TRABAJO FIN DE GRADO

GRADO EN INGENIERÍA INFORMÁTICA

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

Autor
David González Durán

Director
Pablo Mesejo Santiago

Mentor
Javier Venema Rodríguez

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

—

Granada, mes de 2025

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

David González Durán

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

Resumen

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

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

David González Durán

Keywords: Keyword1, Keyword2, Keyword3, ...

Abstract

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

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

Fdo: David González Durán

Granada, a X de mes de 202.

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

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

Informan:

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

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

Los directores:

Pablo Mesejo Santiago **Javier Vénema Rodríguez**

Agradecimientos

Poner aquí agradecimientos...

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Experimentación

1.1. Protocolo de validación experimental

Como se ha comentado anteriormente, se han proporcionado los datos ya divididos en conjunto de entrenamiento (*train*) y de test, para evitar problemas asociados al *data snooping*. [1]. Al proporcionar las particiones predefinidas, se garantiza que no haya contaminación entre los datos de entrenamiento y test, manteniendo así la validez de las métricas obtenidas en el test.

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

En la Figura 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 sexo [2].

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

[1] El ***data snooping*** ocurre cuando información del conjunto de test se filtra, directa o indirectamente, en el proceso de entrenamiento del modelo, lo que puede llevar a una sobreestimación del rendimiento y a modelos que generalizan pobremente en datos nuevos

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

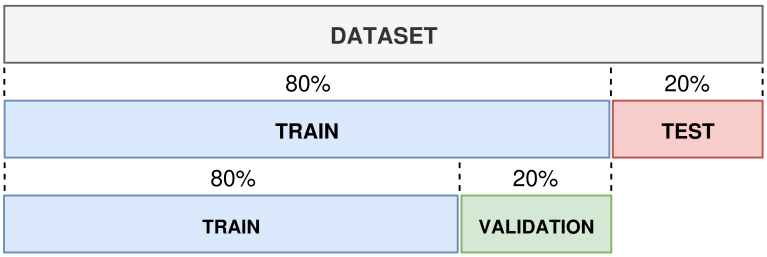


Figura 1.1: Diagrama de división del *dataset* en *train*, *validation* y *test*. Elaboración propia.

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.

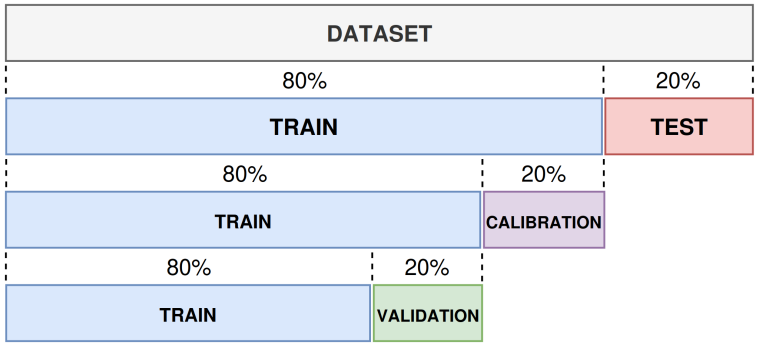


Figura 1.2: Diagrama de división del *dataset* en *train*, *validation*, *calibration* y *test*. Elaboración propia.

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. Experimentos propuestos

1.2.1. Comparativa de métodos para la estimación de edad

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

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

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

  para predecir la predicción puntual, límite inferior y límite superior, respectivamente.
- **Método 'CQR'**: Este modelo implement

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

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

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

Los métodos propuestos son:

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

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

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

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

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

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

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

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

Partiendo de esto, los métodos propuestos son:

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

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

1.3. Entrenamiento de los modelos

1.3.1. Preparación de los datos de entrenamiento

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

448×224 píxeles para homogenizar las entradas del modelo [3]. Se ha establecido un tamaño de *batch* de 32, tras encontrar preeliminarmente un equilibrio entre regularización y buen ritmo de aprendizaje. Y también se ha realizado *data augmentation* en el conjunto de entrenamiento, introduciendo transformaciones aleatorias en cada época para simular condiciones de posicionamiento del paciente y de la máquina e iluminación ligeramente variables:

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

1.3.2. Adaptación de la red para la estimación de edad

Como se venía anticipando en el anterior capítulo, adaptaremos la arquitectura del modelo ResNeXt50 para el problema de regresión. El tamaño de las imágenes de entrada no modifica la arquitectura del modelo, pues el extractor 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 un capa *batch normalization*, una capa de *dropout* y una capa completamente conectada (FC), con una activación ReLU entre ambos bloques. La primera capa FC contiene 4.096 neuronas, la segunda 512, y finalmente se incluye una capa de salida de una sola neurona. Esta configuración ha sido seleccionada siguiendo la recomendación de los tutores, quienes cuentan con experiencia previa en el trabajo con este conjunto de datos.

Añadir un dibujo con el cambio de cabecera (AGOSTO)

Los componentes clave del *pipeline* de entrenamiento son:

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

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

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

- Optimizador AdamW [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 entrenan con tasas más altas. Esta asignación se ha realizado mediante una progresión exponencial, que varía desde 1.5e-4 en los bloques más profundos hasta 1.5e-2 en los más superficiales. Este enfoque busca preservar el conocimiento útil de las capas inferiores y permitir una adaptación más rápida en las superiores.

La política OneCycle se ha aplicado individualmente a cada grupo de capas, haciendo que cada uno siga un ciclo de una sola fase: el *learning rate* comienza en un valor inicial bajo, aumenta progresivamente durante las primeras épocas (*warm-up*), y desciende de forma suave hasta un valor final aún menor [4]. 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.

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

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

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

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

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

Dado que la tarea de estimación de mayoría de edad guarda una estrecha relación con la estimación de edad continua, se ha optado por reutilizar el 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.

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

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

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

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

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

1.4. Métricas usadas en los experimentos

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

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

1.4.1. Métricas para regresión

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- El **tamaño de intevalo medio (*mean inteval width*, MIW)** mide qué tan amplios son en promedio los intervalos predichos.

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

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

- La ***mean interval score*** **(MIS)** [5] trata de unificar en una sola métrica el *trade-off* cobertura vs. amplitud del intervalo. Su expresión es la siguiente:

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

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

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

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

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

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

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

1.4.2. Métricas para clasificación

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

- La **matriz de confusión** es una herramienta fundamental que permite visualizar el rendimiento de modelos de clasificación, tanto binarios como multiclase. Esta muestra una tabla con tantas columnas y filas como clases haya. En un eje, se representan las clases reales (etiquetas verdaderas), y en el otro eje, las clases predichas por el modelo. Cada celda de la matriz indica la cantidad de ejemplos que pertenecen a una clase real específica y que han sido clasificados como una clase predicha específica (véase la Figura 1.3). Idealmente, los valores se concentrarían en la diagonal principal, lo que indicaría que las predicciones coinciden con los valores reales. Prácticamente todas las métricas y visualizaciones parten de la información ofrecida en esta matriz.

- La **cobertura empírica (*empirical coverage*)**, de forma análoga a la regresión, mide la proporción de veces que la etiqueta verdadera está contenida dentro del conjunto predicho.

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

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

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

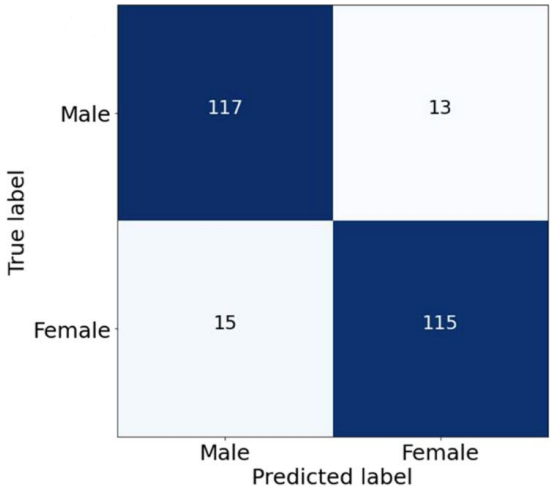


Figura 1.3: Matriz de confusión para la estimación de sexo según el modelo *random forest* propuesto en [6].

