisión del modelo base y el de obtener mejores intervalos conformales es común y está alineado**. Esta sinergia entre el modelo y el método conformal es también crucial a la hora de su implementación.

En términos prácticos, **la mayoría de las técnicas de predicción conformal presentan la ventaja de no requerir un reentrenamiento completo del modelo**, siempre que se disponga de suficientes ejemplos para la calibración, distintos de los empleados en el entrenamiento o la validación. En caso de no contar con este volumen de datos, resulta necesario reentrenar el modelo tras realizar una nueva partición del conjunto de datos que reserve un subconjunto específico para la calibración. Existen, no obstante, métodos que sí implican modificaciones en la arquitectura y reentrenamiento o, como la *Conformalized Quantile Regression*. Este enfoque requiere incorporar nuevas salidas —de la *Quantile Regression*— y entrenar nuevamente el modelo bajo una función de pérdida adaptada. Sin embargo, cabe destacar que este proceso de ajuste resulta relativamente poco costoso, dado que el modelo suele converger en pocas épocas a partir de uno base ya entrenado con una única salida.

La principal contribución de la predicción conformal reside en **proporcionar una medida rigurosa de incertidumbre a través del conjunto predicho**. Algunas técnicas miden la incertidumbre del conjunto completo, de modo que todos los ejemplos reciben conjuntos de predicción del mismo tamaño. En estos casos, la incertidumbre no se refleja en la variabilidad del tamaño del conjunto, sino en la frecuencia con que dicho conjunto contiene o no el valor o clase verdadero. Así, los tamaños de los conjuntos conformales son constantes, lo que no deja de ser una aproximación muy cercana al

análisis del error tradicional: se garantiza que, en promedio, el error se mantenga bajo un umbral prefijado. Sin embargo, los métodos conformales adaptativos entrañan un mayor potencial, ya que ajustan dinámicamente el tamaño del conjunto de predicción en función de la dificultad de predecir cada instancia. De este modo, ejemplos en los que el modelo está más seguro tienden a recibir conjuntos más pequeños, mientras que en aquellos en los que la predicción es más incierta, los conjuntos se amplían para mantener la garantía de cobertura. Esta adaptatividad permite capturar mejor tanto la incertidumbre epistémica (p.ej., los intervalos de edad más amplios por escasez de datos en individuos de edad avanzada) como la estocástica (p.ej., los intervalos amplios por la mayor variabilidad fisiológica en edades avanzadas), reflejando así de manera más fiel la heterogeneidad de los datos. Esta adaptatividad no siempre consigue conjuntos más pequeños, pero aproxima más a una cobertura condicional.

En cuanto a los **costes de implementar la predicción conformal**, estos se concentran en dos aspectos principales:

- **Reserva de datos para calibración:** destinar una fracción del conjunto de entrenamiento puede degradar ligeramente el rendimiento del modelo en la predicción puntual. No obstante, en el caso analizado, el volumen de datos fue suficiente para que la retención del 20% apenas afectara los resultados: en el problema de estimación de edad, el error absolute medio apenas se resentía un 3.5% (de 1.16 en métodos no conformales a 1.20 en ICP); o en el problema de estimación de la mayoría de edad, la exactitud solo variaba un 0.63% de media (de 87.63% en métodos no conformales al 87% en los conformales). En contextos con conjuntos de datos reducidos, este aspecto puede volverse más problemático, por lo que resultaría recomendable explorar estrategias alternativas de predicción conformal como *Jackknife+*.
- **Incorporación del proceso de calibración e inferencia conformal:** la calibración supone una fase adicional, y en algunos casos (con los métodos APS, RAPS y SAPS) la inferencia conformal también un mayor coste en tiempo que la inferencia puntual. Aun así, en el problema de estimación de edad, los tiempos de calibración han supuesto en media menos de un 6% de los tiempos totales de modelado, lo que evidencia que la sobrecarga computacional introducida por este método es mínima comparada con el coste global del proceso, especialmente cuando se emplean modelos de ML complejos.

Cabe destacar que, si bien todos los métodos conformales garantizan teóricamente la cobertura marginal, en la práctica exhiben diferencias sustanciales:

- Las garantías de cobertura estadística no garantizan la cobertura empírica al nivel deseado sobre el conjunto de datos nuevo. La elección de la función de no conformidad es crucial, pues determina la eficiencia de los intervalos: funciones mal calibradas pueden producir intervalos excesivamente amplios o poco informativos, mientras que elecciones adecuadas permiten intervalos más ajustados sin perder la validez. Nuestros resultados para el problema de clasificación de edad, en igualdad de condiciones (mismo dataset de calibración y modelo), exhiben métodos que logran desde un 94.29% de cobertura hasta un 95.42% para una cobertura global deseada del 95%.
- Algunas variantes tienden a generar intervalos o conjuntos inestables, sobrecubriendo ciertos grupos e infracubriendo otros, mientras que otros

ofrecen un mayor equilibrio, reduciendo la sobrecobertura en los primeros en favor de una cobertura más homogénea entre subpoblaciones.

En resumen, y atendiendo a los métodos analizados en este trabajo, se pueden establecer algunas conclusiones claras sobre estos:

- Para problemas de regresión, CQR se consolida como la opción más robusta. Sus intervalos muestran una cobertura muy cercana a la nominal y mantienen tasas de cobertura empírica consistentes tanto en diferentes tamaños muestrales como en subpoblaciones definidas por edad cronológica y sexo. Esto lo convierte en un método confiable y estable en diversos escenarios.
- En problemas de clasificación, no es tan claro:
 - LAC es la opción predeterminada: fácil de implementar y eficiente, logra valores muy cercanos a la cobertura marginal con conjuntos de predicción de tamaño muy moderado.
 - MCM es una buena opción en casos con pocas clases, donde es prioritario lograr la cobertura requerida en todas ellas.
 - SAPS es la mejor opción para problemas con muchas clases, poniéndo énfasis en la adaptatividad, logra tasas de cobertura marginal muy sólidas y conjuntos de predicción de tamaño muy estable.

