MINISTERIO DE CIENCIA E INNOVACIÓN

Cofinanciado por la Unión Europea

AGENCIA ESTATAL DE INVESTIGACIÓN

MEMORIA CIENTÍFICO-TÉCNICA DE PROYECTOS INDIVIDUALES
Convocatoria 2021 - «Proyectos de Generación de Conocimiento»

AVISO IMPORTANTE - La memoria no podrá exceder de 20 páginas. Para rellenar correctamente esta memoria, lea detenidamente las instrucciones disponibles en la web de la convocatoria. Es obligatorio rellenarla en inglés si se solicita 100.000 € o más (en costes directos).

IMPORTANT – The research proposal cannot exceed 20 pages. Instructions to fill this document are available in the website. If the project cost is equal or greater than 100.000 €, this document must be filled in English.

1. DATOS DE LA PROPUESTA – *PROPOSAL DATA*

IP 1 (Nombre y apellidos)**: Álvaro Sánchez Romero**

TÍTULO DEL PROYECTO (ACRÓNIMO): Sesgo de género en la clasificación de imágenes: El análisis

TITLE OF THE PROJECT (ACRONYM): Gender Bias in Image classification: An analysis

2. ANTECEDENTES, ESTADO ACTUAL Y JUSTIFICACIÓN DE LA PROPUESTA - *BACKGROUND, CURRENT STATUS AND JUSTIFICATION OF THE PROPOSAL*

2.1 Antecedentes

La detección y el análisis de las caras es un desafío en los problemas de visión artificial. La arquitectura tradicional para el reconocimiento de caras se basa en varias etapas. La primera, la detección de las caras en las imágenes, después la alineación de las caras basándose en puntos clave de las caras, continuando por el procesado de las caras y finalizando con extracción de características para poder por último clasificar. Algunos modelos basados en redes convolucionales profundas, unifican este proceso [1] y algunos otros, además, se centran en la clasificación de estas imágenes faciales por género [2].

Las arquitecturas de estos modelos se basan en el Transfer Learning recortando capas de un modelo preentrenado y ajustándolas a los datos con los que se va trabajar [3]. Estos modelos utilizan implementaciones de arquitecturas nuevas de CNN que consiguen mejores resultados basándose en bloques de Inception que capturan características locales para diferentes tamaños de imagen y conexiones residuales que aceleran el entrenamiento de las redes, permitiendo entrenar modelos más profundos y con mejores resultados [4, 5, 6].

Los principales datasets en relación con la predicción de género de imágenes faciales son:

- **FERET** [7] es una base de datos formada por 14126 fotos a color de caras de 1199 individuos distintos, es en general una de las bases de datos más utilizadas en el ámbito del reconocimiento de caras y en concreto en la clasificación por género [8, 9, 10, 11], debido en parte a que es históricamente relevante, al haber sido publicada originalmente en 1996, cuando las bases de datos de esta magnitud escaseaban.

Por ello, ha sido elegido como uno de los objetos de nuestro estudio. Además, existe un dataset derivado llamado GENDER-FERET, utilizado en [2], que contiene 946 imágenes de caras en blanco y negro y que en teoría está equilibrado específicamente para la clasificación por género, de manera que comprobaremos si este equilibrio funciona para evitar el sesgo, y si es así qué podemos aprender de él.

- **UTKFace** es otra base de datos de imágenes de caras muy utilizada [13, 14, 15] y que resulta interesante debido a la gran cantidad de imágenes que contiene (más de 20000) y a que estas son muy variadas, tanto respecto a los sujetos de las imágenes, que son de todos los géneros, etnias y edades, como respecto a las propias características de las imágenes, como la iluminación, el color y resolución. Utilizaremos este dataset en el trabajo debido a que la gran variedad en las imágenes añaden ruido y lo acercan más a un caso de uso real, frente a otros datasets que utilizan fotografías "de estudio".
- **FairFace** [16] es una base de datos de imágenes de caras balanceada en cuanto a género, raza y edad, utilizada en [5]. Comprobaremos si este equilibrio en varias características funciona para reducir el sesgo en la clasificación por género, y si es así qué podemos aprender de él. Además se usa t-SNE, es un algoritmo estocástico de representación de datos de alta dimensionalidad como pueden ser en nuestro caso imágenes. Se tiende a usar previamente PCA para el mejor funcionamiento del algoritmo. En la imagen se puede ver que los datos son separables en su mayoría y están bien agrupados, mientras que con otras representaciones de otros datasets que se analizan en el artículo no se produce esto.

Dentro de estos conjuntos de datos o cualquier otro pueden aparecer sesgos, es decir, condiciones que provocarán que las predicciones de un modelo de aprendizaje automático sean peores y cometan errores. Estos sesgos pueden venir por errores humanos al etiquetar o generar los datos (ruido) o por la propia naturaleza de los mismos. Dependiendo del tipo de sesgo estos podrán ser eliminados de una manera o de otra, siendo nuestro objetivo crear una metodología que permita generar un conjunto de imágenes que tengan el menor sesgo posible. Dado que queremos hacer las predicciones más justas y precisas posibles es fundamental eliminar los sesgos. Para lograr esto, el primer paso para ello es el detectarlos.

Algunos posibles sesgos o causas de sesgos son fácilmente detectables mediante una exploración de los datos. Es común tener datasets con clases muy desbalanceadas, lo que puede dar lugar a errores en la predicción. Pero hay otros sesgos que son más difíciles de detectar y no pueden ser directamente hallados mediante exploración, por ello se han creado métricas específicas para la detección de sesgos, como es el caso de la Relevance Mass Accuracy, Relevance Rank Accuracy y Area over the perturbation curve [17, 18, 19]. Estas métricas derivadas de la evaluación de mapas de atribución [20] (útiles en imágenes al explicar la influencia de los píxeles) pueden medir cómo de representativos son estos teniendo en cuenta el sesgo. Existen otras métricas para medir sesgos pero están especializadas en el derivado de algoritmos (Selection Rate [3]).

Las principales causas de sesgos debidos a los datos son:

- **Sesgo por desbalanceo de datos:** Una gran mayoría de los sesgos que se presentan en nuestros datos son debidos a la existencia de unos datos que no representan cuantitativamente la realidad o que tienen falta de representación para ciertos valores. Esta descompensación suele verse como un desbalanceo entre

clases [22, 23, 24, 25] en el que una de ellas está poco representada, generando que se tienda a predecir más la clase más representada y se obtengan malos resultados para la que tiene pocas instancias. Pese a que existen algoritmos menos sensibles al desbalanceo de clases esto sigue siendo un gran problema.

Una solución muy usada en clasificación de imágenes es la técnica de data augmentation mediante la cual se generan nuevas instancias de las antiguas a partir de rotaciones y operaciones en las anteriores imágenes. A parte de esto es muy común el uso de Generative Adversarial Networks (GANs) para la creación sintética de datos como por ejemplo imágenes de gente que no existe.

Por otro lado, no debemos olvidar que para obtener la mejor predicción posible no solo las clases deben estar bien representadas sino cualquier atributo [3, 5]. Por ejemplo, de poco sirve si nuestro dataset tiene un buen número de imágenes de hombres y mujeres si no hemos representado adecuadamente a gente de distintas etnias. Cada etnia puede tener diferentes rasgos diferenciales entre géneros por lo que un dataset en el que prácticamente solo haya gente blanca no va a funcionar bien cuando aparezca gente de otras razas.

