



*ugr* | Universidad  
de Granada

TRABAJO FIN DE GRADO  
GRADO EN INGENIERÍA INFORMÁTICA

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

**Autor**  
David González Durán

**Director**  
Pablo Mesejo Santiago

**Mentor**  
Javier Venema Rodríguez



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE  
TELECOMUNICACIÓN

—  
Granada, mes de 2025



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

David González Durán

**Palabras clave:** palabra\_clave1, palabra\_clave2, palabra\_clave3, ...

## Resumen

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



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

David González Durán

**Keywords:** Keyword1, Keyword2, Keyword3, ...

## **Abstract**

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



---

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

Fdo: David González Durán

Granada, a X de mes de 202.



---

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

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

**Informan:**

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

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

**Los directores:**

Pablo Mesejo Santiago

Javier Vénema Rodríguez



# **Agradecimientos**

Poner aquí agradecimientos...



# Índice general

<b>1. Materiales y métodos</b>	<b>1</b>
1.1. Conjunto de datos disponibles . . . . .	1
1.2. Problemas propuestos . . . . .	4
1.2.1. Problema de estimación de edad . . . . .	4
1.2.2. Problema de estimación de mayoría de edad . . . . .	4
1.2.3. Problema de clasificación de edad . . . . .	5
1.3. Métodos propuestos . . . . .	5
1.3.1. Arquitectura empleada . . . . .	5
1.3.2. Regresión cuantílica . . . . .	6
1.3.3. Métodos de predicción conformal para regresión . . . . .	8
1.3.4. Calibración de probabilidades en clasificación . . . . .	11
1.3.5. Métodos de predicción conformal para clasificación . . . . .	11
1.4. Métricas . . . . .	18
1.4.1. Métricas para regresión . . . . .	18
1.4.2. Métricas para clasificación . . . . .	21
<b>2. Experimentación</b>	<b>25</b>
2.1. Protocolo de validación experimental . . . . .	25
2.2. Preprocesado de los datos . . . . .	27
2.3. Esquema general de los experimentos realizados . . . . .	27
2.3.1. Problema de estimación de edad . . . . .	28
2.3.2. Problema de clasificación de mayoría de edad . . . . .	30
2.3.3. Problema de clasificación de edad . . . . .	32
2.3.4. Tests estadísticos . . . . .	34

---

2.4.	Experimentación para la estimación de edad . . . . .	35
2.4.1.	Entrenamiento de los modelos . . . . .	35
2.4.2.	Resultados . . . . .	37
2.5.	Experimentación para la clasificación de mayoría de edad . .	48
2.5.1.	Entrenamiento de los modelos . . . . .	48
2.5.2.	Resultados . . . . .	48
2.6.	Experimentación para la estimación de edad como problema de clasificación . . . . .	54
2.6.1.	Entrenamiento de los modelos . . . . .	54
2.6.2.	Resultados . . . . .	54

# Índice de figuras

1.1.	Histograma de edad de los individuos del conjunto de datos disponible. . . . .	3
1.2.	Gráficas de densidad y de caja de edad por sexo de los individuos del conjunto de datos disponible. . . . .	3
1.3.	Esquema visual del modelos de regresión propuesto. . . . .	5
1.4.	Visualización de la función de pérdida <i>pinball</i> para cada valor de error. . . . .	8
1.5.	Matriz de confusión para la estimación de sexo según el modelo <i>random forest</i> propuesto en [15]. . . . .	21
2.1.	Diagrama de división del <i>dataset</i> en <i>train</i> , <i>validation</i> y <i>test</i> . .	26
2.2.	Diagrama de división del <i>dataset</i> en <i>train</i> , <i>validation</i> , <i>calibration</i> y <i>test</i> . . . . .	26
2.3.	Esquema de experimentación para la estimación de edad. .	29
2.4.	Esquema de experimentación para la clasificación de mayoría de edad. . . . .	31
2.5.	Esquema de experimentación para la clasificación de edad. .	33
2.6.	Curva de aprendizaje de uno de los modelos para el método ICP. . . . .	37
2.7.	Gráfica de dispersión de la Cobertura empírica frente a la Amplitud media del intervalo de predicción. . . . .	41
2.8.	Histogramas del amplitud del intervalo de predicción con diferenciación por cobertura, correspondientes a los modelos QR y CQR. . . . .	44
2.9.	Gráficos de líneas comparativos de la cobertura empírica y la amplitud media del intervalo de predicción por edad cronológica para los diferentes métodos evaluados. . . . .	47

2.10. Gráfica de dispersión Cobertura empírica - Tamaño Medio de Conjunto de Predicción. . . . .	51
2.11. Matrices de confusión conformal correspondientes a los métodos ‘base’, LAC y MCM. . . . .	53
2.12. Cobertura empírica y tamaño medio del conjunto de predicción obtenidos por cada método de predicción a lo largo de las distintas ejecuciones. . . . .	56
2.13. Gráfica de dispersión Cobertura empírica - Tamaño Medio de Conjunto de Predicción. . . . .	57
2.14. Mapa de calor de cobertura empírica en base al tamaño del conjunto por cada método de predicción a lo largo de las distintas ejecuciones. . . . .	59
2.15. Gráficos de líneas comparativos de la cobertura empírica y el tamaño medio del conjunto de predicción por edad cronológica para los diferentes métodos evaluados. . . . .	62

# Índice de tablas

1.1.	Lista de instituciones participantes en la recolección de los datos e imágenes dentales utilizados en el trabajo. . . . .	2
1.2.	Comparativa de métodos propuestos de CP para problemas de regresión. . . . .	12
2.1.	Error absoluto medio y error cuadrático medio obtenidos por cada método de predicción a lo largo de distintas ejecuciones. . . . .	38
2.2.	Resultados de la prueba <i>post-hoc</i> de Tukey HSD para MAE entre pares de métodos. . . . .	39
2.3.	Resultados de la prueba <i>post-hoc</i> de Tukey HSD para MSE entre pares de métodos. . . . .	39
2.4.	Cobertura empírica y amplitud media del intervalo de predicción obtenidos por cada método de predicción a lo largo de distintas ejecuciones. . . . .	40
2.5.	Resultados de la prueba <i>post-hoc</i> de Games-Howell para la amplitud media del intervalo de predicción entre pares de métodos. . . . .	42
2.6.	Resultados de las predicciones obtenidas por los modelos para el problema de estimación de edad en cada ejecución. . . . .	43
2.7.	Cobertura empírica del intervalo de predicción obtenida por cada método de predicción para distintas franjas de amplitud de intervalos. . . . .	45
2.8.	Exactitud, sensibilidad y especificidad obtenidos por cada método de predicción a lo largo de distintas ejecuciones. . . . .	49
2.9.	Cobertura empírica y tamaño medio del conjunto de predicción obtenidos por cada método de predicción a lo largo de las distintas ejecuciones. . . . .	50
2.10.	Resultados de la prueba <i>post-hoc</i> de Tukey HSD para la cobertura empírica entre pares de métodos. . . . .	51

---

2.11. Resultados de la prueba <i>post-hoc</i> de Tukey HSD para la cobertura empírica entre pares de métodos. . . . .	55
2.12. Resultados de la prueba <i>post-hoc</i> de Tukey HSD para el tamaño medio del conjunto de predicción entre pares de métodos. . . . .	57

# Capítulo 1

## Materiales y métodos

### 1.1. 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 1.1), obtenidas con distintos modelos de máquinas de rayos X<sup>1</sup>. Este conjunto de datos ha sido proporcionado por Panacea Cooperative Research, empresa *spin-off* de la Universidad de Granada.

Este *dataset* incluye:

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

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

Se ha ignorado el campo “sample”, dado que se trata de una asignación sesgada y no representa necesariamente una clasificación fiable del origen poblacional de los individuos. Por tanto, este campo no se emplea en el análisis ni en el entrenamiento de los modelos, centrándose exclusivamente en las variables de edad, sexo e imagen.

---

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

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

Tabla 1.1: Lista de instituciones participantes en la recolección de los datos e imágenes dentales utilizados en el trabajo.

En el *dataset* hay un total de 10 739 ejemplos, de los que 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 1.1 se observa que el número de ejemplos por edad se mantiene relativamente constante desde los 14 hasta los 21 años, a partir de los cuales disminuye progresivamente, con una representación notablemente menor en los grupos de 24, 25 y 26 años.

En la Figura 1.2 podemos comprobar cómo en términos relativos la distribución de edad por sexo es muy similar, compartiendo ambas prácticamente el mismo rango de edades y patrones de dispersión, sin observarse diferencias sustanciales en la mediana ni en la forma general de las distribuciones.

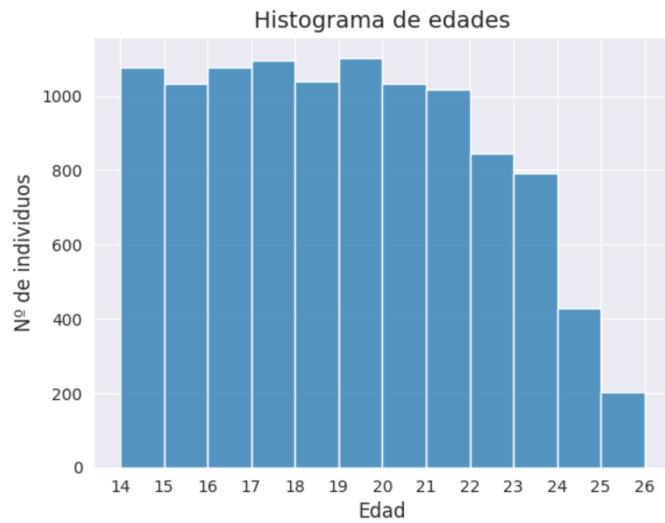


Figura 1.1: Histograma de edad de los individuos del conjunto de datos disponible.

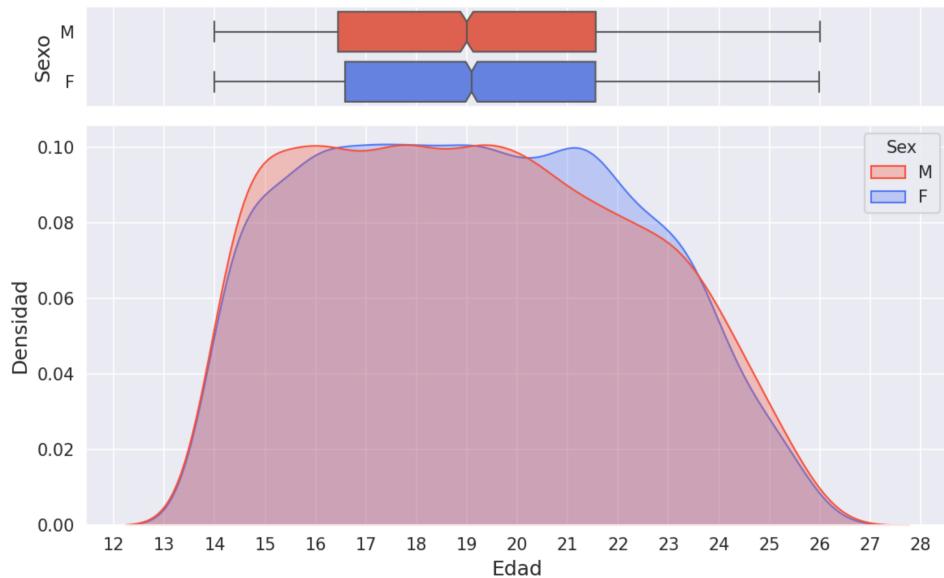


Figura 1.2: Gráficas de densidad y de caja de edad por sexo de los individuos del conjunto de datos disponible.

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

para evaluar su posible impacto en el rendimiento y generalización de los modelos entrenados.

Se proporcionan los datos ya divididos en *train* —con un 80% de los individuos— y *test* —con el 20% restante—, con la intención de que puedan ser utilizados para entrenar y evaluar modelos de predicción. La división de ambos conjuntos se hizo de forma estratificada, de lo que se asume que la distribución será igual en ambos datasets.

## 1.2. 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;
2. estimación de la edad legal planteada como un problema de clasificación binaria (mayor o menor de 18 años);
3. estimación de la edad legal resuelta como un problema de clasificación multiclas; y
4. estimación del sexo como problema de clasificación binario.

### 1.2.1. Problema de estimación de edad

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 1.3). Inicialmente se consideró incluir el sexo como metadato adicional en el modelo; sin embargo, se descartó tras observar de manera preliminar que no tenía un impacto significativo en el rendimiento del modelo, además de que su exclusión simplifica la arquitectura.

