FACULTAD DE CIENCIAS ECONÓMICAS

ÉTICA EN LA INTELIGENCIA ARTIFICIAL: DISCRIMINACIÓN Y SESGO EN LOS PROCESOS DE CONTRATACIÓN

Autor: Jaime Blanco Ledesma

5º E3-Analytics

Tutor: María Reyes Calderón Cuadrado

Madrid

Marzo 2025

Resumen

La inteligencia artificial está escalando posiciones en todos los sectores empresariales como banca, energía, educación, seguros, medicina, debido a sus muchas ganancias en eficiencia, mejoras de productividad y reducción de costes. No obstante, su tendencia al sesgo, la discriminación, y la carencia de contexto están provocando ciertos problemas de gobernanza y la necesidad de crear sistemas de eliminación de sesgos, de evitar plagios, etc. Se trata de analizar el estado de la cuestión en distintos sectores, y con distintos elementos éticos: justicia, equidad, libertad, seguridad, etc. Este trabajo, en concreto, estudia la discriminación y el sesgo presentes en los procesos de contratación de empresas e instituciones.

Palabras clave

Inteligencia Artificial, sesgo, discriminación, algoritmo, procesos de contratación, machine learning, desigualdades

Abstract

Artificial Intelligence is climbing positions in all business sectors such as banking, energy, education, insurance, and medicine, due to its many efficiency gains, productivity improvements and cost reduction. However, its tendency to bias, discrimination, and the lack of context are causing certain governance problems and the need to create systems for eliminating bias, avoiding plagiarism, etc. The aim is to analyze the state of the question in different sectors, and with different ethical elements: justice, equity, freedom, security, etc. This work studies the discrimination and bias present in the hiring processes of companies and institutions.

Key words

Artificial Intelligence, bias, discrimination, algorithms, hiring process, machine learning, inequality

ÍNDICE

1. INTRODUCCIÓN

Las actividades de gestión de recursos humanos cuentan con varias tareas rutinarias que consumen mucho tiempo, y otras que también están sujetas a la percepción, subjetividad o sesgos humanos. Por ambas razones, se considera un terreno donde el uso de la inteligencia artificial (IA) puede generar mucha potencia (Rodgers et al., 2023). Entre ellas, destaca el reclutamiento y selección de personas, un proceso que ayuda a las empresas a identificar y atraer a candidatos calificados. La IA puede automatizar gran parte del proceso de reclutamiento y, ayudar a mejorar la selección de candidatos a través de plataformas algorítmicas personalizables, clasificar currículums, programar entrevistas, proporcionar retroalimentación, etc. (Kot et al., 2021).

No obstante, siendo la "materia prima" personas, la empresa debe analizar cómo se utiliza la IA en esta área, tanto desde el punto de vista técnico y de eficiencia, como desde su impacto social, organizativo y ético. Por ejemplo, el estudio de Sebastian Kot encontró que el reclutamiento y la calidad basados en IA afectan significativamente la reputación del empleador (Kot et al., 2021).

Mientras existen ventajas innegables en la automatización, también hay inconvenientes serios, destacando el sesgo y la discriminación que pueden surgir a partir de algoritmos mal diseñados.

El objetivo de este trabajo es analizar el impacto de la inteligencia artificial en uno de los procesos de gestión de recursos humanos: los procesos de contratación, tratando de identificar las causas y consecuencias de estos sesgos, así como proponer estrategias para mitigarlos. Además, se pretende fomentar una discusión sobre la necesidad de implementar prácticas éticas y regulaciones adecuadas en el uso de IA en el ámbito laboral.

Este TFG se desarrolla de la siguiente manera: después de este apartado introductorio, se analizará la definición y los tipos de IA, sus ventajas e inconvenientes, los dilemas éticos en torno a su uso, y la muy variada regulación que existe en torno al concepto. El tercer apartado aborda como nace la discriminación y el sesgo en el uso de la IA y los algoritmos en los procesos de contratación, los tipos de sesgo y su impacto. En el apartado cuarto se realizará una aproximación práctica al asunto, utilizando un *dataset* real para analizar, mediante el uso de herramientas de *machine learning* (a menor escala que

aquellas que en la práctica utilizan las empresas), como los distintos parámetros afectan a que la IA se decante por un candidato u otro. Analizados los resultados obtenidos, se estudiarán las estrategias que se utilizan para el diseño ético y transparente de los algoritmos que vayan a ser utilizados en los procesos de contratación.

