



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

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

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 doble titulación de Ingeniería Informática y Administración y Dirección de Empresas de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada**, con DNI 32071015E, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Grado en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: David González Durán

Granada, a 22 de agosto de 2025.

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

D. **Javier Vénema Rodríguez**, 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

En primer lugar, quiero agradecer a mis padres por todo el apoyo incondicional que me han brindado estos años, para facilitarme en la transición a la vida lejos de mi ciudad, por mantenerme económicamente, por motivarme con la carrera. Sin ellos,

Quiero agradecer a mi novio Juan por apoyarme en el último tramo de mis estudios,

También a mis amigos más cercanos estos últimos años, María y Javier,

También quiero dar las gracias a la Delegación de Estudiantes de Ingenierías Informática y Telecomunicaciones, que ...

Y, por último, gracias a Pablo y Javier por la ...

Índice general

1. Experimentación	1
1.1. Protocolo de validación experimental	1
1.2. Preprocesado de los datos	3
1.3. Esquema general de los experimentos realizados	3
1.3.1. Problema de estimación de edad	4
1.3.2. Problema de clasificación de mayoría de edad	6
1.3.3. Problema de clasificación de edad	8
1.3.4. Tests estadísticos	10
1.4. Experimentación para la estimación de edad	11
1.4.1. Entrenamiento de los modelos	11
1.4.2. Resultados	13
1.5. Experimentación para la clasificación de mayoría de edad . .	25
1.5.1. Entrenamiento de los modelos	25
1.5.2. Resultados	25
1.6. Experimentación para la clasificación de edad	31
1.6.1. Entrenamiento de los modelos	31
1.6.2. Resultados	31
2. Conclusiones y trabajos futuros	41
2.1. Conclusiones	41
2.2. Trabajos futuros	43
A. Problema de clasificación de sexo	61

B. Comparación de resultados de estimación de edad y clasificación de edad	63
C. Intervalos de valores razonables	65

Índice de figuras

1.1.	Diagrama de división del <i>dataset</i> en <i>train</i> , <i>validation</i> y <i>test</i>	2
1.2.	Diagrama de división del <i>dataset</i> en <i>train</i> , <i>validation</i> , <i>calibration</i> y <i>test</i>	2
1.3.	Esquema de experimentación para la estimación de edad.	5
1.4.	Esquema de experimentación para la clasificación de mayoría de edad.	7
1.5.	Esquema de experimentación para la clasificación de edad.	9
1.6.	Curva de aprendizaje de uno de los modelos para el método ICP.	13
1.7.	Gráfica de dispersión de la Cobertura empírica frente a la Amplitud media del intervalo de predicción.	17
1.8.	Histogramas del amplitud del intervalo de predicción con diferenciación por cobertura, correspondientes a los modelos QR y CQR.	20
1.9.	Mapa de calor de cobertura empírica en base a la amplitud del intervalo de predicción por cada método de predicción a lo largo de las distintas ejecuciones.	22
1.10.	Gráficos de líneas comparativos de la cobertura empírica y la amplitud media del intervalo de predicción por edad cronológica para los diferentes métodos evaluados.	24
1.11.	Gráfica de dispersión Cobertura empírica - Tamaño Medio de Conjunto de Predicción.	28
1.12.	Matrices de confusión conformal correspondientes a los métodos ‘base’, LAC y MCM.	30
1.13.	Cobertura empírica y tamaño medio del conjunto de predicción obtenidos por cada método de predicción a lo largo de las distintas ejecuciones.	32

1.14. Gráfica de dispersión Cobertura empírica - Tamaño Medio de Conjunto de Predicción.	33
1.15. Mapa de calor de cobertura empírica en base al tamaño del conjunto por cada método de predicción a lo largo de las distintas ejecuciones.	36
1.16. Gráficos de líneas comparativos de la cobertura empírica y el tamaño medio del conjunto de predicción por edad cronológica para los diferentes métodos evaluados.	39
C.1. Ejemplo de intervalo de confianza para la media poblacional.	66
C.2. Ejemplo de intervalo de credibilidad para la media poblacional.	67
C.3. Intervalos de predicción (95 % de confianza) construidos con CQR para estimación de edad.	68

Índice de tablas

1.1.	Error absoluto medio y error cuadrático medio obtenidos por cada método de predicción a lo largo de distintas ejecuciones.	14
1.2.	Resultados de la prueba <i>post-hoc</i> de Tukey HSD para MAE entre pares de métodos.	15
1.3.	Resultados de la prueba <i>post-hoc</i> de Tukey HSD para MSE entre pares de métodos.	15
1.4.	Cobertura empírica y amplitud media del intervalo de predicción obtenidos por cada método de predicción a lo largo de distintas ejecuciones.	16
1.5.	Resultados de la prueba <i>post-hoc</i> de Games-Howell para la amplitud media del intervalo de predicción entre pares de métodos.	18
1.6.	Resultados de las predicciones obtenidas por los modelos para el problema de estimación de edad en cada ejecución.	19
1.7.	Exactitud, sensibilidad y especificidad obtenidos por cada método de predicción a lo largo de distintas ejecuciones.	26
1.8.	Cobertura empírica y tamaño medio del conjunto de predicción obtenidos por cada método de predicción a lo largo de las distintas ejecuciones.	27
1.9.	Resultados de la prueba <i>post-hoc</i> de Tukey HSD para la cobertura empírica entre pares de métodos.	28
1.10.	Resultados de la prueba <i>post-hoc</i> de Tukey HSD para la cobertura empírica entre pares de métodos.	34
1.11.	Resultados de la prueba <i>post-hoc</i> de Tukey HSD para el tamaño medio del conjunto de predicción entre pares de métodos.	34
B.1.	Cobertura empírica y	64

Capítulo 1

Experimentación

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

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

En la Figura 1.1 podemos ver la división del *dataset* planteada. Cabe comentar que la división se ha realizado de forma estratificada en base a la edad y el sexo².

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

²La estratificación se realizó en intervalos de medio año de edad y por sexo; por ejemplo, una instancia con edad 17.7 y sexo masculino se etiquetó como “17.5_M”, o una de edad 18.2 y sexo femenino como “18.0_F”.

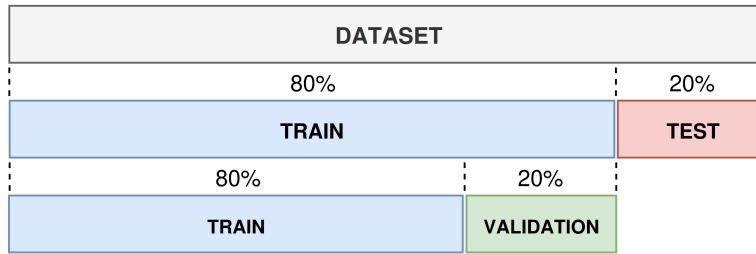


Figura 1.1: Diagrama de división del *dataset* en *train*, *validation* y *test*.

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 [1]. 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 [2] (que recomienda dedicar entre un 10 % y 30 % de datos de entrenamiento a calibración), tal y como se muestra en la Figura 1.2.

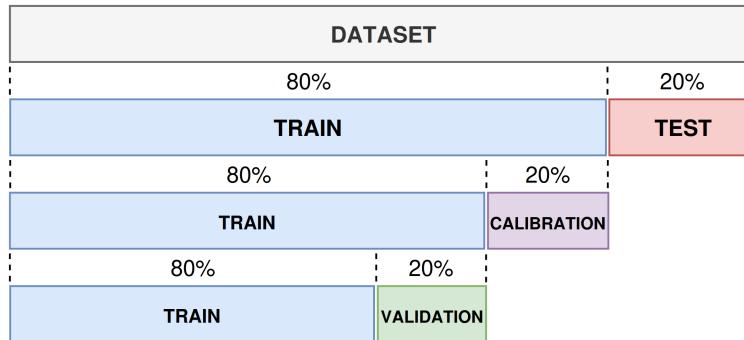


Figura 1.2: Diagrama de división del *dataset* en *train*, *validation*, *calibration* y *test*.

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.

1.2. 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 448×224 píxeles para homogenizar las entradas del modelo³. También se ha realizado *data augmentation* en el conjunto de entrenamiento, introduciendo transformaciones aleatorias en cada época para simular condiciones de posicionamiento del paciente y de la máquina o iluminación ligeramente variable:

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

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

1.3. 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

³El 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.

1.3.1. 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 ± 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.