Al margen del marco étnico debemos pensar en cualquier otro grupo menos representado, desde ancianos a gente del colectivo LGBT. Cuánto mejor representemos a estos grupos mejores serán nuestros resultados y a más países y contextos podrá ser aplicado nuestros datos.

- **Sesgo debido a técnicas de generación sintética de datos:** Pese a la utilidad de estas técnicas, principalmente GANs, tienen el inconveniente que se puede producir un sesgo entre los datos reales y los sintéticos [12]. Una opción para tratar esto es el uso de redes adversarias como FESGAN que generen nuevas identidades con distintas expresiones y a las que le añadamos un distinto tipo de back-propagation (RDBP) al clasificar que disminuya las variaciones dentro de la misma clase. Otro problema es la generación de discrepancia entre instancias de atributos por realizar data augmentation o usar GANs [15]. Usando redes siamesas podemos minimizar las discrepancias en la distribución de variables de ambas fuentes anteriormente mencionadas.
- **Sesgo por etiquetado de datos:** Al necesitar una cantidad tan ingente de datos nuestros algoritmos de aprendizaje automático, no siempre los etiquetados son correctos [21]. Estos errores previos al entrenamiento son fundamentales a evitar dado que si entrenamos nuestro modelo con datos erróneos vamos a obtener predicciones basadas en datos erróneos (Garbage in, garbage out). Estos pueden ser errores de anotación de atributos hasta clases. Hemos de considerar que los errores de etiquetado pueden ser debido a personas o a posibles etiquetados automáticos.

  Un caso importante a tener en cuenta en la predicción de género es que en prácticamente ningún dataset ni paper se tiene en cuenta que hay gente que no se identifica con ningún género [3]. Una posibilidad sería añadir un etiquetada como Otro o No binario, para funcionar de forma adecuada con esta gente.

- **Sesgo debido al ruido:** Es muy probable que en nuestro conjunto de datos tengamos datos, en este caso, imágenes con ruido. Como es normal en el mundo real no todos nuestros datos de aprendizaje van a ser perfectos o tener una definición perfecta y eso puede provocar que nuestros modelos aprendan peor. Para atajar problemas de ruido se puede usar URNet [26] que permite dar diferentes pesos a cada instancia de entrenamiento según la confianza que tengamos en estas. Esto puede servir también para atajar otros sesgos como desbalanceo de datos o problemas de etiquetado.
- **Sesgo contextual:** Este tipo de sesgo es aquel que hace que nuestras predicciones empeoren al tener en cuenta el contexto de la imagen [27, 28, 29]. Por ejemplo, si entrenamos con imágenes de hombres haciendo deporte, se producen errores al predecir mujeres que están realizando algún deporte. Debido a esta problemática se puede hacer preprocesamiento de los datos para detectar la cara o el cuerpo.

Por otro lado, al margen de los distintos sesgos derivados de los datos explicados anteriormente, también existen sesgos derivados de los modelos. Por ejemplo se pueden producir sesgos debido a la creación de particiones para el conjunto de validación [30], por underfitting al crear un modelo demasiado general que no ha aprendido lo suficiente [31] o por sesgo computacional al tener pocos datos (few-shot learning [32])... Otros sesgos de modelos están específicamente relacionados con ciertos modelos como redes neuronales convolucionales (CNNs [33, 34]) y su forma de aprender (shape bias e inductive bias [35]).

Un elemento fundamental en la Inteligencia Artificial en la actualidad, es la explicabilidad de las predicciones. Es importante saber por qué se ha realizado una predicción, o qué variables han influido positiva o negativamente por múltiples motivos. Por ejemplo, para entender mejor las predicciones y qué hace nuestro modelo, detectar sesgos, mejorar el rendimiento… Si no podemos garantizar que un modelo es justo o no tiene sesgos, su inserción en herramientas para realizar actividades automáticas, como determinar si un banco debe dar un crédito a una persona, estaríamos dando un servicio que podría ser malo. Esto es debido a que el aprendizaje automático tiene un poder enorme, no deja de ser una máquina que puede cometer errores, y la supervisión humana mediante técnicas de explicabilidad es la solución perfecta.

En la actualidad se utilizan una cantidad enorme de diferentes tipos de modelos, desde árboles de decisión a redes neuronales. En un árbol de decisión es evidente cómo han contribuido las variables en una predicción (modelos de caja blanca) mientras que en las redes neuronales no debido a su complejidad (modelos de caja negra).

Con el avance del deep learning, cada vez se usan más modelos de caja negra por lo tanto es importante hallar técnicas para entender esta clase de modelos [36]. Existen dos tipos de técnicas principales de explicabilidad:

- **Métodos agnósticos del modelo global:** son aquellos que independientemente de la entrada dan un resumen de cómo afectan los valores de los atributos a la predicción. Destacan los Partial Dependence Plots que indican cómo influye una variable a una predicción. Estos tienen el problema de que asumen independencia

de variables, para solventar esto hay otras técnicas como Accumulated Local Effects (ALE).

- **Métodos agnósticos del modelo local:** son aquellos que a partir de una predicción de un algoritmo indican cómo han influido las variables para obtener esa salida. Las dos más usadas son LIME y SHAP. La primera gracias a investigaciones se ha descubierto que es un "subconjunto" de la segunda por lo que SHAP siempre rendirá igual o mejor que LIME. El problema de SHAP es que tiene unos tiempos de computación muy elevados, y no siempre será interesante tardar mucho más para obtener un resultado muy similar a LIME.

Referencias nuevas(luego se organizan):

[1] Ranjan, R., Patel, V., & Chellappa, R. (2019). HyperFace: A Deep Multi-Task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(1), 121-135.

[2] Dwivedi, N., & Singh, D. (2018). Review of Deep Learning Techniques for Gender Classification in Images. In Harmony Search and Nature Inspired Optimization Algorithms (Advances in Intelligent Systems and Computing, pp. 1089-1099). Singapore: Springer Singapore.

[3] Wu, W., Protopapas, P., Yang, Z., & Michalatos, P. (2020). Gender Classification and Bias Mitigation in Facial Images. 12th ACM Conference on Web Science, 106-114.

[4] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2016).

[5] Krishnan, A., Almadan, A., & Rattani, A. (2020). Understanding Fairness of Gender Classification Algorithms Across Gender-Race Groups. 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), 1028-1035.

[6] Islam, M., Tasnim, N., & Baek, J. (2020). Human Gender Classification Using Transfer Learning via Pareto Frontier CNN Networks. Inventions (Basel), 5(2), 16.

[7] Phillips, P., Hyeonjoon Moon, Rizvi, S., & Rauss, P. (2000). The FERET evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(10), 1090-1104.

[8] Vallimeena, P., Gopalakrishnan, U., Nair, B., & Rao, S. (2019). CNN Algorithms for Detection of Human Face Attributes - A Survey. 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 576-581.

[9] Dammak, S., Mliki, H., & Fendri, E. (2021). Gender effect on age classification in an unconstrained environment. Multimedia Tools and Applications, 80(18), 28001-28014.