1.5. Resultados

1.5.1. Resultados para la estimación de edad

Análisis de métricas para la estimación de edad puntual

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 solo unas centésimas en las métricas evaluadas. Sin embargo, el análisis de varianza (ANOVA) [5] de un factor reveló diferencias estadísticamente significativas entre los modelos para ambas métricas: MAE ($F(3, 28) = 20.38$, $p < 0.001$) y MSE ($F(3, 28) = 15.09$, $p < 0.001$). Para identificar qué pares de modelos presentaban diferencias significativas, se aplicó la prueba *post-hoc* de comparaciones múltiples de Tukey (véanse las Tablas 1.2 y 1.3). Los resultados indicaron lo siguiente:

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

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

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

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

Tabla 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. Punto como separador decimal.

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

Análisis de métricas para la estimación de edad interválica

A continuación, la Tabla 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 inferiores al nivel deseado (alrededor del 88-89 % frente al 95 % nominal), lo que indica una infracobertura sistemática. Esto ocurre porque ni la heurística del método 'base' ni las regiones generadas por la regresión cuantílica en QR cuentan con garantías teóricas de cobertura estadística.

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

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

Tabla 1.2: Resultados de la prueba *post-hoc* de Tukey HSD para MAE entre pares de modelos. Se muestran la diferencia media entre grupos, el valor p ajustado, el intervalo de confianza al 95 % y si la diferencia es estadísticamente significativa ($\alpha = 0.05$).

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

Tabla 1.3: Resultados de la prueba *post-hoc* de Tukey HSD para MSE entre pares de modelos. Se muestran la diferencia media entre grupos, el valor p ajustado, el intervalo de confianza al 95 % y si la diferencia es estadísticamente significativa ($\alpha = 0.05$).

les. Esta relación de compromiso o *trade-off* entre cobertura y amplitud de los intervalos —típico en la predicción interválica— se visualiza claramente en la Figura 1.4, donde se observa una alta correlación entre la cobertura empírica y el tamaño del intervalo de predicción.

Sin embargo, CQR presenta unas amplitudes promedias de intervalo significativamente más reducidas que ICP, logrando ambos métodos coberturas muy similares. De hecho, en la Tabla 1.5 apreciamos cómo CQR logra significativamente menores valores de *interval score* que ICP, indicando que CQR tiene un mejor equilibrio entre cobertura y tamaño del intervalo.

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

Apliqué ANOVA y Tukey aquí, pero no resultaron muy útiles ya que no tienen en cuenta la relación de correlación entre cobertura empírica y tamaño medio del intervalo

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

Tabla 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. Punto como separador decimal.

Análisis de la cobertura en base al tamaño del intervalo

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

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

En la Figura 1.5 se presentan los histogramas de la amplitud de los intervalos de predicción para dos modelos representativos, uno QR y otro CQR. En cada caso, se diferencia visualmente la proporción de instancias cuya predicción cubre el valor real de aquellas en las que no lo hace. Es notable en ambas figuras la presencia de dos grupos principales de instancias: uno más reducido, asociado a intervalos más estrechos, y otro más numeroso, correspondiente a intervalos de mayor amplitud. Respecto a la cobertura, el

