Bias-detection-in-age-recognition-from-photos

Marian Aguilar Tavier, Jennifer de la Caridad Sánchez Santana, Katherine Rodríguez Rodríguez, Reinaldo Cánovas Gamón

1 Introducción

La elección de investigar y desarrollar métodos para detectar sesgos en modelos de detección de edad a partir de fotografías faciales se fundamenta en una serie de consideraciones cruciales relacionadas con la ética, la equidad y la precisión en la tecnología de Inteligencia Artificial (IA). Este tema es particularmente relevante en el contexto actual, donde la IA se integra cada vez más en diversos aspectos de nuestra vida cotidiana, desde la identificación facial hasta la personalización de servicios y productos. Sin embargo, la falta de diversidad y representatividad en los conjuntos de datos utilizados para entrenar estos modelos puede conducir a sesgos sistemáticos que afectan negativamente tanto a los usuarios como a la sociedad en general.

Uno de los pilares fundamentales de la IA es su potencial para mejorar la vida humana y resolver problemas complejos. Sin embargo, la ausencia de diversidad en los conjuntos de datos puede resultar en modelos que no solo son inexactos sino que también pueden reforzar estereotipos y discriminación. Por ejemplo, si un modelo de detección de edad está sesgado hacia ciertos grupos demográficos, podría dar lugar a errores sistemáticos en la estimación de la edad, lo que podría tener implicaciones serias en contextos como la inspección de menores de edad, la determinación de la elegibilidad para ciertos beneficios o incluso en la selección de empleados. La detección y corrección de estos sesgos es, por lo tanto, una responsabilidad ética crucial en el desarrollo de tecnologías de IA.

La precisión en la detección de edad es fundamental para muchas aplicaciones,

desde la medicina hasta la publicidad y la seguridad. Los sesgos en la estimación de la edad pueden llevar a conclusiones erróneas que tienen un impacto directo en las vidas de las personas. Por ejemplo, una mala estimación de la edad puede resultar en la exclusión de individuos de oportunidades basadas en supuestos incorrectos sobre su edad. La investigación en este campo busca garantizar que los modelos de IA sean justos y precisos para todos los usuarios, independientemente de su origen étnico, género o edad.
Investigar y abordar los sesgos en los modelos de detección de edad también contribuye al avance de la ciencia de la IA. La comprensión de cómo los sesgos se introducen y perpetúan en los modelos de IA es esencial para desarrollar mejores prácticas y herramientas que ayuden a construir sistemas más justos y equitativos. Este conocimiento puede ser transferido a otras áreas de la IA, contribuyendo a la construcción de tecnologías que beneficien a toda la sociedad.
Finalmente, seleccionar este tema ofrece una oportunidad única para innovar y contribuir a la mejora continua de la tecnología de IA. Al centrarse en la detección de sesgos en los modelos de detección de edad, se abren nuevas vías de investigación y desarrollo que pueden llevar a avances significativos en la precisión, la eficacia y la aceptación social de la IA. Este enfoque no solo beneficia a la comunidad científica y técnica sino que también tiene el potencial de mejorar la calidad de vida de millones de personas al asegurar que las tecnologías de IA sean accesibles, justas y efectivas para todos.

2 Estado del Arte

El estado actual del arte en la mitigación de sesgos en modelos, incluidos aquellos que se utilizan para predecir la edad a partir de fotografías faciales, presenta una gama diversa y compleja de técnicas y enfoques. Estos métodos abarcan desde la manipulación de los conjuntos de datos antes del entrenamiento hasta la modificación de los modelos y la post-procesamiento de las predicciones. Cada uno de estos enfoques tiene sus propias ventajas y desventajas, y su efectividad puede variar dependiendo del tipo específico de sesgo que se esté tratando de mitigar.
Técnicas de Preprocesamiento
El preprocesamiento de los datos es una etapa crucial en la preparación de los conjuntos de datos para el entrenamiento de modelos de IA. Incluye la limpieza de los datos, la normalización, la selección de características y la generación de nuevas características. Para mitigar el sesgo, los investigadores han explorado varias técnicas, como:
-Balanceo de clases: Ajustar la representación de las clases minoritarias en el conjunto de datos para evitar sesgos hacia las clases majoritarias.
-Generación de datos sintéticos: Crear nuevas instancias de datos que representen mejor la diversidad de la población objetivo, especialmente útil cuando los datos reales son escasos o sesgados.
-Selección de características: Eliminar o modificar características que puedan