En consecuencia, **del mismo modo que las predicciones puntuales exigen un análisis de error, las conformales requieren una evaluación sistemática de la cobertura empírica y la comparación entre métodos**. Este análisis debe identificar discrepancias entre la cobertura nominal y la real, detectar subpoblaciones sistemáticamente infracubiertas o sobrecubiertas, y examinar el tamaño de los intervalos. Unos intervalos excesivamente amplios carecen de utilidad práctica; unos demasiado estrechos, comprometen las garantías de cobertura. **El desafío inmediato reside en avanzar hacia métodos que se aproximen a la cobertura condicional, cerrando la brecha entre teoría y práctica.**

6.2. Valoración del trabajo realizado

En relación con los objetivos planteados en la introducción, todos han sido cumplidos de manera satisfactoria. Se ha realizado un análisis detallado de las distintas técnicas de predicción conformal y del estado del arte en la cuantificación de la incertidumbre aplicada a problemas de estimación de edad. Asimismo, se implementaron con éxito diversas variantes de predicción conformal inductiva (disponibles en el repositorio [esdavide2910/tfg-bioprofile-uncertainty](#)), tanto para tareas de regresión como de clasificación, aplicándolas a un caso de estimación del perfil biológico. Para garantizar una comparación justa, dichas técnicas se contrastaron con aproximaciones heurísticas de predicción interválica y basada en conjuntos. Los resultados obtenidos han permitido evidenciar tanto el potencial como las limitaciones de estos métodos en el contexto específico de la estimación de edad a partir de radiografías maxilofaciales.

A lo largo del desarrollo de este trabajo se ha evidenciado la adquisición y consolidación de competencias clave en el ámbito académico. En primer lugar, se ha fortalecido la gestión de recursos académicos, desde la búsqueda de fuentes fiables y actualizadas

hasta la correcta aplicación de normas de citación, junto con la comprensión crítica de documentos especializados. Cientos de referencias a trabajos del ámbito del aprendizaje automático y de la antropología forense han servido de base para construir un marco teórico sólido y fundamentar adecuadamente las decisiones metodológicas adoptadas. Esta competencia, unida a los conocimientos básicos de programación y al manejo de modelos avanzados de *Machine Learning*, ha resultado esencial para la implementación de las técnicas de predicción conformal. Asimismo, he reforzado la capacidad de elaborar y maquetar la memoria del proyecto con un formato claro, estructurado y profesional. Por último, destaco la competencia desarrollada en la gestión de un proyecto individual, que ha requerido planificación, organización del tiempo y toma de decisiones autónomas —aunque en algunos casos guiadas por los tutores— orientadas a alcanzar los objetivos planteados.

6.3. Trabajos futuros

Aún queda por explorar un **análisis de estas herramientas con otros conjuntos de datos**, preferiblemente más equilibrados y con mayor diversidad en los rangos de edad, para un análisis más rico.

Una de las virtudes más destacables de la predicción conformal es su inherente flexibilidad, que permite mejorar sus capacidades mediante su integración sinérgica con otros marcos metodológicos. Esta versatilidad abre la vía para el desarrollo de sistemas de *Machine Learning* más robustos y confiables. En concreto, su potencial se puede ampliar en varias direcciones:

- **Integración con otros paradigmas de cuantificación de la incertidumbre:** La predicción conformal puede combinarse con métodos como *Monte Carlo Dropout* (como en [108]) o la regresión cuantílica con modelos *ensembles* para generar intervalos conformales que no solo garantizan una cobertura marginal, sino que también se benefician de una estimación de incertidumbre más afinada.
- **Combinación con técnicas de detección de datos fuera de distribución (*Out-of-Distribution*, OOD) y explicabilidad (*Explainable Artificial Intelligence*, XAI):** La unión de estas áreas es fundamental para construir sistemas confiables. La detección de OOD es relevante dado que la suposición más importante realizada por la predicción conformal es que los datos son intercambiables y, por tanto, pertenecientes a una misma distribución subyacente. Cuando esta premisa fundamental se viola, las garantías de cobertura estadística dejan de ser válidas. Este mecanismo actuaría como un sistema de alerta temprana, identificando ejemplos para los cuales las predicciones, y sus intervalos de incertidumbre asociados, deben ser tratados con precaución.

Por su parte, las técnicas de explicabilidad (XAI) aportan transparencia al proceso, permitiendo comprender las razones detrás de la incertidumbre cuantificada. Por ejemplo, XAI puede revelar si un intervalo amplio se debe a la escasez de datos de entrenamiento en una región específica (incertidumbre epistémica) o a una alta variabilidad inherente en la característica objetivo (incertidumbre aleatoria). Esta distinción es crucial para tomar decisiones informadas: en el primer caso, se podría resolver recopilando más datos específicos, mientras que en el segundo, la incertidumbre es inherente al problema. Así, la combinación de predicción conformal, detección de OOD y XAI no solo identifica cuándo una predicción es incierta,

sino también por qué lo es, permitiendo a los expertos (p.ej., antropólogos forenses) evaluar la credibilidad de los intervalos conformales en contextos de toma de decisiones críticas.

Por último, el **aprovechamiento de la información experta del dominio para mejorar el análisis de la cobertura** representa una dirección fundamental. De esta forma, al potencial técnico de la herramienta le acompañe una interpretación significativa y rigurosa, permitiendo a los antropólogos forenses realizar un análisis mucho más rico y preciso, analizando tendencias y certidumbres por subpoblaciones, y trasladando así los resultados abstractos del modelo a conclusiones prácticas y accionables en contextos forenses reales.