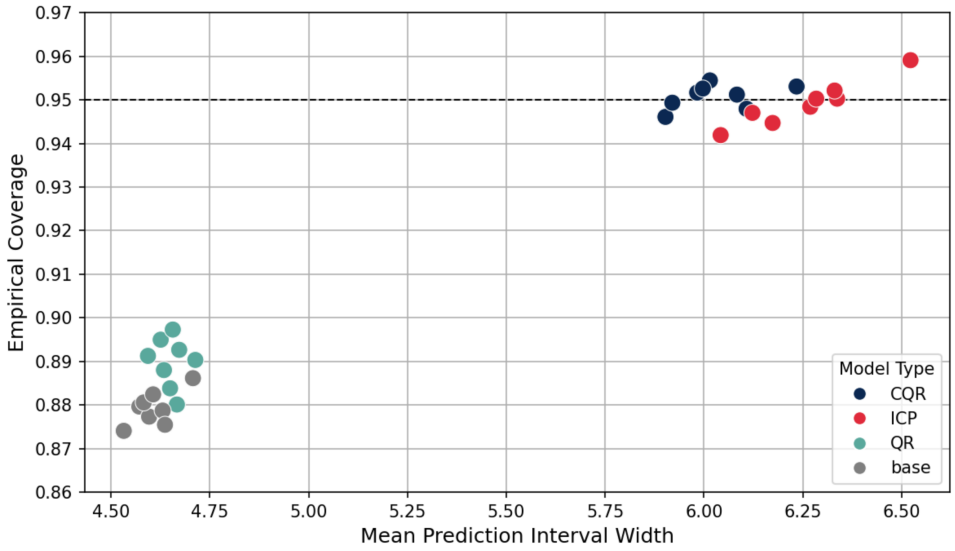


Figura 1.4: Gráfica de dispersión *Empirical Coverage-Mean Interval Width*. Elaboración propia.

modelo QR presenta valores inferiores, lo cual es consistente con su cobertura marginal, que ya se encontraba por debajo del 89 %. En cuanto al ratio entre cobertura e incobertura, este parece mantenerse relativamente estable a lo largo de los distintos rangos de amplitud del intervalo. Sin embargo, para un análisis más detallado y específico sobre cómo varía la cobertura en función del tamaño del intervalo, observemos la información desglosada en la Tabla 1.6.

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

En la Tabla 1.6 se ofrece información detallada sobre la cobertura empírica alcanzada por cada método de predicción (en todas sus ejecuciones) en función de diferentes rangos de amplitud del intervalo de predicción. Esta desagregación permite analizar si existe una relación entre el tamaño del intervalo y la capacidad del modelo para cubrir el valor real.

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

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

No estoy entrando a hacer valoraciones de lo grave o leve que sea que la cobertura se reduzca de un 95 a un 93.4, porque entiendo que aquí entraría mi subjetividad, y

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

Tabla 1.5: Resultados de las predicciones obtenidas por los modelos para el problema de estimación de edad en cada ejecución. Punto como separador decimal.

1.5.2. Análisis de la cobertura en base a la edad cronológica

Por último, se ha analizado la cobertura en base a la edad real de los individuos. En la Tabla 1.7

Se ha escogido el mejor modelo en cada método

En la figura 1.4 se

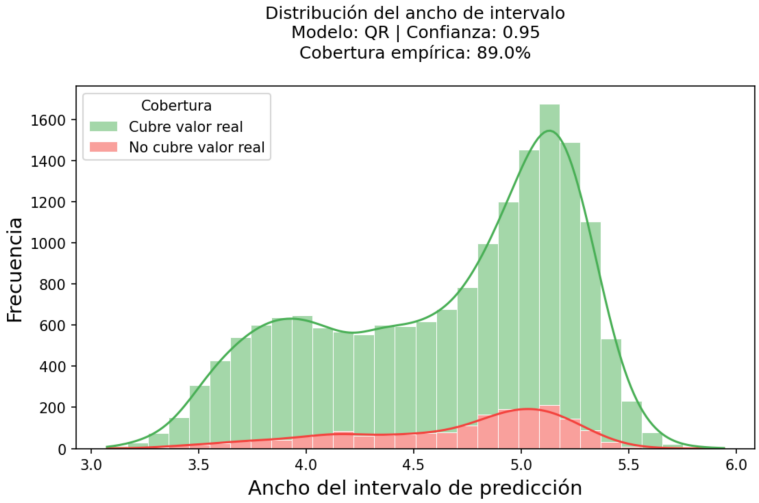
se aprecia el *trade-off* entre cobertura y tamaño, concretamente

¿Por qué las predicciones interválicas de ancho intermedio son las que menos cobertura ofrecen?