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

### 1.2.2. Problema de estimación de mayoría de edad

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

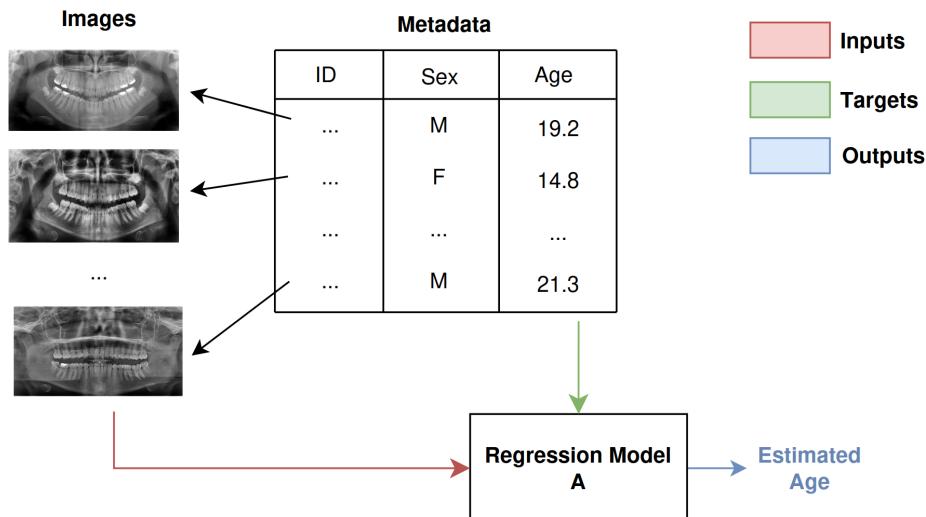


Figura 1.3: Esquema visual del modelos de regresión propuesto. El modelo solo tiene radiografías maxilofaciales como entrada.

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”.

### 1.2.3. Problema de clasificación de edad

Se propone un problema de estimación de edad, pero planteado como problema de clasificación multiclase, 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.

No sé muy bien qué título poner para el problema, estoy entre “clasificación de mayoría de edad” o “estimación/valoración de mayoría de edad” (age assessment). He dejado clasificación para dejar claro que es un problema de este tipo pero en AF se suele usar más bien el segundo término por lo que he podido ver.

## 1.3. Métodos propuestos

### 1.3.1. Arquitectura empleada

El primer problema propuesto es el de estimación de edad. Partiremos de un planteamiento muy simple: imágenes bidimensionales de las radiografías

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

Como modelo, empleamos una CNN, dado su buen desempeño en tareas de visión por computador. Específicamente, utilizamos la arquitectura ResNeXt50 [1] preentrenada en Imagenet [2] 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.

### 1.3.2. Regresión cuantílica

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

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

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

En redes neuronales, esta regresión requiere de:

- Definir una capa de salida con múltiples neuronas, una por cada cuantil deseado ( $\hat{q}_\tau$ ). 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* [3]. La **función de pérdida pinball** es una generalización de la función de pérdida  $L1$ <sup>2</sup>, 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 1.4 ilustra cómo la pérdida penaliza de forma desigual los errores positivos y negativos. Mientras que la pérdida  $L1$  se centra en ajustar la mediana (cuantil 0.5), la pérdida pinball permite dirigir una salida del modelo en cualquier cuantil deseado. Esto es especialmente útil cuando se desea modelar distribuciones asimétricas y capturar diferentes percentiles de la variable de salida, en lugar de asumir una distribución de errores simétrica, como la normal.

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

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

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

---

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

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

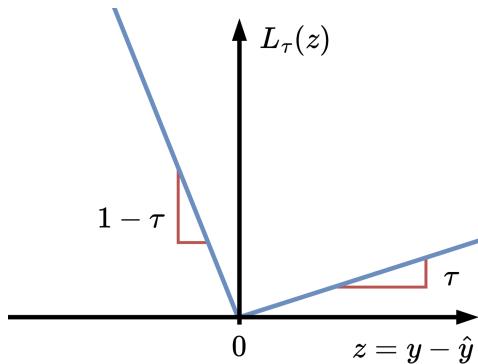


Figura 1.4: Visualización de la función de pérdida *pinball* para cada valor de error. Adaptado de la Figura 1 de [4]. Esta concretamente muestra la función de pérdida para un cuantil cercano a cero, ya que es más permisivo con los errores positivos que con los negativos, lo cual empujará sus predicciones hacia la parte inferior de la distribución objetivo.

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 [4]. 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.

### 1.3.3. Métodos de predicción conformal para regresión

Todos los métodos propuestos en este trabajo son *split calibration*, es decir, los datos de entrenamiento se dividen en dos subconjuntos: entrenamiento y calibración. No hemos implementado métodos *cross-calibration* como [5] 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 [6] 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:

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

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:

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

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

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

Este método de CP presenta varias ventajas:

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

Sin embargo, también presenta importantes limitaciones:

- **Intervalo simétrico y no adaptativo:** El intervalo es simétrico, además de tener siempre el mismo ancho ( $2q_{1-\alpha}$ ), no permitiendo adaptarse a la incertidumbre específica de cada predicción.

- **Sensibilidad a datos ruidosos o OOD:** Si el conjunto de calibración contiene *outliers* o viola el supuesto de intercambiabilidad, el umbral  $q_{1-\alpha}$  puede inflarse, generando intervalos excesivamente conservadores. Tampoco detecta heterocedasticidad automáticamente.

### *Conformalized Quantile Regression (CQR)*

Como su nombre indica, este método se realiza sobre la regresión cuantílica. La CQR [4] 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 [4], que incluye la calibración de ambas colas para obtener intervalos asimétricos [7]. 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:

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

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  $R$ :

$$\begin{aligned} \delta_{lower\alpha} &= Quantile_{\lceil(1-\alpha)(1+1/n)\rceil}(\{A_{lower}(x_i, y_i)\}_{i=1}^n) \\ \delta_{upper\alpha} &= Quantile_{\lceil(1-\alpha)(1+1/n)\rceil}(\{A_{upper}(x_i, y_i)\}_{i=1}^n) \end{aligned}$$

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

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

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

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

En la Tabla 1.2 observamos un cuadro comparativo de los distintos métodos propuestos de CP.

#### 1.3.4. Calibración de probabilidades en clasificación

#### 1.3.5. Métodos de predicción conformal para clasificación

##### *Least-Ambiguous set-valued Classifiers (LAC)*

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

LAC [8] 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** [9], y se calcula como la unidad menos la probabilidad de la clase verdadera<sup>3</sup> o, lo que es lo mismo, la suma de valores de probabilidad de todas las clases salvo la correspondiente a la etiqueta verdadera:

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

donde  $\hat{\pi}_{y_i}(x_i)$  es la probabilidad para la clase de la etiqueta verdadera<sup>4</sup>.

---

<sup>3</sup>Se 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.

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

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

Tabla 1.2: Comparativa de métodos propuestos de CP para problemas de regresión.

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:

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

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:

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

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

Algunas propiedades de este método son:

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

### Mondrian Confidence Machine (MCM)

(MCM) [10] 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,

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

MCM segmenta el conjunto de calibración por clase y calcula las puntuaciones de no conformidad y los umbrales de decisión de forma independiente para cada una.

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

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

### Adaptive Prediction Sets (APS)

APS [11], como sugiere su nombre, tiene como objetivo generar conjuntos de predicción adaptativos, cuyo tamaño se ajusta dinámicamente en función de la incertidumbre del modelo para cada muestra. De este modo, se busca que las predicciones sean más informativas y reflejen con mayor precisión la confianza del modelo.

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

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

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

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

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

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 multiclase, donde las puntuaciones de no conformidad de APS son más permisivas que las de LAC, ya que reconocen que un modelo puede identificar características comunes entre varias clases y generar valores probabilísticos repartidos. No existe incertidumbre cuando la puntuación probabilística más alta corresponde a la clase verdadera. Por tanto, APS penaliza menos los casos en que la clase correcta está entre las más probables, aunque no necesariamente en primer lugar.

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

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

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:

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

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

Sin embargo, en la práctica se ha observado que esta adaptabilidad lleva **conjuntos de predicción más grandes en promedio** [11, 12]. Este fenómeno es un *trade-off* inherente al intentar **aproximar la cobertura condicional** sin asumir distribuciones subyacente, que analizaremos en profundidad con nuestros datos en la experimentación.

### Regularized Adaptive Prediction Sets (RAPS)

RAPS [12] 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.

- $\lambda$ , 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  $\lambda$  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:

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

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:

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

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).

### Sorted Adaptive Prediction Sets (SAPS)

SAPS [13] 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, SAPS ordena las clases por probabilidad decreciente y aplica un esquema adaptativo de umbral basado en la posición en el ranking, en lugar de en el valor absoluto de la probabilidad. De esta forma, se evita el efecto negativo de las colas largas y se prioriza la inclusión de clases con mayor relevancia relativa según el modelo.

Por completar (JULIO)

Al enfocarse en esta información comprimida y más representativa, SAPS logra generar conjuntos más compactos, sin comprometer la garantía de cobertura estadística que exige el marco de predicción conforme.

Primero, en su planteamiento, y a diferencia de los métodos anteriores, no define una métrica de no conformidad, sino más bien una métrica de confianza,

## 1.4. Métricas

### 1.4.1. Métricas para regresión

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

Por una parte, las métricas destinadas a las predicciones puntuales se basan fundamentalmente en medir el error entre el valor real ( $y_i$ ) y el 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.

- La **mean interval score** [14] trata de unificar en una sola métrica el *trade-off* cobertura vs. amplitud del intervalo. Su expresión es la siguiente:

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

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  $\alpha$ , 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.

### 1.4.2. Métricas para clasificación

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

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 1.5). 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.

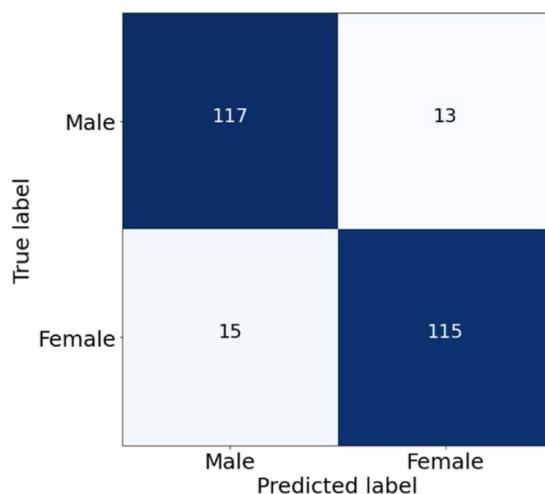


Figura 1.5: Matriz de confusión para la estimación de sexo según el modelo *random forest* propuesto en [15].

- 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.

- Se denomina **violación de la cobertura empírica** (*empirical coverage violation*) a la magnitud en que la cobertura empírica no alcanza el nivel de cobertura teórico deseado  $1 - \alpha$ , definida como:

$$ECV = \max \{0, (1 - \alpha) - EC\}$$

Este valor cuantifica cuánto se desvía el método de la garantía nominal <sup>5</sup> cuando no se alcanza la cobertura esperada. Una violación igual a cero indica que el método cumple o supera el nivel de cobertura deseado.

Esta métrica se suele calcular sobre subconjuntos específicos del conjunto de instancias para evaluar la cobertura condicional, es decir, la calidad de la cobertura dentro de subpoblaciones del dominio.

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

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

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

- **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:**

---

<sup>5</sup>Se denomina garantía de cobertura nominal al nivel de cobertura garantizado estadísticamente.

- **Gráfica de barras de cobertura en base al tamaño del conjunto.** Esta gráfica muestra la cobertura empírica lograda para cada tamaño del conjunto de predicción. Esto permite visualizar en qué tamaños de conjunto el modelo tiende a infracubrir o sobrecubrir más en cuanto a cobertura.



# Capítulo 2

## Experimentación

### 2.1. 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*<sup>1</sup>. 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 2.1 podemos ver la división del *dataset* planteada. Cabe comentar que la división se ha realizado de forma estratificada en base a la edad y el sexo<sup>2</sup>.

---

<sup>1</sup>El *data snooping* ocurre cuando información del conjunto de test se filtra, directa o indirectamente, en el proceso de entrenamiento del modelo, lo que puede llevar a una sobreestimación del rendimiento y a modelos que no generalizan adecuadamente ante datos nuevos.

<sup>2</sup>La 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”.