introducir sesgos en el modelo.
Procesamiento y Posprocesamiento
Durante el entrenamiento y la inferencia, los modelos de IA pueden introducir sesgos adicionales debido a la naturaleza de los algoritmos y los parámetros de modelado. Para abordar estos sesgos, se han propuesto varias técnicas:
-Regularización: Introducir restricciones en el modelo para evitar el sobreajuste y la introducción de sesgos.
-Optimización de hiperparámetros: Ajustar los parámetros del modelo para mejorar su rendimiento y reducir el sesgo.
-Post-procesamiento: Aplicar reglas o modelos adicionales después de la inferencia para ajustar las predicciones finales y mitigar el sesgo.
Cada una de estas técnicas de mitigación de sesgos puede tener consecuencias secundarias, como:
-Complejidad aumentada: Implementar técnicas de mitigación de sesgos puede complicar el modelo, haciéndolo más difícil de entender y mantener.
-Pérdida de precisión: Algunas técnicas pueden afectar negativamente la precisión del modelo, especialmente si se aplican de manera excesiva o incorrecta.
-Pérdida de datos importantes: La manipulación de los datos para mitigar el sesgo puede resultar en la pérdida de información valiosa que podría ser crucial para el rendimiento del modelo.
La predicción de la edad a partir de fotografías faciales es un problema complejo que involucra múltiples tipos de sesgos, incluidos el sesgo de raza, género y edad. Varios enfoques han sido explorados para abordar este desafío, incluyendo:
-Modelos de aprendizaje profundo: Utilizar redes neuronales convolucionales (CNNs) y otros modelos de aprendizaje profundo para extraer características faciales relevantes para la estimación de la edad.
-Ensamblado de modelos: Combinar las predicciones de múltiples modelos para mejorar la precisión y reducir el sesgo.
-Técnicas de data augmentation: Generar nuevas imágenes a través de transformaciones (como rotación, zoom, etc.) para aumentar la diversidad del conjunto de datos y mejorar la robustez del modelo.
A pesar de los avances en la detección y mitigación de sesgos, sigue siendo un área activa de investigación. La complejidad inherente de los sesgos y las consecuencias secundarias de las técnicas de mitigación hacen que sea un desafío continuar mejorando la precisión y la equidad de los modelos de IA en la predicción de la edad y otros problemas similares.

3 Propuesta de solución

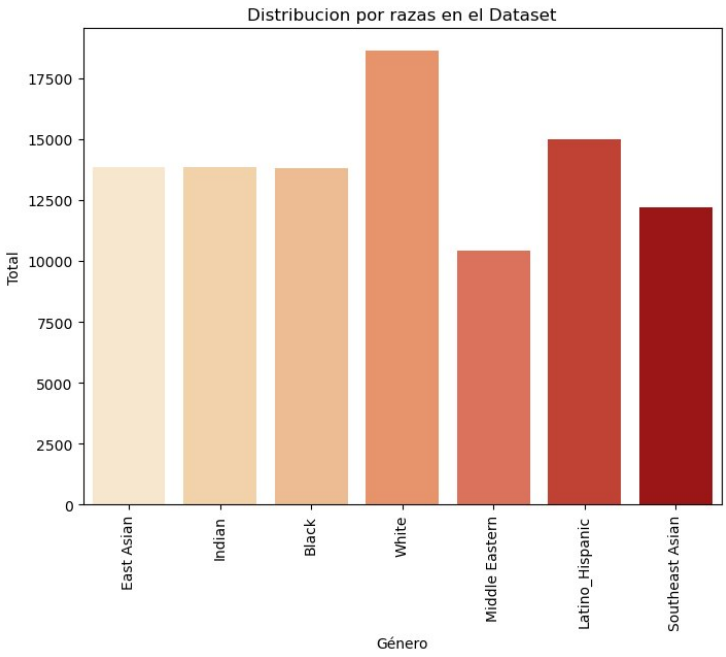
La propuesta de solución aborda varios aspectos críticos en el desarrollo y evaluación de modelos de detección de edad a partir de imágenes, enfocándose en cuatro modelos distintos: ViT-Age-Classifier, Yolov8 con Efficient-NetB0, una red neuronal con Keras, y Vit-B-32 Clip. Cada uno de estos modelos

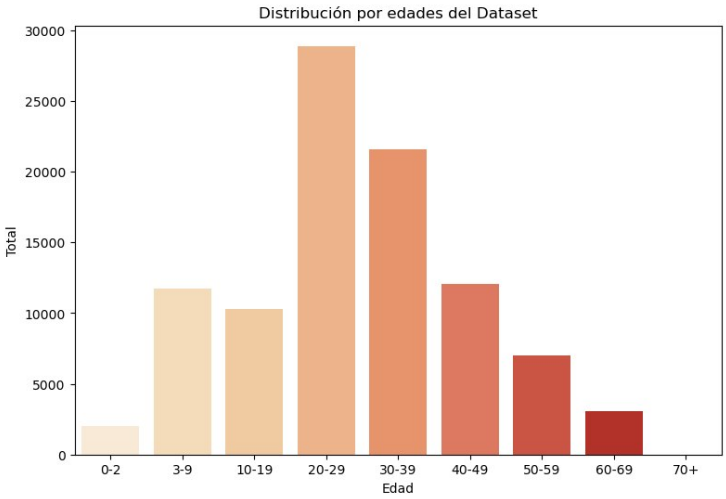
ofrece una perspectiva única sobre cómo abordar el desafío de la estimación precisa de la edad, explorando desde técnicas de transferencia de aprendizaje profundo hasta enfoques de procesamiento de imágenes y redes neuronales convolucionales.
Se hablará sobre el desarrollo de cada modelo en la siguiente sección.