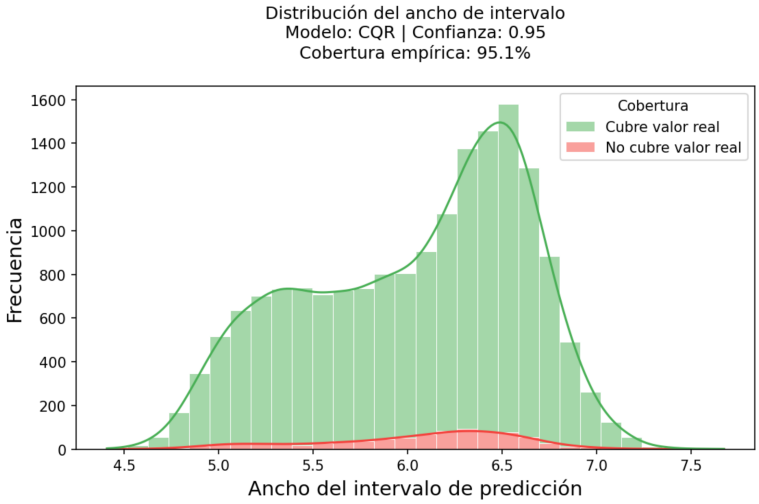
Se podría calibrar CQR por deciles?

Método		**Cobertura Empírica (%)**			
		base	**ICP**	**QR**	**CQR**
Amplitud del intervalo	$[4.0, 4.5)$	–	–	91.58	100
	$[4.5, 5.0)$	87.93	–	92.85	96.32
	$[5.0, 5.5)$	–	–	88.45	96.47
	$[5.5, 6.0)$	–	–	85.62	94.82
	$[6.0, 6.5)$	–	94.78	89.43	93.40
	$[6.5, 7.0)$	–	95.91	97.14	96.09
	$[7.0, 7.5)$	–	–	–	97.37
	$[7.5, 8.0)$	–	–	–	100

Tabla 1.6: Cobertura empírica del intervalo de predicción obtenida por cada método de predicción para distintas franjas de amplitud de intervalos. Nota: Los métodos de intervalos de tamaño fijo (como ICP, en este caso) pueden mostrar varias franjas debido a que los tamaños de intervalo pueden variar ligeramente entre entrenamientos para un mismo método.



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



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

Figura 1.5: Histogramas del amplitud del intervalo de predicción con diferenciación por cobertura, correspondientes a los modelos QR y CQR. Para cada tipo de método se seleccionó el modelo con el mejor *interval score*. La comparación permite visualizar cómo varía la capacidad de cobertura en función del tamaño del intervalo.

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

Tabla 1.7: Cobertura empírica y amplitud media del intervalo de predicción obtenidos por cada método de predicción para distintas edades cronológicas.

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. 7].

[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. 7].

[5] T. Gneiting y A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American statistical Association*, vol. 102, n.º 477, págs. 359-378, 2007. [Citado en pág. 11].

[6] M. A. Bidmos, O. I. Olateju, S. Latiff, T. Rahman y M. E. Chowdhury, "Machine learning and discriminant function analysis in the formulation of generic models for sex prediction using patella measurements," *International Journal of Legal Medicine*, vol. 137, n.º 2, págs. 471-485, 2023. [Citado en pág. 13].

[7] 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.

[8] 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.

[9] 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.

[10] S. Cordner y M. Tidball-Binz, "Humanitarian forensic action — Its origins and future," *Forensic Science International*, vol. 279, págs. 65-71, 2017.

[11] T. Tanaka, "International Humanitarian Law (IHL) and Forensic Document Examination," *Journal of the American Society of Questioned Document Examiners*, vol. 23, n.º 1, 2020.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] H. Lee, S. Tajmir, M. Zissen, B. Yeshiwas, T. Alkasab, G. Choy y S. Do, "Fully Automated Deep Learning System for Bone Age Assessment," *Journal of digital imaging*, vol. 30, págs. 427-441, 2017.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] G. S. Cooper y V. Meterko, "Cognitive bias research in forensic science: A systematic review," *Forensic Science International*, vol. 297, págs. 35-46, 2019.

[32] D. H. Ubelaker y C. M. DeGaglia, "Population variation in skeletal sexual dimorphism," *Forensic Science International*, vol. 278, 407.e1-407.e7, 2017.