[10] Khryashchev, V., Shmaglit, L., Priorov, A., & Shemyakov, A. (2014). Extracting adaptive features for gender classification of human face images. Programming and Computer Software, 40(4), 215-221.

[11] Loo, E., Lim, T., Ong, L., & Lim, C. (2018). The influence of ethnicity in facial gender estimation. 2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA), 187-192.

[12] Yilma, G., Gedamu, K., Assefa, M., Oluwasanmi, A., & Qin, Z. (2021). Generation and Transformation Invariant Learning for Tomato Disease Classification. 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML), 121-128.

[13] Ghildiyal, A., Sharma, S., Verma, I., & Marhatta, U. (2020). Age and Gender Predictions using Artificial Intelligence Algorithm. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 371-375.

[14] Garain, A., Ray, B., Singh, P., Ahmadian, A., Senu, N., & Sarkar, R. (2021). GRA_Net: A Deep Learning Model for Classification of Age and Gender From Facial Images. IEEE Access, 9, 85672-85689.

[15] Yan, Y., Huang, Y., Chen, S., Shen, C., & Wang, H. (2020). Joint Deep Learning of Facial Expression Synthesis and Recognition. IEEE Transactions on Multimedia, 22(11), 2792-2807.

[16] Kimmo Kärkkäinen and Jungseock Joo. 2019. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. https://doi.org/10.48550/ARXIV.1908.04913

[17] Ahmed Osman, Leila Arras, and Wojciech Samek. 2020. Towards Ground Truth Evaluation of Visual Explanations. ArXiv abs/2003.07258 (2020)

[18] Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Muller, K. (2017). Evaluating the Visualization of What a Deep Neural Network Has Learned. IEEE Transaction on Neural Networks and Learning Systems, 28(11), 2660-2673.

[19] Schaaf, N., De Mitri, O., Kim, H., Windberger, A., & Huber, M. (2021). Towards Measuring Bias in Image Classification. In Artificial Neural Networks and Machine Learning – ICANN 2021 (Lecture Notes in Computer Science, pp. 433-445). Cham: Springer International Publishing.

[20] Mengjiao Yang and Been Kim. 2019. BIM: Towards Quantitative Evaluation of Interpretability Methods with Ground Truth. CoRR abs/1907.09701 (2019). arXiv:1907.09701 http://arxiv.org/abs/1907.09701

[21] Han, B., Yun, W., Yoo, J., & Kim, W. (2020). Toward Unbiased Facial Expression Recognition in the Wild via Cross-Dataset Adaptation. IEEE Access, 8, 159172-159181.

[22] Zhao, Y., Yang, J., Du, J., Chen, Z., & Yang, W. (2021). A Lightweight Classifier for Facial Expression Recognition based on Evolutionary SVM Ensembles. 2021 6th International Conference for Convergence in Technology (I2CT), 1-9.

[23] N. M. Nafi and W. H. Hsu, "Addressing Class Imbalance in Image-Based Plant Disease Detection: Deep Generative vs. Sampling-Based Approaches," 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), 2020, pp. 243-248, doi: 10.1109/IWSSIP48289.2020.9145239.

[24] Vowels, M. J., Camgoz, N. C., & Bowden, R. (2020). NestedVAE: Isolating common factors via weak supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9202-9212).

[25] Zhong, Z., Cui, J., Liu, S., & Jia, J. (2021). Improving calibration for long-tailed recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16489-16498).

[26] Li, J., Song, Y., Zhu, J., Cheng, L., Su, Y., Ye, L., . . . Han, S. (2021). Learning From Large-Scale Noisy Web Data With Ubiquitous Reweighting for Image Classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(5), 1808-1814.

[27] Hendricks, L., Burns, K., Saenko, K., Darrell, T., & Rohrbach, A. (2018). Women Also Snowboard: Overcoming Bias in Captioning Models. In Computer Vision – ECCV 2018 (Lecture Notes in Computer Science, pp. 793-811). Cham: Springer International Publishing.

[28] L. F. d. Araujo Zeni and C. Rosito Jung, "Real-Time Gender Detection in the Wild Using Deep Neural Networks," 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2018, pp. 118-125, doi: 10.1109/SIBGRAPI.2018.00022.

[29] Jiang, L., Zhang, J., & Deng, B. (2020). Robust RGB-D Face Recognition Using Attribute-Aware Loss. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10), 2552-2566.

[30] Chansong, D., & Supratid, S. (2021). Impacts of Kernel Size on Different Resized Images in Object Recognition Based on Convolutional Neural Network. 2021 9th International Electrical Engineering Congress (iEECON), 448-451.

[31] Regina Lourdhu Suganthi, S., Hanumanthappa, M., & Kavitha, S. (2018). Event Image Classification using Deep Learning. 2018 International Conference on Soft-computing and Network Security (ICSNS), 1-8.

[32] Li, Z., & Mu, K. (2021). Integrating Task Information into Few-Shot Classifier by Channel Attention. In Knowledge Science, Engineering and Management (Lecture Notes in Computer Science, pp. 137-148). Cham: Springer International Publishing.

[33] Siniosoglou, I., Argyriou, V., Bibi, S., Lagkas, T., & Sarigiannidis, P. (2021, August). Unsupervised ethical equity evaluation of adversarial federated networks. In The 16th International Conference on Availability, Reliability and Security (pp. 1-6).

[34] Aka, O., Burke, K., Bäuerle, A., Greer, C., & Mitchell, M. (2021). Measuring Model Biases in the Absence of Ground Truth.

[35] Malhotra, G., Evans, B., & Bowers, J. (2020). Hiding a plane with a pixel: Examining shape-bias in CNNs and the benefit of building in biological constraints. Vision Research (Oxford), 174, 57-68.

[36] Molnar, C. (2020). Interpretable machine learning. Lulu. com.

2.2 Estado actual

En la actualidad se están tratando de desarrollar nuevas técnicas en el campo de la Inteligencia Artificial Explicable, tanto globales como locales. Se está tratando de aplicar a cualquier tipo de datos desde estructurados a no estructurados. Al igual que se utilizan mapas de atribución de los píxeles de una imagen a una predicción, se está también investigando cómo aplicar esto a vídeos para obtener una mayor estabilidad en la contribución de cada píxel. Por la parte de los sesgos hemos detectado que se está tratando de desarrollar datasets de calidad como FairFace, que tengan un menor grado de sesgo y así obtener mejores resultados. Lo que hemos notado es que falta un análisis genérico de los sesgos presentes en datos y algoritmos además de técnicas o metodologías para reducirlos.

Además, puesto que, a pesar de su potencial, la Inteligencia Artificial Explicable es un campo aún en su infancia, no hemos podido encontrar ninguna investigación que trate de aplicarla a problemáticas relacionadas con el sesgo de género.

2.3 Justificación de la propuesta

En los últimos 20 años ha existido una tendencia creciente en cuanto a la cantidad de información generada que no parece que vaya a atenuarse en un futuro próximo. En la actualidad, la cantidad de datos generados es demasiado grande como para ser procesados manualmente, esto hace necesario el uso de herramientas que sean capaces de automatizar estos procesos y aquí es donde cobra importancia la inteligencia artificial y más concretamente el aprendizaje automático. Este proceso de automatización debe tener un alto nivel de confianza para asegurar la buena utilización de los datos, consiguiendo algoritmos que recreen de manera fiel la realidad, evitando errores que puedan tener consecuencias graves en los distintos campos en los que se utilicen, a su vez, se deben evitar sesgos que puedan provocar desigualdades en la toma de decisiones.