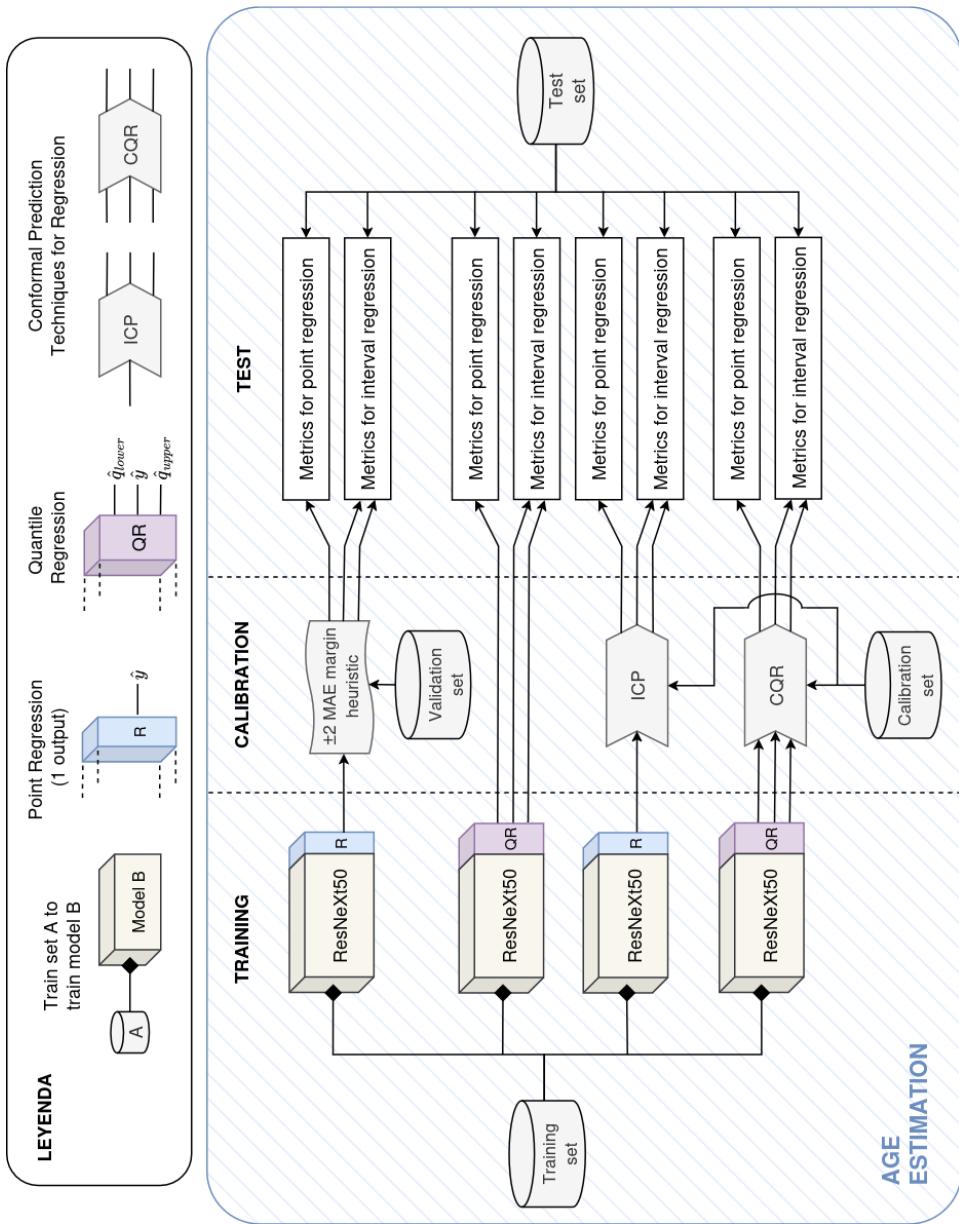


Figura 1.3: Esquema de experimentación para la estimación de edad. Cada modelo se entrena por separado. ‘R’ se refiere a ‘Rregresión puntual’ (de una sola neurona de salida), ‘QR’ a ‘Quantile Regression’ , ‘ICP’ a ‘Inductive Conformal Prediction’ y ‘CQR’ a Conformalized Quantile Regression.

1.3.2. 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.

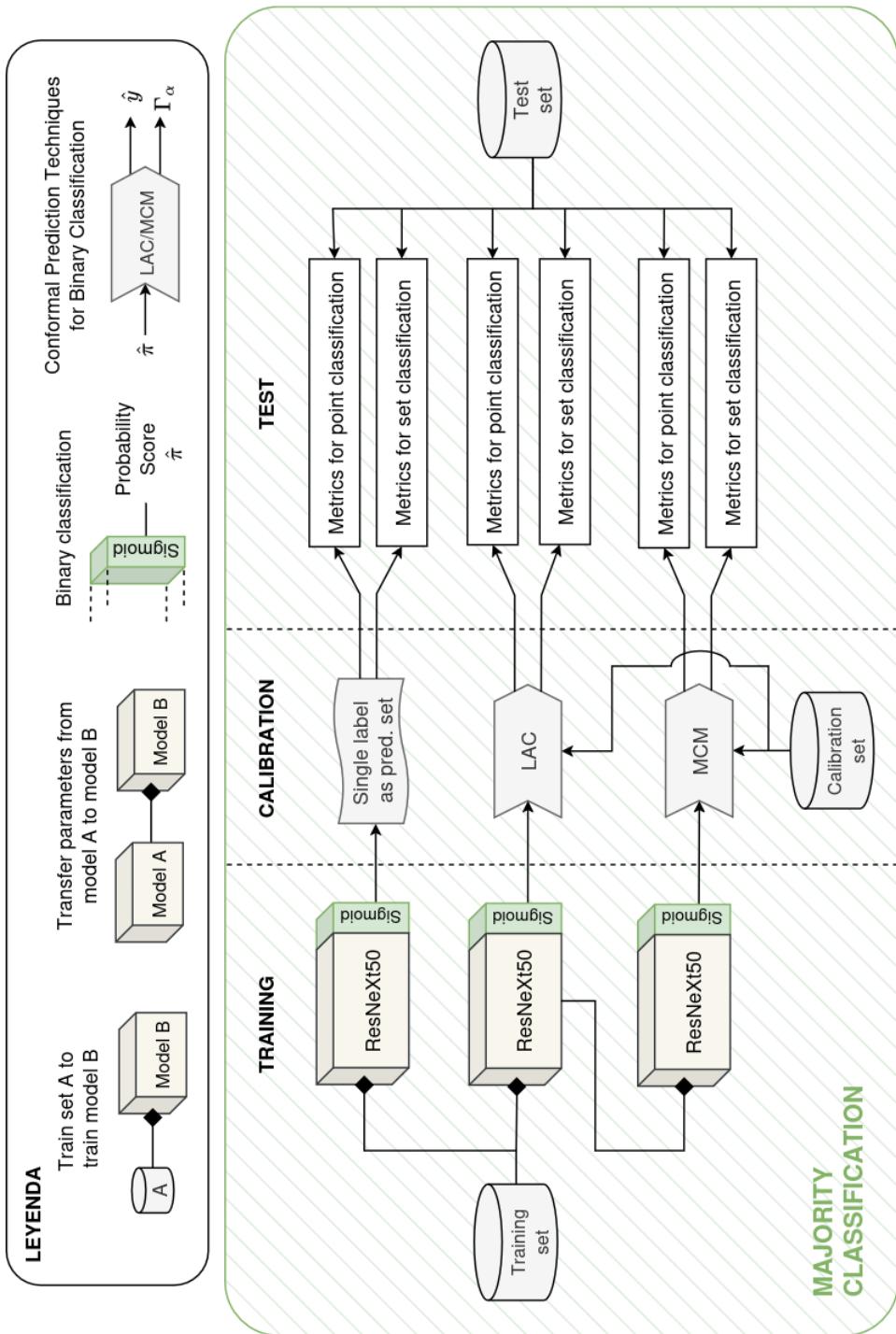


Figura 1.4: Esquema de experimentación para la clasificación de mayoría de edad.

1.3.3. Problema de clasificación de edad

Para el problema de clasificación de edad, se ha empleado la técnica de calibración de probabilidades *Temperature Scaling* para ajustar las salidas del modelo de clasificación multiclase, con el objetivo de mejorar la calidad de las probabilidades utilizadas durante la fase de inferencia conformal. Esta calibración probabilística se realiza antes del *softmax*. Se ha optado por utilizar el conjunto de validación para llevar a cabo dicha calibración de probabilidades, dado que, aunque no es el enfoque más riguroso —ya que lo ideal sería dividir el conjunto de calibración en dos subconjuntos independientes, uno para la calibración de probabilidades y otro para la calibración conformal— esta estrategia mostró buenos resultados en la práctica. Esto se debe a que el conjunto de validación empleado era suficientemente representativo y permitió obtener probabilidades calibradas de manera adecuada. Esta calibración probabilística no afecta a la variabilidad entre modelos con los mismos parámetros, dado que el algoritmo es determinista y produce resultados consistentes para un mismo conjunto de datos y parámetros.

Los métodos propuestos para este problema son:

- **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, con componente aleatoria para tamaños de conjunto de predicción más ajustados. El modelo será exactamente el mismo que el de LAC para este mismo problema.
- **Método ‘RAPS’:** Implementa la técnica RAPS para CP, también con componente aleatoria. El modelo será exactamente el mismo que el de LAC para este mismo problema.
- **Método ‘SAPS’:** Implementa la técnica SAPS para CP. Usará el mismo modelo que LAC.

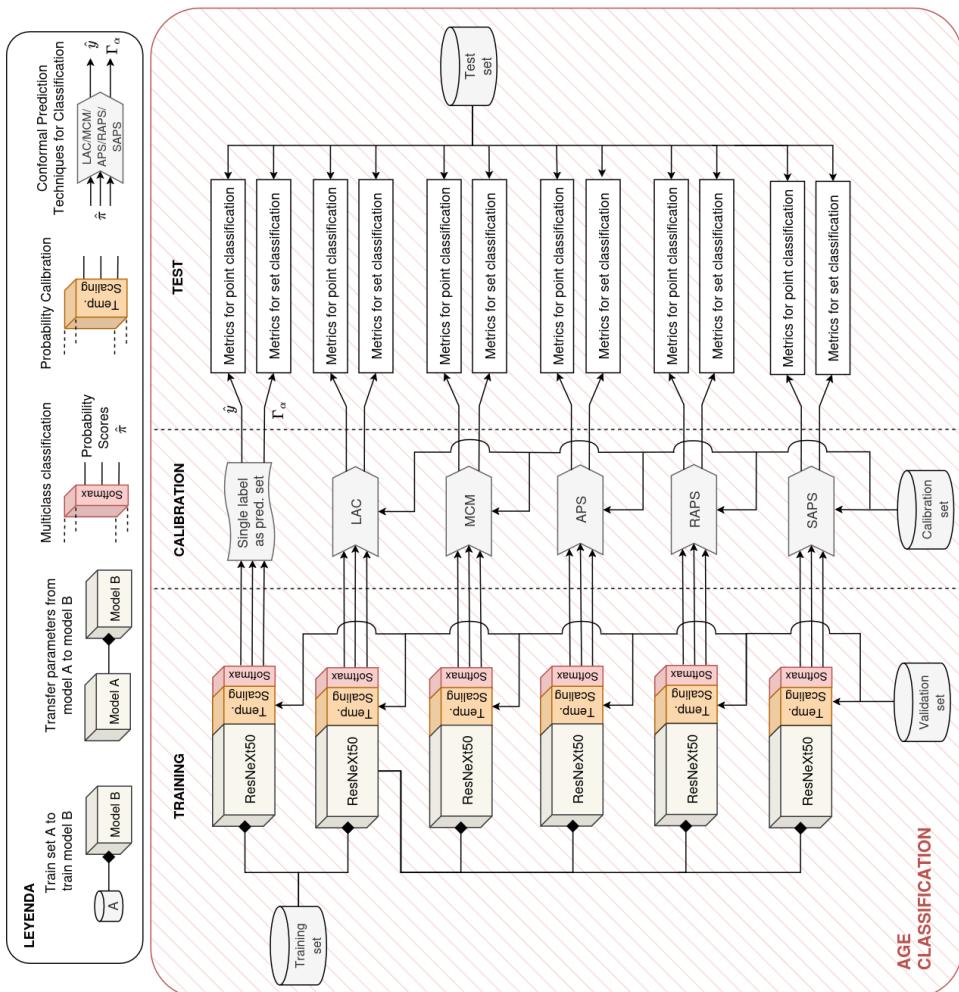


Figura 1.5: Esquema de experimentación para la clasificación de edad. Se recuerda que LAC, MCM, APS, RAPS y SAPS refieren a los métodos de CP para clasificación presentados en el anterior capítulo; y *Temp. Scaling* hace referencia a *Temperature Scaling*.