2. MARCO TEÓRICO

2.1. Definición y tipología de Inteligencia Artificial

No existe una definición generalmente aceptada de inteligencia artificial. Algunos lo equiparan con herramientas que usan algoritmos programados, aunque estos ya existían antes que la IA (Russell & Norvig, 2010). Otros lo asocian con herramientas que imitan capacidades complejas humanas, y otros, más precisos, con herramientas que imitan con habilidades intelectuales de las personas incluyendo el aprendizaje. Los puristas señalan que esas herramientas son aún muy simples, y que no son más que un antecesor de la IA, que confluirían en la idea de sistemas que despliegan comportamiento inteligente analizando su entorno y tomando acciones con cierto grado de autonomía para lograr objetivos concretos (Davidson, 2024; Sheikh et al, 2023).

El Parlamento europeo (2024) emplea la siguiente definición: "AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments". Por su parte, las autoridades norteamericanas la describen como: "a machine-based system that can, for a given set of human-defined objectives, make pre-dictions, recommendations, or decisions influencing real or virtual environments. AI systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action" (15 U.S.C. 9401(3), 2023, p.3).

En todo caso, podemos entenderlo como la conjunción de: una herramienta artificial, es decir, hecha por el hombre como copia de lo natural y con una base física y con data como input, y un cierto grado de inteligencia, entendida como la habilidad de aprender, entender y hacer juicios o expresar opiniones basadas en un cierto razonamiento (Cf. Sheikh et al, 2023).

Los avances recientes en el campo de la IA y la rápida expansión del ecosistema de datos han abierto una nueva senda en la IA: la AI generativa o LLM-chatbot, grandes modelos de lenguaje con excepcional capacidad para comprender, generar y manipular el lenguaje humano. ChatGPT desarrollado por OpenAI es un ejemplo de LLM-chatbot capaz de

entablar diálogos con humanos y responder consultas. Emplea algoritmos de aprendizaje profundo para predecir la siguiente palabra en una secuencia dada en función de un gran corpus de textos con más de 300 mil millones de palabras y complementado con retroalimentación humana (Dam et al, 2024).

En el estudio de la IA se han identificado distintos tipos, en atención a su capacidad y a su funcionalidad, clasificándose de esta manera en función del nivel de desarrollo y de las implicaciones que potencialmente pueden llegar a tener.

En primer lugar, como hemos dicho, se puede clasificar la IA en función de su capacidad. La primera categoría sería la IA "estrecha" o IA débil, que es aquella que puede realizar tareas que se deben definir de manera específica para que la IA lo entienda, y que una vez bien definida, es muy probable que lo haga con una eficiencia considerablemente mayor al humano. Algunos ejemplos de esta categoría son Alexa, Google Assistant, Siri o los sistemas de recomendación y anuncios en el *e-commerce* (IBM Data and AI team, 2023). Se trata, por lo tanto, de plataformas de IA muy útiles para la realización de tareas claramente definidas, pero no son capaces de pensar u opinar como lo haría un humano.

La segunda categoría en función de la capacidad es la IA general, que sería aquella capaz de actuar como un humano, esto es, pensar, razonar y aprender en nuevas situaciones pudiendo adaptarse a cualquier entorno. Todavía no se ha perfeccionado este tipo de IA y sigue siendo un objetivo dentro del campo de la IA (Khan, 2021).

El tercer y último tipo es la super inteligencia artificial, que sería aquella capaz de superar al ser humano en el razonamiento tomando decisiones estratégicas con una creatividad y precisión mucho más alta que la de los seres humanos, despertando este tipo de inteligencia una serie de cuestiones o incluso problemas éticos y riesgos en torno a la posibilidad de su desarrollo (Amita, 2024).

Como hemos dicho, también se puede clasificar la IA en atención a su funcionalidad. La primera categoría en esta clasificación sería la IA reactiva o *reactive AI* que es la más básica, en la que la máquina está diseñada para responder a cuestiones muy concretas, no teniendo capacidad de almacenamiento y estudio de datos pasados para el desarrollo y formación de sus respuestas. Un ejemplo de esta sería, *Deep Blue*, un ordenador creado por IBM en 1997 para derrotar en ajedrez al jugador Gary Kasparov, que pese a tener la capacidad de analizar todas las variantes en esa partida, no era capaz de recurrir a datos de

partidas anteriores para hacerlo de manera más eficiente (IBM Data and AI team, 2023).

La segunda categoría respecto a la funcionalidad es, la IA con memoria limitada, que supera a la anterior en que tiene memoria, que pese a ser limitada, le permite aprender en atención a datos recopilados en tareas y consultas anteriores, siendo un gran ejemplo los vehículos autónomos que recopilan información en tiempo real para mejorar la conducción (Amita, 2024).