Estudiando el estado del arte, observamos que hay una gran cantidad de causas posibles que crean sesgo en el reconocimiento de imágenes faciales, interfiriendo con el objetivo de los modelos de aprendizaje automático, haciendo que se obtengan peores resultados y que haya diferentes minorías infrarrepresentadas que no recibirán un trato justo a la hora de usar estos algoritmos.

Teniendo esto en mente, nuestros objetivos son, que las diferentes aplicaciones que se realicen en un futuro que utilicen reconocimiento facial consigan mejores resultados en cualquier tipo de población, centrándonos especialmente en el sesgo de género. Para ello, podemos enumerar tres hipótesis:

Hipótesis 1: Es posible mejorar las bases de datos de imágenes faciales usadas en el reconocimiento facial evitando discriminación.

Hipótesis 2: Es posible mejorar los algoritmos de reconocimiento de imágenes de manera que eviten sesgos.

Hipótesis 3: Es posible explicar las decisiones que toman los algoritmos de aprendizaje automático, permitiendo así entender los errores con la intención de conseguir soluciones más óptimas.

Para demostrar estas hipótesis dividiremos el proyecto en tareas diferenciadas que se explicarán con más detalle más adelante pero que podemos resumir aquí las tres que se centran en investigación y desarrollo:

1. En primer lugar llevaremos a cabo un estudio de los datos utilizados hasta ahora y obtendremos las claves, tanto positivas como negativas, de estos datasets, permitiéndonos realizar un informe con pautas para la creación de datasets con el menor nivel de sesgo posible y creando, al final de esta etapa, un dataset usando nuestra propia metodología.
2. En segundo lugar realizaremos lo mismo que en el apartado anterior, pero centrándonos en los métodos de reconocimiento facial, creando un informe con una metodología que permita mejorar los algoritmos que se utilicen y de nuevo, creando un algoritmo propio que compararemos con los mejor posicionados en la actualidad.
3. Por último, asumiendo que las pautas que demos obtengan una fiabilidad alta pero sin llegar al 100%, desarrollaremos una herramienta de inteligencia artificial explicable que permita entender cuáles son las causas que hacen que las predicciones del algoritmo sean erróneas. Esto posibilitará analizar si el error es debido a algún tipo de sesgo dando la posibilidad de mejorar la aplicación.

Vistos los problemas que se intentan resolver con este proyecto, lo enmarcamos bajo la prioridad temática “CULTURA, CREATIVIDAD Y SOCIEDAD INCLUSIVA”, en concreto el uso de las metodologías antes explicadas permitirá la creación de aplicaciones mucho más inclusivas y menos injustas con la ciudadanía. En la tabla 1 observamos matriz DAFO podemos ver nuestros puntos fuertes y débiles.

Fortalezas	**Debilidades**
● Incorporación de persona experta en sociología al equipo ● Gran formación en reconocimiento de imágenes y biométrico ● Estudio profundo del estado del arte en reconocimiento de imágenes y sesgos ● Capacidad formativa ● Disminución de la opacidad de los algoritmos de aprendizaje automático	● Equipo de trabajo muy homogéneo ● Leyes de protección de datos restrictivas ● Poca experiencia en investigación del equipo ● Literatura escasa en explicabilidad de los algoritmos
Oportunidades	**Amenazas**
● Creciente necesidad de una sociedad más inclusiva ● Apuesta por la automatización de procesos en la actualidad ● Incremento de financiación privada ● Reutilización de datasets públicos ● Poner a la universidad de vanguardia en XAI	● Incurrir en violaciones de la Ley General de Protección de Datos ● Posible aparición de competencia durante el desarrollo debido a la duración del proyecto

Tabla 1. Matriz DAFO del proyecto

3. OBJETIVOS, METODOLOGÍA Y PLAN DE TRABAJO - *OBJECTIVES, METHODOLOGY AND WORK PLAN*

Los objetivos del proyecto se dividen en tres categorías:

1. Estudiar y proponer una solución en el problema derivado por sesgo producido debido a los datos en los algoritmos de reconocimiento de imágenes enfocados a caras, en concreto en el reconocimiento de género.
    1.1. Estudiar por qué se produce el sesgo de género en el reconocimiento de imágenes faciales debido a los datos, en crudo y tras un preprocesado de los mismos.
    1.2. Crear pautas para la generación de un dataset centrado en el reconocimiento de personas que minimice el sesgo de género.
    1.3. Obtener un dataset de personas balanceado en género, que trate personas de diferentes edades, orígenes étnicos y daños estéticos.
2. Estudiar y proponer una solución en el problema derivado por sesgo producido debido a los algoritmos de aprendizaje automático de reconocimiento de imágenes enfocados a caras, en concreto en el reconocimiento de género.
    2.1. Estudiar porque se produce el sesgo de género en el reconocimiento de imágenes faciales debido al aprendizaje de los algoritmos.
    2.2. Crear pautas para el entrenamiento de un algoritmo de reconocimiento del género de las personas que minimice el sesgo de género.
    2.3. Obtener un algoritmo pre-entrenado de reconocimiento del género de las personas que minimice el sesgo de género.
3. Estudiar y desarrollar una herramienta que mediante técnicas de inteligencia artificial explicable ayude a entender la predicciones de los algoritmos de reconocimiento de imágenes y observar si existe un sesgo en los mismos.
    3.1. Estudiar el estado del arte de la inteligencia artificial explicable en los algoritmos de reconocimiento de imágenes.
    3.2. Desarrollar una herramienta que mediante técnicas de inteligencia artificial explicable permita reconocer en qué partes, bien del dataset o bien de los algoritmos se produzca un sesgo.
    3.3. Estudiar el sesgo en diferentes conjuntos de datos de imágenes faciales mediante la herramienta, para problemas de detección de género.
    3.4. Estudiar el sesgo en diferentes algoritmos de reconocimiento de imágenes mediante la herramienta, para problemas de detección de género.

Para satisfacer los objetivos anteriores se ha diseñado un plan de trabajo de cinco paquetes.

- Paquete de trabajo 1: Análisis de los datos. Este paquete de trabajo se relaciona con el objetivo 1 y sus objetivos específicos. Así mismo este paquete de trabajo se divide en tres tareas, cada una relacionada con uno de los objetivos específicos:
    - Tarea 1.1: Estudio del estado del arte de los sesgos en los datos. Esta tarea dará como entregable el informe del estado del arte de los sesgos en los datos.
    - Tarea 1.2: Diseño de una metodología para generación de datasets no sesgados. Que dará como entregable el informe sobre metodología de la creación de un dataset no sesgado.
    - Tarea 1.3: Generación del dataset siguiendo la metodología. Que dará como entregable un dataset de imágenes faciales que minimice el sesgo de género.