4 Experimentación y resultados

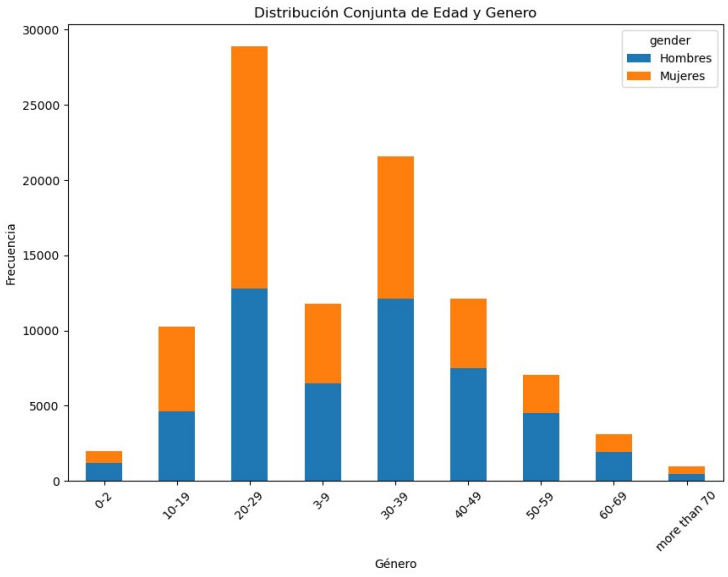
4.1 Análisis del dataset

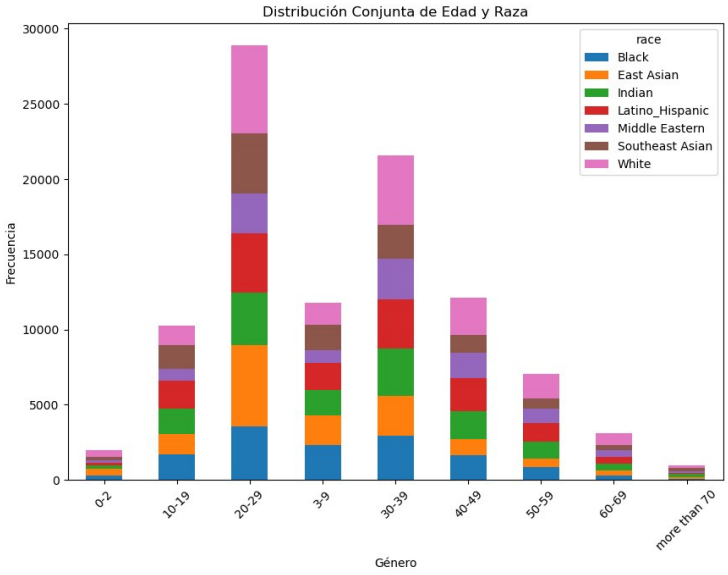
El conjunto de datos FairFace representa un hito significativo en el avance de la inteligencia artificial y el aprendizaje automático, enfocado en corregir y mitigar los sesgos sistemáticos presentes en los conjuntos de datos faciales tradicionales. Esta iniciativa surge en respuesta a las críticas y preocupaciones acerca de la falta de diversidad y representatividad en los conjuntos de datos faciales existentes, los cuales han sido señalados por perpetuar estereotipos y discriminación en la aplicación de tecnologías de IA.