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, 35, 36, 40].
- [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 pág. 2].
- [7] M. Skinner, D. Alempijevic y M. Djuric-Srejic, “Guidelines for International Forensic Bio-archaeology Monitors of Mass Grave Exhumations,” *Forensic Science International*, vol. 134, n.^o 2, págs. 81-92, 2003. [Citado en pág. 2].
- [8] J. A. Sanchis-Gimeno, J. Iglesias-Bexiga, M. E. Schwab, G. López-García, E. Ariza, A. Calpe, M. Mezquida, S. Nalla e I. Ercan, “Identification success rates in the post-Spanish Civil War mass graves located in the cemetery of Paterna, Spain: Meta-research on 15 mass graves with 933 subjects,” *Forensic Science International*, vol. 361, págs. 112-122, ago. de 2024. [Citado en pág. 2].
- [9] M. Baeta, C. Núñez, S. Cardoso, L. Palencia-Madrid, L. Herrasti, F. Etxeberria y M. M. de Pancorbo, “Digging up the recent Spanish memory: genetic identification of human remains from mass graves of the Spanish Civil War and posterior dictatorship,” *Forensic Science International: Genetics*, vol. 19, págs. 272-279, 2015. [Citado en pág. 2].
- [10] V. Ataliva, N. F. Bahamondes, C. M. Suárez y B. Rosignoli, “Arqueología Forense y prácticas genocidas del Cono Sur americano: reflexionando desde los confines,” *Revista de Arqueología Americana*, vol. 41, págs. 403-441, jun. de 2024. [Citado en pág. 2].

- [11] T. Tanaka, "International Humanitarian Law (IHL) and Forensic Document Examination," *Journal of the American Society of Questioned Document Examiners*, vol. 23, n.º 1, 2020. [Citado en pág. 2].
- [12] T. Thompson y S. Black, *Forensic Human Identification: An Introduction*, 1.^a ed. Taylor & Francis, 2006. [Citado en pág. 2].
- [13] D. Higgins, A. B. Rohrlach, J. Kaidonis, G. Townsend y J. J. Austin, "Differential Nuclear and Mitochondrial DNA Preservation in Post-Mortem Teeth with Implications for Forensic and Ancient DNA Studies," *PLoS One*, vol. 10, n.º 5, págs. 1-17, 2015. [Citado en pág. 2].
- [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. 2].
- [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. 2].
- [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. 3].
- [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. 3].
- [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. 3].
- [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. 3].
- [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. 3, 36, 37].
- [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. 4].
- [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. 4].
- [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. 4, 38].

- [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. 4].
- [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. 4].
- [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. 4].
- [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. 4].
- [29] 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 Computing and Applications*, vol. 35, págs. 5987-5998, 2022. [Citado en págs. 4, 25, 39].
- [30] 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," *International Journal of Legal Medicine*, vol. 138, págs. 1741-1757, 2024. [Citado en pág. 4].
- [31] J. R. Kim, W. H. Shim, H. M. Yoon, S. H. Hong, J. S. Lee, Y. A. Cho 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. 4, 41].
- [32] 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. 4].
- [33] H. Lee, S. Tajmir, M. Zissen, B. Yeshivas, 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. 4].
- [34] K. Imaizumi, S. Usui, K. Taniguchi, Y. Ogawa, T. Nagata, K. Kaga, H. Hayakawa y S. Shiotani, "Development of an age estimation method for bones based on machine learning using post-mortem computed tomography images of bones," *Forensic Imaging*, vol. 26, pág. 200 477, 2021. [Citado en pág. 4].
- [35] 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. 4].
- [36] L. Ferrante y R. Cameriere, "Statistical methods to assess the reliability of measurements in procedures for forensic age estimation," *International Journal of Legal Medicine*, vol. 123, n.º 4, págs. 277-283, 2009. [Citado en pág. 4].
- [37] Ministerio del Interior de España, "Informe anual sobre personas desaparecidas 2025," Ministerio del Interior, inf. téc., 2025. [Citado en págs. 4, 5].

- [38] 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. 5].
- [39] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2024,” Fiscalía General del Estado, Madrid, España, inf. téc., 2024. [Citado en pág. 5].
- [40] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2019,” Fiscalía General del Estado, Madrid, España, inf. téc., 2019. [Citado en pág. 5].
- [41] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2016,” Fiscalía General del Estado, Madrid, España, inf. téc., 2016. [Citado en pág. 5].
- [42] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2013,” Fiscalía General del Estado, Madrid, España, inf. téc., 2013. [Citado en pág. 5].
- [43] 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. 5].
- [44] M. V. Tidball-Binz y S. M. Cordner, “Humanitarian forensic action: A new forensic discipline helping to implement international law and construct peace,” *Wiley Interdisciplinary Reviews: Forensic Science*, vol. 4, n.º 1, e1438, 2021. [Citado en pág. 5].
- [45] Comisión Europea, *Guía de uso del ECTS*, es, Consultado el 28/08/2025, 2015. URL: https://education.ec.europa.eu/sites/default/files/document-library-docs/ects-users-guide_es.pdf. [Citado en pág. 6].
- [46] ERI Economic Research Institute. “Engineer Artificial Intelligence Salary in Spain.” Consultado el 28/08/2025. URL: <https://www.erieri.com/salary/job/engineer-artificial-intelligence/spain>. [Citado en pág. 11].
- [47] A. Turing, “I.—Computing Machinery and Intelligence,” *Mind*, vol. LIX, n.º 236, págs. 433-460, 1950. [Citado en pág. 12].
- [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. 12].
- [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. 12, 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. 12, 15].
- [51] J. R. Quinlan, “Induction of decision trees,” *Machine Learning*, vol. 1, págs. 81-106, 1986. [Citado en pág. 12].
- [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, 13, 15, 19, 25].
- [54] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Heidelberg: Springer-Verlag, 2006. [Citado en págs. 13, 14].

- [55] E. Alpaydin, *Introduction to Machine Learning*, 2nd. The MIT Press, 2010. [Citado en pág. 13].
- [56] Y. LeCun, Y. Bengio y G. Hinton, “Deep Learning,” *Nature*, vol. 521, págs. 436-44, 2015. [Citado en pág. 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, 20-22, 26, 115].
- [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. 17].
- [65] R. Szeliski, *Computer Vision: Algorithms and Applications*, 1st. Berlin, Heidelberg: Springer-Verlag, 2010. [Citado en págs. 18, 20, 24, 25].
- [66] L. Bottou, “Large-Scale Machine Learning with Stochastic Gradient Descent,” *Proceedings of the International Conference on Computational Statistics (COMPSTAT)*, págs. 177-186, 2010. [Citado en pág. 19].
- [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. 19].
- [68] I. Goodfellow, Y. Bengio y A. Courville, *Deep Learning*. MIT Press, 2016. [Citado en pág. 19].
- [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. 21].
- [70] C. Kiourt, G. Pavlidis y S. Markantonatou, “Deep learning approaches in food recognition,” en *Machine learning paradigms: advances in deep learning-based technological applications*, Springer, 2020, págs. 83-108. [Citado en pág. 23].
- [71] 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. 24].