En tercer lugar, tenemos la teoría de la mente que serían las máquinas capaces de procesar pensamientos, intenciones e incluso emociones humanas, lo que es de gran utilidad en el campo de la robótica social y la salud mental humana.

Por último, nos encontramos con el nivel de autoconciencia que sería aquel en el que la máquina no solo comprende lo que le rodea, sino que también es consciente de sí mismo y tiene objetivos propios de desarrollo (Amita, 2024).

Los tipos de IA que existen hoy están siendo utilizados en campos como la economía, la seguridad o la medicina, mejorando notablemente la eficiencia, pero generando dudas en torno a la eticidad en el uso de los datos (IBM Data and AI team, 2023).

La posibilidad de que se desarrolle una superinteligencia o una IA autoconsciente tiene una serie de implicaciones éticas que resaltan la importancia de que el humano mantenga siempre el control y la responsabilidad sobre la máquina. También es necesario entender el uso limitado y consciente que debe hacerse de estas plataformas de cara a que el humano no vaya poco a poco perdiendo su capacidad analítica y de estudio y se quede en un segundo plano (Amita, 2024).

La IA estrecha ya está presente en nuestro día a día y ha encajado de manera más o menos adecuada, pero el desarrollo tecnológico de una IA general o una superinteligencia (que llegará tarde o temprano en mi opinión) requiere del desarrollo de un marco regulatorio y ético que no relegue al ser humano a un segundo plano, de manera que este siga siendo el actor principal.

2.2. Ventajas e inconvenientes

Según avanza la inteligencia artificial, esta reporta innumerables beneficios para la facilitación de procesos que esta puede realizar de manera mucho más precisa y eficiente, pero los avances más punteros están despertando riesgos y cuestiones éticas sobre hasta

donde van a ser capaces de llegar estás maquinas en un futuro.

Entrando en materia, la ventaja más clara que ha traído la IA consigo es la mejora de la precisión y la eficiencia en procesos repetitivos y rutinarios. En estos procesos la máquina minimiza errores humanos, optimiza la posterior toma de decisiones y lo hace todo de manera más eficiente (China, C.R. 2023). En la empresa esto permite el análisis de datos en tiempo real para sacar conclusiones sobre tendencias de mercado, personalización de productos y servicios, reducir costes operativos, y mejorar la productividad en campos como la atención al cliente.

Otros beneficios discurren por la vía de la mejora de la calidad de vida general de las personas. Citando alguno de los ejemplos que utiliza un informe de la Universidad de Virginia Tech, la IA ha mejorado la asistencia prestada a personas con discapacidad, mejora los tratamientos médicos de los enfermos, e incluso hace el tráfico de vehículos más eficiente mediante el análisis de datos a tiempo real (Losey, D. et al. 2023).

La posibilidad de que la IA asuma las tareas repetitivas permite al ser humano centrarse en las decisiones estratégicas y creativas que verdaderamente generan impacto. Otro beneficio es la capacidad de análisis en tiempo real de datos como tendencias de consumo y sociales que permitan ofrecer experiencias personalizadas al consumidor (Cortes, M., 2025).

Entrando en los inconvenientes, uno de los mayores riesgos que plantea la IA es que esta se construye a base de algoritmos, y estos algoritmos utilizan grandes volúmenes de datos. El problema surge cuando en estos volúmenes de datos de entrenamiento y test, existen ya los sesgos, de manera que cuando se entrene al algoritmo, este consolide o incluso magnifique el sesgo ya presente en la sociedad. Este problema, está particularmente presente en el uso de la IA en los procesos de contratación laboral, que es el tema específico que nos ocupa en este trabajo, y en otros ámbitos como por ejemplo, el uso de la IA en la concesión de préstamos bancarios. Es una realidad que, en ambos casos, existen sesgos sociales por los que se favorece a unos y no a otros, lo que posteriormente repercute en el *dataset* utilizado por los algoritmos para su toma de decisiones (Duggal, N., 2025).

Otro de los inconvenientes que genera la IA es su impacto en el mundo laboral. Si bien es verdad, que se están generando empleos en el sector del análisis de datos, el desarrollo de software y en el mundo de la informática en general, la IA es capaz de realizar

trabajos repetitivos y rutinarios de manera mucho más eficiente y precisa que un ser humano, afectando a sectores como la producción, la logística o las finanzas (China, C.R. 2023). Es necesario entender la importancia de que según se desarrollan las maquinas inteligentes, estas deben complementar el trabajo humano, ya que una sustitución masiva, generaría unas desigualdades sociales que podrían traducirse en una crisis económica de gran magnitud. Otro inconveniente complementario es la dificultad de implementación de la IA a nivel empresarial por los altos costes que supondría una integración efectiva y ordenada, además de la falta de expertos en el campo de la IA que hay en el mundo laboral actualmente (Cortes, M., 2025).