1.3.4. 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).

1.4. Experimentación para la estimación de edad

1.4.1. 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 [3]. 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* [4].

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 menor⁴. Esta estrategia permite acelerar la convergencia en las fases iniciales del entrenamiento y afinar los pesos. En las etapas finales, mejorando tanto la estabilidad como el rendimiento del modelo.

Esta combinación entre *learning rates* discriminativos y la política de un solo ciclo permite acelerar la convergencia en las primeras etapas del entre-

⁴Se han mantenido los parámetros por defecto del método OneCycle en PyTorch. Con esta configuración, cada grupo de capas comienza con una tasa de aprendizaje equivalente al 4 % del valor máximo asignado. Durante aproximadamente el 30 % inicial de las épocas, esta tasa crece de forma progresiva, y posteriormente decrece hasta alcanzar el 0.01 % del learning rate máximo.

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 1.6 se puede ver la curva de aprendizaje de uno de los modelos entrenados.

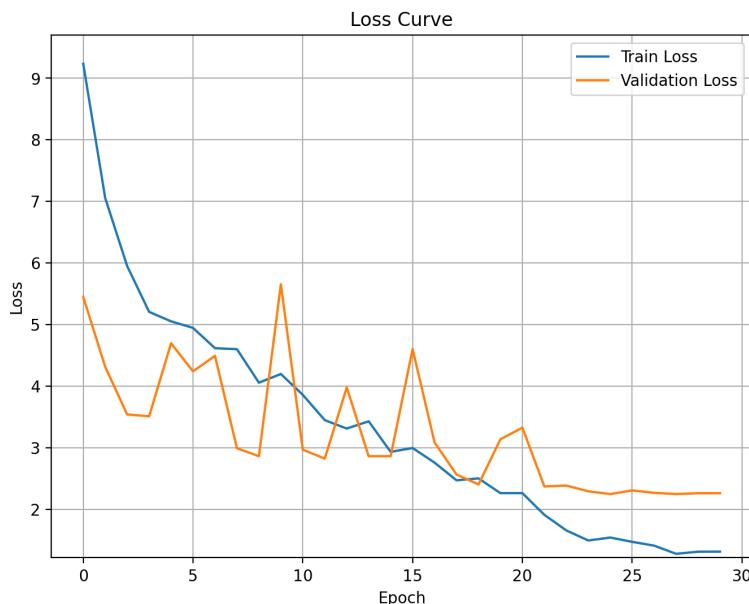


Figura 1.6: Curva de aprendizaje de uno de los modelos para el método ICP. En color azul se muestran las pérdidas obtenidas en el conjunto de entrenamiento, mientras que en color naranja se representan las correspondientes al conjunto de validación. Se observa una convergencia alrededor de la época 25.

1.4.2. Resultados

Análisis de métricas para la estimación puntual de edad

La Tabla 1.1 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 1.2 y 1.3). 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	1.16	1.20	1.16	1.18	2.36	2.48	2.39	2.44

Tabla 1.1: 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 1.2: 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 1.3: 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 ($\alpha = 0.05$).

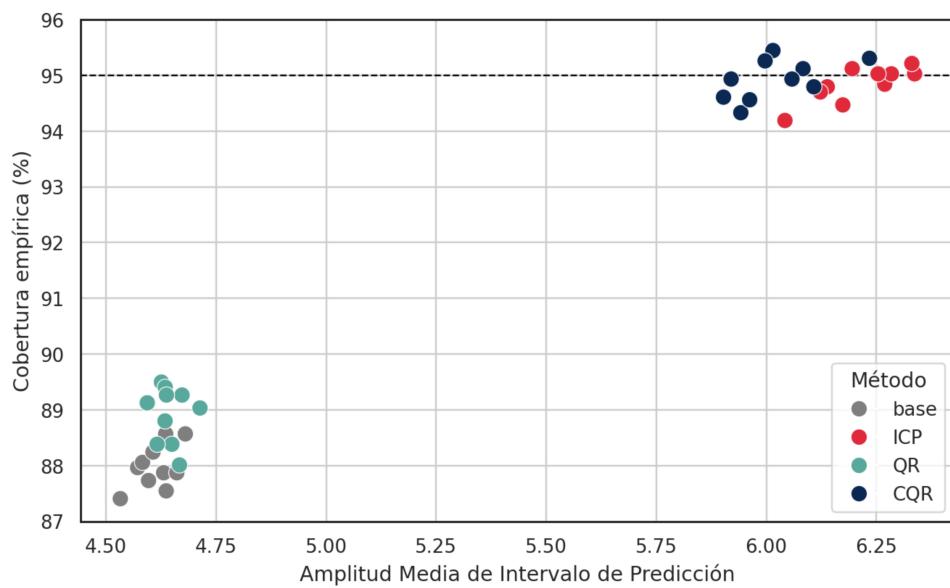
Análisis de métricas para la estimación interválica de edad

A continuación, la Tabla 1.4 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	94.84	88.92	94.93	4.61	6.21	4.64	6.02

Tabla 1.4: 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 1.7, donde se observa una alta correlación entre la cobertura empírica y el tamaño del intervalo de predicción.



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 ANOVA⁵, 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 1.5) 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 1.5: 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 1.6 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-

⁵Se 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.

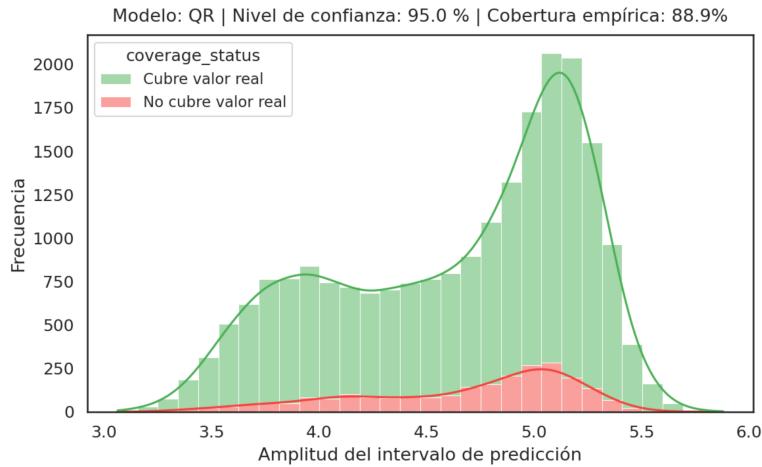
Ejecución	Mean Interval Score			
	base	ICP	QR	CQR
Ejecución 1	9.16	8.17	8.48	8.02
Ejecución 2	8.93	8.21	8.72	8.04
Ejecución 3	8.90	8.24	8.86	7.85
Ejecución 4	8.69	8.00	8.59	7.98
Ejecución 5	8.88	8.27	8.82	7.89
Ejecución 6	8.93	8.19	8.46	8.01
Ejecución 7	8.81	8.19	8.96	7.85
Ejecución 8	8.88	8.03	8.80	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	7.97

Tabla 1.6: 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.

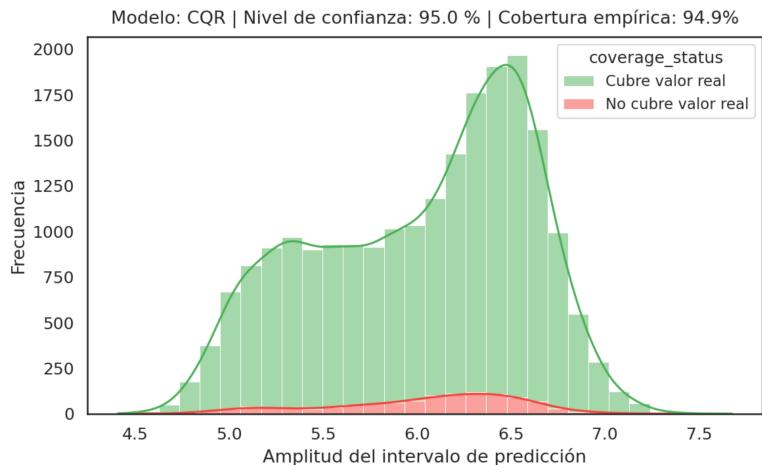
za, de forma que todos los intervalos lograrían cubrir al nivel de confianza deseado los valores reales.

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 1.8 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 1.8: 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 Figura 1.9.

En esta figura 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 93.14 %.

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 1.10 muestra la evolución de la cobertura empírica y el ancho medio de los intervalos de predicción en función de la edad cronológica⁶.

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

⁶Parte entera o suelo de la edad real.