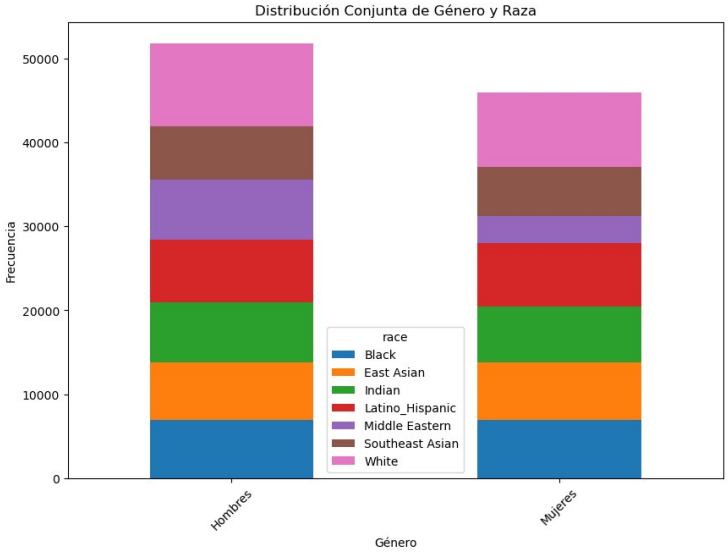




FairFace se distingue por su enfoque holístico en la inclusión y diversidad, incorporando una amplia gama de características demográficas y expresivas en sus imágenes. Con un total de 108,500 fotografías, el conjunto de datos se divide en categorías de género (Mujer y Hombre) y rasas (incluyendo East Asian, Indian, Black, White, Middle Eastern, Latino Hispanic, y Southeast Asian), además de rangos de edad específicos ('0-2', '3-9', '10-19', '20-29', '30-39', '40-49', '50-59', '60-69', '70 y más'). Este diseño meticuloso busca reflejar la riqueza y variedad de la humanidad en el ámbito digital, ofreciendo así una base más justa y equitativa para el entrenamiento de modelos de visión por computadora.







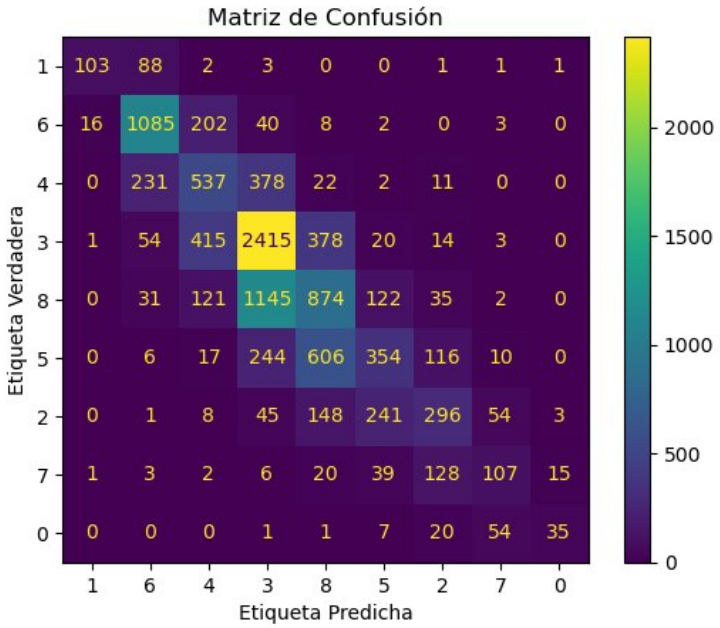
A pesar de los esfuerzos por maximizar la representatividad y minimizar el sesgo, FairFace reconoce que aún persisten desequilibrios en la distribución de las edades. Mientras que el rango de edad '20-29' constituye aproximadamente el 30 porciento del conjunto de datos, los rangos de edad superiores a 70 años representan apenas el 0.98 porciento. Este desequilibrio en la representación de las edades plantea desafíos particulares en términos de sesgo y precisión de

los modelos entrenados con FairFace.
Para abordar este desafío, se han explorado diversas estrategias de mitigación del sesgo. Entre ellas, se consideró la posibilidad de eliminar las entradas de las clases más sobrerrepresentadas, una decisión que, aunque podría parecer intuitiva, resultaría en la pérdida de una cantidad significativa de datos. En lugar de ello, se optó por aplicar transformaciones a las imágenes del rango de edad más afectado, ajustando así la distribución de las edades en el conjunto de datos de manera que se aproxime más a la distribución real de la población. Esta elección subraya la importancia de un enfoque cuidadoso y considerado en la gestión de sesgos en los conjuntos de datos de IA, especialmente cuando se trata de características demográficas tan sensibles como la edad. Al hacerlo, FairFace no solo contribuye a la mejora de la precisión y la justicia en la tecnología de IA sino que también establece un precedente valioso para futuras iniciativas en el campo de la diversidad y la inclusión en la inteligencia artificial.

