Detección de sesgos en la estimación de edad, género y etnicidad a partir de fotos

Marian Aguilar Tavier, Jennifer de la C. Sánchez, Katherine Rodríguez Rodríguez, Reinaldo Cánovas.

Resumen

Este proyecto se basa en la estimación de la edad, género y etnicidad de una persona a partir de su foto facial. En este se implementan varios algoritmos de Machine Learning para lograr este objetivo y además se utilizan diferentes métricas de detección de sesgos en los modelos. Para la experimentación se utiliza el dataset FairFace de HuggingFace y luego de un intensivo análisis del mismo nos percatamos de que este presenta falta de representación para ciertas clases en específico. Además se emplean varias métricas de sesgo como equalized odds, disparate impact, disparate impact ratio, entre otras.

1. Introducción

A lo largo de la historia, el avance tecnológico ha transformado profundamente nuestra vida cotidiana, simplificando tareas que antes requerían mucho esfuerzo y tiempo. Este progreso no solo ha hecho nuestras rutinas más ágiles y eficientes, sino que también ha abierto la puerta a logros que en otro tiempo hubieran parecido inimaginables. En este contexto, la inteligencia artificial (IA) y las técnicas de machine learning han desempeñado un papel fundamental. Estas tecnologías no solo automatizan procesos, sino que amplían las fronteras del conocimiento humano, impulsando innovaciones en campos como la medicina, la seguridad, la educación y la investigación científica. De esta forma se puede obtener información muy valiosa a partir de simples imágenes, textos, sonidos, etc.
En el contexto de las imágenes los diferentes algoritmos de Machine Learning le han permitido a empresas, hospitales, e incluso autoridades tomar decisiones importantes a partir de toda la información que reporta una simple imagen. Por la importancia que tiene para la sociedad, en este proyecto de Machine Learning hemos decidido implementar un modelo que sea capaz de estimar a través de una foto de una persona, su edad, género y etnicidad. Además tenemos como objetivo determinar si existen o no sesgos en la estimación de estos atributos, así como el impacto negativo que tienen los mismos sobre los modelos.

2. Estado del arte

Al adentrarnos en el estado del arte referente a nuestro problema, comprobamos que a pesar de existir diferentes modelos de Machine Learning cuyo objetivo es lograr una alta precisión en la estimación de edad, género y etnicidad, comprobamos que el mayor problema a la hora de entrenar un modelo en particular es encontrar un dataset cuya distribución por grupos de edad, raza y sexo sea homogénea y no existan clases con una menor representación que otras, ya que para estos grupos el modelo tiende a equivocarse con mayor frecuencia que en el caso de los grupos con una mayor representación. Según el estado del arte consultado FairFace resultó ser uno de los datasets con mayor equilibrio en cuanto a la distribución por clases y este fue el que se utilizó en este proyecto.
El resumen del estado del arte consultado se encuentra en Google Sheets cuyo link es: Estado del arte en Google Sheets.

3. Propuesta de solución

Para la estimación de la edad, el sexo y la etnicidad se han implementando 5 modelos de Machine Learning: ViT-Age-Classifier, YOLOv8 con EfficientNetB0, una red neuronal con Keras, el modelo de Fair-Face master y Vit-B-32 Clip. Cada uno de estos modelos realiza las estimaciones desde un enfoque diferente, algunos de ellos son modelos ya entrenados, especializados en estas estimaciones, mientras que otros son modelos entrenados desde cero o con ajustes de hiperparámetros. Se analizan las ventajas y desventajas de cada modelo y se analiza además la posible existencia de sesgo en ellos.

4. Análisis del dataset utilizado

Como se mencionó anteriormente para la experimentación se utilizó el dataset FairFace de HuggingFace.
FairFace, a diferencia de otros datasets como UTKFace, intenta mitigar los sesgos existentes hacia determinados grupos étnicos en los datasets relacionados con rostros de personas y lo cual provoca que la mayoría de los modelos los discriminen.
Cuenta con exactamente 108 501 imágenes faciales, tanto de mujeres como de hombres, así como 7 grupos étnicos distintos: White, Black, Indian, East Asian, Southeast Asian, Middle East, y Latino.