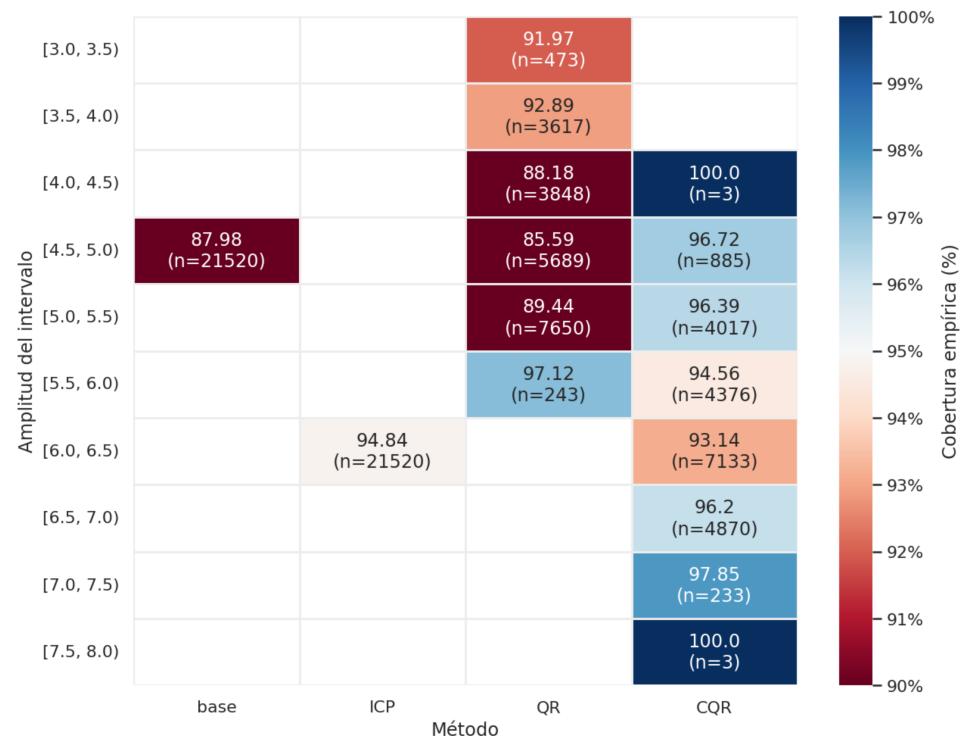


Figura 1.9: Mapa de calor de cobertura empírica en base a la amplitud del intervalo de predicción por cada método de predicción a lo largo de las distintas ejecuciones. Se especifica entre paréntesis el número de instancias clasificadas en cada franja de amplitud de intervalo. La escala de colores está centrada en la cobertura nominal (0.95): los valores por debajo de este umbral se representan en tonos rojos, los superiores en tonos azules, y el blanco indica una cobertura empírica equivalente a la nominal.

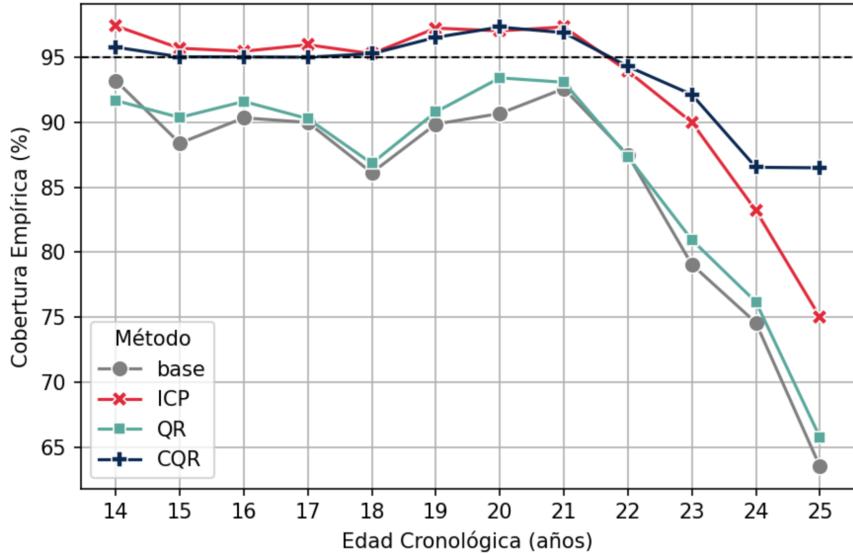
años, lo que coincide con la reducción en la cobertura predictiva. Esto causaría incertidumbre epistémica.

- Mayor variabilidad fisiológica en adultos jóvenes: A medida que aumenta la edad, los individuos suelen presentar una mayor diversidad en sus características biológicas debido a la acumulación de factores ambientales y estilos de vida [5, 6]. En este caso, esta incertidumbre sería estocástica, ya que es inherente al sistema.

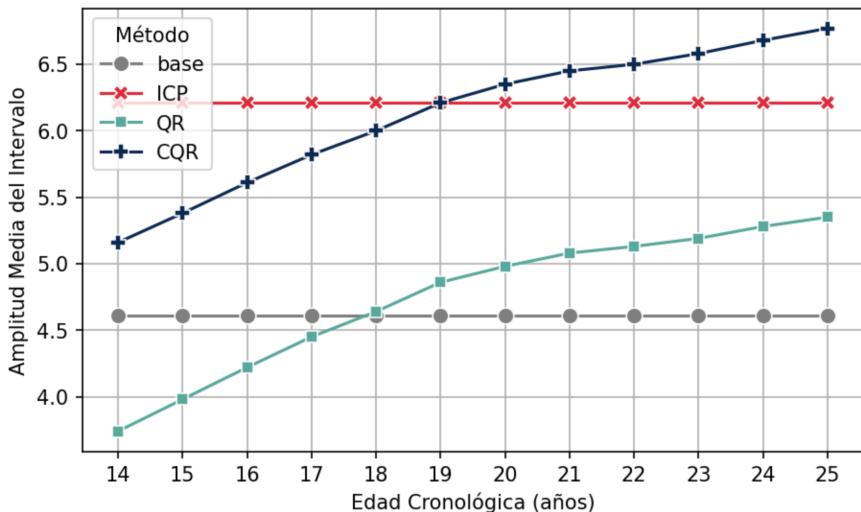
Discusión de resultados

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

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



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

1.5. Experimentación para la clasificación de mayoría de edad

1.5.1. 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.

1.5.2. Resultados

Análisis de métricas para la clasificación puntual de mayoría de edad

En la Tabla 1.7 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 conformales⁷, 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

⁷Comprobado 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.

26 1.5. Experimentación para la clasificación de mayoría de edad

el desequilibrio de clases. Como resultado, tienden a favorecer la clase mayoritaria (≥ 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	87.63	87.00	89.47	89.47	84.83	83.22

Tabla 1.7: 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 1.8 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 1.11.

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

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 1.8: 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.

28 1.5. Experimentación para la clasificación de mayoría de edad

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 1.9: 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.

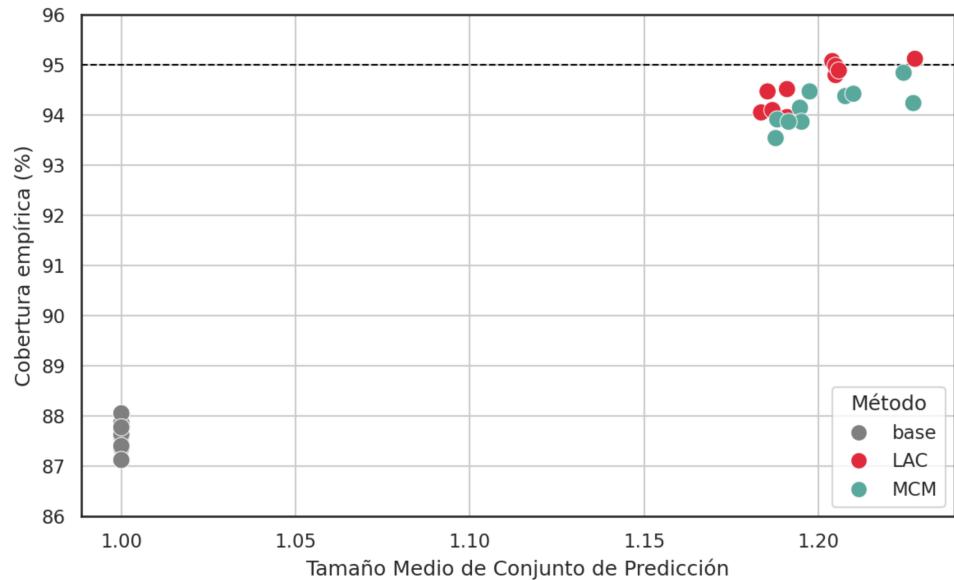


Figura 1.11: Gráfica de dispersión Cobertura empírica - Tamaño Medio de Conjunto de Predicción.

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 1.12 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.

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

Discusión de resultados

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

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

30 1.5. Experimentación para la clasificación de mayoría de edad

		Conjunto predicho			Cobertura
		{<18}	{≥18}	{<18,≥18}	
Etiqueta	<18	84.43	15.17	–	
	≥18	10.53	89.47	–	

(a) base

		Conjunto predicho			Cobertura
		{<18}	{≥18}	{<18,≥18}	
Etiqueta	<18	68.28	6.58	25.15	
	≥18	4.63	78.99	16.37	

(b) LAC

		Conjunto predicho			Cobertura
		{<18}	{≥18}	{<18,≥18}	
Etiqueta	<18	76.86	3.77	19.37	
	≥18	7.18	72.00	20.82	

(c) MCM

Figura 1.12: 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.

1.6. Experimentación para la clasificación de edad

1.6.1. 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 extracto 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 multiclas mutuamente excluyente. Esta función compara la distribución de probabilidad predicha por el modelo con la distribución real codificada como etiqueta única, y penaliza fuertemente las asignaciones erróneas. Su formulación es robusta, ampliamente utilizada y permite una interpretación probabilística directa de la salida del modelo cuando se combina con una capa de activación *softmax* al final.