- [72] A. Zhang, Z. C. Lipton, M. Li y A. J. Smola, *Dive into Deep Learning*, 2021. [Citado en pág. 24].
- [73] 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. 24].
- [74] 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. 24].
- [75] 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. 25].
- [76] 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. 25].
- [77] S. Arora, Z. Li y K. Lyu, “Theoretical analysis of auto rate-tuning by batch normalization,” *arXiv preprint arXiv:1812.03981*, 2018. [Citado en pág. 25].
- [78] 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. 26].
- [79] 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. 26, 27].
- [80] 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. 27, 30].
- [81] 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. 27, 28].
- [82] 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. 27].
- [83] 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. 28].
- [84] 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. 28].
- [85] 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. 28].

- [86] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial Intelligence*, vol. 267, págs. 1-38, 2019. [Citado en pág. 28].
- [87] 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. 107161, 2022. [Citado en pág. 28].
- [88] 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. 28].
- [89] 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. 111902, 2023. [Citado en pág. 28].
- [90] 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. 105846, 2025. [Citado en pág. 28].
- [91] C. E. Rasmussen, “Gaussian processes in machine learning,” en *Summer School on Machine Learning*, Springer, 2003, págs. 63-71. [Citado en pág. 28].
- [92] R. M. Neal, *Bayesian learning for neural networks*. Springer Science y Business Media, 2012, vol. 118. [Citado en pág. 28].
- [93] C. Blundell, J. Cornebise, K. Kavukcuoglu y D. Wierstra, “Weight uncertainty in neural network,” en *International Conference on Machine Learning*, PMLR, 2015, págs. 1613-1622. [Citado en pág. 28].
- [94] Y. Gal y Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” en *International Conference on Machine Learning*, PMLR, 2016, págs. 1050-1059. [Citado en pág. 28].
- [95] D. Opitz y R. Maclin, “Popular ensemble methods: An empirical study,” *Journal of Artificial Intelligence Research*, vol. 11, págs. 169-198, 1999. [Citado en pág. 28].
- [96] H. Papadopoulos, K. Proedrou, V. Vovk y A. Gammerman, “Inductive confidence machines for regression,” en *European Conference on Machine Learning*, Springer, 2002, págs. 345-356. [Citado en págs. 29, 32, 33, 47].
- [97] 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. 29, 32, 50, 117].
- [98] 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. 29, 31].
- [99] V. Vovk, A. Gammerman y G. Shafer, *Algorithmic learning in a random world*. Springer, 2005, vol. 29. [Citado en pág. 29].
- [100] 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. 29].

- [101] 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. 30].
- [102] 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. 30].
- [103] 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. 30].
- [104] 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. 30].
- [105] V. Vovk, “Cross-conformal predictors,” *Annals of Mathematics and Artificial Intelligence*, vol. 74, n.º 1, págs. 9-28, 2015. [Citado en pág. 32].
- [106] R. F. Barber, E. J. Candes, 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 págs. 32, 47].
- [107] Y. Romano, E. Patterson y E. Candès, “Conformalized quantile regression,” *Advances in Neural Information Processing Systems*, vol. 32, 2019. [Citado en págs. 32, 46-48, 117].
- [108] 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. 20939-20948. [Citado en págs. 32, 101].
- [109] 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. 32, 53, 54, 117].
- [110] 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. 32, 50, 54, 55, 117].
- [111] 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. 32].
- [112] 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. 32].
- [113] 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. 32].
- [114] 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. 32].
- [115] 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. 35].

- [116] 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. 35].
- [117] L. Scheuer y S. Black, *Developmental Juvenile Osteology*, 1.^a ed. Academic Press, 2000. [Citado en pág. 35].
- [118] J. Adserias-Garriga, *Age estimation: a multidisciplinary approach*. Academic Press, 2019. [Citado en págs. 35, 36].
- [119] 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. 35].
- [120] 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. 36].
- [121] 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. 36].
- [122] 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 págs. 36, 83].
- [123] L. Scheuer y S. Black, *The juvenile skeleton*, 1.^a ed. Elsevier, 2004. [Citado en págs. 36, 83].
- [124] 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. 36].
- [125] 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. 36].
- [126] 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. 36, 40].
- [127] 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. 36].
- [128] 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. 36].
- [129] 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. 36].
- [130] 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. 38].

- [131] 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. 38].
- [132] 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. 38].
- [133] 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. 38].
- [134] 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. 38].
- [135] D. Stern, T. Ebner, H. Bischof, S. Grassegger, T. Ehamer y M. Urschler, “Fully automatic bone age estimation from left hand MR images,” en *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2014, págs. 220-227. [Citado en pág. 39].
- [136] 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. 39, 42].
- [137] 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. 40].
- [138] J. E. Buikstra, “Standards for data collection from human skeletal remains,” *Arkansas Archaeological Survey Research Series*, vol. 44, pág. 44, 1994. [Citado en pág. 40].
- [139] 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. 45].
- [140] 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. 45].
- [141] 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. 46].
- [142] 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. 48].
- [143] 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. [Citado en pág. 49].
- [144] J. Huang, H. Xi, L. Zhang, H. Yao, Y. Qiu y H. Wei, “Conformal prediction for deep classifier via label ranking,” *arXiv preprint arXiv:2310.06430*, 2023. [Citado en págs. 50, 56].