Para Gao y Qian los desafíos o inconvenientes más notables son los relativos a la seguridad y privacidad de los datos recopilados de los clientes, siendo mucha de esta información de categoría sensible. Otro de los inconvenientes para estos autores es la eliminación de trabajos de naturaleza repetitiva si bien, esto se compensaría con la creación de nuevos trabajos en el ámbito de la supervisión y la mejora continua de sistemas. La clave para estos autores es equilibrar la eficiencia operativa con un enfoque ético y sostenible (Gao, Z. & Qian, Q., 2022).

Otro de los inconvenientes según los autores del informe de Virginia Tech antes citado, es la capacidad que tiene la IA de influir en la toma de decisiones de los humanos. Los algoritmos desarrollados específicamente para hacer recomendaciones pueden influir en la percepción que tenga el ser humano sobre aspectos como la política o el consumo. Todo despierta una ingente necesidad de regulación en sectores como la publicidad, los medios de comunicación o la educación, donde el uso de IA sin regulación podrá afectar a la autonomía de las personas (Losey, D. et al. 2023).

Para Pujari y Multani, el mayor inconveniente que trae consiga la IA es la privacidad y la seguridad de datos. Según estos autores una IA de suficiente nivel requiere de un volumen de datos tan grande que en la mayoría de los casos esta contendrá datos que afecten a la privacidad de los usuarios. Esto sumado a una falta de regulación clara en ciertos países, ha llevado a el uso indebido de ciertas categorías de datos que se entrometen manifiestamente en la privacidad del individuo (Pujari, V., et al, 2020).

El último inconveniente que debe mencionarse es el impacto medioambiental que ha tenido el desarrollo de la IA en la sociedad. El impacto en la huella de carbono de las

cantidades de energía y los recursos computacionales necesarios es enorme, habiendo aumentado el consumo energético en los centros de procesamiento de datos y de desarrollo de sistemas en los últimos años (Duggal, N. 2025). Todo lleva a la necesidad de implantar sistemas de uso de energías renovables y el desarrollo de sistemas de IA que optimicen el uso energético.

Con todo la IA ya aporta innumerables beneficios a la sociedad en materia de industria y calidad de vida, pero también trae consigo una serie de inconvenientes y riesgos en materia de equidad, privacidad, sostenibilidad y empleo. Así el objetivo de los países y sus empresas es maximizar los beneficios al mismo tiempo que se limitan los inconvenientes, siendo vital para ello una regulación flexible y clara como veremos ahora.

2.3. Regulación

Con el desarrollo exponencial de la IA, el tema ha cobrado una importancia fundamental en la agenda legislativa de países y organizaciones internacionales con el objetivo de definir marcos regulatorios flexibles, pero al mismo tiempo claros y limitativos de cara a garantizar la protección de valores democráticos, derechos fundamentales y la protección general del ser humano ante los riesgos que trae consigo la IA. Para el análisis de la regulación, estudiaré los distintos niveles territoriales empezando por España, pasando por Europa y terminando por la regulación existente a nivel internacional.

En primer lugar, en España debemos destacar como el hito más relevante a nivel regulatorio la creación de la Agencia para la Supervisión de la Inteligencia Artificial o AESIA. Está fue creada en 2023 con la publicación del Real Decreto 729/2023, y tiene como misión principal supervisar el desarrollo y el uso de la IA en territorio nacional, de cara a asegurar la protección de los derechos fundamentales y promover la transparencia en la implementación de los sistemas (Agencia Española de Supervisión de Inteligencia Artificial [AESIA], 2023).

La AESIA es la encargada de coordinar con el resto de los organismos nacionales y europeos relacionados, el cumplimiento del Reglamento Europeo de la IA, que posteriormente analizaremos, así como el marco legislativo español que regula el uso de la IA en las Administraciones Públicas y en el sector privado, y que impone una serie de normas de supervisión y auditoria, para garantizar la equidad y un cierto nivel de confianza en estos sistemas.

A nivel europeo, se ha desarrollado el Reglamento UE 2024/1689, más conocido como Reglamento de inteligencia artificial, que fue definitivamente publicado en 2024 y que establece un marco legislativo estricto pero flexible imponiendo una serie de normas en función del nivel de riesgo de cada sistema de IA. En el nivel inaceptable se encuentran los sistemas de manipulación cognitiva o de vigilancia masiva mediante reconocimiento facial en espacios públicos, siendo ambas prácticas prohibidas por la norma (Diario Oficial de la Unión Europea, 2024).