Cuenta además con 9 grupos de edades: ('0-2', '3-9', '10-19', '20-29', '30-39', '40-49', '50-59', '60-69', '70 y más').

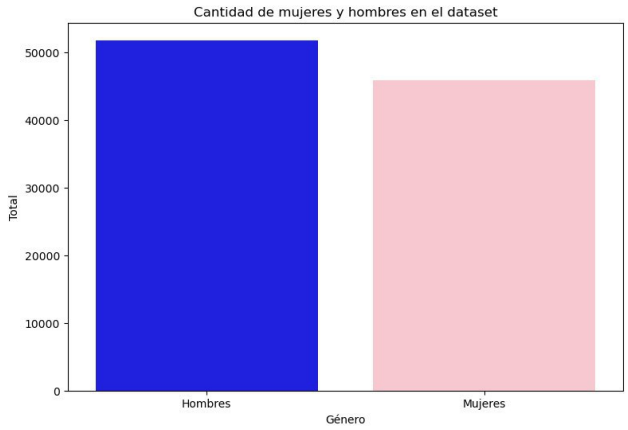

Figura 1: Hombres vs Mujeres.

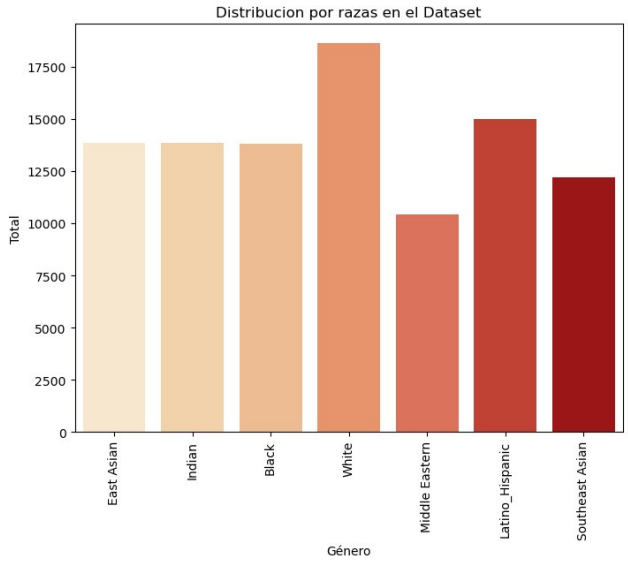

Figura 2: Distribución del dataset por grupos étnicos.

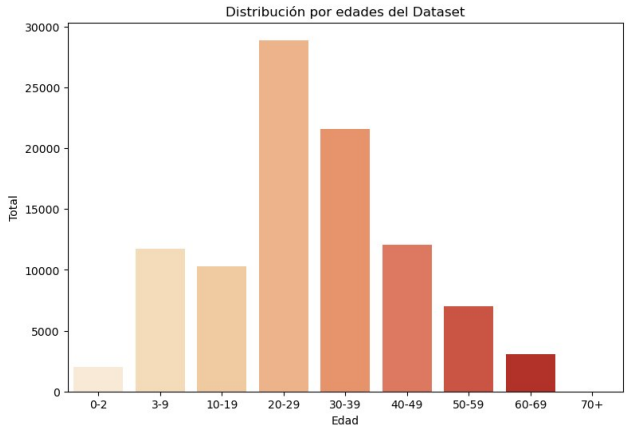

Figura 3: Edades en el dataset.

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 % del conjunto de datos, los rangos de edad superiores a 70 años representan apenas el 0.98 %.