- [145] U. Johansson, H. Linusson, T. Löfström y H. Boström, “Model-agnostic non-conformity functions for conformal classification,” en *2017 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2017, págs. 2072-2079. [Citado en pág. 51].
- [146] Mindful Modeler. “Week #1: Getting Started With Conformal Prediction For Classification.” Consultado el 31/08/2025. URL: <https://mindfulmodeler.stack.com/p/week-1-getting-started-with-conformal>. [Citado en pág. 51].
- [147] V. Vovk, D. Lindsay, I. Nouretdinov y A. Gammerman, “Mondrian confidence machine,” *Technical Report*, 2003. [Citado en pág. 52].
- [148] 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. 61].
- [149] 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. 61].
- [150] 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. 68].
- [151] 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. 70].
- [152] C. E. Agbangba, E. S. Aide, H. Honfo y R. G. Kakai, “On the use of post-hoc tests in environmental and biological sciences: A critical review,” *Heliyon*, vol. 10, n.º 3, 2024. [Citado en pág. 70].
- [153] I. Loshchilov y F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017. [Citado en pág. 73].
- [154] 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. 73].
- [155] A. Heinrich, “Accelerating computer vision-based human identification through the integration of deep learning-based age estimation from 2 to 89 years,” *Scientific Reports*, vol. 14, pág. 4195, 2024.
- [156] 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.
- [157] 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.
- [158] J. Postels, M. Segu, T. Sun, L. Sieber, L. Van Gool, F. Yu y F. Tombari, “On the practicality of deterministic epistemic uncertainty,” *arXiv preprint arXiv:2107.00649*, 2021.
- [159] J. G. Sam Lau y D. Nolan, *Cross Validation*, Consultado el 26/05/2025, 2023. URL: https://learningds.org/ch/16/ms_cv.html.
- [160] J. R. Berrendero. “Materiales del libro de Estadística.” Consultado el 2 de junio de 2025. URL: <https://verso.mat.uam.es/~joser.berrendero/libro-est/>. [Citado en pág. 115].

- [161] A. Charpentier. “Confidence vs. Credibility Intervals.” Consultado el 21 de agosto de 2025. URL: <https://freakonometrics.hypotheses.org/18117>.
- [162] J. Vermorel. “Quantile Regression,” LOKAD Quantitive Supply Chain, visitado 2 de jun. de 2025. URL: <https://www.lokad.com/quantile-regression-time-series-definition/>.
- [163] R. Koenker, *Quantile Regression* (Econometric Society Monographs). Cambridge University Press, 2005.
- [164] S. T. Tokdar y J. B. Kadane, “Simultaneous linear quantile regression: a semi-parametric Bayesian approach,” *Bayesian Analysis*, vol. 7, n.º 1, págs. 51-72, 2012.
- [165] J. Feldman y D. Kowal, “Bayesian Quantile Regression with Subset Selection: A Posterior Summarization Perspective,” *arXiv preprint arXiv:2311.02043*, 2023.
- [166] R. Luo y Z. Zhou, “Conformal thresholded intervals for efficient regression,” en *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, págs. 19 216-19 223. [Citado en pág. 117].

Apéndice A

Intervalos de valores razonables

En este apartado diferenciaremos los tipos de intervalos de valores nos permiten cuantificar la variabilidad de los resultados y, por tanto, la incertidumbre de la medición realizada.

- El **intervalo de confianza (IC)** es una herramienta común de la estadística frecuentista, que permite estimar un rango de valores tal que podamos confiar en que contiene al valor verdadero de un parámetro poblacional desconocido θ (p.ej., la media) [160].

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

Es importante aclarar un malentendido común: un intervalo de confianza con nivel 95 % para un parámetro θ no significa que exista un 95 % de probabilidad de que θ esté dentro del intervalo calculado a partir de una muestra específica. En realidad, el 95 % se refiere a la frecuencia con la que, si muestreásemos muchas veces los datos, los intervalos construidos a partir de esas muestras incluirían al valor verdadero de θ en aproximadamente el 95 % [60] (véase la Figura A.1).

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

La diferencia radica en que, a diferencia del intervalo de confianza, que parte de que θ es un parámetro fijo desconocido y los datos son tratados como aleatorios, el enfoque bayesiano fija los datos(ya que son conocidos) y el parámetro θ lo trata como aleatorio (ya que es desconocido) [60].

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

- El **intervalo de predicción (*prediction interval*)** es radicalmente diferente a los intervalos previos. Trata de predecir un valor futuro de una observación, no determinar un parámetro poblacional. Existen numerosos métodos, con y sin necesidad de conocer la distribución de los datos.