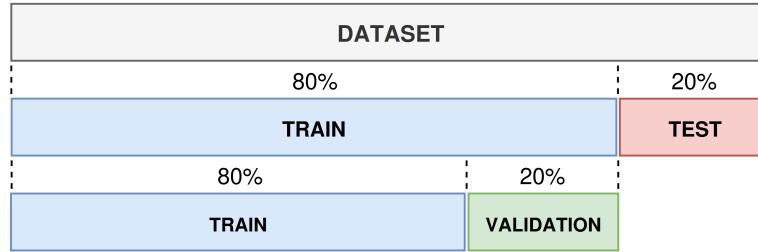


Figura 2.1: Diagrama de división del *dataset* en *train*, *validation* y *test*.

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, algo especialmente relevante dado que los problemas abordados están jerárquicamente relacionados (la clasificación de sexo y mayoría de edad se deriva directamente de la clasificación de mayoría de edad, que a su vez se deriva de la estimación de edad).

Sin embargo, al emplear métodos de calibración o predicción conformal, si usamos los mismos datos de entrenamiento para la calibración, las probabilidades o intervalos de predicción tenderán a ser optimistas, pues el modelo ha sido entrenado con esos datos [16]. 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 [17] (que recomienda dedicar entre un 10 % y 30 % de datos de entrenamiento a calibración), tal y como se muestra en la Figura 2.2.

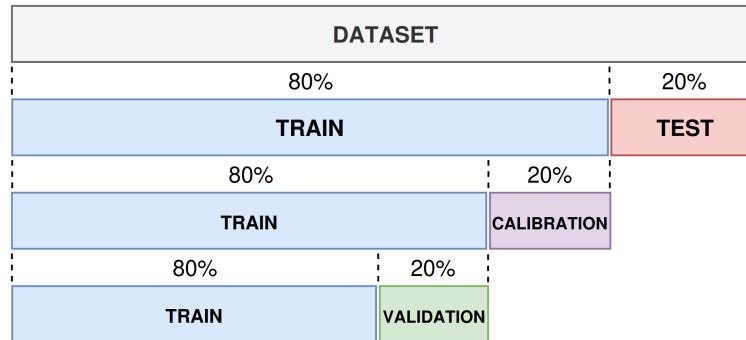


Figura 2.2: Diagrama de división del *dataset* en *train*, *validation*, *calibration* y *test*.

Para una comparativa más justa entre los métodos que usan CP y los

que no, se utilizará la siguiente estrategia: los métodos que no emplean CP seguirán el esquema tradicional de división de datos (en entrenamiento, 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.

## 2.2. 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  $448 \times 224$  píxeles para homogenizar las entradas del modelo<sup>3</sup>. 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.

## 2.3. 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

---

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

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.

### **2.3.1. Problema de estimación de edad**

Para el problema de estimación de edad se han propuesto los siguientes cuatro métodos:

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

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

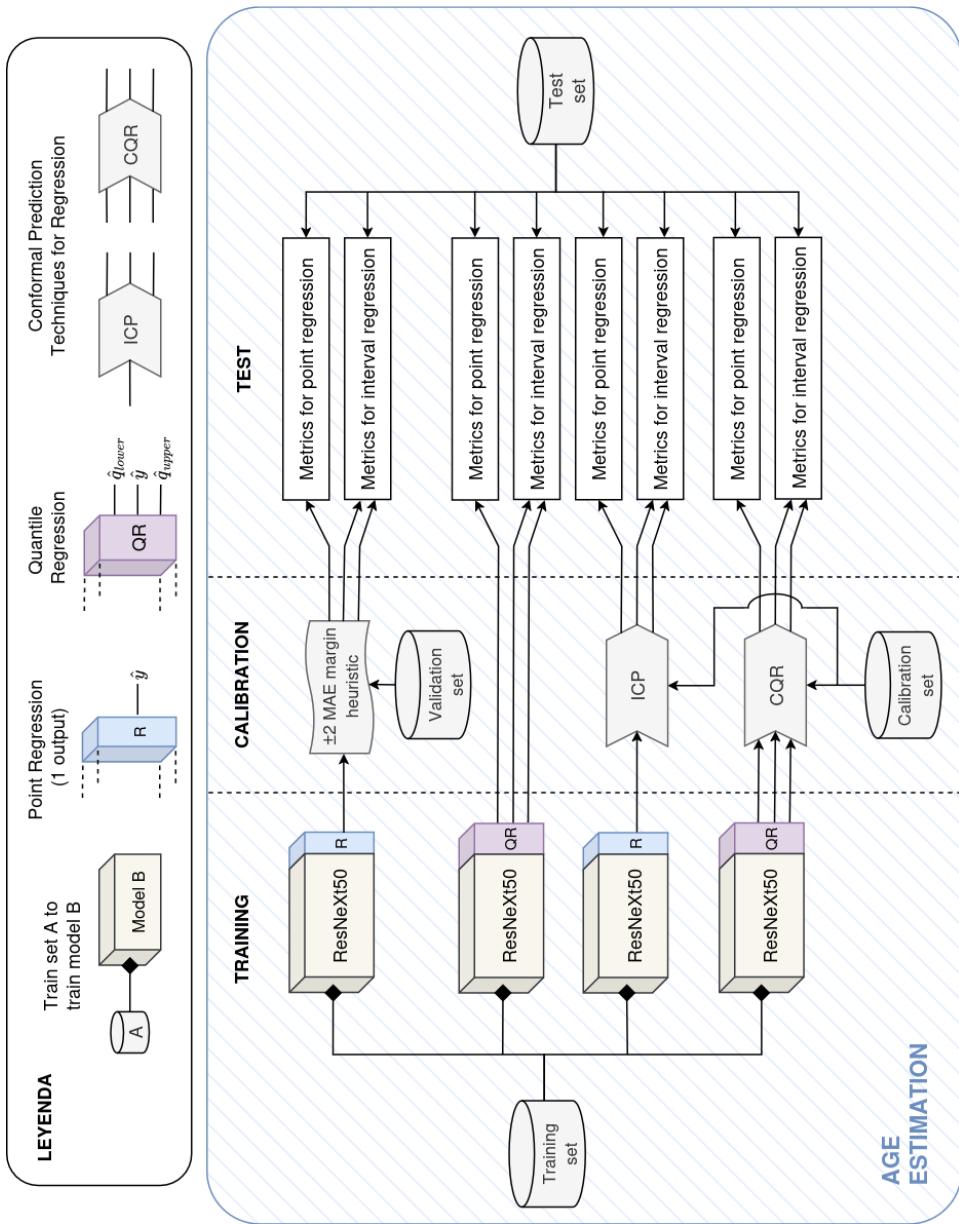


Figura 2.3: Esquema de experimentación para la estimación de edad. Cada modelo se entrena por separado. ‘R’ se refiere a ‘Rregresión puntual’ (de una sola neurona de salida), ‘QR’ a ‘Quantile Regression’ , ‘ICP’ a ‘Inductive Conformal Prediction’ y ‘CQR’ a Conformalized Quantile Regression.

### 2.3.2. Problema de clasificación de mayoría de edad

Respecto al problema de clasificación de mayoría de edad, se han propuesto los siguientes tres métodos:

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

No se han implementado 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.

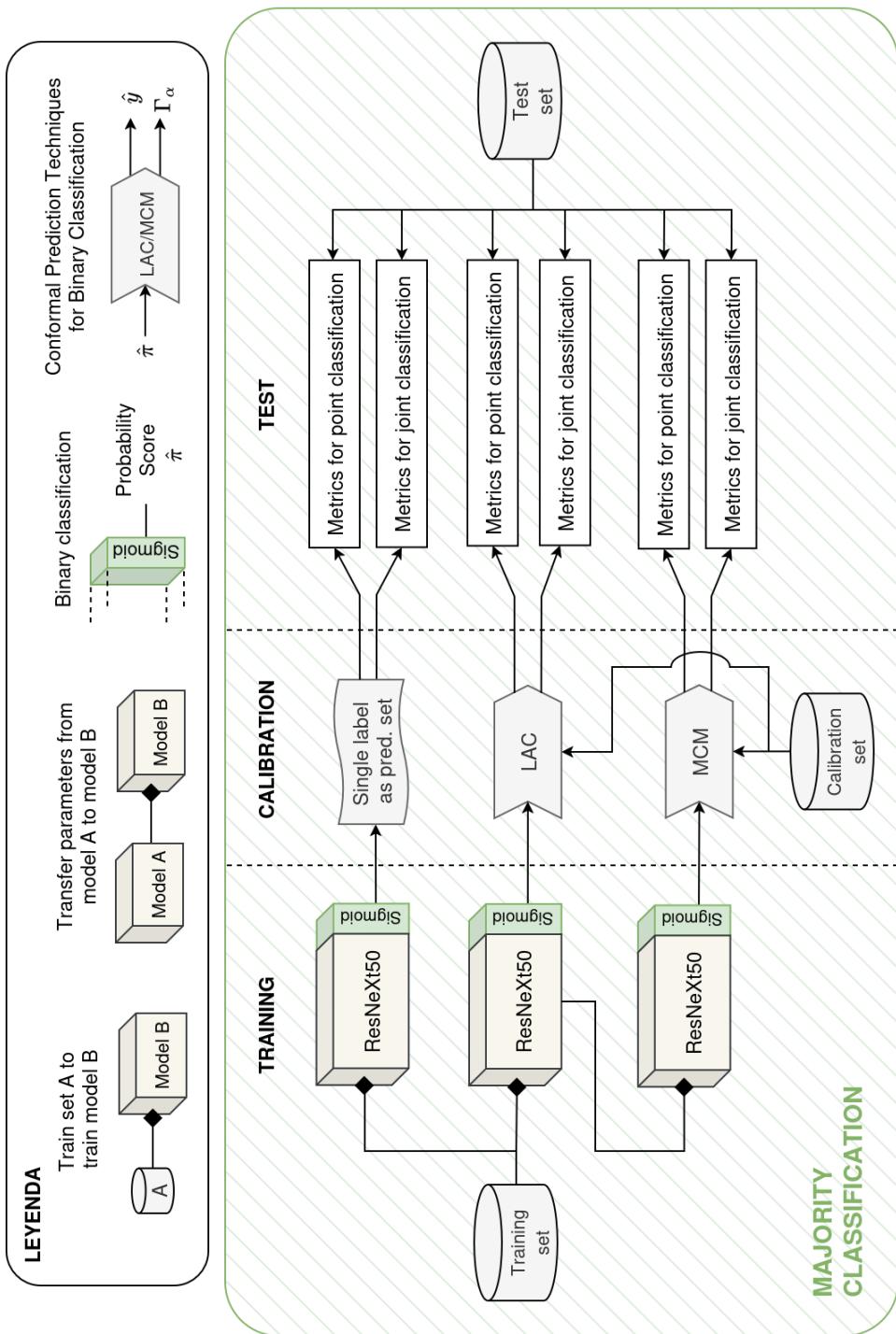


Figura 2.4: Esquema de experimentación para la clasificación de mayoría de edad.