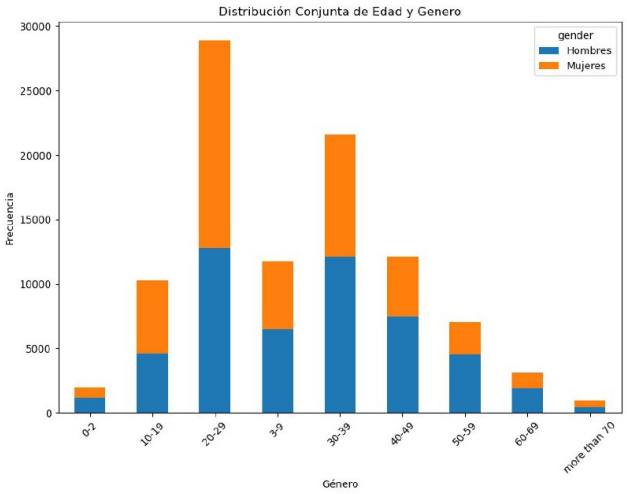

Figura 4: Distribución del dataset teniendo en cuenta sexo por edad.

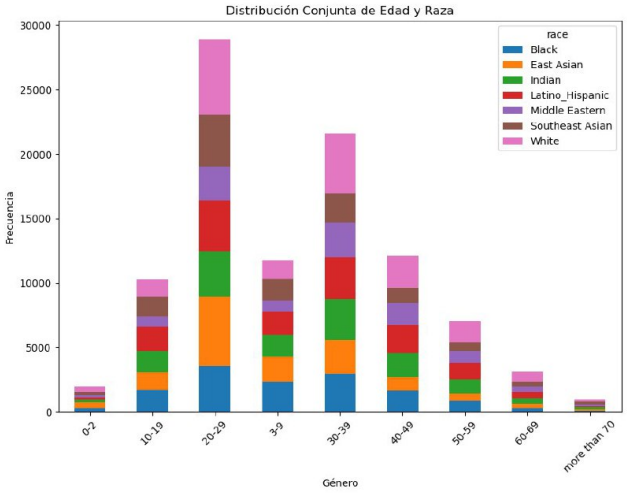

Figura 5: Distribución del dataset teniendo en cuenta el grupo étnico al que pertenecen por edad.

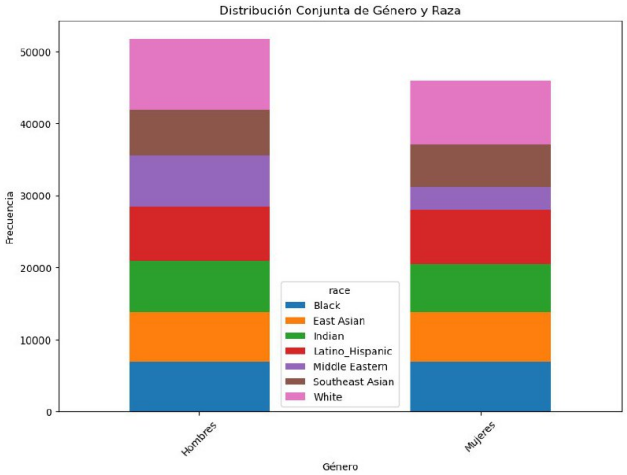

Figura 6: Distribución del dataset teniendo en cuenta el género por grupo étnico.

5. Experimentación y resultados

5.1 Modelos utilizados para la estimación

- ViT Age Clasiffier:
  Utilizamos este modelo proporcionado por nateraw en Hugging Face. Este modelo es un transformer de visión (ViT) al que se le ha realizado 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 %. Si analizamos la precisión por grupos de edad nos percatamos de 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.
  Por lo que un enfoque a seguir en este caso sería hacer un sobremuestreo en las clases con menor representación y reentrenar el modelo con estos nuevos datos.