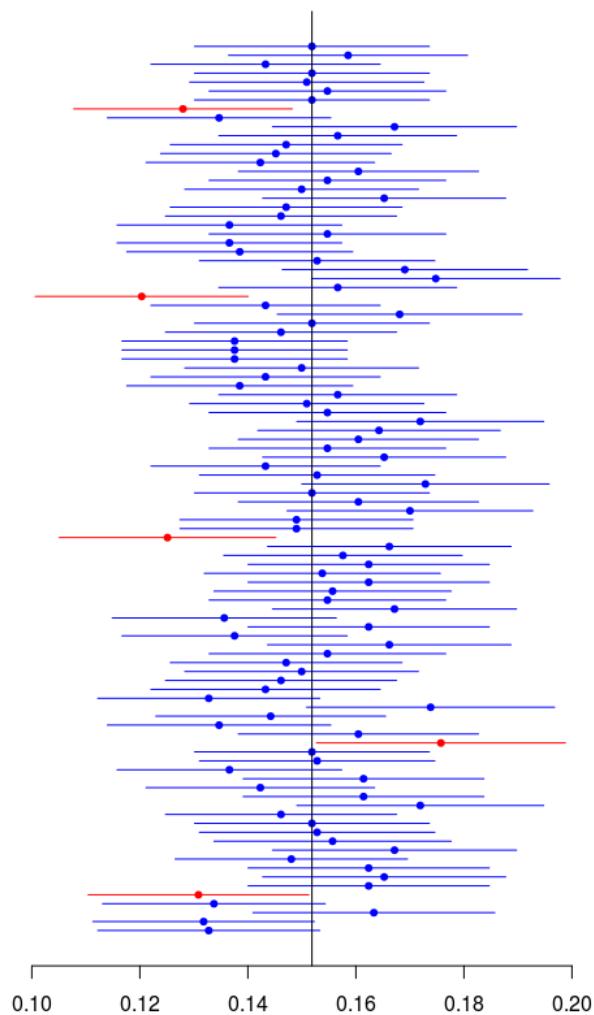


Figura A.1: Ejemplo de intervalos de confianza para la media poblacional. La interpretación correcta del nivel de confianza (95 % en este caso) es: *Si repitiéramos el proceso de muestreo y construcción de intervalos muchas veces, aproximadamente el 95 % de ellos contendrían el verdadero valor de la media poblacional.* En esta simulación, la media real conocida es 0.153, y podemos ver que la mayoría de los intervalos la capturan, mientras que unos pocos (generalmente alrededor del 5 %) no lo logran. En general, se suele pedir uno solo de estos intervalos, calculado con toda la muestra disponible, aunque la media poblacional podrá estar o no contenida, pero es desconocido.

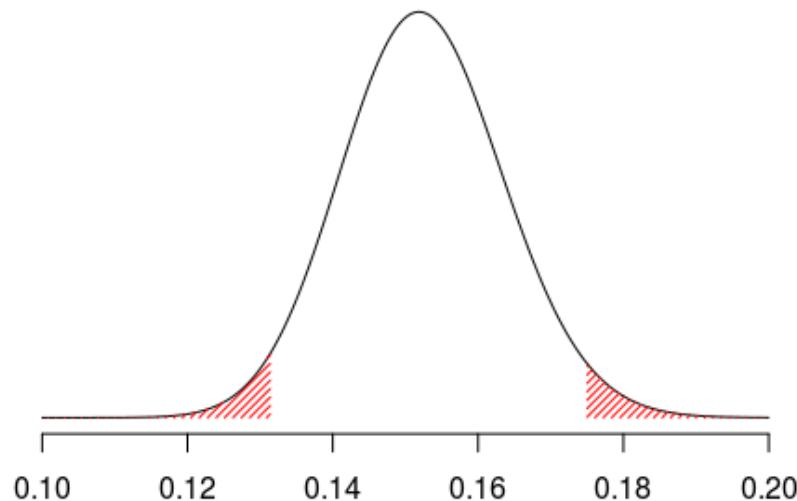
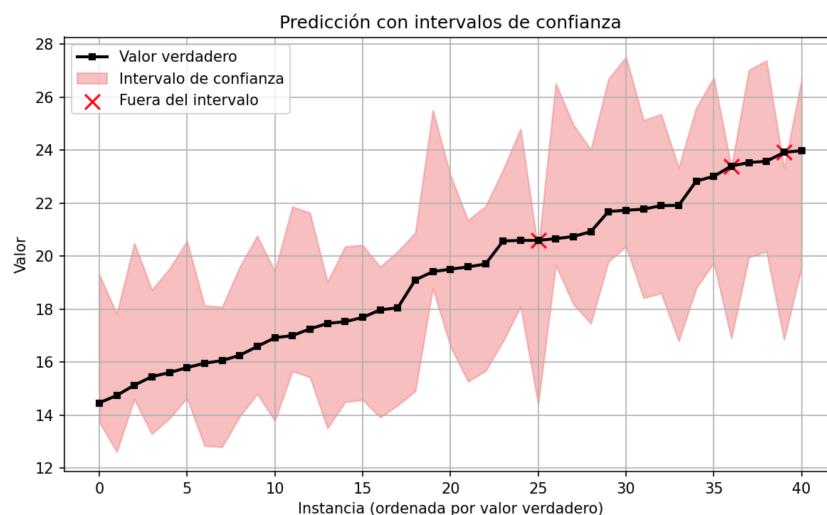


Figura A.2: Ejemplo de intervalo de credibilidad para la media poblacional. La interpretación correcta es: *Con un 95 % de probabilidad, el valor verdadero está dentro del intervalo.* En esta simulación, la media real conocida es 0.153, y podemos observar que efectivamente este valor está contenido en el intervalo.

El enfoque explorado en este trabajo es la predicción conformal, que ha demostrado ser eficaz en contextos donde los supuestos clásicos (normalidad, homocedasticidad) no se cumplen [107], y es actualmente el enfoque más robusto para la construcción de intervalos de predicción en aplicaciones modernas de ML [97, 107, 109, 110, 166]. La predicción conformal tiene una interpretación frecuentista: $1 - \alpha$ intervalos producidos cubren el verdadero valor (véase la Figura A.3).



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

Apéndice B

Problema de estimación de sexo

Se propone un problema adicional, para la estimación de sexo a partir de los imágenes de las radiografías maxilofaciales. Para este problema de clasificación binaria identificamos dos clases: sexo masculino (M) y sexo femenino (F).

En la Tabla B.1 se recogen las métricas obtenidas por los diferentes métodos. La exactitud apenas se reduce en medio punto porcentual entre los métodos conformales y el método ‘base’, en la línea de resultados infírmamente peores en los primeros.

Por otro lado, la cobertura del método ‘base’, al solo incluir la etiqueta de la clase más probable en el conjunto de predicción, presenta una cobertura empírica igual a la exactitud, del 88.57 %. LAC y MCM sí logran una cobertura del 95 %, con un 18 % de las instancias indeterminadas (con ambas etiquetas en el conjunto de predicción) en promedio, en ambos casos. Finalmente, en la gráfica de dispersión de la Figura B.1 se observa que no existen diferencias relevantes entre los métodos conformales, cuyos resultados se solapan en el diagrama.