1.6.2. 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 1.13 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.

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 1.13: 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.

Para facilitar la interpretación, en la Figura 1.14 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.

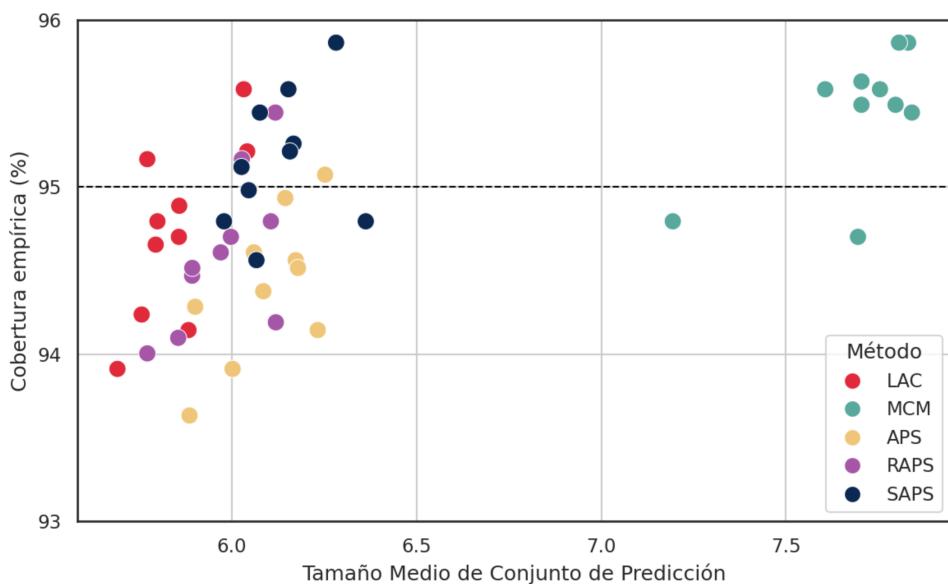


Figura 1.14: Gráfica de dispersión Cobertura empírica - Tamaño Medio de Conjunto de Predicción. No se incluyen los puntos del método ‘base’ dado que estos

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 1.10 y 1.11.

Modelo 1	Modelo 2	Dif. media	Valor <i>p</i>	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 1.10: 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 ($\alpha = 0.05$).

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 1.11: 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 ($\alpha = 0.05$).

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 %, 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 1.15 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.

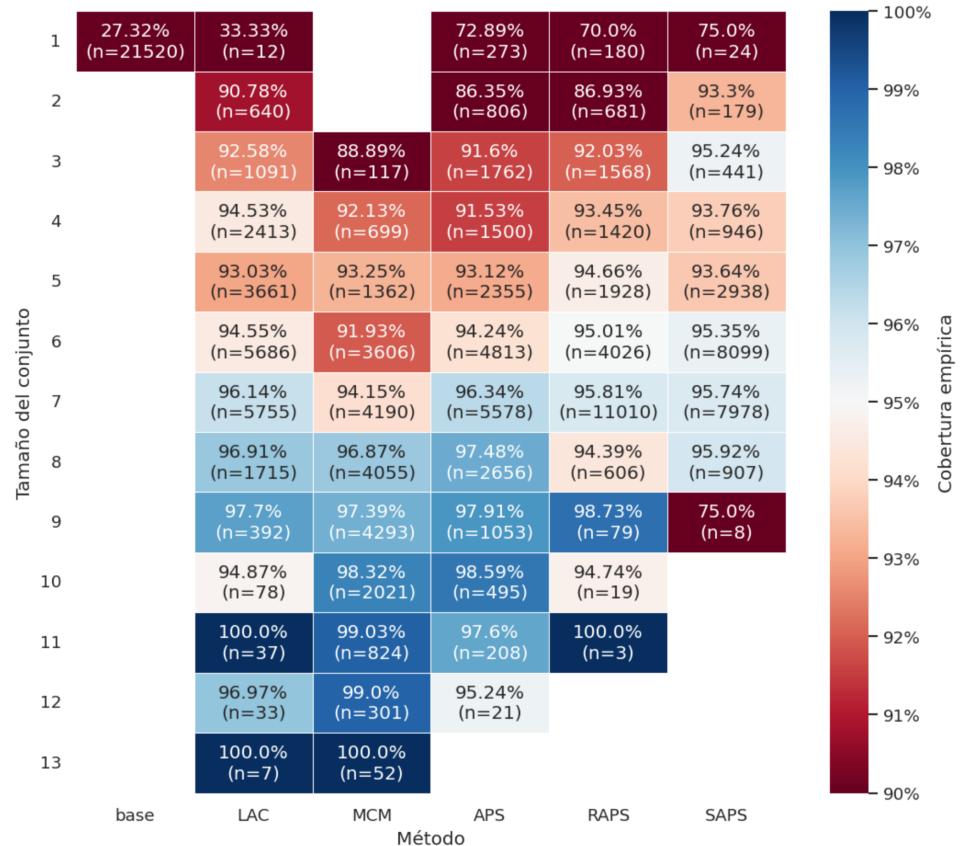


Figura 1.15: Mapa de calor de cobertura empírica en base al tamaño del conjunto por cada método de predicción a lo largo de las distintas ejecuciones. Se especifica entre paréntesis el número de instancias clasificadas en cada franja de amplitud de intervalo. La escala de colores está centrada en la cobertura nominal (0.95): los valores por debajo de este umbral se representan en tonos rojos, los superiores en tonos azules, y el blanco indica una cobertura empírica equivalente a la nominal.

- **Sobre cobertura como síntoma de desequilibrio:** la presencia de sobre cobertura en determinados tamaños implica, inevitablemente, infracobertura en otros. Cuando este patrón se repite y la sobre cobertura se concentra en conjuntos de gran tamaño, suele indicar que el método está “compensando” un mal ajuste en los conjuntos pequeños, lo cual resulta indeseable. En contextos prácticos, esto significa sacrificar precisión en situaciones de alta confianza para inflar artificialmente los resultados en escenarios menos exigentes.

Y ahora, centrándonos en los métodos:

- **MCM:** genera conjuntos de predicción muy conservadores, con un gran número de etiquetas. Presenta infracobertura para instancias cuyo conjunto de predicción contiene entre 3 y 7 etiquetas, y sobre cobertura 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 sobre cobertura 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 sobre cobertura 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 sobrecobertura 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 1.16 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.

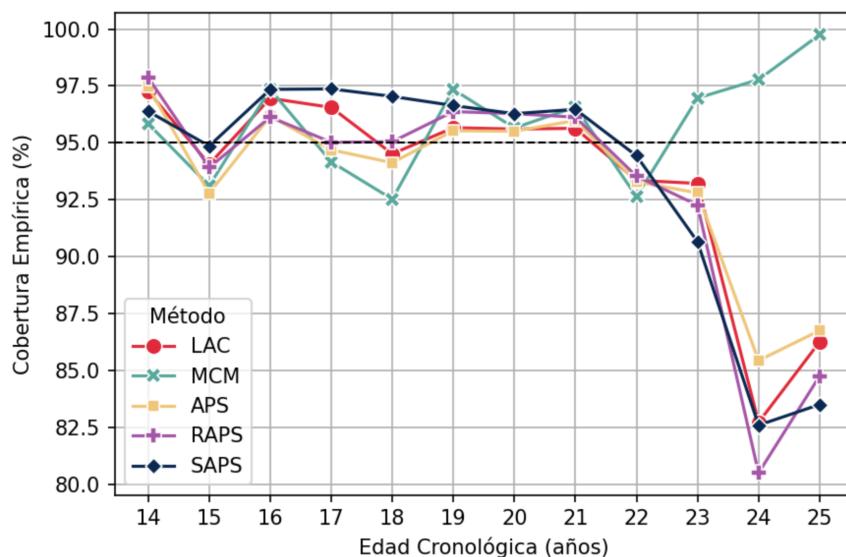
Entre los métodos, se identifican algunos patrones destacables:

- **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, alcanzando coberturas muy cercanas al 95 %, y en la mayoría de casos superándolo.

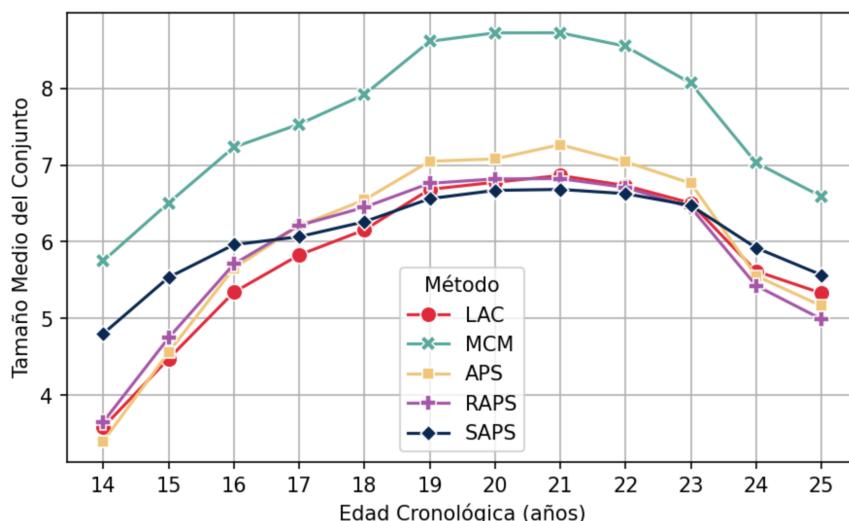
El resto de métodos adaptativos y el método LAC presentan tamaño medio de conjunto muy variables para las distintas edades cronológicas, Su cobertura fluctúa alrededor de SAPS, si bien para la mayoría de edades ligeramente por debajo.