[33] S. Aja-Fernández, R. de Luis-García, M. Martín-Fernández y C. Alberola-López, "A computational TW3 classifier for skeletal maturity assessment. A Computing with Words approach," *Journal of Biomedical Informatics*, vol. 37, n.º 2, págs. 99-107, 2004.

[34] 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.

[35] J. Venema, D. Peula, J. Irurita y P. Mesejo, "Employing deep learning for sex estimation of adult individuals using 2D images of the humerus," *Neural Comput & Applic*, vol. 35, págs. 5987-5998, 2022.

[36] S. Park, S. Yang, J. Kim, J. Kang, J. Kim, K. Huh, S. Lee, W. Yi y M. Heo, "Automatic and robust estimation of sex and chronological age from panoramic radiographs using a multi-task deep learning network: a study on a South Korean population," *Int J Legal Med*, vol. 138, págs. 1741-1757, 2024.

[37] K. Imaizumi, S. Usui, K. Taniguchi, Y. Ogawa, T. Nagata, K. Kaga, H. Hayakawa y S. Shiotani, "Development of an age estimation method for bones based on machine learning using post-mortem computed tomography images of bones," *Forensic Imaging*, vol. 26, pág. 200 477, 2021.

[38] M. Štepanovský, Z. Buk, A. Pilmann Kotěrová, J. Brůžek, Š. Bejdová, N. Techataweewan y J. Velemínská, "Application of machine-learning methods in age-at-death estimation from 3D surface scans of the adult acetabulum," *Forensic science international*, vol. 365, pág. 112 272, 2024.

[39] A. Heinrich, "Accelerating computer vision-based human identification through the integration of deep learning-based age estimation from 2 to 89 years," *Sci Rep*, vol. 14, pág. 4195, 2024.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] D. D. Martin, D. Deusch, 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] J. E. Buikstra, "Standards for data collection from human skeletal remains," *Arkansas archaeological survey research series*, vol. 44, pág. 44, 1994.

[56] 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.

[57] 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.

[58] 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.

[59] A. Schmeling, R. B. Dettmeyer, E. Rudolf, V. Vieth y G. Geserick, "Forensic Age Estimation," *Deutsches Arzteblatt international*, vol. 113, n.º 4, págs. 44-50, 2016.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] D. Stern, T. Ebner, H. Bischof, S. Grassegger, T. Ehammer 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.

[67] Ministerio del Interior de España, "Informe anual sobre personas desaparecidas 2025," Ministerio del Interior, inf. téc., 2025.

[68] 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.

[69] American Anthropological Association. "What is Anthropology?" Consultado el 01/04/2025, American Anthropological Association. URL: https://americananthro.org/learn-teach/what-is-anthropology/.

[70] S. N. Byers y C. A. Juarez, *Introduction to Forensic Anthropology*, 6.ª ed. Routledge, 2023.

[71] T. Thompson y S. Black, *Forensic Human Identification: An Introduction*, 1.ª ed. Taylor & Francis, 2006.

[72] L. Scheuer y S. Black, *The juvenile skeleton*, 1.ª ed. Elsevier, 2004.

[73] 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.

[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: https://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] V. Nemani, L. Biggio, X. Huan, Z. Hu, O. Fink, A. Tran, Y. Wang, X. Zhang y C. Hu, "Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial," *Mechanical Systems and Signal Processing*, vol. 205, pág. 110 796, 2023.

[97] 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.

[98] 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.

[99] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez y F. Herrera, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Information fusion*, vol. 99, pág. 101 805, 2023.

[100] 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.

[101] 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.

[102] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, págs. 1-38, 2019.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] 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.

[108] 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.

[109] 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.

[110] 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.

[111] 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.

[112] 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.

[113] 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.

[114] Y. LeCun, Y. Bengio y G. Hinton, "Deep Learning," *Nature*, vol. 521, págs. 436-44, 2015.

[115] 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.

[116] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari y U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Computer methods and programs in biomedicine*, vol. 226, pág. 107 161, 2022.

[117] P. Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. USA: Penguin Books Limited, 2015.