4.2 Desarrollo de los modelos

ViT-Age-Classifier[6]:
Utilizamos este modelo proporcionado por nateraw en Hugging Face. Este modelo es un transformer de visión (ViT) que al cual se la ha hecho fine-tuning para clasificar la edad de una persona a partir de su rostro.
Con este modelo, la tarea de estimación de edad tiene un accuracy general de 53 porciento.
Si analizamos, la precision por grupos de edad nos percatamos que el modelo presenta una mayor eficiencia para los grupos de edad de 0-2 y de 3-9, mientras que para los otros grupos de edad el accuracy decae significativamente. Esto nos indica que el modelo empleado puede estar generalizando mejor para ciertos grupos de edad, en comparación con otros. Además como se analizó previamente, la distribución de los datos no es homogénea, sin embargo, se observa que los grupos con mayor representación tienen una menor precisión. Se debera de hacer un sobremuestreo en las clases con menor representación y reentrenar el modelo con los datos.
Yolov8 con Efficient-NetB0[2]:
Para este modelo utilizamos los modelos ya entrenados de Yolo y Efficient-Net B0, de acuerdo con el estado del arte(poner referencia).
Este modelo consta de dos modelos fusionados: se utiliza yolo para la detección de rostros, y Efficient-Net B0 para la estimación de edad.
Se ha utilizado un modelo preentrenado utilizando yolov8, que está especializado en la detección de rostros a partir de una foto. En el caso de EfficientNet-B0, se utilizó por ser el modelo más ligero con 237 capas pero solo 5.3 millones de parámetros si lo comparamos con los demás modelos de Efficient-Net.
Para adaptar el modelo para la tarea de estimación, al igual que en la bibliografía mencionada anteriormente se eliminó el clasificador original en el top y se sustituyó por una capa de global average pooling, una capa de batch

normalization, una de dropout(con un pequeño dropout rate de 0.2) y una capa de salida de una sola neurona con activación lineal. Luego se reentrenó EfficienNet-B0 con las nuevas capas sobre el conjunto de entrenamiento.
Para poder procesar la gran cantidad de imágenes del conjunto de entrenamiento se procesaron por batch (lotes) y se reajustó el tamaño de la foto a 224x224 como tamaño predeterminado de la capa de entrada. Sin embargo, con este tamaño de foto, todavía nos resultaba imposible procesarlo debido a elevado consumo de memoria que este implicaba, por lo que se reajustó la foto nuevamente para un tamaño de 96x96.
Esta práctica de por sí no es recomendable en el caso de modelos como EfficientNet, y se confirmó con los resultados obtenidos. El hecho de cambiar el tamaño de la imagen de 448x448 a 96x96 provoca que se pierda demasiada información de la foto, y por consiguiente los resultados en este modelo en cuanto a accuracy llega a duras penas a un 11 porciento.



Red neuronal con Keras
En este modelo realizamos la estimación de la edad así como del sexo de la persona.
Utilizamos la biblioteca de Python keras para todo esto.
Para la edad empleamos una capa Conv2d de 32 canales con un kernel de tamanno 3x3 y una función de activación relu, con un input shape de 96X96x3 por problemas de hardware. Luego se pasa por una capa de MaxPooling2D

con un tamanno de kernel de 2x2, luego otra capa de Conv2D con 64 canales, otra de Maxpooling, una 3era capa de de Conv2D, otra de MaxPooling2D, una capa de aplanamientonaca capa densa, otra de dropout y otra densa con una function de activación relu. Luego se compila con la función de pérdida del MSE, el optimizador de Adam y un learning rate de 0.0001.
Para la prediccion del sexo se utiliza prácticamente el mismo modelo, la única diferencia es que la última capa densa utiliza una función de activación sigmoid y la función de pérdida binary-crossentropy, ya que son valores binarios.

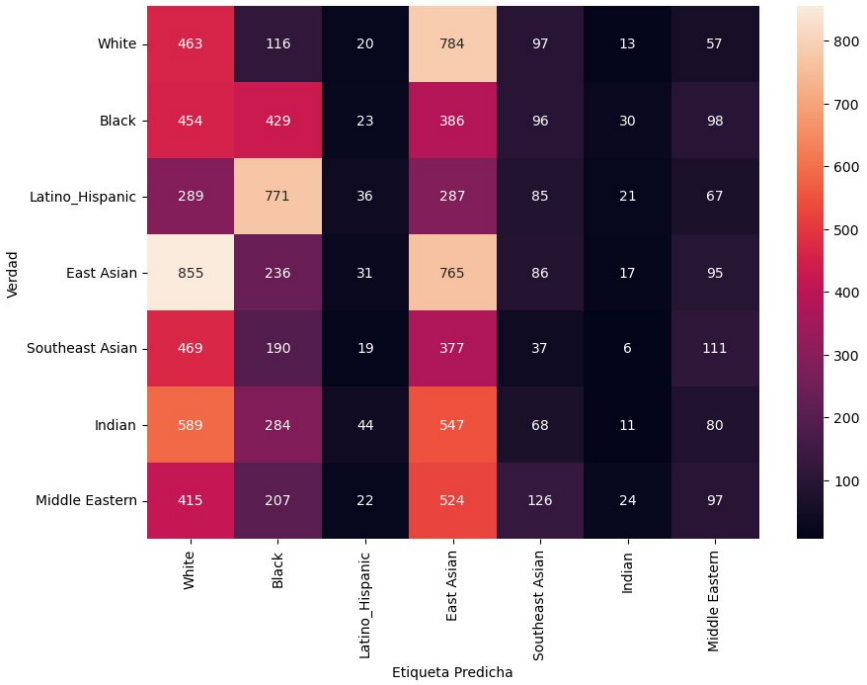
	precision	recall	f1-score	support
0	0.55	0.18	0.27	199
1	0.66	0.11	0.18	1356
2	0.17	0.07	0.10	1181
3	0.34	0.14	0.20	3300
4	0.23	0.40	0.29	2330
5	0.18	0.54	0.27	1353
6	0.21	0.17	0.19	796
7	0.26	0.03	0.06	321
8	0.86	0.05	0.10	118
9	0.00	0.00	0.00	0
accuracy			0.23	10954
macro avg	0.35	0.17	0.16	10954
weighted avg	0.31	0.23	0.21	10954