Discusión de resultados

En definitiva, para este problema, LAC y SAPS se perfilan como los métodos más equilibrados, ya que ambos se aproximan a la cobertura nomi-



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

nal y mantienen una adecuada relación entre cobertura y tamaño medio de los conjuntos.

LAC sobresale por su sencillez de implementación y por generar conjuntos de tamaño moderado, que si bien alcanzan una cobertura ligeramente inferior a la nominal, resultan muy eficientes en términos prácticos.

SAPS, por su parte, se caracteriza por producir conjuntos de predicción con tamaños poco variables, ofreciendo una mayor consistencia. Además, presenta una mayor adaptatividad respecto al tamaño, con tasas de cobertura más estables a lo largo de los diferentes tamaños del conjunto.

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

Capítulo 2

Conclusiones y trabajos futuros

2.1. Conclusiones

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

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

En términos prácticos, **la mayoría de las técnicas de predicción conformal presentan la ventaja de no requerir un reentrenamiento completo del modelo**, siempre que se disponga de suficientes ejemplos adicionales para la calibración, distintos de los empleados en el entrenamiento o la validación. En caso de no contar con este volumen de datos, resulta necesario reentrenar el modelo tras realizar una nueva partición del conjunto de datos que reserve un subconjunto específico para la calibración. Existen,

En este apartado no he metido abreviaciones, por si alguien lee las conclusiones directamente tras la introducción, pero no creo que sea la mejor manera de hacerlo. ¿Tal vez debería añadir una hoja con abreviaciones al final del trabajo?

He tratado de introducir el mayor número posible de referencias al problema de estimación del perfil biológico aquí, aunque siento que tal vez puede quedar un poco pobre, pero me estaba quedando muy largo el apartado.

no obstante, métodos que sí implican modificaciones en el entrenamiento, como la *Conformalized Quantile Regression*. Este enfoque requiere incorporar nuevas salidas —de la *Quantile Regression*— y entrenar nuevamente el modelo bajo una función de pérdida adaptada. Sin embargo, cabe destacar que este proceso de ajuste resulta relativamente poco costoso, dado que el modelo suele converger en pocas épocas.

La principal contribución de la predicción conformal reside en **proporcionar una medida rigurosa de incertidumbre a través del conjunto predicho**. Algunas técnicas miden la incertidumbre del conjunto completo, de modo que todos los ejemplos reciben conjuntos de predicción del mismo tamaño. En estos casos, la incertidumbre no se refleja en la variabilidad del tamaño del conjunto, sino en la frecuencia con que dicho conjunto contiene o no el valor o clase verdadero. Así, los tamaños de los conjuntos conformales son constantes, lo que no deja de ser una aproximación muy cercana al análisis del error tradicional: se garantiza que, en promedio, el error se mantenga bajo un umbral prefijado. Sin embargo, **los métodos conformales adaptativos entrañan un mayor potencial, ya que ajustan dinámicamente el tamaño del conjunto de predicción en función de la dificultad de predecir cada instancia**. De este modo, ejemplos en los que el modelo está más seguro tienden a recibir conjuntos más pequeños, mientras que en aquellos en los que la predicción es más incierta, los conjuntos se amplían para mantener la garantía de cobertura. Esta adaptatividad permite capturar mejor tanto la incertidumbre epistémica (p.ej., los intervalos de edad más amplios por escasez de datos en individuos de edad avanzada) como la estocástica (p.ej., los intervalos amplios por la mayor variabilidad fisiológica en edades avanzadas), reflejando así de manera más fiel la heterogeneidad de los datos y proporcionando información más útil para la toma de decisiones.

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

- **Reserva de datos para calibración:** destinar una fracción del conjunto de entrenamiento puede degradar el rendimiento del modelo. No obstante, en el caso analizado, el volumen de datos fue suficiente para que la retención del 20% apenas afectara los resultados. En contextos con conjuntos de datos reducidos, este aspecto puede volverse más problemático, por lo que resultaría recomendable explorar estrategias alternativas de predicción conformal como *Jackknife+*.
- **Incorporación del proceso de calibración e inferencia conformal:** este paso introduce un coste adicional, pero comparado con los tiempos de entrenamiento e inferencia de un modelo de ML —especialmente en redes profundas—, suele ser marginal.

Cabe destacar que, si bien todos los métodos conformales garantizan teóricamente la cobertura marginal, en la práctica exhiben diferencias sustanciales. **Algunas variantes tienden a generar intervalos demasiado conservadores (sobrecubriendo en ciertos grupos) mientras que otras ofrecen un mejor equilibrio**, reduciendo la sobrecobertura en favor de una cobertura más homogénea entre subpoblaciones.

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

En definitiva, se puede sugerir que los métodos de predicción conformal emergen como una herramienta prometedora para enriquecer los análisis de estimación del perfil biológico en antropología forense. Su principal valor reside en su capacidad para proporcionar una cuantificación rigurosa y probabilística de la incertidumbre inherente a la predicción de variables como la edad o el sexo, sin requerir, en la mayoría de los casos, de costes computacionales prohibitivos.

2.2. Trabajos futuros

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

- **Integración con otros paradigmas de cuantificación de la incertidumbre:** La predicción conformal puede combinarse con métodos como *Monte Carlo Dropout* (como en [7]) o los *ensembles* de modelos para generar intervalos conformales que no solo garantizan una cobertura marginal, sino que también se benefician de una estimación de incertidumbre más afinada.
- **Aprovechamiento de información experta del dominio:** El marco conformal es agnóstico al modelo, pero no al problema. Puede ser adaptado para incorporar conocimiento experto y restricciones propias del ámbito de aplicación (p.ej., correlaciones biológicas conocidas en la

estimación de la edad). Se podría añadir, por tanto, un postprocesado para refinar los conjuntos de predicciones conformales, asegurando que no solo sean estadísticamente válidos, sino también biológica y contextualmente plausibles.

- **Combinación con técnicas de detección de datos fuera de distribución (*Out-of-Distribution*, OOD) y explicabilidad (*Explainable Artificial Intelligence*, XAI):** La unión de estas áreas es fundamental para construir sistemas confiables. La detección de OOD es relevante dado que la suposición más importante realizada por la predicción conformal es que los datos son intercambiables y, por tanto, pertenecientes a una misma distribución subyacente. Cuando esta premisa fundamental se viola, las garantías de cobertura estadística dejan de ser válidas. Este mecanismo actuaría como un sistema de alerta temprana, identificando ejemplos para los cuales las predicciones, y sus intervalos de incertidumbre asociados, deben ser tratados con precaución.

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

Bibliografía

- [1] 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. 2].
- [2] 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. 2].
- [3] I. Loshchilov y F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017. [Citado en pág. 12].
- [4] 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. 12].
- [5] 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. 23].
- [6] L. Scheuer y S. Black, *The juvenile skeleton*, 1.ª ed. Elsevier, 2004. [Citado en pág. 23].
- [7] D. Bethell, S. Gerasimou y R. Calinescu, “Robust uncertainty quantification using conformalised Monte Carlo prediction,” en *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, págs. 20939-20948. [Citado en pág. 43].
- [8] J. A. Sanchis-Gimeno, J. Iglesias-Bexiga, M. E. Schwab, G. López-García, E. Ariza, A. Calpe, M. Mezquida, S. Nalla e I. Ercan, “Identification success rates in the post-Spanish Civil War mass graves located in the cemetery of Paterna, Spain: Meta-research on 15 mass graves with 933 subjects,” *Forensic Science International*, vol. 361, págs. 112-122, ago. de 2024.
- [9] M. Baeta, C. Núñez, S. Cardoso, L. Palencia-Madrid, L. Herrasti, F. Etxeberria y M. M. de Pancorbo, “Digging up the recent Spanish memory: genetic identification of human remains from mass graves of the Spanish Civil War and posterior dictatorship,” *Forensic Science International: Genetics*, vol. 19, págs. 272-279, 2015.

- [10] V. Ataliva, N. F. Bahamondes, C. M. Suárez y B. Rosignoli, “Arqueología Forense y prácticas genocidas del Cono Sur americano: reflexionando desde los confines,” *Revista de Arqueología Americana*, vol. 41, págs. 403-441, jun. de 2024.
- [11] S. Cordner y M. Tidball-Binz, “Humanitarian forensic action — Its origins and future,” *Forensic Science International*, vol. 279, págs. 65-71, 2017.
- [12] 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.
- [13] D. Higgins, A. B. Rohrlach, J. Kaidonis, G. Townsend y J. J. Austin, “Differential Nuclear and Mitochondrial DNA Preservation in Post-Mortem Teeth with Implications for Forensic and Ancient DNA Studies,” *PLoS One*, vol. 10, n.º 5, págs. 1-17, 2015.
- [14] K. E. Latham y J. J. Miller, “DNA Recovery and Analysis from Skeletal Material in Modern Forensic Contexts,” *Forensic Sciences Research*, vol. 4, n.º 1, págs. 51-59, 2018.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] 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.

- [21] 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.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] H. Lee, S. Tajmir, M. Zissen, B. Yesiwas, 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.
- [28] 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.
- [29] A. Ross y S. Williams, "Ancestry Studies in Forensic Anthropology: Back on the Frontier of Racism," *Biology*, vol. 10, n.º 7, pág. 602, 2021.
- [30] A. Ross y M. Pilloud, "The need to incorporate human variation and evolutionary theory in forensic anthropology: A call for reform," *American Journal of Physical Anthropology*, vol. 176, n.º 4, págs. 672-683, 2021.
- [31] S. Nakhaeizadeh, I. E. Dror y R. M. Morgan, "Cognitive bias in forensic anthropology: Visual assessment of skeletal remains is susceptible to confirmation bias," *Science & Justice*, vol. 54, n.º 3, págs. 208-214, 2014.