- Paquete de trabajo 2: Análisis de los algoritmos. Se relaciona con el objetivo 2 y sus objetivos específicos. Está dividido en tres tareas, cada una relacionada con un objetivo específico.
    - Tarea 2.1: Estudio del estado del arte de los sesgos producidos en los algoritmos de reconocimiento de imágenes. Cuyo entregable es el informe del estado del arte de los algoritmos de reconocimiento de imágenes y los sesgos producidos en los mismos.
    - Tarea 2.2: Diseño de una metodología de un algoritmo de reconocimiento de imágenes que minimice el sesgo. Que dará como entregable el informe sobre metodología de la creación de un algoritmo de reconocimiento de imágenes que minimice el sesgo.
    - Tarea 2.3: Generación del algoritmo de reconocimiento de imágenes faciales sin sesgo que detecte el género, siguiendo la metodología. Cuyo entregable será el algoritmo desarrollado.
- Paquete de trabajo 3: Desarrollo de herramienta para la explicabilidad. Se relaciona con el objetivo 3 y está dividido en 2 tareas, la primera relacionada con el objetivo específico 3.1 y la segunda con los objetivos 3.2, 3.3 y 3.1.
    - Tarea 3.1: Estudio del estado del arte y la explicabilidad en la predicción de los algoritmos. Esta tarea dará como entregable el informe del estado del arte y la explicabilidad en la predicción de los algoritmos.
    - Tarea 3.2: Implementación de una herramienta que analice la explicabilidad de los algoritmos. Así mismo esta tarea está dividida en dos subtareas:
        - Tarea 3.2.1: Desarrollo de la herramienta. Que dará como entregable una herramienta que analice la explicabilidad de la predicción de diferentes algoritmos que se le pasen.
        - Tarea 3.2.2: Estudio de datos. Que dará como entregable un informe del estudio de diferentes datasets utilizando la herramienta, así como del dataset desarrollado en la tarea 1.3.
        - Tarea 3.2.3: Estudio de algoritmos. Que dará como entregable un informe del estudio de diferentes algoritmos utilizando la herramienta, así como del algoritmo desarrollado en la tarea 2.3.
- Paquete de trabajo 4: Difusión y Comunicación. Este paquete de trabajo es el encargado de la difusión, comunicación y explotación del proyecto y se divide en tres tareas.
    - Tarea 4.1: Planificación de difusión y comunicación. Tiene como entregable el plan de difusión y el plan de comunicación.
    - Tarea 4.2: Difusión Web y redes sociales. Tiene tres subtareas la difusión inicial, la difusión intermedia y la difusión final y tiene como entregable la web y las redes sociales donde se realizará la comunicación.
    - Tarea 4.3: Estrategia de transferencia. Tiene tres subtareas, la transferencia del paquete de trabajo 1, la transferencia del paquete de trabajo 2 y la transferencia del paquete de trabajo 3. Su entregable es el plan de explotación.
- Paquete de trabajo 5: Gestión de proyecto. Este paquete de trabajo es el encargado de la gestión del proyecto como su nombre indica y está dividido en tres tareas.
    - Tarea 5.1: Planificación del proyecto. Que incluye tres subtareas: el estudio de la gestión de riesgos, el estudio de la gestión de calidad y el estudio de la gestión de documentación, cada una con un entregable que es el informe de dicho estudio. Por último el entregable de la tarea es el plan de proyecto.
    - Tarea 5.2: Seguimiento y control técnico. Su entregable será el informe de seguimiento y control técnico.
    - Tarea 5.3: Seguimiento y control económico. Su entregable será el informe de seguimiento y control económico.

Este plan de trabajo consta de 745 días laborables, que es un total de 1089 días naturales y teniendo en cuenta que tres años son 1095 días naturales. Teniendo en cuenta que la jornada laboral es de lunes a viernes 40 horas semanales se puede concluir que está pensado con una duración de tres años. En la figura 1 se puede apreciar el diagrama de Gantt del proyecto donde en rojo se remarcan las tareas críticas.

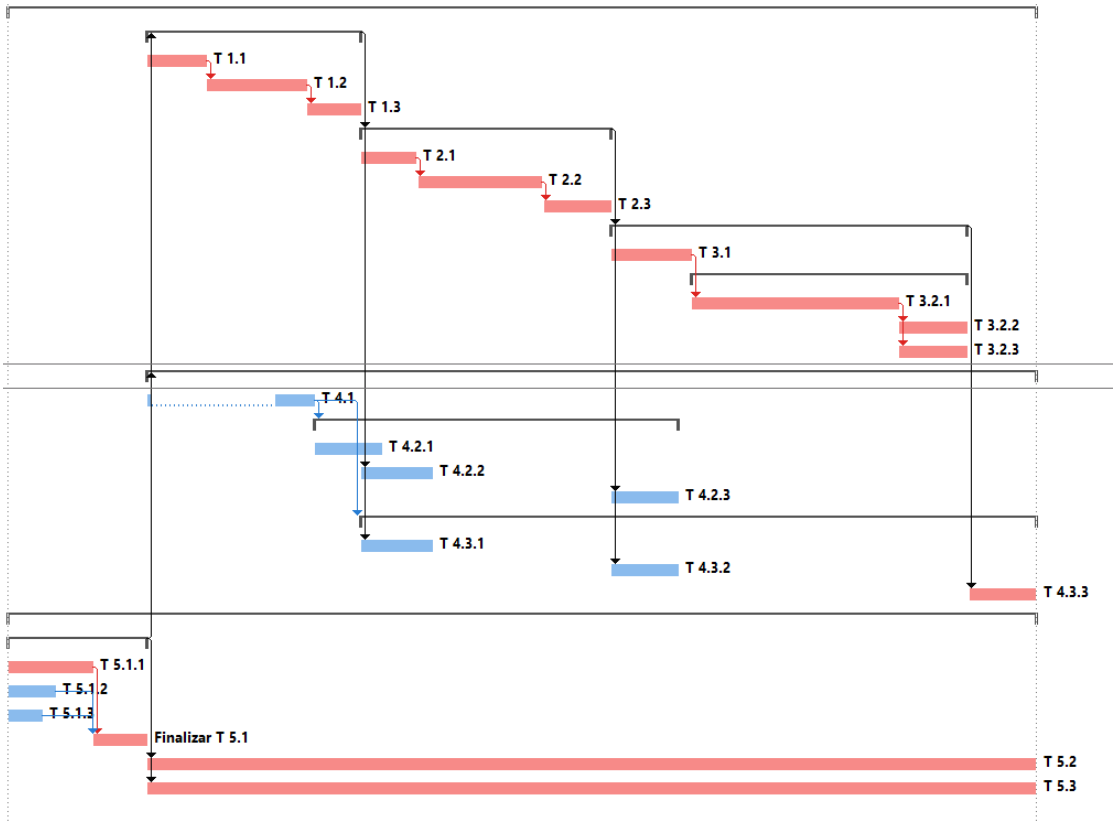

Figura 1. Cronograma del proyecto

Se definen cuatro hitos a lo largo de todo el proyecto con sus respectivas revisiones. Estos se definen al final de cada paquete de trabajo comprobando que se han conseguido los objetivos fijados para dicho paquete de trabajo y tras el desarrollo de la herramienta para la explicabilidad tras la tarea 3.2.1.