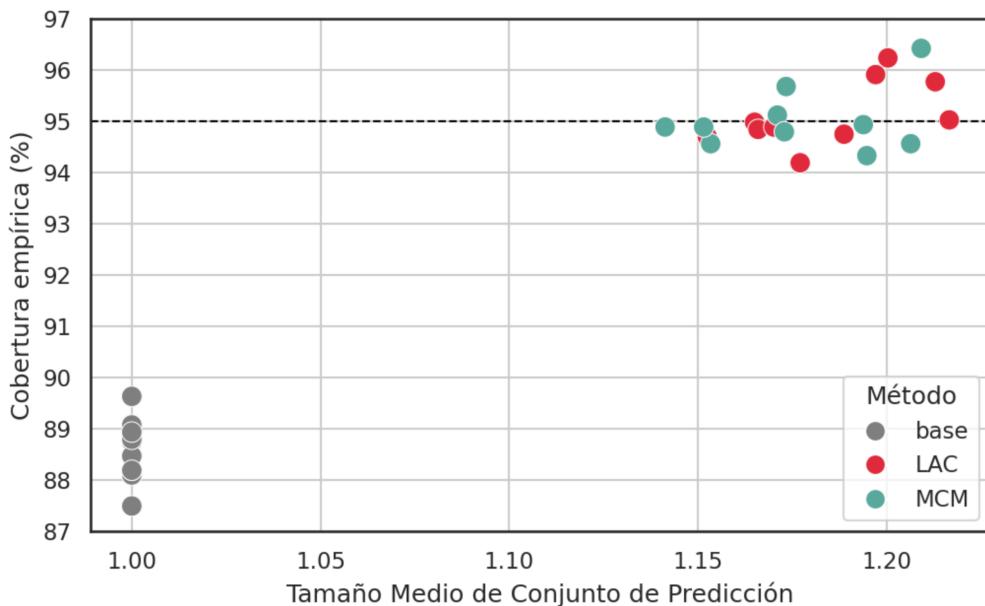


Figura B.1: Gráfica de dispersión de la cobertura empírica frente al tamaño medio de conjunto de predicción.

A continuación, en la Figura B.2 se analiza la cobertura en función del tamaño del conjunto. Se observa que aproximadamente un 18 % de las instancias presenta un

Método	Exactitud (%)		Cobertura Empírica (%)			Tamaño Medio del Conjunto		
	base	CP	base	LAC	MCM	base	LAC	MCM
Ejecución 1	88.75	87.04	88.75	94.19	94.33	1.00	1.18	1.19
Ejecución 2	89.08	89.03	89.08	96.24	96.42	1.00	1.20	1.21
Ejecución 3	89.64	88.15	89.64	94.98	94.56	1.00	1.16	1.15
Ejecución 4	88.10	88.75	88.10	94.84	94.89	1.00	1.17	1.15
Ejecución 5	88.20	88.29	88.20	95.91	95.68	1.00	1.20	1.17
Ejecución 6	88.48	87.04	88.48	94.75	94.93	1.00	1.19	1.19
Ejecución 7	88.80	86.99	88.80	95.03	94.56	1.00	1.22	1.21
Ejecución 8	87.50	87.83	87.50	94.89	95.12	1.00	1.17	1.17
Ejecución 9	88.94	88.48	88.94	94.70	94.89	1.00	1.15	1.14
Ejecución 10	88.20	88.57	88.20	95.77	94.80	1.00	1.21	1.17
Media	88.57	88.02	88.57	95.13	95.02	1.00	1.18	1.18

Tabla B.1: Valores de exactitud, cobertura empírica y tamaño medio del conjunto obtenidas por cada método de predicción a lo largo de 10 ejecuciones, así como la media final para cada método entre todas las ejecuciones. ‘CP’ se refiere a los métodos conformales empleados: LAC y MCM (se recuerda que es el mismo modelo para todos los métodos conformales y, por ello, presentan las mismas predicciones puntuales). Se marca en negrita las medias con mejor marca para cada métrica: mayor valor en exactitud, valores próximos a 95 % en cobertura empírica y mínimo valor en tamaño medio del conjunto.

conjunto de predicción indeterminado, el cual, como es obvio, alcanza una cobertura del 100 %. El 82 % restante de las instancias obtiene una cobertura cercana al 94 %, tanto con el método LAC como con MCM.

En definitiva, el método ‘base’ logra: 88.57 % de las instancias bien clasificadas, y el 11.43 % mal clasificadas. Frente a esto, los métodos conformales logran: 77 % de las instancias bien clasificadas, 5 % mal clasificadas, y un 18 % indeterminadas (véase la Figura B.3). De esta forma, se ha conseguido reducir la proporción de instancias mal clasificadas de un 11.43 % a un 5 %, aunque a cambio se generan instancias indeterminadas que, en el método base, habrían sido clasificadas correctamente.

Y, por último, vamos a analizar la cobertura en base al sexo y la edad, a partir de las Figuras B.4 y B.5. Se pueden observar varias tendencias en el método ‘base’:

- Se dan más errores de clasificación en las edades más jóvenes que en las avanzadas, de lo que se puede inferir que existe mayor incertidumbre en edades jóvenes. Probablemente este efecto sea consecuencia de que los rasgos sexuales maxilofaciales aún no están lo suficientemente marcados en edades jóvenes.
- Hay más instancias mal clasificadas en varones que en hembras, para prácticamente todas las edades. Esto indica que también existe mayor incertidumbre en individuos de sexo masculino, debido a que el desarrollo de caracteres sexuales secundarios faciales se completa más tarde que en las hembras.

No obstante, estas dos fuentes de incertidumbre son abordadas por los métodos LAC y MCM mediante el uso de conjuntos indeterminados, los cuales permiten representar de