- [32] G. S. Cooper y V. Meterko, “Cognitive bias research in forensic science: A systematic review,” *Forensic Science International*, vol. 297, págs. 35-46, 2019.
- [33] D. H. Ubelaker y C. M. DeGaglia, “Population variation in skeletal sexual dimorphism,” *Forensic Science International*, vol. 278, 407.e1-407.e7, 2017.
- [34] S. Aja-Fernández, R. de Luis-García, M. Martín-Fernández y C. Alberola-López, “A computational TW3 classifier for skeletal maturity assessment. A Computing with Words approach,” *Journal of Biomedical Informatics*, vol. 37, n.º 2, págs. 99-107, 2004.
- [35] D. Štern, C. Payer y M. Urschler, “Automated age estimation from MRI volumes of the hand,” *Medical Image Analysis*, vol. 58, pág. 101 538, 2019.
- [36] 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.
- [37] 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.
- [38] K. Imaizumi, S. Usui, K. Taniguchi, Y. Ogawa, T. Nagata, K. Kaga, H. Hayakawa y S. Shiotani, “Development of an age estimation method for bones based on machine learning using post-mortem computed tomography images of bones,” *Forensic Imaging*, vol. 26, pág. 200 477, 2021.
- [39] 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.
- [40] 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.
- [41] 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.

- [42] 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.
- [43] 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.
- [44] 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.
- [45] 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.
- [46] 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.
- [47] 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.
- [48] 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.
- [49] 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.
- [50] 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.
- [51] 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.
- [52] 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.

- [53] 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.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] J. E. Buikstra, “Standards for data collection from human skeletal remains,” *Arkansas archaeological survey research series*, vol. 44, pág. 44, 1994.
- [58] H. H. de Boer, S. Blau, T. Delabarde 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.
- [59] 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.
- [60] 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.
- [61] 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.
- [62] 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.
- [63] 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.

- [64] 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.
- [65] 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.
- [66] 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.
- [67] 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.
- [68] 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.
- [69] Ministerio del Interior de España, “Informe anual sobre personas desaparecidas 2025,” Ministerio del Interior, inf. téc., 2025.
- [70] 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.
- [71] American Anthropological Association. “What is Anthropology?” Consultado el 01/04/2025, American Anthropological Association. URL: <https://americananthro.org/learn-teach/what-is-anthropology/>.
- [72] S. N. Byers y C. A. Juarez, *Introduction to Forensic Anthropology*, 6.^a ed. Routledge, 2023.
- [73] T. Thompson y S. Black, *Forensic Human Identification: An Introduction*, 1.^a ed. Taylor & Francis, 2006.
- [74] L. Scheuer y S. Black, *Developmental Juvenile Osteology*, 1.^a ed. Academic Press, 2000.
- [75] J. Adserias-Garriga, *Age estimation: a multidisciplinary approach*. Academic Press, 2019.

- [76] 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/>.
- [77] 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.
- [78] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2024,” Fiscalía General del Estado, Madrid, España, inf. téc., 2024.
- [79] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2019,” Fiscalía General del Estado, Madrid, España, inf. téc., 2019.
- [80] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2016,” Fiscalía General del Estado, Madrid, España, inf. téc., 2016.
- [81] Fiscalía General del Estado, “Memoria de la Fiscalía General del Estado 2013,” Fiscalía General del Estado, Madrid, España, inf. téc., 2013.
- [82] A. Turing, “I.—COMPUTING MACHINERY and INTELLIGENCE,” *Mind*, vol. LIX, n.º 236, págs. 433-460, 1950.
- [83] A. L. Samuel, “Some Studies in Machine Learning Using the Game of Checkers,” *IBM Journal of Research and Development*, vol. 3, n.º 3, págs. 210-229, 1959.
- [84] F. Rosenblatt, “The perceptron: a probabilistic model for information storage and organization in the brain,” *Psychological Review*, vol. 65(6), págs. 386-408, 1958.
- [85] W. S. McCulloch y W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics*, vol. 5, n.º 4, págs. 115-133, 1943.
- [86] J. R. Quinlan, “Induction of decision trees,” *Machine learning*, vol. 1, págs. 81-106, 1986.
- [87] D. E. Rumelhart, G. E. Hinton y R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, págs. 533-536, 1986.
- [88] 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>.

- [89] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever y R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research*, vol. 15, n.º 56, págs. 1929-1958, 2014.
- [90] J. Tompson, R. Goroshin, A. Jain, Y. LeCun y C. Bregler, *Efficient Object Localization Using Convolutional Networks*, 2015. URL: <http://arxiv.org/abs/1411.4280>.
- [91] 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>.
- [92] L. Bottou, “Large-Scale Machine Learning with Stochastic Gradient Descent,” *Proc. of COMPSTAT’2010*, págs. 177-186, 2010.
- [93] S. Ioffe y C. Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, 2015. URL: <https://arxiv.org/abs/1502.03167>.
- [94] S. Santurkar, D. Tsipras, A. Ilyas y A. Madry, *How Does Batch Normalization Help Optimization?* 2019. URL: <https://arxiv.org/abs/1805.11604>.
- [95] S. Arora, Z. Li y K. Lyu, *Theoretical Analysis of Auto Rate-Tuning by Batch Normalization*, 2018. URL: <https://arxiv.org/abs/1812.03981>.
- [96] 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.
- [97] 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. 110796, 2023.
- [98] 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.
- [99] 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.
- [100] 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. 101805, 2023.

- [101] 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.
- [102] 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.
- [103] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial intelligence*, vol. 267, págs. 1-38, 2019.
- [104] 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.
- [105] 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.
- [106] 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.
- [107] R. Foygel Barber, E. J. Candes, A. Ramdas y R. J. Tibshirani, “The limits of distribution-free conditional predictive inference,” *Information and Inference: A Journal of the IMA*, vol. 10, n.º 2, págs. 455-482, 2021.
- [108] 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.
- [109] 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.
- [110] 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.
- [111] 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.

- [112] 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.
- [113] 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.
- [114] 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.
- [115] J. Postels, M. Segu, T. Sun, L. Sieber, L. Van Gool, F. Yu y F. Tombari, “On the practicality of deterministic epistemic uncertainty,” *arXiv preprint arXiv:2107.00649*, 2021.
- [116] Y. Gal y Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” en *international conference on machine learning*, PMLR, 2016, págs. 1050-1059.
- [117] D. Opitz y R. Maclin, “Popular ensemble methods: An empirical study,” *Journal of artificial intelligence research*, vol. 11, págs. 169-198, 1999.
- [118] Y. LeCun, Y. Bengio y G. Hinton, “Deep Learning,” *Nature*, vol. 521, págs. 436-44, 2015.
- [119] 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.
- [120] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari y U. R. Acharya, “Application of explainable artificial intelligence for health-care: A systematic review of the last decade (2011–2022),” *Computer methods and programs in biomedicine*, vol. 226, pág. 107161, 2022.
- [121] P. Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. USA: Penguin Books Limited, 2015.
- [122] S. Russell y P. Norvig, *Artificial Intelligence: A Modern Approach*, 4rd. Prentice Hall Press, 2021.
- [123] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Heidelberg: Springer-Verlag, 2006.
- [124] E. Alpaydin, *Introduction to Machine Learning*, 2nd. The MIT Press, 2010.
- [125] P. J. Werbos, *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*. USA: Wiley-Interscience, 1994.

- [126] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. [Citado en págs. 65, 67].
- [127] R. Szeliski, *Computer Vision: Algorithms and Applications*, 1st. Berlin, Heidelberg: Springer-Verlag, 2010.
- [128] R. M. Neal, *Bayesian learning for neural networks*. Springer Science & Business Media, 2012, vol. 118.
- [129] A. Zhang, Z. C. Lipton, M. Li y A. J. Smola, *Dive into Deep Learning*, 2021.
- [130] I. Goodfellow, Y. Bengio y A. Courville, *Deep Learning*. MIT Press, 2016.
- [131] V. Vovk, A. Gammerman y G. Shafer, *Algorithmic learning in a random world*. Springer, 2005, vol. 29.
- [132] Red Hat, *Deep learning*, Consultado el 10/05/2025, 2023. URL: <https://www.redhat.com/es/topics/ai/what-is-deep-learning>.
- [133] Code World, *Understanding ML & DL in python*, Consultado el 19/05/2025, 2022. URL: <https://codeworld.tistory.com/2>.
- [134] NVIDIA, *Convolutional Neural Network*, Consultado el 21/05/2025, 2025. URL: <https://www.nvidia.com/en-eu/glossary/convolutional-neural-network/>.
- [135] 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>.
- [136] J. G. Sam Lau y D. Nolan, *Cross Validation*, Consultado el 26/05/2025, 2023. URL: https://learningds.org/ch/16/ms_cv.html.
- [137] 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.
- [138] 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.
- [139] 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.

- [140] 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.
- [141] 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.
- [142] 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.
- [143] C. E. Rasmussen, “Gaussian processes in machine learning,” en *Summer school on machine learning*, Springer, 2003, págs. 63-71.
- [144] C. Blundell, J. Cornebise, K. Kavukcuoglu y D. Wierstra, “Weight uncertainty in neural network,” en *International conference on machine learning*, PMLR, 2015, págs. 1613-1622.
- [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.
- [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.
- [147] J. R. Berrendero. “Materiales del libro de Estadística.” Consultado el 2 de junio de 2025. URL: <https://verso.mat.uam.es/~joser.berrendero/libro-est/>. [Citado en pág. 65].
- [148] A. Charpentier. “Confidence vs. Credibility Intervals.” Consultado el 21 de agosto de 2025. URL: <https://freakonometrics.hypotheses.org/18117>.
- [149] 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.
- [150] J. Gama, “A survey on learning from data streams: current and future trends,” *Progress in Artificial Intelligence*, vol. 1, págs. 45-55, 2012.
- [151] J. Vermorel. “Quantile Regression,” LOKAD Quantitive Supply Chain, visitado 2 de jun. de 2025. URL: <https://www.lokad.com/quantile-regression-time-series-definition/>.