**2.3.3. Problema de clasificación de edad**

Para el problema de clasificación de edad, se ha empleado la técnica de calibración de probabilidades *Platt Scaling* para ajustar las salidas del modelo de clasificación multiclas, 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:

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

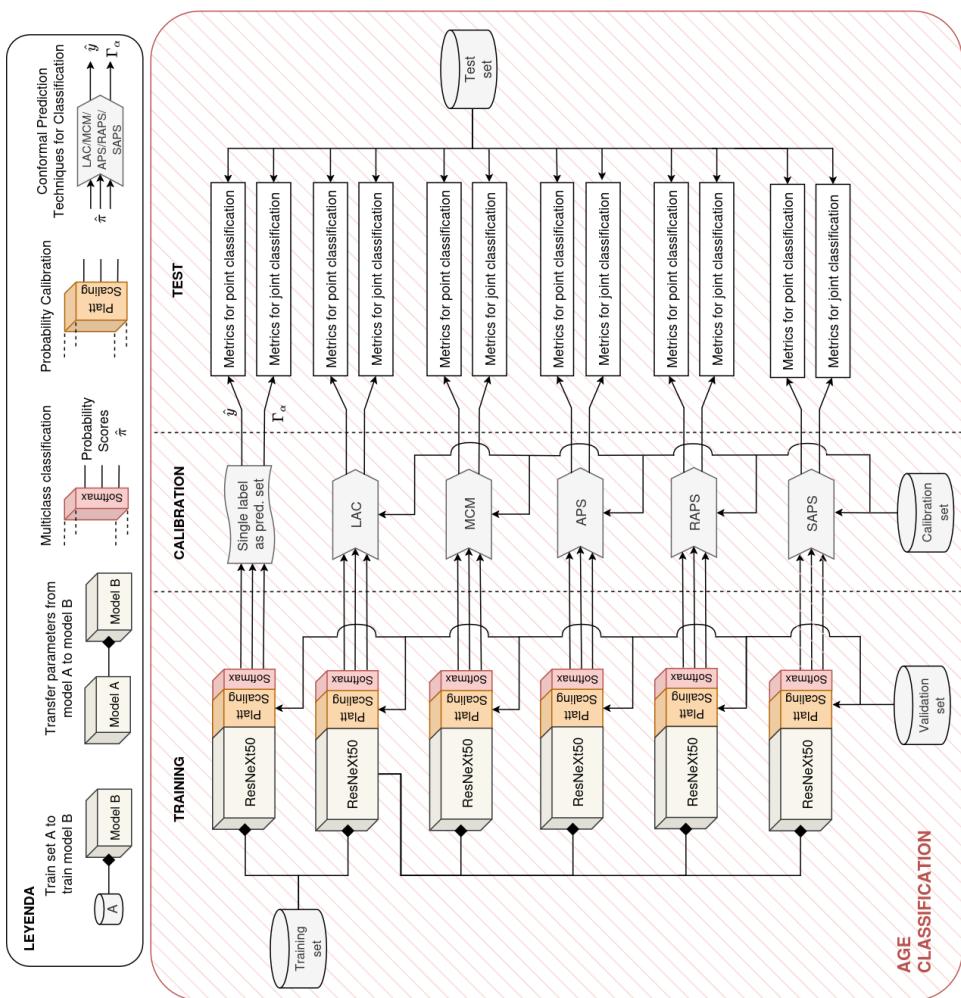


Figura 2.5: Esquema de experimentación para la clasificación de edad. Se recuerda que LAC, MCM, APS, RAPS y SAPS refieren a los métodos de CP para clasificación presentados en el anterior capítulo.

**2.3.4. 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:

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.
3. **Kruskal-Wallis + Dunn:** Si no se cumple la normalidad, se opta por un enfoque no paramétrico. El test de Kruskal-Wallis compara medianas entre grupos y no requiere que los datos sigan una distribución normal. Cuando se detectan diferencias significativas, se realizan comparaciones por pares con el test de Dunn, aplicando corrección de Bonferroni para controlar el error tipo I. Esta estrategia asegura la validez estadística incluso cuando los supuestos paramétricos no se cumplen.

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 parámetricas) 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 I mediante correcciones apropiadas según la prueba utilizada (Tukey HSD, Games-Howell o Dunn con Bonferroni).

## 2.4. Experimentación para la estimación de edad

### 2.4.1. Entrenamiento de los modelos

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

La salida aplanada pasa por dos bloques densos consecutivos, cada uno compuesto por una capa *batch normalization*, una capa de *dropout* y una capa completamente conectada (FC), con una activación ReLU entre ambos bloques. La primera capa FC contiene 4096 neuronas, la segunda 512, y finalmente se incluye una capa de salida de una sola neurona. Esta configuración ha sido seleccionada siguiendo la recomendación de los tutores, quienes cuentan con experiencia previa en el trabajo con este conjunto de datos.

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

Añadir un dibujo con el cambio de cabecera (AGOSTO)

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 [18]. 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* [19].

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 menor<sup>4</sup>. 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 entre-

---

<sup>4</sup>Se 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.

namiento, 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.

En la Figura 2.6 se puede ver la curva de aprendizaje de uno de los modelos entrenados.

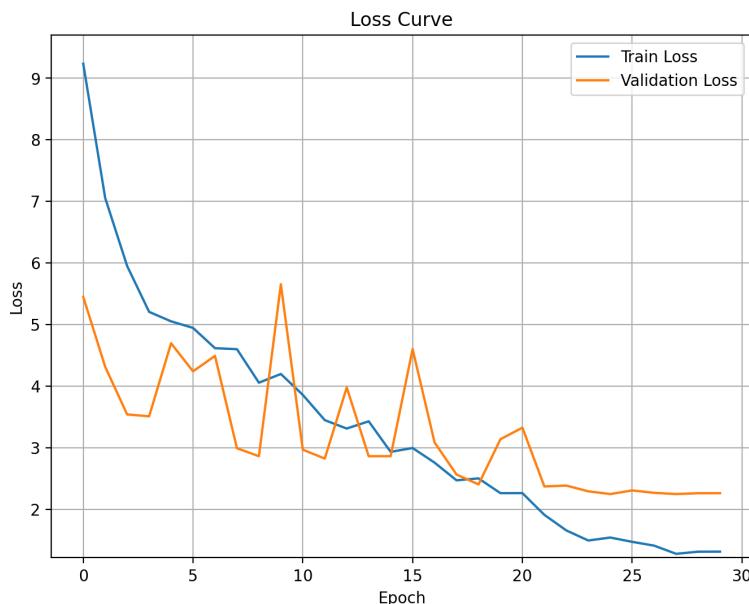


Figura 2.6: Curva de aprendizaje de uno de los modelos para el método ICP. En color azul se muestran las pérdidas obtenidas en el conjunto de entrenamiento, mientras que en color naranja se representan las correspondientes al conjunto de validación. Se observa una convergencia alrededor de la época 25.

#### 2.4.2. Resultados

##### Análisis de métricas para la estimación puntual de edad

La Tabla 2.1 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 reveló diferencias significativas entre métodos tanto en el MAE ( $F(3, 36) = 27.754, p < 0.001$ ) como el MSE ( $F(3, 36) = 17.284, p < 0.001$ ), confirmadas mediante ANOVA bajo el cumplimiento de todos los supuestos: 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 de Tukey HSD (véanse las Tablas 2.2 y 2.3). Los resultados identificaron los siguientes patrones:

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	<b>1.16</b>	1.20	<b>1.16</b>	1.18	<b>2.36</b>	2.48	2.39	2.44

Tabla 2.1: 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.

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

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 2.2: Resultados de la prueba *post-hoc* de Tukey HSD para MAE 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 2.3: Resultados de la prueba *post-hoc* de Tukey HSD para MSE 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 2.4 presenta las métricas sobre las predicciones interválicas de los métodos. A primera vista, se observan diferencias marcadas entre los métodos conformales y no conformales en las métricas de cobertura empírica y amplitud del intervalo. En particular, los métodos no conformales ('base' y QR) muestran coberturas notablemente inferiores al nivel deseado (alrededor del 88-89 % frente al 95 % nominal), lo que indica una infracobertura sistemática. Esto ocurre porque ni la heurística del método 'base' ni las regiones generadas por la regresión cuantílica en QR cuentan con garantías teóricas de cobertura estadística.

Ejecución	Cobertura Empírica (%)				Amplitud Media del Intervalo			
	base	ICP	QR	CQR	base	ICP	QR	CQR
Ejecución 1	87.41	94.47	89.03	95.31	4.53	6.17	4.71	6.23
Ejecución 2	87.96	94.84	89.27	94.80	4.57	6.27	4.67	6.11
Ejecución 3	87.73	95.03	88.38	95.45	4.60	6.34	4.65	6.02
Ejecución 4	88.06	94.19	89.50	94.61	4.58	6.04	4.63	5.90
Ejecución 5	87.87	95.03	89.13	94.93	4.63	6.28	4.59	5.92
Ejecución 6	88.57	94.80	89.41	94.33	4.68	6.14	4.63	5.94
Ejecución 7	88.24	95.21	88.80	95.26	4.61	6.33	4.63	6.00
Ejecución 8	87.55	94.70	88.01	95.12	4.64	6.12	4.67	6.08
Ejecución 9	87.87	95.03	88.38	94.93	4.66	6.25	4.62	6.06
Ejecución 10	88.57	95.12	89.27	94.56	4.64	6.20	4.64	5.96
Media	87.98	<b>94.84</b>	88.92	<b>94.93</b>	4.61	<b>6.21</b>	4.64	<b>6.02</b>

Tabla 2.4: 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 cercana o superior al 95 %.

En contraste, los métodos conformales (ICP y CQR) sí logran coberturas próximas al valor nominal, tal como se espera dada su fundamentación estadística. Esta mayor cobertura, sin embargo, tiene un coste en cuanto a la amplitud del intervalo, que es mayor en estos métodos. Esta relación de compromiso o *trade-off* entre cobertura y amplitud de los intervalos —típico en la predicción interválica— se visualiza claramente en la Figura 2.7, donde se observa una alta correlación entre la cobertura empírica y el tamaño del intervalo de predicción.

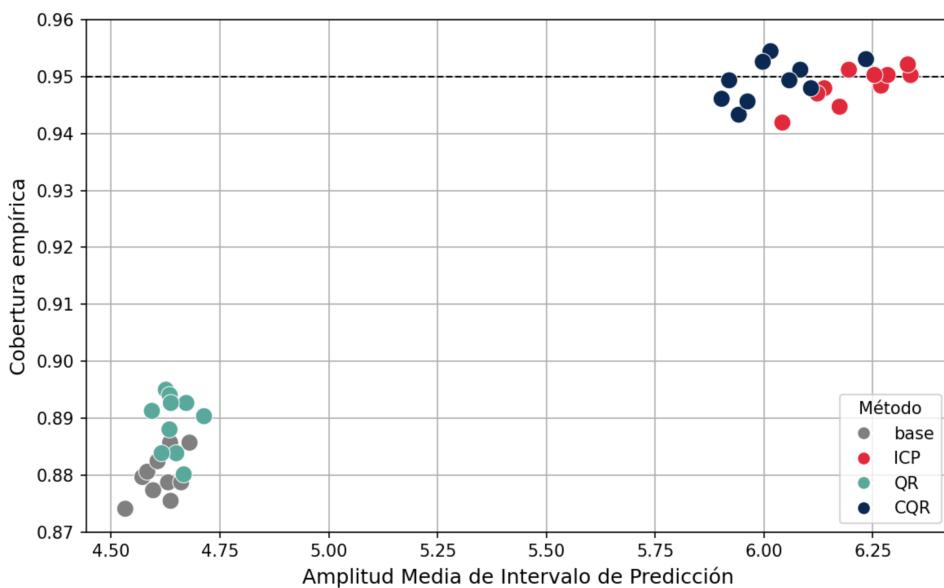


Figura 2.7: Gráfica de dispersión de la Cobertura empírica frente a la Amplitud media del intervalo de predicción. Existe una relación de compromiso entre la cobertura y la amplitud de los intervalos: al aumentar una, generalmente también lo hace la otra, y viceversa. Los métodos más eficaces son aquellos 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.

CQR presenta unas amplitudes medias de intervalo significativamente más reducidas que ICP, logrando ambos métodos coberturas muy similares. Esta diferencia significativa entre amplitudes de intervalo se ha comprobado estadísticamente mediante un test Welch ANOVA<sup>5</sup>, que mostró diferencias globales significativas entre los métodos ( $F(3, 18.62) = 1240.15, p < 0.001$ ). Posteriormente, las comparaciones por pares mediante Games-Howell (véase la Tabla 2.5) confirmaron que CQR tiene intervalos significativamente más estrechos que ICP, así como que también se diferencia significativamente de otros métodos como QR y base. Estas pruebas permiten concluir que, aunque la cobertura empírica sea similar, CQR consigue reducir la amplitud del intervalo de predicción de manera estadísticamente significativa frente a ICP y otros métodos.

Modelo 1	Modelo 2	Dif. media	Valor $p$	IC 95 %	Signif.
base	ICP	-1.6012	<0.0001	[-1.6739, -1.5286]	Sí
base	QR	-0.0310	0.323	[-0.0680, 0.0061]	No
base	CQR	-1.4090	<0.0001	[-1.4850, -1.3328]	Sí
ICP	QR	1.5703	<0.0001	[1.4993, 1.6412]	Sí
ICP	CQR	0.1923	0.002	[0.0993, 0.2854]	Sí
QR	CQR	-1.3780	<0.0001	[-1.4524, -1.3035]	Sí

Tabla 2.5: 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 ( $\alpha = 0.05$ ).

De hecho, en la Tabla 2.6 apreciamos cómo CQR logra significativamente menores valores de *interval score* que ICP, indicando que CQR tiene un mejor equilibrio entre cobertura y tamaño del intervalo. En consecuencia, CQR se perfila como una opción más ventajosa, con garantías de cobertura e intervalos de predicción ajustados.