En un nivel inferior, encontramos los sistemas calificados por la norma como de "alto riesgo" que se refiere a aquellos utilizados en ámbitos sociales críticos como la educación o la justicia, debiendo estos cumplir una serie de requisitos muy estrictos a nivel de evaluación de impacto, supervisión y transparencia (Parlamento Europeo, 2023). Por último, otro de los hitos significativos marcados por la norma es la creación de la Oficina Europea de Inteligencia Artificial que se encarga de supervisar la correcta implementación en los países miembros de la UE de las normas impuestas en el reglamento, así como de garantizar la armonización de las normas en los distintos países, para garantizar un mismo nivel de cumplimiento en todos ellos.

Por último, a nivel internacional, nos encontramos el escenario regulatorio más desafiante por su disgregación. Cada país adopta un enfoque distinto en torno al tema. Así, por ejemplo, mientras que la UE adopta un enfoque integral y tendiendo a estricto, que además afecta a todos los países miembros, otros países que por no pertenecer a la UE tengan la libertad de adoptar el enfoque que consideren oportuno, han optado por dotar a sus normas de IA de mayor flexibilidad. Según el *Washington International Law Journal*, la situación se puede dividir entre aquellos países que han optado por un enfoque normativo horizontal como Canadá, Brasil y la UE y aquellos que han optado por una regulación más concreta y atomizada como Estados Unidos, Israel o Reino Unido (Park, S., 2024). Esto conlleva un escenario muy disgregado a nivel global en el que los países no pueden operar de acuerdo con las mismas condiciones y estándares, lo que ha llevado a que ciertos organismos como la ONU propongan la creación de un organismo internacional que supervise de manera unitaria y armonizada el uso de la IA, comparable a el organismo que ya se creó para la armonización normativa de la energía atómica con la Agencia Internacional de la Energía Atómica (Liu, J. 2024).

El faro que guía la regulación europea e internacional en materia de IA es la protección de los derechos fundamentales. Informes como el elaborado por el *Journal Risk Research* estudian como la protección de los derechos como la privacidad, la seguridad o la no discriminación, ha sido el tema principal en torno al cual ha girado toda la redacción del Reglamento europeo de la IA o *AI Act (Kusche, I., 2024).* Como hemos mencionado antes la UE ha optado por clasificar las áreas de uso de IA en función del riesgo imponiendo mayores salvaguardas en áreas especialmente sensibles como la contratación laboral o la justicia, entre otras.

A nivel internacional ha cobrado especial importancia en torno al debate de la regulación, asuntos como los conflictos militares o los ciberataques. De acuerdo con el *Nordic Journal of International Law*, en supuestos escenarios de guerra y conflictos geopolíticos, es necesaria la incorporación de una regulación clara, estricta e incluso prohibitiva en la toma de decisiones militares y de ciberseguridad (Arvidsson, M. & Noll, G., 2023). A nivel de privacidad este informe también resalta las preocupaciones en torno a la privacidad de las personas por el flujo transfronterizo de datos que lleva de la mano el uso de la IA, obligando a organismos internacionales a tomar medidas en torno a este tema, también.

Con todo, en los distintos niveles, la situación regulatoria en torno a la IA pretende alcanzar un equilibrio entre flexibilidad y protección del individuo, algo en lo que se está avanzando con la creación de la AESIA en España, la publicación del Reglamento europeo de IA, y que sigue representando un desafió a nivel internacional, donde no existe todavía una regulación armonizada aplicable en todos los países.

3. DISCRIMINACIÓN Y SESGO EN INTELIGENCIA ARTIFICIAL

3.1. Origen del sesgo en los algoritmos

El principal tema que se estudia en este trabajo es el sesgo en los algoritmos de aprendizaje automático que se utilizan para construir la IA. Estos algoritmos se elaboran con datos reales que constituyen la base de entreno de estos y que por ser datos provenientes del mundo real conllevan una serie de sesgos sociales presentes en nuestro contexto desde hace mucho tiempo. El hecho de que en un *dataset,* el algoritmo encuentre correlación entre que se rechazara a un determinado candidato para una oferta de empleo por su raza o sexo, por ejemplo, se traduce en sesgos que afectan crucialmente a la IA. La consecuencia principal es que de esta manera el uso de la IA perpetúa estas desigualdades sociales e incluso en algunas ocasiones crean nuevas desigualdades en base a una discriminación no justificada, lo que tiene mayor relevancia incluso en los campos ya mencionados como el empleo, la seguridad o la justicia (García-Campos, J. et al, 2022).