Para la investigación se va a seguir la metodología del método científico. El método científico consta de siete etapas. Las siete etapas del método científico entonces son: 1º antecedentes; 2º duda, pregunta; 3º hipótesis; 4º diseño experimental; 5º experimentación; 6º Análisis de datos; 7º conclusiones. Este se va aplicar tres veces en cada uno de los paquetes de trabajo. En cada paquete se va a llevar a cabo las dos primera etapas durante la primera tarea de cada paquete de trabajo y las siguientes cinco etapas en la segunda tarea de cada paquete de trabajo.

Durante el desarrollo de la tarea 3.2.1 para el desarrollo de la herramienta, se va seguir la metodología ágil SCRUM, siendo el Product Owner el IP del proyecto, Álvaro Sánchez Romero, el rol de Scrum Master lo desempeñará Salvador Martín Barcia y los miembros del Scrum Team serán Andoni Aizpuru Andrés, Nicolás Serrano Salas y el doctorando realizando la tesis.

No obstante si dividimos la capacidad del equipo investigador tenemos como Investigador Principal, a Álvaro Sánchez Romero; como Director Técnico, a Salvador Martín Barcia; como Director de Gestión, a Nicolás Serrano Salas y como Líder del paquete de trabajo 4, Andoni Aizpuru Andrés. Los cuatro investigadores miembros del Área de Ciencias de la

Computación e Inteligencia Artificial (CCIA) ejecutan simultáneamente y de manera coordinada todas las tareas detalladas en el cronograma. Así mismo y como se explica en el apartado 6 se incorporará al equipo de investigación un estudiante de doctorado a la investigación mediante un contrato predoctoral que acompañará a los investigadores del equipo así como seguirá el programa de formación previsto. Por último para las tareas 1.2 y 3.2.2 se contará con una persona con un doctorado en sociología para el análisis y estudio de los datos. Su objetivo es dar un punto de vista experto respecto a si se tiene una buena representación de los mismos o si se está teniendo algún tipo de sesgo de etiquetado.

El plan de trabajo se llevará a cabo completando cada uno de los tres primeros paquetes de trabajo secuencialmente tras acabar la tarea 5.1. Como observamos en el diagrama de Gantt de la figura 1 y en el diagrama PERT de la figura 4 existe un camino crítico empezando por la tarea 5.1 y continuando por los paquetes de trabajo 1, 2 y 3 terminando en la tarea 4.3.3 que es la transferencia de la tarea 3. Aunque se observa también un camino crítico dependiente de las tareas 5.2 y 5.3 en estas no es necesario crear un plan de contingencia pues son realmente tareas de seguimiento del proyecto.

El riesgo principal que se puede encontrar en el proyecto, con una gravedad alta pero con una baja probabilidad de ocurrencia es el retraso en alguna de las tareas del camino crítico. En el caso de que alguna de las tareas se retrase se tiene una semana de holgura antes de la finalización de los tres años para poder recuperarse. Si el retraso es mayor de una semana se puede tomar tiempo de las 10 semanas dedicadas a la transferencia del paquete de trabajo 3, o desarrollarla paralelamente mientras se finalizan la tareas 3.2.2 y 3.2.3 pues la transferencia de dicho paquete se puede empezar realmente finalizada la tarea 3.2.1 aunque no es lo óptimo.

Durante el desarrollo del proyecto algunos recursos hardware y software deberán ser financiados. Una estación de trabajo de 1.650€ para el estudiante de doctorado y cinco ordenadores portátil de 800€ que serán utilizados por los investigadores, el estudiante y la persona experta en sociología y durante los viajes a diferentes congresos. A estos seis equipos se les deberá añadir el coste del entorno de desarrollo integrado con un costo de 1.100€ cada uno. Así mismo se alquilará un equipo de computación en la nube de altas prestaciones por un precio de 1.756,16€ al mes durante las tareas 1.2, 1.3, 2.2 y 2.3 para el cómputo y entrenamiento de los distintos algoritmos y datos.

Como se ha indicado anteriormente se realizará un contrato predoctoral durante los tres años del proyecto con un coste de 27.000€ al año y un contrato de una persona experta en sociología con un coste de 30.000€ al año. El número total de horas que trabajarán tanto de los investigadores como de las personas y servicios que serán contratados se puede observar en la figura 2.

En relación con gastos derivados de la propiedad intelectual se destinarán 6.000€ para la suscripción a bases de datos de artículos científicos y compra de artículos científicos. Además se destinarán 5000€ para gastos de adquisición de conjuntos de datos o imágenes complementarias.

Se dedicarán 1.000€ por persona para cada uno de los viajes previstos. Se realizarán cuatro viajes donde acudirán un investigador y el estudiante de doctorado para la presentación de proyectos en congresos y dos viajes a conferencias donde viajarán los cuatro investigadores acompañados del estudiante. También se dedicarán 3.500€ al mes para estancias en el extranjero y 2.800€ al mes para estancias nacionales del estudiante doctoral duplicando la ayuda otorgada en «BOE» núm. 96, de 22/04/2021.

Por último se presupuesta 50€ para gastos de cartelería en la presentación de artículos en congresos así como 1000€ para la compra de cualquier otro material inventariable.

Nombre	Comienzo	Fin	Trabajo restante
IP	jue 02/06/22	lun 19/05/25	5.960 horas
Investigador 1	jue 02/06/22	lun 19/05/25	5.960 horas
Investigador 2	jue 02/06/22	lun 19/05/25	5.960 horas
Investigador 3	jue 02/06/22	lun 19/05/25	5.960 horas
Doctorando	jue 02/06/22	vie 07/03/25	5.560 horas
Sociologa	mar 27/12/22	vie 07/03/25	1.720 horas
Computacion en la Nube	mar 27/12/22	lun 26/02/24	1.723,2 horas
Estancia Nacional	jue 08/06/23	vie 04/08/23	320 horas
Estancia Internacional	mar 27/02/24	mar 21/05/24	480 horas

Figura 2. Horas de trabajo de los diferentes recursos

El coste total del proyecto teniendo en cuenta los costes de subcontratación, costes de instrumental y material inventariable, gastos derivados de la propiedad intelectual y gastos derivados de viajes y suministros es de un total de 174.444,61€ la división total se puede encontrar en formato tabular en la tabla 2.

Recurso	**Precio**	**Cantidad**	**Total**
Doctorando	27.000€/a	5.560 horas	72.173,08€
Sociólogo	30.000€/a	1.720 horas	24.807,69€
Computación nube	1.756,16€/ms	1.732,2 horas	18.913,84€
Estancia Nacional	2.800€/ms	320 horas	5.600€
Estancia Internacional	3.500€/ms	480 horas	10.500€
Estación de trabajo	1.650€/u	1u	1.650€
Portátil	800€/u	5u	4.000€
IDE	1.100€/u	6u	6.600€
Artículos	6.000€	1	6.000€
Imágenes	5.000€	1	5.000€
Viajes	1.000€/p	18p	18.000€
Cartelería	50€/u	4u	200€
Otros	1.000€	1	1.000€
Total			174.444,61€