### Análisis de la cobertura en base al tamaño del intervalo

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

<sup>5</sup>Se aplicaron estos tests porque los datos mostraban normalidad en los residuos (Shapiro-Wilk,  $p > 0.8$ ), pero no cumplían homocedasticidad (Levene,  $p < 0.01$ ), lo que hace inapropiado un ANOVA clásico y justifica el uso de Welch ANOVA y Games-Howell.

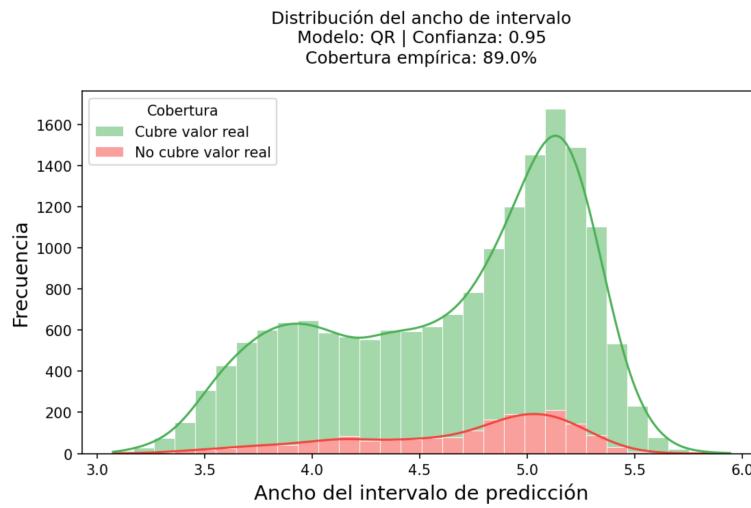
Ejecución	Mean Interval Score			
	base	ICP	QR	CQR
Ejecución 1	9.16	8.17	8.48	8.02
Ejecución 2	8.93	8.21	8.72	8.04
Ejecución 3	8.90	8.24	8.86	7.85
Ejecución 4	8.69	8.00	8.59	7.98
Ejecución 5	8.88	8.27	8.82	7.89
Ejecución 6	8.93	8.19	8.46	8.01
Ejecución 7	8.81	8.19	8.96	7.85
Ejecución 8	8.88	8.03	8.80	7.91
Ejecución 9	8.89	7.99	8.96	7.92
Ejecución 10	8.62	8.07	8.56	8.20
Media	8.85	8.14	8.72	<b>7.97</b>

Tabla 2.6: Resultados de las predicciones obtenidas por los modelos para el problema de estimación de edad en cada ejecución. Se marca en negrita la mejor marca en la métrica media.

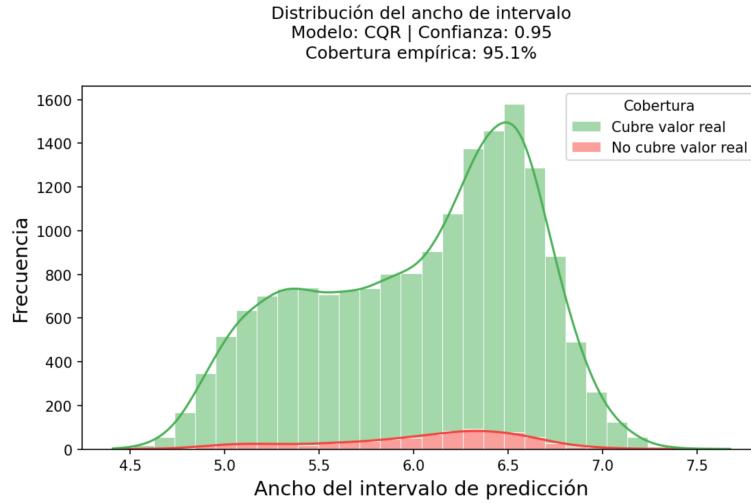
za, de forma que todos los intervalos lograrían cubrir al nivel de confianza deseado los valores reales.

En el peor de los escenarios, los intervalos más estrechos tenderían a infracubrir (es decir, no contienen 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.

En la Figura 2.8 se presentan los histogramas de la amplitud de los intervalos de predicción para dos modelos representativos, uno QR y otro CQR. En cada caso, se diferencia visualmente la cantidad de instancias cuya predicción cubre el valor real de aquellas en las que no lo hace. Es notable en ambas figuras la presencia de dos grupos principales de instancias: uno más reducido, asociado a intervalos más estrechos, y otro más numeroso, correspondiente a intervalos de mayor amplitud. Respecto a la cobertura, el modelo QR presenta valores inferiores, lo cual es consistente con su cobertura marginal, que ya se encontraba por debajo del 89 %. En cuanto al ratio entre cobertura e incobertura, este parece mantenerse relativamente estable a lo largo de los distintos rangos de amplitud del intervalo. Sin embargo, para un análisis más detallado y específico sobre cómo varía la cobertura en función



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



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

Figura 2.8: Histogramas de amplitud del intervalo de predicción con diferenciación por cobertura, correspondientes a los modelos QR y CQR. Para cada tipo de método se ha seleccionado el modelo con el mejor *interval score*. La comparación permite visualizar cómo varía la capacidad de cobertura en función del tamaño del intervalo.

del tamaño del intervalo, observemos la información desglosada en la Tabla 2.7.

En la Tabla 2.7 se ofrece información detallada sobre la cobertura empírica alcanzada por cada método de predicción (en todas sus ejecuciones) en función de diferentes rangos de amplitud del intervalo de predicción. Esta desagregación permite analizar si existe una relación entre el tamaño del intervalo y la capacidad del modelo para cubrir el valor real.

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

Llama la atención que se logra sobrecobertura tanto en los intervalos más estrechos como en los más amplios, a costa de una infracobertura en los intervalos de amplitud intermedia, concretamente entre 5.5 y 6.5 años, siendo especialmente más bajas en el último medio tramo, donde la cobertura alcanza un 93.24 %.

Amplitud del intervalo	Cobertura Empírica (%)			
	base	ICP	QR	CQR
[3.0, 3.5)	—	—	91.97	—
[3.5, 4.0)	—	—	92.89	—
[4.0, 4.5)	—	—	88.18	100
[4.5, 5.0)	87.99	—	85.59	96.72
[5.0, 5.5)	—	—	89.44	96.39
[5.5, 6.0)	—	—	97.12	94.56
[6.0, 6.5)	—	94.84	—	93.14
[6.5, 7.0)	—	—	—	96.20
[7.0, 7.5)	—	—	—	97.85
[7.5, 8.0)	—	—	—	100

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

Tabla 2.7: Cobertura empírica del intervalo de predicción obtenida por cada método de predicción para distintas franjas de amplitud de intervalos. Nota: Estos cálculos se han realizado sobre todas las instancias predichas para cada método en las 10 ejecuciones o entrenamientos realizados. Cabe recordar que la cobertura objetivo es del 95 %.

### Análisis de la cobertura en base a la edad cronológica

Por último, se ha analizado la cobertura en base a la edad real de los individuos, ya que resulta crucial identificar posibles sesgos en el desempeño del modelo a lo largo de esta variable. La Figura 2.9 muestra la evolución de la cobertura empírica y el ancho medio de los intervalos de predicción en función de la edad cronológica<sup>6</sup>.

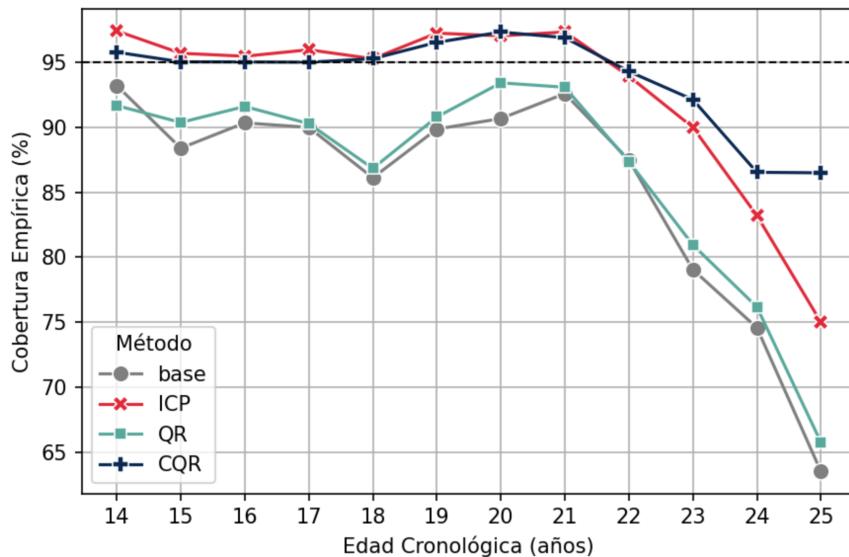
Se observa que todos los métodos tienden a reducir su cobertura conforme aumenta la edad cronológica de los individuos. Esta disminución es especialmente notable a partir de los 22 años, afectando incluso al método CQR, el método hasta ahora con la cobertura más robusta.

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

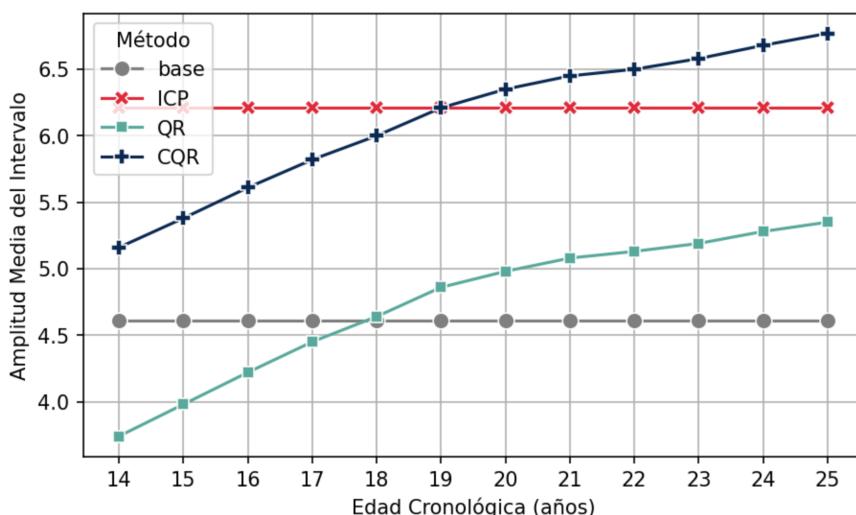
- Escasez de ejemplos en edades avanzadas: El conjunto de datos presenta una disminución en el número de muestras a partir de los 23 años, lo que coincide con la reducción en la cobertura predictiva.
- Mayor variabilidad fisiológica en adultos jóvenes: A medida que aumenta la edad, los individuos suelen presentar una mayor diversidad en sus características biológicas debido a la acumulación de factores ambientales y estilos de vida [20, 21].

---

<sup>6</sup>Parte entera (suelo) de la edad real.



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



(b) Gráfico de líneas de amplitud media del intervalo de predicción para cada método en función de la edad cronológica entera de los individuos. Esta gráfica muestra cómo cambia el tamaño de intervalo con la edad.

Figura 2.9: Gráficos de líneas comparativos de la cobertura empírica y la amplitud media del intervalo de predicción por edad cronológica para los diferentes métodos evaluados.

## **48 2.5. Experimentación para la clasificación de mayoría de edad**

### **2.5. Experimentación para la clasificación de mayoría de edad**

#### **2.5.1. 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 extractor 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.

#### **2.5.2. Resultados**

##### **Análisis de métricas para la clasificación puntual de mayoría de edad**

En la Tabla 2.8 se presentan las métricas que evalúan el rendimiento del modelo de clasificación en sus predicciones de una sola etiqueta. El método ‘base’ obtiene una exactitud (*accuracy*) significativamente superior que los métodos conformales<sup>7</sup>, 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

---

<sup>7</sup>Comprobado estadísticamente mediante 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.

el desequilibrio de clases. Como resultado, tienden a favorecer la clase mayoritaria ( $\geq 18$ ), lo que incrementa los falsos positivos y reduce los verdaderos negativos.

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	<b>87.63</b>	87.00	<b>89.47</b>	<b>89.47</b>	<b>84.83</b>	83.22

Tabla 2.8: 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 los mismas predicciones puntuales). Se marca en negrita la media con mejor valor para cada métrica.

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

La Tabla 2.9 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 2.10.