Los sesgos se originan de diversas maneras como veremos a continuación, pero quizás uno de los más relevantes es la selección de datos que se utilizan para entrenar el algoritmo como hemos dicho antes. En primer lugar, existe la posibilidad de que estos datos tiendan a unos resultados que reflejan los prejuicios históricos contra determinados sexos o razas, de manera que el algoritmo aprenda de esos patrones, y los replique o incluso los magnifique (Gines I. Fabrellas, A., 2024). En segundo lugar, cabe la posibilidad de que el *dataset* no sea lo suficientemente representativo de un sector social o geográfico determinado siendo aplicable solo a aquel en el que fue construido (Cruzado, M., 2024). Esto ocurriría por ejemplo si se elaborara un *dataset* con inputs del mercado laboral en Ghana, y posteriormente este se utilizará para entrenar algoritmos y sistemas de IA que serán utilizados en Finlandia. Existe una falta de representatividad absoluta lo que provocará que el algoritmo este sesgado incorrectamente y que no tenga mucha utilidad. Es, por lo tanto, esencial elegir bien los datos que se utilizan para la estructuración de sistemas de IA de cara a asegurar su utilidad, precisión y lo más importante, que no estén sesgados.

Otro factor clave en el origen del sesgo es el diseño del algoritmo, ya que, muchos de estos son creados de acuerdo con el sistema de “caja negra” o *blackbox* que significa que la decisión mostrada al usuario no tenga una explicación de base o que esta no sea transparente. Esto conlleva un problema aun mayor que es no solo que se confirma la

existencia del sesgo si no que no se puede saber de dónde viene para corregirlo (Flores, A.J., 2024). Un ejemplo de esta opacidad es la utilización de este tipo de algoritmos en el ámbito judicial donde se puede llegar a conclusiones discriminatorias sin que se pueda identificar el origen del sesgo ni responsabilidades por la toma de una decisión concreta (Moine, M.B., et al, 2022).

Los sesgos también se originan por la selección de parámetros que utilizan los seres humanos en la construcción del algoritmo. Si por ejemplo el desarrollador de un sistema de aprendizaje de IA decide incluir la raza, el sexo o el código postal, esto puede dar lugar a decisiones discriminatorias por razones socioeconómicas si un determinado barrio con cierto código postal está asociado a una comunidad marginalizada y discriminada históricamente (Silberg, J. & Manyika, J., 2019). Esto se hace particularmente evidente en equipos desarrolladores de IA en los que falta diversidad de razas, género u otros condicionantes, ya que puede ser que los parámetros escogidos, estén sesgados por las opiniones y la percepción de un grupo social concreto.

Otra de las maneras en las que se origina el sesgo es mediante retroalimentación. Los sistemas de aprendizaje automático no solo aprenden de los datos iniciales sino también de las predicciones propias que va realizado a medida que se van utilizando. De esta manera, incluso un *dataset* que casi no contenga sesgos iniciales, puede llegar a estar muy sesgado si después de muchas iteraciones, ha magnificado el pequeño y casi imperceptible sesgo que existía en los datos iniciales. Un ejemplo de esto son los sistemas de recomendación de consumo en los que determinadas preferencias se fortalecen, las nuevas tendencias en consumo no se aprecian y se acaban generando “burbujas informativas” (Silberg, J. & Manyika, J., 2019). En el campo judicial esto se ha probado en que ciertos sistemas que predicen la posibilidad de reincidencia de los exconvictos penalizan de manera desproporcionada y poco precisa a determinadas minorías raciales en base a patrones históricos de condenas (Cruzado, M., 2024).

El origen de los sesgos en sistemas de aprendizaje queda evidenciado de distinta manera en función del sector. En el ámbito laboral, algunos sistemas de IA descartan automáticamente candidaturas que no coinciden con los patrones históricos de éxito dentro de las empresas u organismos públicos (Moine, M.B., et al, 2022). En el sector financiero, se utiliza mucho la IA para la concesión de créditos bancarios y en muchos casos se penaliza de

manera sesgada e injusta a determinados sectores socioeconómicos en base a prejuicios y no se estudia realmente la capacidad real de pago (Silberg, J. & Manyika, J., 2019). En el sector de la salud los algoritmos de diagnóstico se han probado menos precisos en determinadas minorías raciales, resultando en un servicio deficiente (Cruzado, M., 2024).

Las soluciones para proteger la creación del sesgo van desde la auditoría de los algoritmos hasta la formación de equipos de desarrollo más diversos, si bien ante la falta de regulación legal y la oposición de las empresas a revelar los detalles de sus modelos propietarios de IA, esto resulta complejo (Flores, A.J., 2024). En la práctica el sesgo se erradica eliminado los parámetros y variables que pueden dar lugar a mayores problemas, mediante el uso de datos más diversos y representativos, y mediante la corrección del sesgo en las fases iniciales de entrenamiento del algoritmo (Silberg, J. & Manyika, J., 2019).