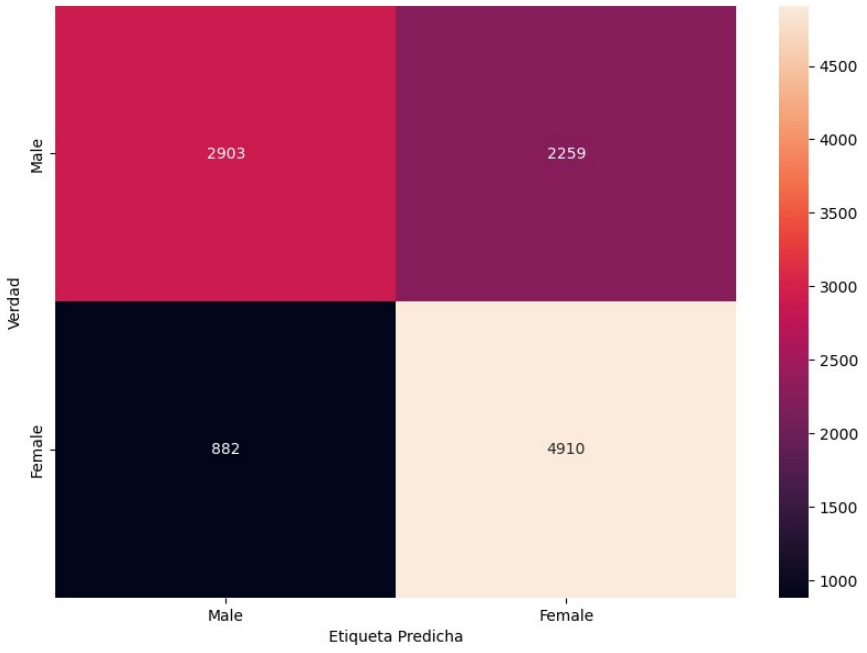
Con este modelo el accuracy sin aplicarle ninguna técnica de mitigación de sesgo al conjunto de entrenamiento es de 30 porciento para la edad y 73 porciento para el sexo.
Dado el problema con respecto a la falta de cómputo para aplicar oversampling en los grupos con menor distribución en el dataset hemos aplicado un método de mitigación de sesgo in-processing. En este caso hemos utilizado la ponderación por clases, o sea, se calculó para cada grupo de edad su pesos correspondiente en el modelo, que se calcula como la cantidad de datos en el dataset entre la cantidad de datos que corresponden a su categoría, o sea: Frecuencia relativa = (Cantidad de datos en la categoría) / (Cantidad total de datos en el dataset) [5]
Estos pesos se le pasan al modelo que se entrena, de esta forma, el modelo le dará más peso a los grupos menos representados y discriminará a aquellos que tienen una mayor cantidad de personas.
Con esta modificación el modelo obtiene una accuraccy de 23 porciento, o sea que empeora, sin embargo tiene un mejor rendimiento para el grupo de edad de más de 70 annos, por lo que puede ser que este realizando overfiting en ese grupo en particular, para las clases uno y dos también mejora, sin embargo por cuestiones de tiempo no es posible reentrenar el modelo para intentar mejorarlo, también hay que tener en cuenta el tamaño reducido de la foto.
Vit-B-32 Clip[3]:
Hemos utilizado este modelo para predecir la edad a partir de la foto de una