Se observa que 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. Por otro lado, aunque los métodos LAC y MCM muestran tamaños medios del conjunto de predicción muy similares, LAC alcanza una cobertura significativamente superior en prácticamente

## **50 2.5. Experimentación para la clasificación de mayoría de edad**

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 2.9: 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.

todas las ejecuciones. Esto se confirma estadísticamente mediante un test ANOVA:  $F(2, 27) = 1097.68, p < 0.001$ , cumpliendo todos los supuestos necesarios: normalidad (Shapiro-Wilk,  $p > 0.5$ ) y homocedasticidad (Levene  $p > 0.18$ ).

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 2.10: 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 ( $\alpha = 0.05$ ).

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.

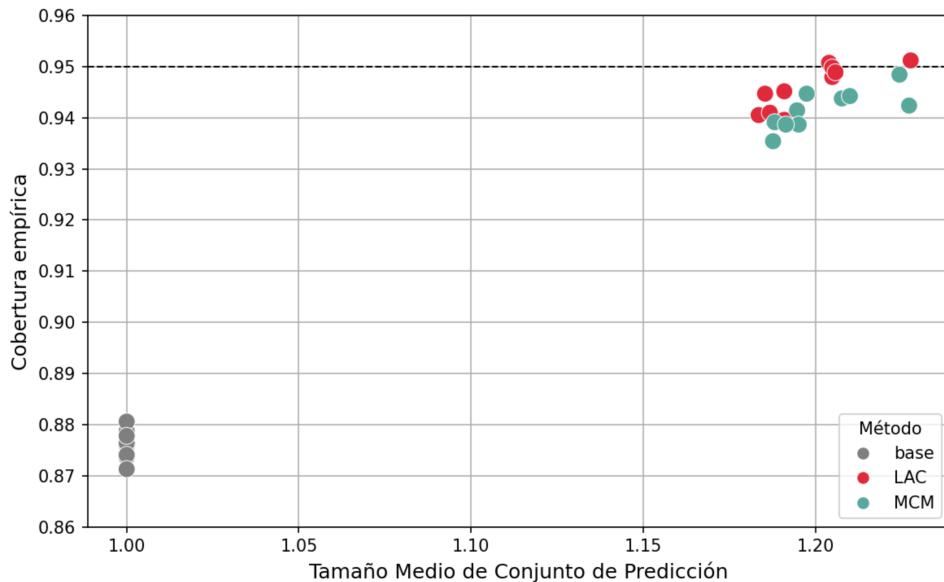


Figura 2.10: Gráfica de dispersión Cobertura empírica - Tamaño Medio de Conjunto de Predicción.

## **52 2.5. Experimentación para la clasificación de mayoría de edad**

### **Análisis de la cobertura en base a la clase**

Ahora analizaremos la cobertura en cada clase mediante las matrices de confusión obtenidas por cada método. En la Figura 2.11 se recogen las matrices de confusión conformales —normalizadas por el número de instancias total de cada etiqueta real— de los diferentes métodos.

La cobertura empírica de una clase se define como la proporción de instancias en las que la etiqueta verdadera está presente dentro del conjunto de predicción generado. Para calcularla, se suman las proporciones de instancias cuyo conjunto de predicción incluye la etiqueta real, considerando únicamente aquellas instancias pertenecientes a la clase en cuestión. Por ejemplo, la cobertura empírica para la clase ‘menor de 18’ corresponde a la suma de las proporciones de instancias que contienen la etiqueta ‘menor de 18’ en su conjunto de predicción, siendo su etiqueta real ‘menor de 18’.

Respecto al método ‘base’, cabe señalar que la cobertura para cada clase coincide con las métricas clásicas de sensibilidad y especificidad, ya que el conjunto de predicción contiene siempre una única etiqueta y no se emplea ningún ajuste adicional para calibrar la confianza.

Resulta llamativo que el método LAC muestre infracobertura en las instancias de menores de 18 años y sobrecobertura en aquellas de 18 años o más, mientras que en el caso del método MCM ocurre lo contrario, lo que querría indicar que LAC es más fiable para estimaciones en población adulta, mientras que MCM ofrecería mejores resultados en población menor de edad.

Si la prioridad en la predicción fuera maximizar la cobertura de los 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 sobrecobertura en el resto de la población.

Este último párrafo no aporta mucho, pero si lo borro no habré discutido nada sobre el método ‘base’.

		Conjunto predicho			Cobertura
		{<18}	{≥18}	{<18,≥18}	
Etiqueta real	<18	84.43	15.17	–	<b>84.43</b>
	≥18	10.53	89.47	–	<b>89.47</b>

(a) base

		Conjunto predicho			Cobertura
		{<18}	{≥18}	{<18,≥18}	
Etiqueta real	<18	68.28	6.58	25.15	<b>93.43</b>
	≥18	4.63	78.99	16.37	<b>95.36</b>

(b) LAC

		Conjunto predicho			Cobertura
		{<18}	{≥18}	{<18,≥18}	
Etiqueta real	<18	76.86	3.77	19.37	<b>96.23</b>
	≥18	7.18	72.00	20.82	<b>92.82</b>

(c) MCM

Figura 2.11: Matrices de confusión conformal correspondientes a los métodos ‘base’, LAC y MCM. En cada celda, el valor indica la proporción de instancias que se obtiene un determinado conjunto de predicción dada una determinada etiqueta verdadera. Se recomienda leer horizontalmente, dado que estos valores están normalizados en esta dimensión. Todos los valores están expresados en porcentaje.

## 2.6. Experimentación para la estimación de edad como problema de clasificación

### 2.6.1. Entrenamiento de los modelos

Dado que este es un problema directamente derivado del primer problema de estimación de edad como regresión, 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 25 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 clasificación 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.

### 2.6.2. Resultados

En este caso no se han analizado en profundidad 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 han probado 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

La Tabla 2.12 presenta las métricas sobre el conjunto de predicción de los métodos. Se observa, como era de esperar una cobertura muy baja para el método ‘base’ como se podía venir augurando por las mismas razones

anteriormente descritas para la clasificación puntual. Por ello, ignoraremos este método de ahora en adelante.

Para facilitar la interpretación, en la Figura 2.13 se representan gráficamente estos valores, lo que permite apreciar también la relación de *trade-off* entre las métricas. En particular, se observa que los métodos LAC y los adaptativos forman una nube de puntos claramente separada de la correspondiente a MCM, el cual ofrece una cobertura empírica ligeramente superior, aunque a costa de un tamaño medio del conjunto de predicción considerablemente mayor. Esto probablemente se deba a que el MCM 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 para el presente problema y ha sido descartado del análisis estadístico.

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:  $p = 0.56$  para la cobertura empírica y  $p = 0.4$  para el tamaño medio) y homocedasticidad (Levene:  $p > 0.9$  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 2.11 y 2.12.

Modelo 1	Modelo 2	Dif. media	Valor $p$	IC 95 %	Signif.
APS	LAC	0.0033	0.394	[-0.002, 0.0087]	No
APS	RAPS	0.002	0.7742	[-0.0035, 0.0074]	No
APS	SAPS	0.0076	0.0037	[0.0021, 0.0131]	Sí
LAC	RAPS	-0.0013	0.919	[-0.0068, 0.0042]	No
LAC	SAPS	0.0043	0.1663	[-0.0012, 0.0098]	No
RAPS	SAPS	0.0056	0.0431	[0.0001, 0.0111]	Sí

Tabla 2.11: 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 ( $\alpha = 0.05$ ).

Los resultados revelan para un 95 % de nivel de confianza:

- Cobertura empírica: SAPS presenta una cobertura ligeramente superior que APS y RAPS, con diferencias medias de 0.76 % y 0.56 %,

Método	Cobertura empírica (%)					
	base	LAC	MCM	APS	RAPS	SAPS
Ejecución 1	26.53	94.66	95.86	94.38	94.47	94.98
Ejecución 2	25.46	94.24	95.45	93.63	94.10	95.12
Ejecución 3	27.51	95.21	95.49	94.61	94.52	95.26
Ejecución 4	27.60	94.89	95.59	94.56	94.80	94.80
Ejecución 5	27.51	95.17	95.86	94.93	95.45	95.21
Ejecución 6	27.74	94.80	94.70	94.52	94.61	95.45
Ejecución 7	28.02	93.91	94.80	94.28	94.01	94.56
Ejecución 8	25.98	95.59	95.49	95.07	95.17	95.86
Ejecución 9	28.39	94.70	95.63	93.91	94.70	95.59
Ejecución 10	28.49	94.14	95.59	94.14	94.19	94.80
Media	27.32	94.73	95.45	94.41	94.60	95.16

(a) Cobertura empírica

Método	Tamaño Medio del Conjunto					
	base	LAC	MCM	APS	RAPS	SAPS
Ejecución 1	1.00	5.79	7.83	6.09	5.89	6.05
Ejecución 2	1.00	5.76	7.84	5.89	5.85	6.03
Ejecución 3	1.00	6.04	7.70	6.06	5.89	6.17
Ejecución 4	1.00	5.86	7.75	6.17	6.11	5.98
Ejecución 5	1.00	5.77	7.81	6.14	6.12	6.16
Ejecución 6	1.00	5.80	7.70	6.18	5.97	6.08
Ejecución 7	1.00	5.69	7.19	5.90	5.77	6.07
Ejecución 8	1.00	6.03	7.80	6.25	6.03	6.28
Ejecución 9	1.00	5.86	7.70	6.00	6.00	6.15
Ejecución 10	1.00	5.88	7.61	6.23	6.12	6.36
Media	1.00	5.85	7.69	6.09	5.97	6.13

(b) Tamaño medio del conjunto de predicción

Figura 2.12: 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.

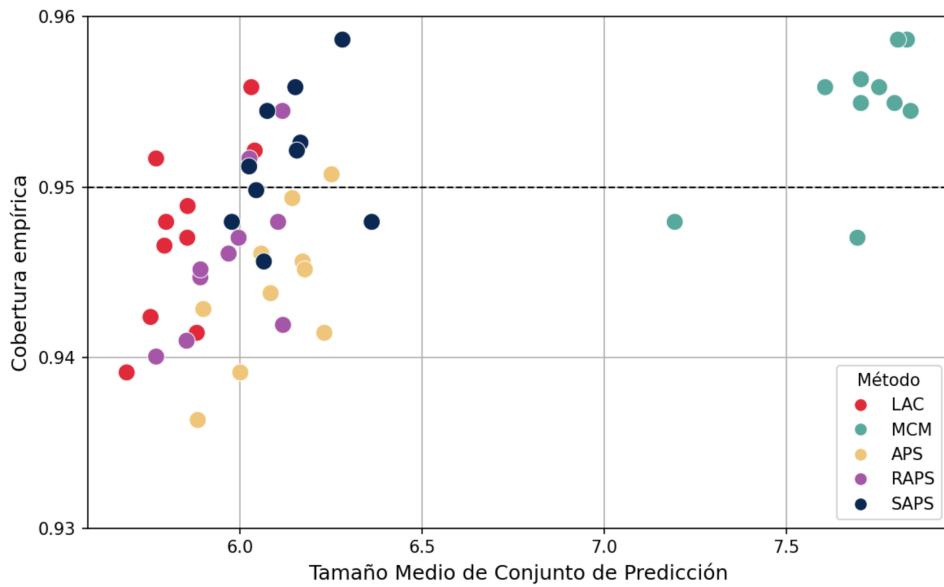


Figura 2.13: Gráfica de dispersión Cobertura empírica - Tamaño Medio de Conjunto de Predicción. No se incluyen los puntos del método ‘base’ dado que estos

Modelo 1	Modelo 2	Dif. media	Valor <i>p</i>	IC 95 %	Signif.
APS	LAC	-0.2435	0.0004	[-0.3892, -0.0978]	Sí
APS	RAPS	-0.1167	0.1551	[-0.2624, 0.029]	No
APS	SAPS	0.0401	0.8802	[-0.1057, 0.1858]	No
LAC	RAPS	0.1268	0.107	[-0.0189, 0.2725]	No
LAC	SAPS	0.2836	0	[0.1378, 0.4293]	Sí
RAPS	SAPS	0.1567	0.031	[0.011, 0.3025]	Sí

Tabla 2.12: 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 ( $\alpha = 0.05$ ).

respectivamente. No se detectaron diferencias significativas entre los demás pares de métodos.

- Tamaño medio del conjunto de predicción: APS genera conjuntos significativamente más grandes que LAC (diferencia de media de 0.2435), y SAPS también supera a RAPS y LAC (diferencia media de 0.1567 y 0.2836, respectivamente). No se detectaron diferencias significativas entre el resto.