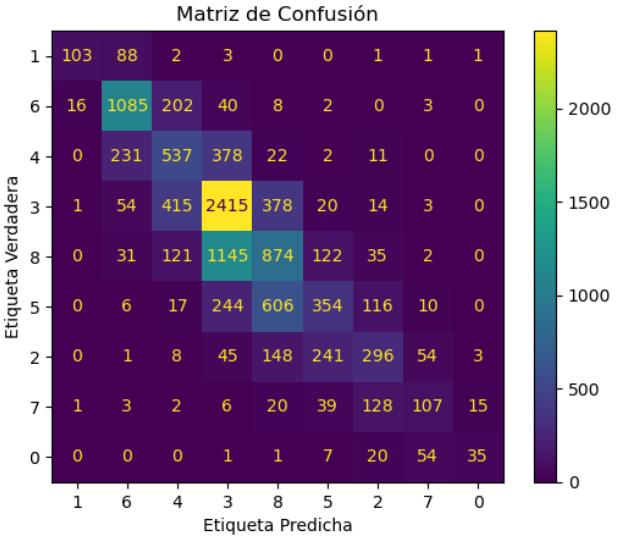

Figura 7: Matriz de confusión del modelo

- Yolov8 con Efficient-NetB0
  Para este modelo utilizamos los modelos ya entrenados de Yolo y Efficient-Net B0, de acuerdo con el estado del arte [1].
  Este modelo consta de dos modelos fusionados: se utiliza Yolov8 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 EfficientNet-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 al 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 %.

- Red neuronal con Keras
  En este modelo no solo realizamos la estimación de la edad, sino que también estimamos el género 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 tamaño 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 Conv2D, otra de MaxPooling2D, una capa de aplanamiento, una 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 predicción 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.

  Con este modelo el accuracy de 30 % para la edad y 73 % 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

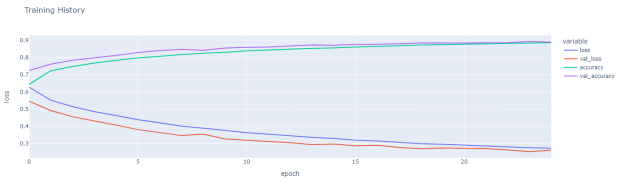

Figura 8: Historial de entrenamiento del modelo para el caso de la edad

se calcula como la cantidad de datos en el dataset entre la cantidad de datos que corresponden a su categoría.
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 accuracy de 23 %, o sea que empeora, sin embargo tiene un mejor rendimiento para el grupo de edad de más de 70 años, 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.
Las métricas en el caso del modelo de estimación de edad luego de los cambios realizados fueron las siguientes:

	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	NaN	NaN	10954
macro avg	0.35	0.17	0.16	10954
weighted avg	0.31	0.23	0.21	10954

- ViT B-32 CLIP
  Hemos utilizado este modelo para predecir la edad.
  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. 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.
  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 % para la edad, con una mayor precisión para los grupos con mayor representación(de 60 % y 59 %), mientras que el accuracy de las minorías no llega al 10 %.

- FairFace master
  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 FairFace.
  La precisión del modelo se ha evaluado en un 30 % para la edad, 70 % para el género y 24 % 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.

5.2 Detección de sesgo

Luego del análisis realizado inicialmente al dataset, nos percatamos de que el sesgo existe desde antes del procesamiento de los datos en el modelo. Teniendo en cuenta esto, hemos seguido varias estrategias de detección de sesgo tanto pre-processing como post-processing.

- Equalized Odds
  Equalized Odds es una métrica que evalúa la equi-

dad entre diferentes grupos o subpoblaciones en un sistema de clasificación. La idea principal es que la métrica debería ser igual para todos los grupos objetivo, independientemente de su distribución en el espacio de características.
Tiene tres objetivos principales:
- Equidad en la precisión positiva: La precisión de clasificar correctamente los miembros del grupo objetivo debería ser igual para todos los grupos.
- Equidad en la precisión negativa: La precisión de clasificar incorrectamente los miembros del grupo objetivo debería ser igual para todos los grupos.
- Equidad en la función de utilidad: El valor esperado de la función de utilidad para el grupo objetivo debería ser igual para todos los grupos.
Al evaluar esta métrica en nuestros modelos buscamos ver qué tan bueno es con respecto a los falsos positivos y falsos negativos, y ver si estas diferencias sobrepasan el grado de significación de 0.1.
En el caso del modelo FairFace master esta métrica arrojó que en el rango de edades de 30 a 39 años es donde menos equidad posee con respecto a los demás rangos.

- Disparate Impact
  La métrica de sesgo Disparate Impact se basa en comparar las tasas de éxito de un modelo para diferentes grupos objetivo. Se mide generalmente como la proporción de la diferencia entre las tasas de éxito para los grupos más desventajados y los mejor aprovechados.