Tabla 2. Precio y unidades de cada recurso

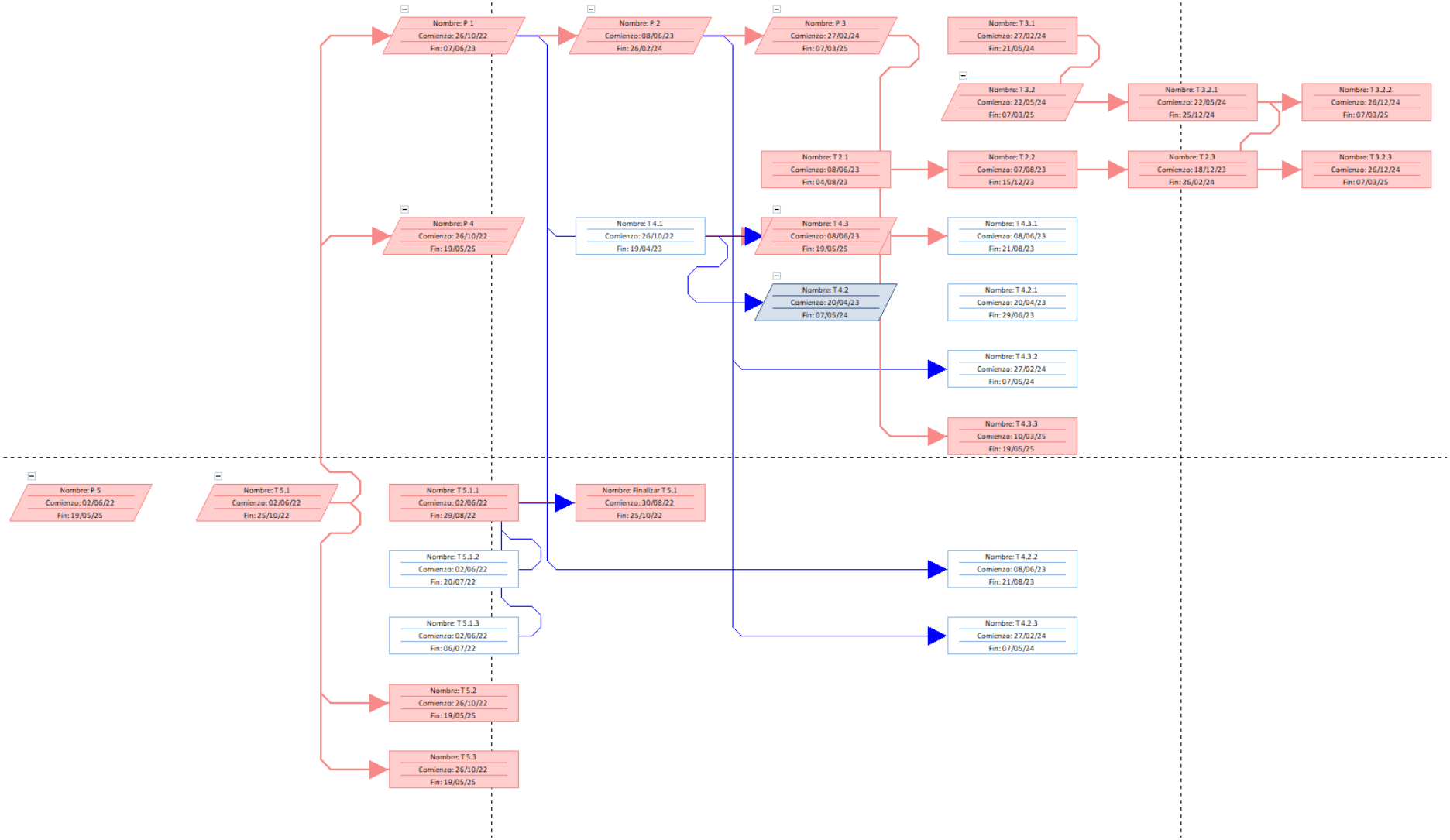

Figura 3. Diagrama PERT del proyecto

4. IMPACTO CIENTÍFICO-TÉCNICO - *SCIENTIFIC-TECHNICAL IMPACT*

Los resultados de este proyecto tendrán un gran impacto en la comunidad investigadora, tanto en los ámbitos más directamente relacionados con el reconocimiento y clasificación de imágenes, como en ámbitos relacionados con la sociología y la ética, donde la relación entre las nuevas tecnologías y el ámbito social es un tema de creciente interés. La creación de un dataset equilibrado de forma que del sesgo en la clasificación de imágenes por género se vea notablemente reducido puede ser de utilidad para investigaciones futuras, además, la creación de unas pautas específicas para generar datasets de estas características permitirá ampliar nuestro dataset equilibrado y podría servir como punto de partida para iniciativas similares que traten otros casos de sesgo por pertenencia a un grupo social, como la creación de datasets insesgados para clasificar imágenes de personas por etnia. Por otro lado, el estudio de formas de reducir el sesgo introducido por los algoritmos en la clasificación, además de mejorar también la calidad y la equidad de la clasificación por género, puede servir como base para resolver problemas de sesgo en otros casos en los que se utilicen algoritmos similares. En esta misma línea, la última fase del proyecto, que consistirá en utilizar inteligencia artificial explicable (XAI), se analizará qué decisiones llevan a los algoritmos a generar el sesgo, lo cual de nuevo no solo tendrá un impacto directo en la comprensión y reducción del sesgo de género y podrá servir como base para investigaciones sobre otros sesgos similares, sino que además podrá ser de utilidad para futuras investigaciones que comparen los procesos de creación de sesgos por el cerebro humano y por la inteligencia artificial.

El plan de difusión se fundamentará en al menos tres artículos en revistas de prestigio especializadas en Aprendizaje Automático y Procesamiento de Imágenes: un primer artículo que estudie el sesgo en datasets y presente un nuevo dataset equilibrado; un segundo artículo que estudie y trate de solucionar el sesgo proveniente de los algoritmos; y un tercero que se centre en la investigación referente a XAI. Los artículos se publicarán en revistas JCR de nivel Q1 (SCImago Journal Rank (SJR)), como Foundations and Trends in Machine Learning y International Journal of Computer Vision. Además, se planea realizar al menos un artículo en revistas especializadas en el impacto social de las nuevas tecnologías que trate en general el sesgo de género y nuestras propuestas para solucionarlo. Conforme el proyecto se desarrolle estos artículos se verán complementados por otras publicaciones referentes por ejemplo a aplicaciones prácticas de nuestros resultados y estudios del estado del arte de los distintos aspectos del proyecto, para completar un total de al menos seis artículos (dos por año). La difusión se verá apoyada por conferencias y ponencias en espacios especializados tanto en el ámbito más relacionado con el género como en el más relacionado con el Aprendizaje Automático y el Procesamiento de Imágenes. Se asistirá a conferencias del máximo nivel (Ránking A+, CORE) como IEEE International Conference on Computer Vision, ACM Multimedia Conference, International Conference on Very Large Databases y National Conference of the American Association for Artificial Intelligence. Además, siendo el tema a tratar de especial interés social, también se realizarán charlas de carácter divulgativo en eventos sobre temáticas sociales orientados al público general, lo cual ayudará a aumentar la concienciación de la ciudadanía al respecto.