Por tanto, de entre todos los métodos seleccionados, basándonos únicamente en las dos métricas de cobertura empírica y tamaño medio del conjunto de predicción, podríamos destacar dos métodos con buena relación cobertura/tamaño medio del conjunto:

- LAC se presenta como la alternativa más equilibrada, ya que, manteniendo una cobertura comparable a la de APS y RAPS, logra un tamaño medio del conjunto de predicción menor. Esto se traduce en salidas más compactas sin pérdida significativa de fiabilidad.
- SAPS alcanza una cobertura empírica ligeramente superior a la de los demás métodos, aunque este incremento viene acompañado de un aumento moderado en el tamaño medio del conjunto de predicción.

#### **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 2.14 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 mayor es el conjunto, más probable es que incluya la clase verdadera.
- **Sobrecobertura como síntoma de desequilibrio:** la presencia de sobrecobertura en determinados tamaños implica, inevitablemente, infracobertura en otros. Cuando este patrón se repite y la sobrecobertura se concentra en conjuntos de gran tamaño, suele indicar que el método está “compensando” un mal ajuste en los conjuntos pequeños, lo cual

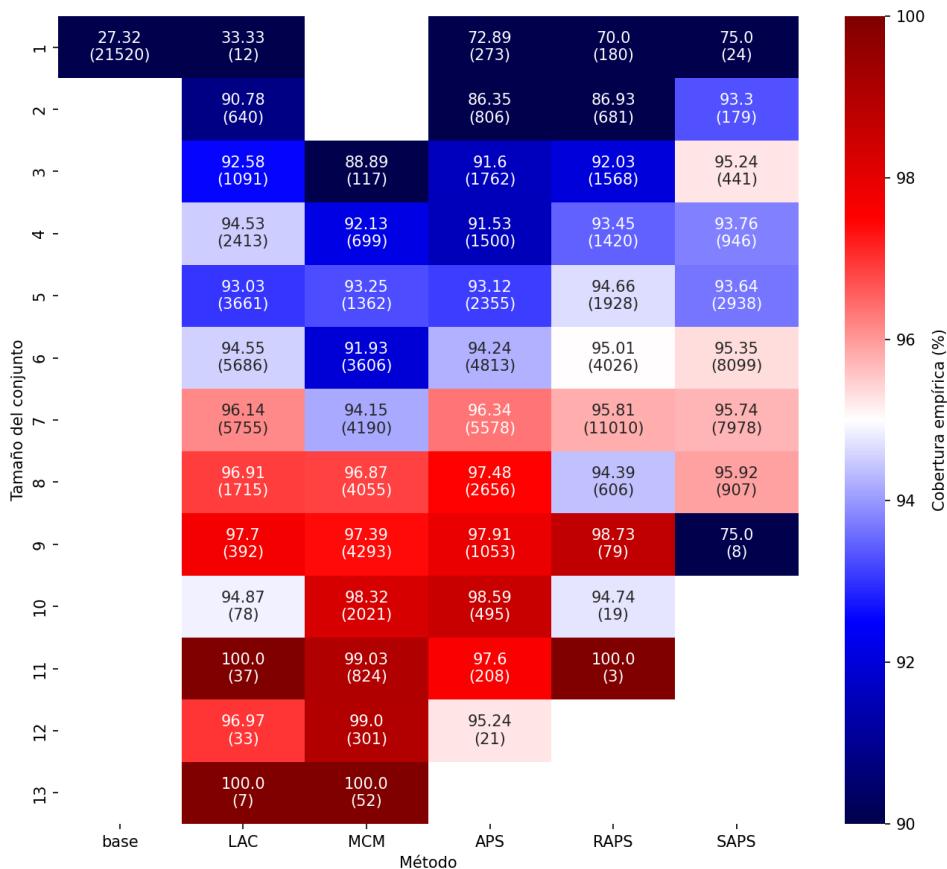


Figura 2.14: Mapa de calor de cobertura empírica en base al tamaño del conjunto por cada método de predicción a lo largo de las distintas ejecuciones. La escala de colores está centrada en la cobertura nominal (0.95): los valores por debajo de este umbral se representan en tonos azules, los superiores en tonos rojos, y el blanco indica una cobertura empírica equivalente a la nominal.

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 muy conservadores, con un gran número de etiquetas. Presenta infracobertura para instancias cuyo conjunto de predicción contiene entre 3 y 7 etiquetas, y sobrecobertura para tamaños de 8 a 13. Su adaptatividad es baja, ya que no ajusta el tamaño del conjunto en función del nivel de incertidumbre de la instancia.
- **LAC:** muestra una alta variabilidad en el tamaño de los conjuntos, que oscilan entre 1 y 13 etiquetas. Registra infracobertura para tamaños de 1 a 6 etiquetas y sobrecobertura para el resto, con la excepción de los conjuntos de 10 etiquetas, donde la cobertura es muy próxima a la nominal.
- **APS:** comparte el patrón de LAC —si bien no presenta conjuntos de predicción de 13 etiquetas—, con infracobertura para tamaños de 6 etiquetas o menos y sobrecobertura para los mayores. Sin embargo, tanto las infracoberturas como las sobre coberturas son más pronunciadas, evidenciando un mayor desequilibrio.
- **RAPS:** muy similar a APS, pero mejorando sus marcas, por lo general aumenta la cobertura de aquellas marcas en las que APS presenta infracobertura, y reduce la cobertura en aquellas en las que presenta sobre cobertura. Además reduce la variabilidad de tamaños del conjunto, concentrando muchas instancias entre 5 y 7 etiquetas, con coberturas muy cercanas al nominal.
- **SAPS:** es el método con mayor estabilidad en el tamaño de los conjuntos de predicción, que varían entre 1 y 9 etiquetas. Presenta las mayores cifras de cobertura empírica para tamaños de conjuntos de predicción menores de 4 etiquetas, si bien siguen infracubriendo.

SAPS ha presentado valores de cobertura más estables para los diferentes tamaños del conjunto de predicción, así como mayor estabilidad en los propios tamaños de los conjuntos, siendo el más equilibrado, sin llegar a ser demasiado conservador ni excesivamente arriesgado. Esto sugiere que SAPS logra un mejor compromiso entre precisión y fiabilidad, manteniendo la cobertura cercana al valor nominal en un rango amplio de tamaños y evitando los extremos de infracobertura pronunciada o sobre cobertura excesiva que presentan otros métodos.

### 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. La Figura 2.15 muestra la relación de la cobertura empírica y el tamaño medio de los conjuntos de predicción con las distintas edades cronológicas en el conjunto de datos.

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 23 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 20-22 años, y posteriormente disminuye en las edades más avanzadas.

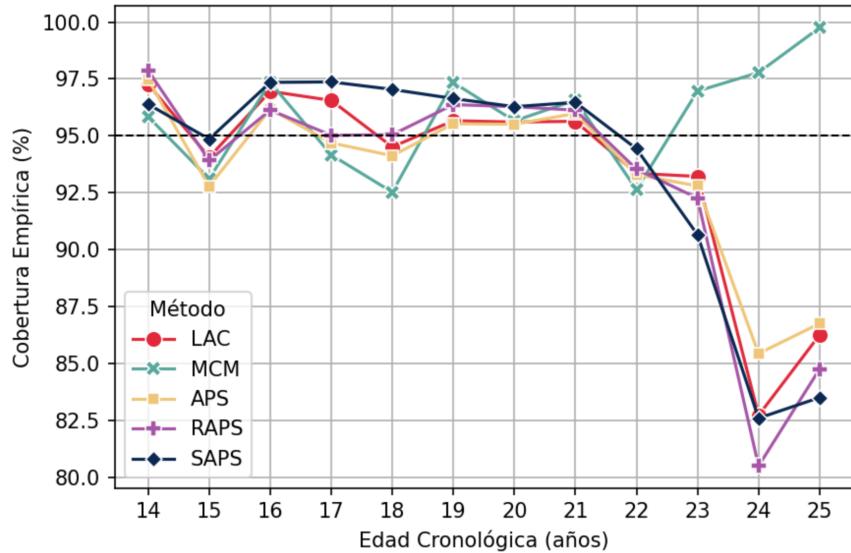
Entre los métodos, se identifican algunos patrones destacables:

- 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.

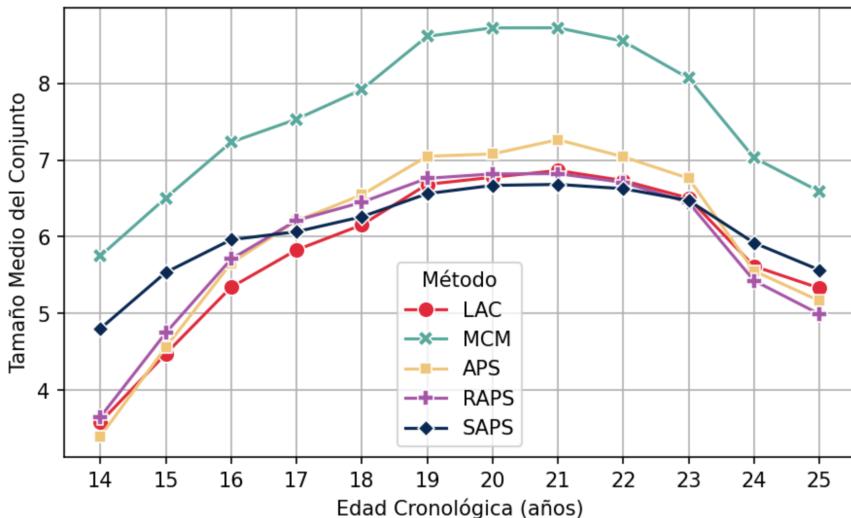
**2.6. Experimentación para la estimación de edad como problema de clasificación**

62

---



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



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

Figura 2.15: Gráficos de líneas comparativos de la cobertura empírica y el tamaño medio del conjunto de predicción por edad cronológica para los diferentes métodos evaluados.

# Bibliografía

- [1] 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. 6].
- [2] 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. 6].
- [3] 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. 7].
- [4] Y. Romano, E. Patterson y E. Candès, “Conformalized quantile regression,” *Advances in neural information processing systems*, vol. 32, 2019. [Citado en págs. 8, 10].
- [5] 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 pág. 8].
- [6] H. Papadopoulos, K. Proedrou, V. Vovk y A. Gammerman, “Inductive confidence machines for regression,” en *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, Springer, 2002, págs. 345-356. [Citado en pág. 9].
- [7] 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. 10].
- [8] M. Sadinle, J. Lei y L. Wasserman, “Least ambiguous set-valued classifiers with bounded error levels,” *Journal of the American Statistical Association*, vol. 114, n.º 525, págs. 223-234, 2019. [Citado en pág. 11].

- [9] U. Johansson, H. Linusson, T. Löfström y H. Boström, “Model-agnostic nonconformity functions for conformal classification,” en *2017 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2017, págs. 2072-2079. [Citado en pág. 11].
- [10] V. Vovk, D. Lindsay, I. Nouretdinov y A. Gammerman, “Mondrian confidence machine,” *Technical Report*, 2003. [Citado en pág. 13].
- [11] 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. 14, 15].
- [12] 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. 15, 16].
- [13] 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 pág. 17].
- [14] 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. 19].
- [15] 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. 21].
- [16] 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. 26].
- [17] 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. 26].
- [18] I. Loshchilov y F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017. [Citado en pág. 36].
- [19] 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. 36].
- [20] D. H. Ubelaker, “Forensic Anthropology: Methodology and Diversity of Applications,” en *Biological Anthropology of the Human Skeleton*. John Wiley & Sons, Ltd, 2018, cap. 2, págs. 43-71. [Citado en pág. 46].
- [21] L. Scheuer y S. Black, *The juvenile skeleton*, 1.<sup>a</sup> ed. Elsevier, 2004. [Citado en pág. 46].

- [22] 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.
- [23] 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.
- [24] 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.
- [25] S. Cordner y M. Tidball-Binz, “Humanitarian forensic action — Its origins and future,” *Forensic Science International*, vol. 279, págs. 65-71, 2017.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] 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.

- [32] 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.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] 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.
- [41] H. Lee, S. Tajmir, M. Zissen, B. Yesilwas, 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.
- [42] 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.