- [152] R. Koenker, *Quantile Regression* (Econometric Society Monographs). Cambridge University Press, 2005.
- [153] 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.
- [154] J. Feldman y D. Kowal, “Bayesian Quantile Regression with Subset Selection: A Posterior Summarization Perspective,” *arXiv preprint arXiv:2311.02043*, 2023.
- [155] 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.
- [156] A. N. Angelopoulos y S. Bates, “A gentle introduction to conformal prediction and distribution-free uncertainty quantification,” *arXiv preprint arXiv:2107.07511*, 2021.
- [157] Scikit-learn-contrib MAPIE developers. “MAPIE: Model-Agnostic Prediction Interval Estimator.” Accessed: 2025-07-06. URL: <https://mapie.readthedocs.io/en/stable/>.
- [158] V. Vovk, “Cross-conformal predictors,” *Annals of Mathematics and Artificial Intelligence*, vol. 74, n.º 1, págs. 9-28, 2015.
- [159] 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. 67].
- [160] V. Vovk, D. Lindsay, I. Nouretdinov y A. Gammerman, “Mondrian confidence machine,” *Technical Report*, 2003.
- [161] Y. Romano, M. Sesia y E. Candès, “Classification with valid and adaptive coverage,” *Advances in neural information processing systems*, vol. 33, págs. 3581-3591, 2020. [Citado en pág. 67].
- [162] 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 pág. 67].
- [163] 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.
- [164] 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.

- [165] Y. Romano, E. Patterson y E. Candès, “Conformalized quantile regression,” *Advances in neural information processing systems*, vol. 32, 2019. [Citado en pág. 67].
- [166] R. Luo y Z. Zhou, “Conformal thresholded intervals for efficient regression,” en *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, págs. 19 216-19 223. [Citado en pág. 67].

Apéndice A

Problema de clasificación de sexo

Apéndice B

Comparación de resultados de estimación de edad y clasificación de edad

Es interesante comparar los resultados obtenidos en los problemas de estimación de edad y clasificación de edad, ya que existe una correspondencia directa entre ambos enfoques, si bien es necesario salvar ciertas distancias conceptuales y asumir limitaciones metodológicas. Esto se debe a que, mientras el primero trata valores como 17.1 y 17.9 como datos numéricamente diferenciados —aunque relativamente próximos—, el segundo los interpreta dentro de una misma categoría, es decir, 17 años, sin capturar la variación interna que existe dentro de la clase. A pesar de ello, dado que existe un número relativamente alto de edades posibles y la variabilidad inherente a la predicción suele estar bastante dispersa en el dominio de predicción, se puede establecer una relación entre ambas tareas.

En la Tabla B.1 se muestran los resultados de los mejores algoritmos en cada modalidad según dos métricas clave: la cobertura empírica y el tamaño medio del conjunto (en clasificación) o la amplitud media del intervalo (en regresión). Se observa que ningún método domina claramente sobre ningún otro; es decir, no existe un algoritmo que consiga simultáneamente una mayor cobertura y un menor tamaño/amplitud media del intervalo predictivo. Esto nos lleva a pensar que los tres algoritmos están sobre una frontera de compromiso de valor similar.

Tendríamos que valorarlo por otras aspectos como adaptatividad en base a tamaño o edad cronológica

Ejecución	Cobertura Empírica (%)			Tamaño Medio del Conjunto		
	CQR	LAC	SAPS	CQR	LAC	SAPS
Ejecución 1	95.31	94.66	94.98	6.23	5.79	6.05
Ejecución 2	94.80	94.24	95.12	6.11	5.76	6.03
Ejecución 3	95.45	95.21	95.26	6.02	6.04	6.17
Ejecución 4	94.61	94.89	94.80	5.90	5.86	5.98
Ejecución 5	94.93	95.17	95.21	5.92	5.77	6.16
Ejecución 6	94.33	94.80	95.45	5.94	5.80	6.08
Ejecución 7	95.26	93.91	94.56	6.00	5.69	6.07
Ejecución 8	95.12	95.59	95.86	6.08	6.03	6.28
Ejecución 9	94.93	94.70	95.59	6.06	5.86	6.15
Ejecución 10	94.56	94.14	94.80	5.96	5.88	6.36
Media	94.93	95.45	95.16	6.02	5.85	6.13

Tabla B.1: Cobertura empírica y

Apéndice C

Intervalos de valores razonables

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

- El **intervalo de confianza (IC)** es una herramienta común de la estadística frecuentista, que permite estimar un rango de valores tal que podamos confiar en que contiene al valor verdadero de un parámetro poblacional desconocido θ (p.ej., la media) [147].

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

Es importante aclarar un malentendido común: un intervalo de confianza con nivel 95 % para un parámetro θ no significa que exista un 95 % de probabilidad de que θ esté dentro del intervalo calculado a partir de una muestra específica. En realidad, el 95 % se refiere a la frecuencia con la que, si muestreásemos muchas veces los datos, los intervalos construidos a partir de esas muestras incluirían al valor verdadero de θ en aproximadamente el 95 % [126] (véase la Figura C.1).

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

La diferencia radica en que, a diferencia del intervalo de confianza, que parte de que θ es un parámetro fijo desconocido y los datos son

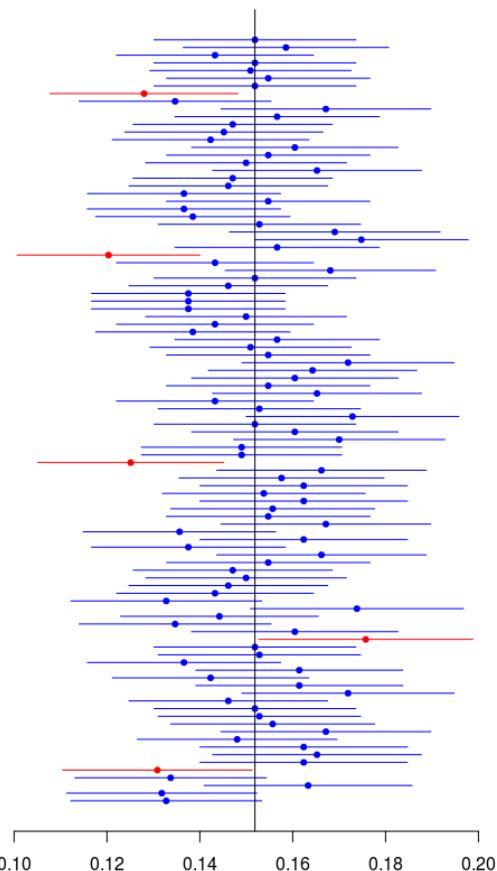


Figura C.1: Ejemplo de intervalos de confianza para la media poblacional. La interpretación correcta del nivel de confianza (95 % en este caso) es: *Si repitiéramos el proceso de muestreo y construcción de intervalos muchas veces, aproximadamente el 95 % de ellos contendrían el verdadero valor de la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

tratados como aleatorios, el enfoque bayesiano fija los datos(ya que son conocidos) y el parámetro θ lo trata como aleatorio (ya que es desconocido) [126].

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

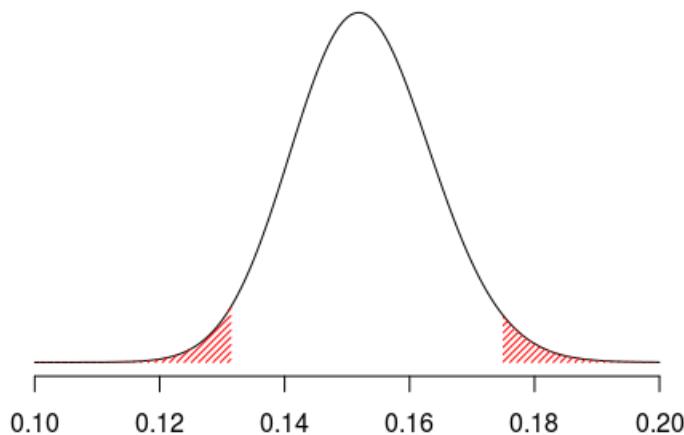


Figura C.2: Ejemplo de intervalo de credibilidad para la media poblacional. La interpretación correcta es: *Con un 95 % de probabilidad, el valor verdadero está dentro del intervalo*. En esta simulación, la media real conocida es 0.153, y podemos observar que efectivamente este valor está contenido en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

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

El enfoque explorado en este trabajo es la predicción conformal, que ha demostrado ser eficaz en contextos donde los supuestos clásicos (normalidad, homocedasticidad) no se cumplen [165], y es actualmente el enfoque más robusto para la construcción de intervalos de predicción en aplicaciones modernas de ML [159, 161, 162, 165, 166]. La predicción conformal tiene una interpretación frecuentista: $1 - \alpha$ intervalos producidos cubren el verdadero valor (véase la Figura C.3).

Como podemos esperar, a más estrecho sea el intervalo que manejemos,

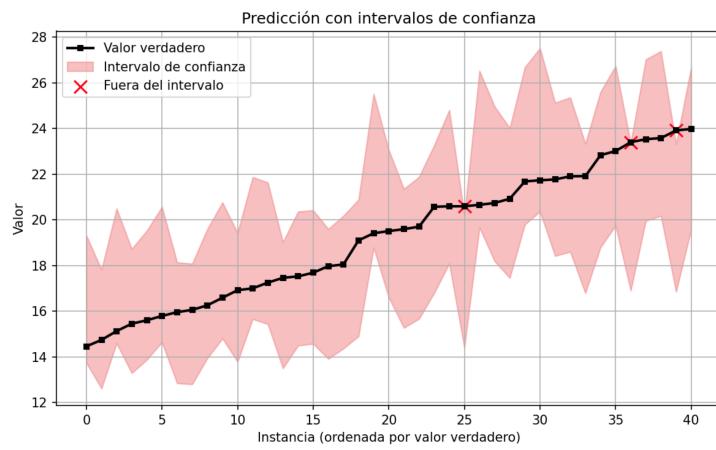


Figura C.3: Intervalos de predicción (95 % de confianza) construidos con CQR para estimación de edad.

más se puede confiar en las predicciones pero no todos los tipos de intervalos revelan la misma información sobre incertidumbre.