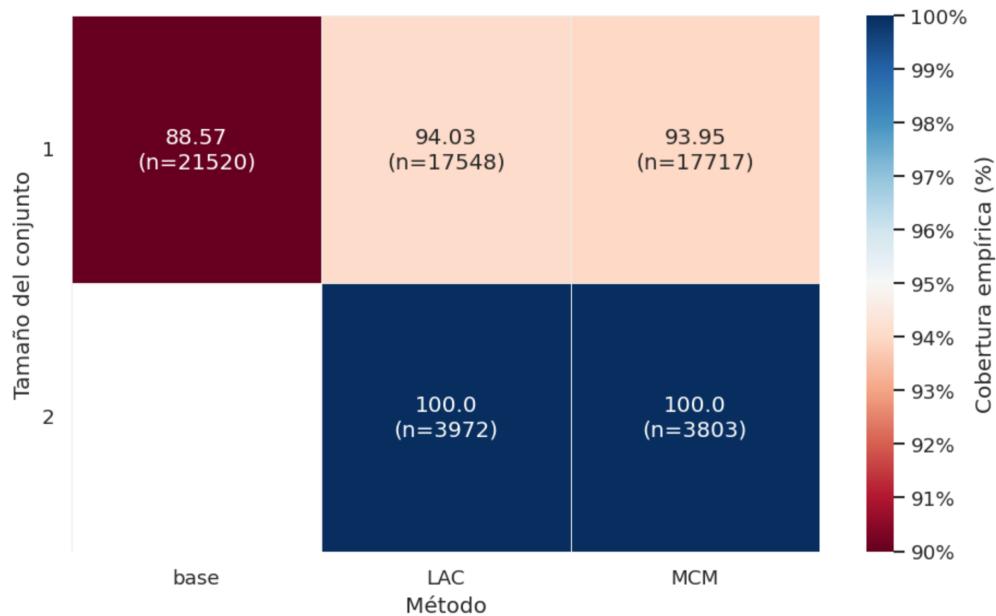


Figura B.2: Mapa de calor de la cobertura empírica en base al tamaño del conjunto por cada método de predicción a lo largo de 10 ejecuciones. Se especifica entre paréntesis el número de instancias con el número de etiquetas en el conjunto de predicción. La escala de colores está centrada en la cobertura nominal (0.95): los valores por debajo de este umbral se representan en tonos rojos, los superiores en tonos azules, y el blanco indica una cobertura empírica equivalente a la nominal.

forma explícita la ambigüedad en la clasificación. De este modo, la cobertura alcanzada por ambos métodos supera el 90 % en prácticamente todas las edades y sexos, y en la mayoría de las combinaciones sexo-edad llega incluso a aproximar o superar el 95 %, manteniendo —aunque en menor medida— las tendencias descritas anteriormente.

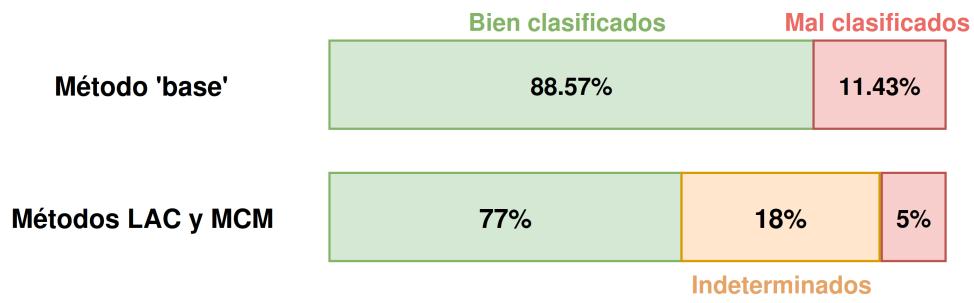


Figura B.3: Proporción de instancias bien clasificadas, mal clasificadas e indeterminadas de los métodos propuestos.

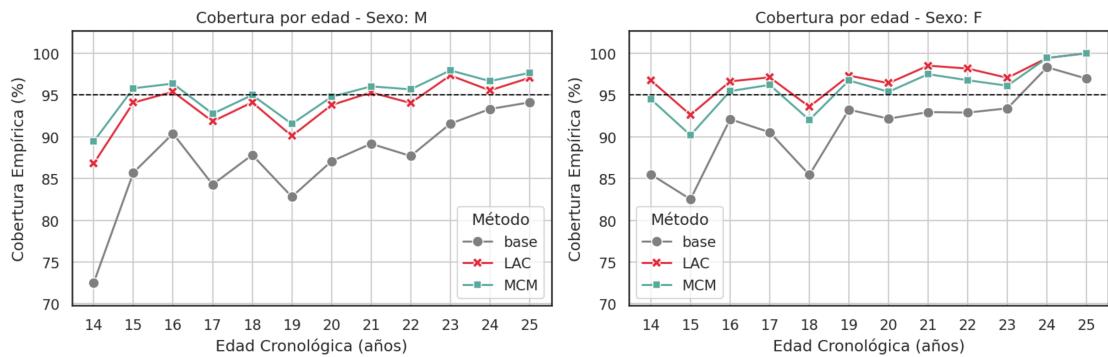


Figura B.4: Diagrama de líneas de la cobertura empírica en base al sexo y la edad cronológica por cada método de predicción a lo largo de 10 ejecuciones.

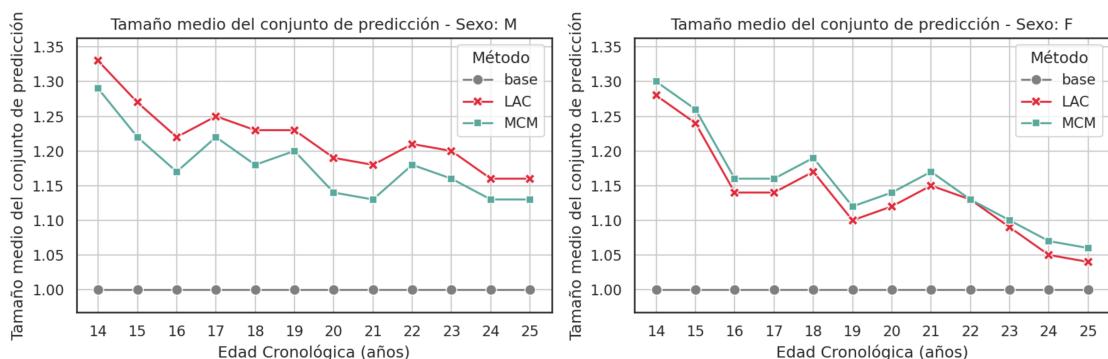


Figura B.5: Diagrama de líneas del tamaño medio de conjunto de predicción en base al sexo y la edad cronológica por cada método de predicción a lo largo de 10 ejecuciones.