5. IMPACTO SOCIAL Y ECONÓMICO - *SOCIAL AND ECONOMIC IMPACT*

Como hemos mencionado en la introducción, una de las motivaciones principales de nuestro proyecto es analizar y mejorar la predicción de género en imágenes, reduciendo tanto sesgos de datos como de los modelos. Esta mejora es posible gracias al uso de metodologías para generar conjuntos de datos de entrenamiento con el menor sesgo posible además de usar el estado del arte en métodos de clasificación de imágenes. Esto lógicamente va a tener un efecto positivo en la clasificación, mejorando los resultados

obtenidos por otras técnicas y conjuntos de datos. El proyecto atraerá a inversores externos suponiendo así un beneficio económico para la universidad. Por si fuera poco, al tratarse de un proyecto tan internacional y ambicioso que pretende eliminar los sesgos debido a la falta de representación de diferentes etnias o colectivos, muchas fuentes diferentes de distintas localizaciones del planeta estarán interesadas.

La llegada de inversiones provocará que se pueda generar empleo para profesionales que continúen con la investigación y mejoras del proyecto, como del mantenimiento de la herramienta para la explicabilidad del sesgo de datos. Además, esto podrá generar empleo de forma externa a la universidad, al aparecer más proyectos relacionados con la ciencia de datos en la administración pública, que puedan ser auditados y con garantías de que realizarán predicciones de calidad y justas. Al aparecer más proyectos así en nuestra administración los ciudadanos nos veremos beneficiados por las mejoras que aporta la inteligencia artificial a nuestras vidas, con la garantía de que el uso de esta será ético y responsable.

Desde el punto de vista social nuestro proyecto ayudará a que colectivos menos favorecidos como mujeres, etnias menos representadas, gente con discapacidad, tengan las mejores predicciones posibles. En la actualidad estos sectores de población reciben peores resultados, lo cual es totalmente injusto y dificulta la aplicación de sistemas con inteligencia artificial en nuestras administraciones. Mediante nuestro proyecto se agilizaría con seguridad la incorporación del aprendizaje automático en entes públicos. Además, estos sesgos que reducen la igualdad se verían mermados, generando así el proyecto un impacto positivo en la igualdad de género y racial de nuestro país. También consideramos que la realización de proyectos para la integración en perspectiva de género fomentará la aparición de nuevas iniciativas e ideas que se retroalimentan, fomentando así el desarrollo en estas áreas.

Para garantizar aún más que las predicciones son correctas y se puedan auditar los modelos añadimos a nuestros sistemas predictivos las técnicas más avanzadas de inteligencia artificial explicable (XAI) que ayudarán a entender a cualquier persona sin expertitud de la materia los motivos por los que se ha tomado una predicción o por qué una predicción es sesgada. Esto se realizará mediante mapas de calor que indiquen la atribución de los píxeles a una clasificación.

6. CAPACIDAD FORMATIVA - *TRAINING CAPACITY*

A causa de la gran envergadura del proyecto se incorporará un alumno de doctorado a la investigación mediante un contrato predoctoral. De esta manera, será necesaria una plaza en la convocatoria de ayudas para contratos predoctorales para la formación de doctores del siguiente año. Este contrato predoctoral será para la formación de personal investigador (FPI). Se planea que el alumno realice su doctorado en la Universidad Autónoma de Madrid (UAM), específicamente en la Escuela Politécnica Superior. El doctorado realizado se ajusta al programa de Doctorado de Ingeniería Informática y Telecomunicaciones ofrecido por la facultad.

El doctorando deberá tener conocimientos avanzados en Ciencia de Datos, específicamente en el área del Deep Learning (Aprendizaje Profundo) orientado a imágenes. A este estudiante se le requerirá experiencia trabajando con algoritmos y técnicas de Inteligencia Artificial Explicable, para ofrecer el mejor apoyo en esa fase de la investigación. Dado que la investigación está centrada en dos áreas diferentes, Inteligencia Artificial Explicable y Deep Learning, se colaborará tanto con el grupo de Aprendizaje Automático como el de BidaLab.

El doctorado realizado por el estudiante estará centrado especialmente en el desarrollo de técnicas para la explicabilidad de algoritmos de clasificación de imágenes y sus sesgos. La tesis será dirigida por el doctor Álvaro Sánchez Romero, mientras que el tutor será Nicolás Serrano Salas. Ambos forman parte del Área de Ciencias de la Computación e Inteligencia Artificial (CCIA). Todos los años se realizarán revisiones del trabajo realizado por el doctorando. Estas revisiones involucrarán los informes realizados por el director y el tutor de tesis y las actividades formativas realizadas por el estudiante, que deberán ser recogidas tras su realización para la supervisión de la investigación.

La oferta del contrato de doctorado estará disponible en el portal de empleabilidad de la Universidad Autónoma de Madrid (https://www.uam.es/uam/investigacion/ofertas-empleo) siguiendo un proceso de selección para la contratación del mejor candidato o candidata.

6.1. Plan de entrenamiento

El plan de entrenamiento del doctorando constará de las siguientes actividades formativas:

- Asistencia a seminarios y conferencias sobre las dos áreas principales del proyecto, la Inteligencia Artificial Explicable y el Deep Learning. Estas podrán ser tanto a nivel nacional como internacional y ayudarán al desarrollo del alumno y su investigación, además de los posibles contactos que el alumno pueda forjar. Se espera que esto mejore las capacidades comunicativas del mismo en el ambiente científico.
- Colaboración en el grupo de investigación gracias a los posibles conocimientos obtenidos mediante su investigación y participación en charlas y talleres. Se espera que el papel del investigador vaya desde el etiquetado de datos y preprocesamiento de los mismos, ayudado por la persona experta en sociología anteriormente mencionada, hasta el desarrollo de modelos de redes neuronales profundas para la predicción de género. También colaborará en el desarrollo de la metodología para la creación de conjuntos de datos sin sesgo y el desarrollo de técnicas de explicabilidad para la clasificación de género.
- Elaboración de artículos científicos en las áreas de investigación mencionadas para su posterior publicación en revistas científicas. Se realizará en colaboración con el resto de integrantes del grupo y se espera como mínimo dos publicaciones anuales. Los avances relacionados con XAI y sesgos serán publicados journals como Foundations and Trends in Machine Learning mientras que los relacionados con Deep Learning irán a otros como International Journal of Computer Vision. Estos journals son de nivel Q1.
- Presentación en congresos a nivel nacional e internacional de los avances logrados en el proyecto junto al resto del equipo.
- Realización de prácticas en grupos de investigación de otras universidades punteras en las áreas de la Inteligencia Artificial Explicable, Deep Learning…

6.2. Contexto científico técnico y formativo

- **Álvaro Sánchez** es profesor asociado en el departamento de Ingeniería Informática de la UAM y pertenece al grupo de Aprendizaje Automático. Ha realizado investigaciones en otras universidades europeas como la University of Bergen (6 meses) o la University of Bristol (4 meses).
- **Nicolás Serrano** es profesor asociado en el departamento de Ingeniería Informática de la UAM y pertenece al grupo de Aprendizaje Automático. Realizó su doctorado en el Massachusetts Institute of Technology en colaboración con la Universidad Autónoma de Madrid.

- **Salvador Martín** es profesor asociado en el departamento de Ingeniería Informática de la UAM y pertenece al grupo BidaLab. Ha realizado investigaciones en otras universidades europeas como la Universidad Paris-Saclay (5 meses) o la Universidad Pompeu Fabra (5 meses).
- **Andoni Aizpuru** es profesor asociado en el departamento de Ingeniería Informática de la UAM y pertenece al grupo BidaLab. Trabajó como investigador postdoctoral en la University College London durante dos años.