  Esta métrica en nuestros modelos la evaluamos para todos los rangos de edad, para saber si la proporción de las tasas de éxito fueron significativamente diferentes (en este caso probamos con un grado de significación del 0.8).
  En el caso del modelo FairFace master esta métrica arrojó que en el rango de edades de 50 a 59 años es donde más diferencia en las tasas de éxito posee con respecto a los demás rangos.

- Label Bias Multi-Class
  Esta métrica se refiere al sesgo en la distribución de las clases objetivo en un problema de clasificación multi-clase. Este tipo de sesgo ocurre cuando las proporciones de las clases objetivo en el conjunto de entrenamiento son muy diferentes de las proporciones reales en el dominio del mundo real. Se mide generalmente como la diferencia entre estas dos distribuciones.

  En nuestro caso con dicha métrica queríamos comprobar que tan homogéneas eran las predicciones con respecto a la raza y el género, y que tanto discriminaban entre un grupo de edades y otro. Al evaluar esta métrica en el modelo FairFace Master, pudimos comprobar que la raza Latino Hispanic es la que posee más significación de diferencia, especialmente en el rango de edades de 30 a 39 años. En cuanto al género, en el rango de 20 a 29 es donde menos homogénea son las predicciones.

- Disparate Impact Ratio
  Disparate Impact Ratio se refiere a la proporción en la que un modelo de clasificación trata desproporcionadamente a ciertas clases objetivo en comparación con otras. Esta métrica mide el grado en el que el modelo predice una clase específica con una probabilidad diferente de lo que se esperaría si el modelo fuera equitativo.

  En nuestro caso se aplicó con respecto a raza y género para ver cuán desproporcionados estaban en cada uno.
  Al evaluar esta métrica en el modelo FairFace Master, nos dice que trata desproporcionadamente a las razas White.

- Test de Chi-cuadrado
  El Test de Chi-cuadrado es una prueba estadística no paramétrica utilizada para determinar si hay una asociación estadísticamente significativa entre variables categóricas. En este contexto, se utiliza para detectar si hay una diferencia significativa en las predicciones del modelo entre diferentes grupos sensibles (como la raza y el género).

  Al evaluar esta métrica en el modelo FairFace Master, nos dice que trata desproporcionadamente a las razas White, Black, Latino_Hispanic y Middle Eastern y a ambos sexos.

6. Conclusiones

El desarrollo de este proyecto nos ha permitido predecir la edad, sexo y género de una persona a partir de una simple imagen facial. Para ello se utilizaron diferentes modelos, algunos de ellos preentrenados y otros entrenados desde cero.
El accuracy para todos estos modelos no fue extramadamente alto, sin embargo, se comprobó que la eficiencia de los modelos no dependen solamente de sus parámetros, sino que los datos utilizados en el entrenamiento desempeñan un papel crucial en el rendimiento del modelo.
Se utilizó FairFace para la experimentación, sin embargo a pesar de ser un dataset que pretende solucionar el problema de poca representación en algunos de los grupos étnicos, se comprobó que hay determinados grupos que tiene una muy baja representación, por lo que el dataset se encuentra de por sí sesgado.
Además se utilizaron diferentes métricas de detección de sesgo, entre ellas disparate impact, label bias multiclass, equalized odds, entre otros.
Se intentó además mitigar el sesgo aplicando técnicas in-processing como la ponderación por clases y no se pudieron aplicar técnicas de oversampling o undersampling por problemas de hardware. 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 delos datos de entrada.
Por lo que se puede afirmar que la estimación de edad, raza y sexo utilizando técnicas de Machine Learning es una tarea bastante compleja; no solo a la hora de elegir correctamente un modelo y ajustarlo, sino también porque hay que tener en cuenta un conjunto de datos que mantenga una distribución bastante equitativa por clases, para garantiza que nuestro modelo aprenda de forma correcta a realizar estimaciones sin discriminar a un grupo en particular.

Referencias

[1] Giovanna Castellano, Berardina De Carolis, Nicola Marvulli, Mauro Sciancalepore, and Gennaro Vessio *Real-Time Age Estimation From Facial Images Using YOLO and EfficientNet.*