La falta de representatividad de los datos, la opacidad de los modelos y la retroalimentación del sesgo son los principales orígenes del sesgo en la IA, siendo fundamental combinar la técnica, la ley y la ética de cara a elaborar modelos verdaderamente justos y útiles.

3.2. Tipos de sesgo

Para analizar los tipos de sesgo, me centraré en lo algoritmos de selección de personal que se utilizan en los procesos de contratación para agilizar los procesos de recursos humanos.

El primero de los tipos, es el sesgo de confirmación que es aquel que se produce cuando los algoritmos reflejan las creencias y preferencias históricas del empleador (la empresa). De esta manera, se priorizan perfiles similares a los que han sido exitosos en el pasado, aumentando la homogeneidad dentro de la empresa y limitando la diversidad (Cruzado, M., 2024). Esto en última instancia, es perjudicial para las empresas que limitan la entrada de nuevos perfiles que podrían realizar aportaciones distintas y aportar puntos de vista novedosos, si bien las empresas suelen tender a priorizar lo seguro y lo que ya ha funcionado en el pasado (de manera errónea en mi opinión) (Gines I. Fabrellas, A., 2024).

En atención a los parámetros pueden existir los siguientes tipos de sesgo: racial, de género, de apariencia, de edad o de preferencia. El sesgo racial, se produce cuando en atención a nombres o códigos postales la IA reconoce ciertos patrones de empleabilidad o estabilidad laboral que desfavorecen a minorías en procesos de selección (Gines I. Fabrellas,

A., 2024). De esta manera, se refuerzan prejuicios históricos en el mercado laboral que hoy no tienen cabida en nuestra sociedad (Cruzado, M., 2024).

El sesgo de género, por otro lado, se produce de la misma manera en aquellos sectores laborales en los que la presencia del hombre o de la mujer ha sido predominante históricamente. Se han dado casos como el de una gran empresa tecnológica en la que se evidenció que el algoritmo de selección penalizaba palabras como "mujer" o "femenino" reduciendo considerablemente las posibilidades de selección del género femenino (Silberg, J. & Manyika, J., 2019).

El sesgo de apariencia surge con el uso de herramientas de video entrevista como *Hirevue* en las que la IA, si está construida incorrectamente o se retroalimenta de datos que encierran prejuicios, puede favorecer las candidaturas de aquellos con determinados rasgos faciales o tonos de piel. Otro ejemplo es el de los candidatos con expresiones más serias frente a aquellos que sonríen más frecuentemente teniendo estos últimos, mayores probabilidades de ser seleccionados en base a un criterio subjetivo que nada tiene que ver con la habilidad profesional del individuo (Moine, M.B., et al, 2022).

El sesgo de edad, por su parte, se produce en las situaciones en las que se da preferencia a candidatos jóvenes bajo la presunción de que tiene mayores habilidades tecnológicas y relativas a la nueva era digital, penalizando, injustamente, a los mayores de 40-45 años que igualmente podrían ser de gran utilidad por su experiencia (Silberg, J. & Manyika, J., 2019).

También existe el sesgo de preferencia, por el que modelos mal entrenados y que incurran en sesgo de este tipo, tienden a seleccionar a candidatos que hayan estado en determinados, colegios, universidades o empresas, favoreciendo de esta manera las carreras de estudio y laborales de aquellos que son más similares a las de casos de empleados exitosos en la empresa (Silberg, J. & Manyika, J., 2019).

Existen distintos tipos de sesgo en función del sector al que nos refiramos, en este caso hemos analizado los existentes en los procesos de contratación, pero, en cualquier caso, el sesgo suele responder a situaciones de injusticia social que no deben perpetuarse ni magnificarse, siendo esencial el correcto entrenamiento, auditoria y corrección de los modelos de IA utilizados.

3.3. Impacto del sesgo en los procesos de contratación.

La IA ha tenido un impacto positivo en los sistemas de reclutamiento en tanto que mejora la eficiencia de los equipos de recursos humanos en la selección. Sin embargo, sin las adecuadas cautelas, la IA resalta más por el impacto negativo que puede tener en los procesos laborales.

El principal de los impactos que hemos comentado antes y que es consecuencia del sesgo, es la homogenización interna de las empresas, decantándose por perfiles similares a los que han tenido éxito, retroalimentando de esta manera una ausencia de diversidad que penaliza la innovación y la creatividad dentro de la empresa (Moine, M.B., et al, 2022).