[118] S. Russell y P. Norvig, *Artificial Intelligence: A Modern Approach*, 4rd. Prentice Hall Press, 2021.

[119] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Heidelberg: Springer-Verlag, 2006.

[120] E. Alpaydin, *Introduction to Machine Learning*, 2nd. The MIT Press, 2010.

[121] P. J. Werbos, *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*. USA: Wiley-Interscience, 1994.

[122] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[123] R. Szeliski, *Computer Vision: Algorithms and Applications*, 1st. Berlin, Heidelberg: Springer-Verlag, 2010.

[124] A. Zhang, Z. C. Lipton, M. Li y A. J. Smola, *Dive into Deep Learning*, 2021.

[125] I. Goodfellow, Y. Bengio y A. Courville, *Deep Learning*. MIT Press, 2016.

[126] V. Vovk, A. Gammerman y G. Shafer, *Algorithmic learning in a random world*. Springer, 2005, vol. 29.

[127] Red Hat, *Deep learning*, Consultado el 10/05/2025, 2023. URL: https://www.redhat.com/es/topics/ai/what-is-deep-learning.

[128] Code World, *Understanding ML & DL in python*, Consultado el 19/05/2025, 2022. URL: https://codeworld.tistory.com/2.

[129] NVIDIA, *Convolutional Neural Network*, Consultado el 21/05/2025, 2025. URL: https://www.nvidia.com/en-eu/glossary/convolutional-neural-network/.

[130] 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/.

[131] J. G. Sam Lau y D. Nolan, *Cross Validation*, Consultado el 26/05/2025, 2023. URL: https://learningds.org/ch/16/ms_cv.html.

[132] 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.

[133] 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.

[134] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li y L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," en *2009 IEEE conference on computer vision and pattern recognition*, IEEE, 2009, págs. 248-255.

[135] 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.

[136] 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.

[137] 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.

[138] 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.

[139] 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.

[140] J. R. Berrendero. "Materiales del libro de Estadística," visitado 2 de jun. de 2025. URL: https://verso.mat.uam.es/~joser.berrendero/libro-est/.

[141] 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.

[142] J. Gama, "A survey on learning from data streams: current and future trends," *Progress in Artificial Intelligence*, vol. 1, págs. 45-55, 2012.

[143] J. Vermorel. "Quantile Regression," LOKAD Quantitive Supply Chain, visitado 2 de jun. de 2025. URL: https://www.lokad.com/quantile-regression-time-series-definition/.

[144] R. Koenker, *Quantile Regression* (Econometric Society Monographs). Cambridge University Press, 2005.

[145] 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.

[146] J. Feldman y D. Kowal, "Bayesian Quantile Regression with Subset Selection: A Posterior Summarization Perspective," *arXiv preprint arXiv:2311.02043*, 2023.

[147] 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.

[148] A. N. Angelopoulos y S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv preprint arXiv:2107.07511*, 2021.

[149] Scikit-learn-contrib MAPIE developers. "MAPIE: Model-Agnostic Prediction Interval Estimator." Accessed: 2025-07-06. URL: https://mapie.readthedocs.io/en/stable/.

[150] 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.

[151] V. Vovk, D. Lindsay, I. Nouretdinov y A. Gammerman, "Mondrian confidence machine," *Technical Report*, 2003.

[152] Y. Romano, M. Sesia y E. Candes, "Classification with valid and adaptive coverage," *Advances in neural information processing systems*, vol. 33, págs. 3581-3591, 2020.

[153] A. Angelopoulos, S. Bates, J. Malik y M. I. Jordan, "Uncertainty sets for image classifiers using conformal prediction," *arXiv preprint arXiv:2009.14193*, 2020.

[154] 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.

[155] Y. Romano, E. Patterson y E. Candès, "Conformalized quantile regression," *Advances in neural information processing systems*, vol. 32, 2019.

[156] D. Bethell, S. Gerasimou y R. Calinescu, "Robust uncertainty quantification using conformalised Monte Carlo prediction," en *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, págs. 20 939-20 948.

[157] R. Luo y Z. Zhou, "Conformal thresholded intervals for efficient regression," en *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, págs. 19 216-19 223.