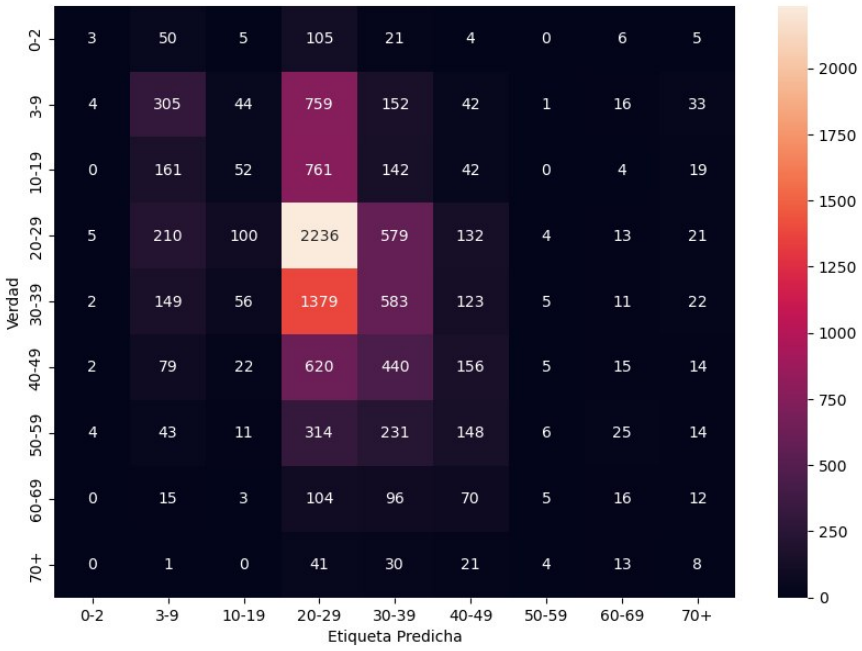
persona.
CLIP, que significa "Contrastive Language-Image Pre-training", es un modelo de aprendizaje automático desarrollado por OpenAI, revolucionando la forma en que interactuamos con imágenes y texto. A diferencia de los modelos de visión artificial tradicionales que se basan en la clasificación de imágenes, CLIP es un modelo multimodal que aprende a comprender la relación entre el texto y las imágenes.
CLIP se entrena con un conjunto masivo de datos de pares de imagen-texto, donde cada par representa una asociación específica entre un texto y una imagen. El objetivo del entrenamiento es aprender una representación común para ambos tipos de datos, permitiendo que el modelo identifique la correspondencia entre imágenes y textos. La clave de CLIP radica en la técnica de entrenamiento contrastivo, donde el modelo se entrena para distinguir pares de imagen-texto correctamente emparejados de aquellos que no lo están.
Esta capacidad de "entender" la relación entre el texto y las imágenes abre un abanico de posibilidades para diversas aplicaciones en el campo del aprendizaje automático. CLIP puede utilizarse para:
• Búsqueda de imágenes: Buscar imágenes relevantes a una consulta de texto, incluso si la consulta no contiene palabras específicas que se encuentren en la descripción de la imagen.
• Generación de subtítulos: Generar subtítulos descriptivos para imágenes, incluso si no hay información de contexto disponible. • Análisis de imágenes: Clasificar imágenes, identificar objetos y escenas, y analizar el contenido de las imágenes de una manera más sofisticada que los modelos de visión artificial tradicionales.
• Interacción hombre-máquina: Permitir que los usuarios interactúen con los sistemas de inteligencia artificial utilizando lenguaje natural, como dar instrucciones a un robot utilizando comandos de texto y imágenes.
CLIP ha demostrado ser un modelo versátil y potente, superando a los modelos de visión artificial tradicionales en diversas tareas. Su capacidad de comprender la relación entre el texto y las imágenes lo convierte en una herramienta esencial para el desarrollo de aplicaciones de aprendizaje automático más intuitivas y eficientes.
En este caso se utilizó el modelo ViT-base-Patch32.
El modelo utiliza una arquitectura de transformador ViT-B/32 como codificador de imagen y utiliza un transformador de autoatención enmascarado como codificador de texto. Estos codificadores están entrenados para maximizar la similitud de los pares (imagen, texto) a través de una pérdida contrastiva.
La implementación original tenía dos variantes: una que usaba un codificador de imágenes ResNet y la otra que usaba un Vision Transformer. Este repositorio tiene la variante con el Vision Transformer.
Para este modelo, se obtuvo una precisión de 39,556 porciento para la edad, con una mayor precisión para los grupos con mayor representación(de 60 y 59 porciento), mientras que el accuracy de las minorías no llega al 10 porciento.

```
Accuracy: 0.395563264560891
Recall: 0.395563264560891
F1-score: 0.40421163561933315
Precision: 0.4455474011120393
              precision    recall  f1-score   support

         0-2       0.00      0.01      0.01       199
       10-19       0.45      0.43      0.44      1181
       20-29       0.60      0.36      0.45      3300
         3-9       0.59      0.84      0.70      1356
       30-39       0.38      0.31      0.34      2330
       40-49       0.36      0.32      0.34      1353
       50-59       0.15      0.13      0.14       796
       60-69       0.21      0.52      0.30       321
more than 70       0.08      0.54      0.13       118

    accuracy                           0.40     10954
   macro avg       0.31      0.39      0.32     10954
weighted avg       0.45      0.40      0.40     10954
```