- [43] A. Ross y S. Williams, "Ancestry Studies in Forensic Anthropology: Back on the Frontier of Racism," *Biology*, vol. 10, n.<sup>o</sup> 7, pág. 602, 2021.
- [44] 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.<sup>o</sup> 4, págs. 672-683, 2021.
- [45] 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.<sup>o</sup> 3, págs. 208-214, 2014.
- [46] G. S. Cooper y V. Meterko, "Cognitive bias research in forensic science: A systematic review," *Forensic Science International*, vol. 297, págs. 35-46, 2019.
- [47] D. H. Ubelaker y C. M. DeGaglia, "Population variation in skeletal sexual dimorphism," *Forensic Science International*, vol. 278, 407.e1-407.e7, 2017.
- [48] S. Aja-Fernández, R. de Luis-García, M. Martín-Fernández y C. Alberola-López, "A computational TW3 classifier for skeletal maturity assessment. A Computing with Words approach," *Journal of Biomedical Informatics*, vol. 37, n.<sup>o</sup> 2, págs. 99-107, 2004.
- [49] D. Štern, C. Payer y M. Urschler, "Automated age estimation from MRI volumes of the hand," *Medical Image Analysis*, vol. 58, pág. 101538, 2019.
- [50] J. Venema, D. Peula, J. Irurita y P. Mesejo, "Employing deep learning for sex estimation of adult individuals using 2D images of the humerus," *Neural Comput & Applic*, vol. 35, págs. 5987-5998, 2022.
- [51] S. Park, S. Yang, J. Kim, J. Kang, J. Kim, K. Huh, S. Lee, W. Yi y M. Heo, "Automatic and robust estimation of sex and chronological age from panoramic radiographs using a multi-task deep learning network: a study on a South Korean population," *Int J Legal Med*, vol. 138, págs. 1741-1757, 2024.
- [52] K. Imaizumi, S. Usui, K. Taniguchi, Y. Ogawa, T. Nagata, K. Kaga, H. Hayakawa y S. Shiotani, "Development of an age estimation method for bones based on machine learning using post-mortem computed tomography images of bones," *Forensic Imaging*, vol. 26, pág. 200477, 2021.

- [53] 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.
- [54] A. Heinrich, “Accelerating computer vision-based human identification through the integration of deep learning-based age estimation from 2 to 89 years,” *Sci Rep*, vol. 14, pág. 4195, 2024.
- [55] L. Porto, L. Lima, A. Franco, D. Pianto, C. Machado y F. Vidal, “Estimating sex and age from a face: a forensic approach using machine learning based on photo-anthropometric indexes of the Brazilian population,” *International journal of legal medicine*, vol. 134(6), págs. 2239-2259, 2020.
- [56] 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.
- [57] 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.
- [58] 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.
- [59] D. D. Martin, D. Deusche, R. Schweizer, G. Binder, H. H. Thodberg y M. B. Ranke, “Clinical application of automated Greulich-Pyle bone age determination in children with short stature,” *Pediatric radiology*, vol. 39, págs. 598-607, 2009.
- [60] D. D. Martin, K. Meister, R. Schweizer, M. B. Ranke, H. H. Thodberg y G. Binder, “Validation of automatic bone age rating in children with precocious and early puberty,” 2011.
- [61] 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.
- [62] 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.

- [63] 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.
- [64] 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.
- [65] 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.
- [66] 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.
- [67] 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.
- [68] N. G. Rao, N. N. Rao, M. Pai y M. Shashidhar Kotian, “Mandibular canine index — A clue for establishing sex identity,” *Forensic Science International*, vol. 42, n.º 3, págs. 249-254, 1989.
- [69] 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.
- [70] J. E. Buikstra, “Standards for data collection from human skeletal remains,” *Arkansas archaeological survey research series*, vol. 44, pág. 44, 1994.
- [71] 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.
- [72] 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.
- [73] 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.

- [74] A. Schmeling, R. B. Dettmeyer, E. Rudolf, V. Vieth y G. Geseck, “Forensic Age Estimation,” *Deutsches Arzteblatt international*, vol. 113, n.º 4, págs. 44-50, 2016.
- [75] M. V. Tidball-Binz y S. M. Cordner, “Humanitarian forensic action: A new forensic discipline helping to implement international law and construct peace,” *WIREs Forensic Science*, 2021.
- [76] 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.
- [77] 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.
- [78] 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.
- [79] 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.
- [80] 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.
- [81] D. Stern, T. Ebner, H. Bischof, S. Grassegger, T. Ehamer y M. Urschler, “Fully automatic bone age estimation from left hand MR images,” en *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2014: 17th International Conference, Boston, MA, USA*, Springer, vol. 17(Pt II), 2014, págs. 220-227.
- [82] Ministerio del Interior de España, “Informe anual sobre personas desaparecidas 2025,” Ministerio del Interior, inf. téc., 2025.
- [83] 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](https://www.mpr.gob.es/servicios/publicaciones/Documents/Exhumaciones_Guerra_Civil_accesible_BAJA.pdf).

- [84] American Anthropological Association. “What is Anthropology?” Consultado el 01/04/2025, American Anthropological Association. URL: <https://americananthro.org/learn-teach/what-is-anthropology/>.
- [85] S. N. Byers y C. A. Juarez, *Introduction to Forensic Anthropology*, 6.<sup>a</sup> ed. Routledge, 2023.
- [86] T. Thompson y S. Black, *Forensic Human Identification: An Introduction*, 1.<sup>a</sup> ed. Taylor & Francis, 2006.
- [87] L. Scheuer y S. Black, *Developmental Juvenile Osteology*, 1.<sup>a</sup> ed. Academic Press, 2000.
- [88] J. Adserias-Garriga, *Age estimation: a multidisciplinary approach*. Academic Press, 2019.
- [89] 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/>.
- [90] 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](https://www.nist.gov/system/files/documents/2018/03/13/swganth_personal_identification.pdf).
- [91] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2024,” Fiscalía General del Estado, Madrid, España, inf. téc., 2024.
- [92] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2019,” Fiscalía General del Estado, Madrid, España, inf. téc., 2019.
- [93] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2016,” Fiscalía General del Estado, Madrid, España, inf. téc., 2016.
- [94] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2013,” Fiscalía General del Estado, Madrid, España, inf. téc., 2013.
- [95] A. Turing, “I.—COMPUTING MACHINERY and INTELLIGENCE,” *Mind*, vol. LIX, n.<sup>o</sup> 236, págs. 433-460, 1950.
- [96] A. L. Samuel, “Some Studies in Machine Learning Using the Game of Checkers,” *IBM Journal of Research and Development*, vol. 3, n.<sup>o</sup> 3, págs. 210-229, 1959.
- [97] F. Rosenblatt, “The perceptron: a probabilistic model for information storage and organization in the brain,” *Psychological Review*, vol. 65(6), págs. 386-408, 1958.

- [98] W. S. McCulloch y W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics*, vol. 5, n.<sup>o</sup> 4, págs. 115-133, 1943.
- [99] J. R. Quinlan, “Induction of decision trees,” *Machine learning*, vol. 1, págs. 81-106, 1986.
- [100] D. E. Rumelhart, G. E. Hinton y R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, págs. 533-536, 1986.
- [101] 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>.
- [102] 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.<sup>o</sup> 56, págs. 1929-1958, 2014.
- [103] 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>.
- [104] 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>.
- [105] L. Bottou, “Large-Scale Machine Learning with Stochastic Gradient Descent,” *Proc. of COMPSTAT’2010*, págs. 177-186, 2010.
- [106] S. Ioffe y C. Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, 2015. URL: <https://arxiv.org/abs/1502.03167>.
- [107] S. Santurkar, D. Tsipras, A. Ilyas y A. Madry, *How Does Batch Normalization Help Optimization?* 2019. URL: <https://arxiv.org/abs/1805.11604>.
- [108] S. Arora, Z. Li y K. Lyu, *Theoretical Analysis of Auto Rate-Tuning by Batch Normalization*, 2018. URL: <https://arxiv.org/abs/1812.03981>.
- [109] V. Nemanic, 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.

- [110] 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.
- [111] 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.
- [112] 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.
- [113] 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.
- [114] A. F. Psaros, X. Meng, Z. Zou, L. Guo y G. E. Karniadakis, “Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons,” *Journal of Computational Physics*, vol. 477, pág. 111 902, 2023.
- [115] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial intelligence*, vol. 267, págs. 1-38, 2019.
- [116] M. Salvi, S. Seoni, A. Campagner, A. Gertych, U. R. Acharya, F. Molinari y F. Cabitza, “Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare,” *International Journal of Medical Informatics*, vol. 197, pág. 105 846, 2025.
- [117] 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.
- [118] 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.
- [119] R. Foygel Barber, E. J. Candès, 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.

- [120] 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.
- [121] 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.
- [122] 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.
- [123] Y. LeCun, Y. Bengio y G. Hinton, “Deep Learning,” *Nature*, vol. 521, págs. 436-44, 2015.
- [124] 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.
- [125] 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.
- [126] P. Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. USA: Penguin Books Limited, 2015.
- [127] S. Russell y P. Norvig, *Artificial Intelligence: A Modern Approach*, 4rd. Prentice Hall Press, 2021.
- [128] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Heidelberg: Springer-Verlag, 2006.
- [129] E. Alpaydin, *Introduction to Machine Learning*, 2nd. The MIT Press, 2010.
- [130] P. J. Werbos, *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*. USA: Wiley-Interscience, 1994.
- [131] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.
- [132] R. Szeliski, *Computer Vision: Algorithms and Applications*, 1st. Berlin, Heidelberg: Springer-Verlag, 2010.
- [133] A. Zhang, Z. C. Lipton, M. Li y A. J. Smola, *Dive into Deep Learning*, 2021.
- [134] I. Goodfellow, Y. Bengio y A. Courville, *Deep Learning*. MIT Press, 2016.

- [135] V. Vovk, A. Gammerman y G. Shafer, *Algorithmic learning in a random world*. Springer, 2005, vol. 29.
- [136] Red Hat, *Deep learning*, Consultado el 10/05/2025, 2023. URL: <https://www.redhat.com/es/topics/ai/what-is-deep-learning>.
- [137] Code World, *Understanding ML & DL in python*, Consultado el 19/05/2025, 2022. URL: <https://codeworld.tistory.com/2>.
- [138] NVIDIA, *Convolutional Neural Network*, Consultado el 21/05/2025, 2025. URL: <https://www.nvidia.com/en-eu/glossary/convolutional-neural-network/>.
- [139] 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>.
- [140] J. G. Sam Lau y D. Nolan, *Cross Validation*, Consultado el 26/05/2025, 2023. URL: [https://learningds.org/ch/16/ms\\_cv.html](https://learningds.org/ch/16/ms_cv.html).
- [141] 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.
- [142] 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.
- [143] 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.
- [144] 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.
- [145] 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](https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf).

- [146] 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](https://www.bipm.org/documents/20126/2071204/JCGM_200_2012.pdf).
- [147] J. R. Berrendero. “Materiales del libro de Estadística,” visitado 2 de jun. de 2025. URL: <https://verso.mat.uam.es/~joser.berrendero/libro-est/>.
- [148] 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.
- [149] J. Gama, “A survey on learning from data streams: current and future trends,” *Progress in Artificial Intelligence*, vol. 1, págs. 45-55, 2012.
- [150] J. Vermorel. “Quantile Regression,” LOKAD Quantitive Supply Chain, visitado 2 de jun. de 2025. URL: <https://www.lokad.com/quantile-regression-time-series-definition/>.
- [151] R. Koenker, *Quantile Regression* (Econometric Society Monographs). Cambridge University Press, 2005.
- [152] S. T. Tokdar y J. B. Kadane, “Simultaneous linear quantile regression: a semiparametric Bayesian approach,” *Bayesian Analysis*, vol. 7, n.º 1, págs. 51-72, 2012.
- [153] J. Feldman y D. Kowal, “Bayesian Quantile Regression with Subset Selection: A Posterior Summarization Perspective,” *arXiv preprint arXiv:2311.02043*, 2023.
- [154] 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.
- [155] A. N. Angelopoulos y S. Bates, “A gentle introduction to conformal prediction and distribution-free uncertainty quantification,” *arXiv preprint arXiv:2107.07511*, 2021.
- [156] Scikit-learn-contrib MAPIE developers. “MAPIE: Model-Agnostic Prediction Interval Estimator.” Accessed: 2025-07-06. URL: <https://mapie.readthedocs.io/en/stable/>.
- [157] V. Vovk, “Cross-conformal predictors,” *Annals of Mathematics and Artificial Intelligence*, vol. 74, n.º 1, págs. 9-28, 2015.
- [158] D. Bethell, S. Gerasimou y R. Calinescu, “Robust uncertainty quantification using conformalised Monte Carlo prediction,” en *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, págs. 20 939-20 948.

- [159] 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.