FairFace master[4]: El modelo FairFace-master es una aplicación del modelo ResNet34, una arquitectura de red neuronal convolucional avanzada diseñada para la clasificación de imágenes. La estructura de 34 capas de ResNet34 le otorga una gran capacidad para aprender representaciones complejas de las imágenes. En este caso, el modelo FairFace-master ha sido modificado quitándole la última capa y agregándole 18 neuronas: dos para género, siete para raza y nueve para los diferentes grupos de edad.
El modelo utiliza la función de activación softmax para calcular las probabilidades de que cada imagen pertenezca a las diferentes clases. Es importante destacar que el modelo ya viene entrenado con sus pesos, lo que significa que ha sido ajustado con un conjunto de datos previamente para reconocer patrones en imágenes de rostros relacionados con género, raza y edad, utilizando el mismo dataset que se utilizó en el paper de referencia del modelo entrenado con Fair Face.
La precisión del modelo se ha evaluado en un 30 porciento para la edad, 70 porciento para el género y 24 porciento para la raza. Es importante tener en cuenta que esta precisión puede variar dependiendo del conjunto de datos utilizado para el entrenamiento y las características específicas de las imágenes. Aunque el modelo comete bastantes errores a la hora de clasificar, en gran parte se lo podemos deber al ajuste del tamaño de las fotos,así como a la escasez de los datos para personas mayores de 70 años.


White
Black
Latino_Hispanic
East Asian
Southeast Asian
Indian
Middle Eastern
Verdad
463 116 20 784 97 13 57
454 429 23 386 96 30 98
289 771 36 287 85 21 67
855 236 31 765 86 17 95
469 190 19 377 37 6 111
589 284 44 547 68 11 80
415 207 22 524 126 24 97
800
700
600
500
400
300
200
100
White
Black
Latino_Hispanic
East Asian
Southeast Asian
Indian
Middle Eastern
Etiqueta Predicha


Male
Female
Verdad
2903
2259
882
4910
4500
4000
3500
3000
2500
2000
1500
1000
Male
Female
Etiqueta Predicha



5 Conclusiones

Para concluir, es evidente que, a pesar de los esfuerzos por maximizar la justicia y la representatividad en el conjunto de datos FairFace, aún persisten sesgos notables, especialmente en el rango de edades superior a 70 años. Este hallazgo es particularmente relevante en un contexto social donde la esperanza de vida está en aumento, y las personas mayores representan una proporción significativa de la población. La observación de sesgos en los resultados finales de varios modelos confirma que la causa de tales sesgos no se limita únicamente a la composición inicial del conjunto de datos; incluso los modelos que logran una precisión notable en la predicción de la edad pueden exhibir sesgos, a veces de manera más pronunciada en ciertos grupos demográficos. Es crucial reconocer que la realización de este proyecto implicó ajustes en el tamaño de las imágenes de entrada tanto durante la fase de entrenamiento como de validación del modelo. Debido a las limitaciones de recursos disponibles, no pudimos utilizar las dimensiones originales de las fotografías, lo que llevó a la pérdida de información visual. Este proceso de reducción de tamaño introduce un tipo específico de sesgo en el entrenamiento y la validación de nuestros modelos, ya que la información perdedida podría haber influido en la precisión y la justicia de las predicciones finales.

Este proyecto destaca la complejidad inherente de abordar los sesgos en los modelos, especialmente en tareas delicadas como la predicción de la edad. No solo es necesario considerar la composición del conjunto de datos y la arquitectura del modelo, sino también factores operativos como el manejo de los datos de entrada. La interacción entre estos componentes subraya la necesidad de un enfoque integral para la construcción de modelos justos y precisos.
Además, este estudio pone de relieve la importancia de la transparencia y la evaluación continua en el desarrollo de tecnologías de Inteligencia Artificial La identificación y documentación de sesgos, así como la adaptación de técnicas para mitigarlos, deben ser partes integrales del ciclo de vida del desarrollo de software de esta. Esto no solo contribuye a la mejora de la calidad y la fiabilidad de los modelos, sino que también promueve la confianza pública en la IA y su potencial para beneficiar a la sociedad.

6 Bibliografía

[1], url = https://docs.google.com/spreadsheets/d/1PdHxuy78tj-h3aw4PhMf7hFHp3hY8UE0ShdxcFEyxxk/edit?gid=0#gid=0
[2], author = Giovanna Castellano, Berardina De Carolis, Nicola Marvulli, Mauro Sciancalepore, title = Real-Time Age Estimation from Facial Images Using YOLO and EfficientNet, url = Real-Time Age Estimation from Facial Images Using YOLO and EfficientNet — Semantic Scholar, year = 2021
[3], url = https://github.com/hbm99/bias-project-ML
[4], url = https://github.com/dchen236/FairFace/blob/master
[5], url = https://drive.google.com/drive/folders/1F-pXfbzWvG-bhCpNsRj6F-xsdjpesiFu?usp=sharing
[6], url = https://huggingface.co/ibombonato/vit-age-classifier