Otro de los impactos principales es el daño que produce la IA sesgada en la reputación y la competitividad en las empresas. Por un lado, las empresas que hacen uso de estos sistemas de selección corren el riesgo de ser demandas por discriminación, resultando en sanciones económicas y legales. La falta de diversidad también afecta a la moral, a la competitividad y al crecimiento dentro de la empresa (Silberg, J. & Manyika, J., 2019).

El ultimo impacto digno de mencionar es la erosión de la confianza en la IA. Si estos sistemas toman decisiones discriminatorias e injustas, los candidatos reducen su confianza en las empresas que los utilizan.

De cara a evitar sanciones, promover la diversidad y aumentar la confianza, las empresas deben implementar sistemas de evaluación, auditoría y corrección que garanticen que los algoritmos funcionan de manera justa y equitativa.

4. CASO DE ESTUDIO

4.1. *Dataset*

Para un mejor entendimiento de cómo pueden las empresas utilizar algoritmos e IA para agilizar y automatizar procesos de contratación, realizaremos un caso práctico utilizando métodos de *machine learning*. El objetivo de este caso de estudio es elaborar una serie de algoritmos de predicción en el programa R-Studio para analizar como ciertas variables contribuyen a la decisión de contratación.

Para hacerlo, se ha utilizado un *dataset* con observaciones reales utilizadas por el algoritmo que utiliza una empresa para valorar las candidaturas. El *dataset* contiene las siguientes categorías:

- Edad: variable de tipo entero contiendo valores del 20 al 50.
- Género: variable de tipo binario en la que 0 corresponde a hombre y 1 a mujer.
- Nivel de educación: variable de tipo categórico en la que 1 corresponde a Grado, 2 a grado avanzado, 3 a máster y 4 a doctorado.
- Años de experiencia: variable de tipo entero que asume valores de 0 a 15.
- Compañías anteriores: variable de tipo entero que asume valores de 1 a 5 y que se refiere al número de empresas anteriores en las que ha trabajo el candidato.
- Distancia desde la compañía: variable de tipo decimal que asume valores desde 1 a 50 y que se refiere a la distancia en kilómetros desde la residencia del candidato a la compañía.
- Resultado entrevista: variable de tipo entero asumiendo valores de 0 a 100 que se refiere a la calificación obtenida por el candidato en entrevista.
- Resultado habilidades: variable de tipo entero que asume valores del 0 al 100 en función de las habilidades técnicas del candidato.
- Resultado personalidad: variable de tipo entero que asume valores del 0 al 100 en función de las habilidades personales del candidato.
- Estrategia de contratación: variable de tipo categórico que asume valores del 1 al 4 en función de la estrategia de contratación adoptada por el equipo de recursos humanos.

- Decisión de contratación: variable objetivo de tipo binario que explica si el candidato fue contratado (1) o no (0).

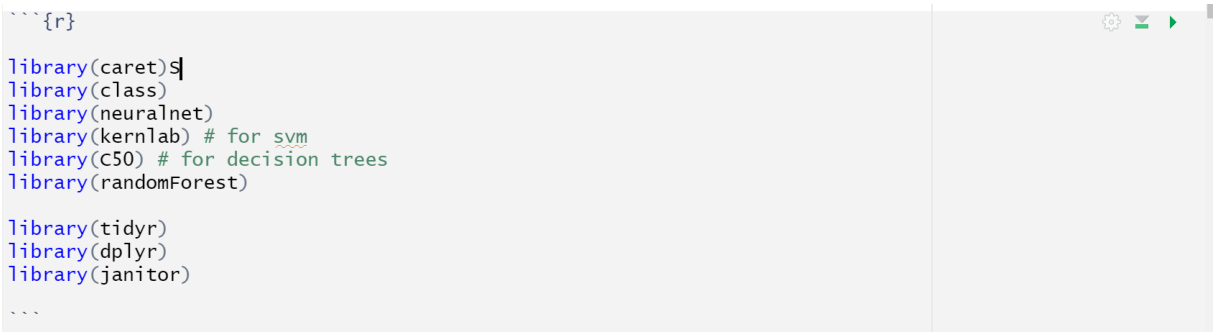
Se trata de un *dataset* con 11 variables y 1500 observaciones que permite un mejor entendimiento de como ciertos atributos del candidato y decisiones en el proceso de contratación influyen en la decisión final del equipo de reclutamiento, permitiendo optimizar los procesos de contratación en distintos contextos corporativos.

4.2. Programación de algoritmos

Entendido el *dataset*, lo hemos utilizado para entrenar los modelos de predicción típicamente utilizados en las prácticas de *machine learning* relacionadas con el aprendizaje supervisado. Los resultados de cada uno de ellos nos permiten entender mejor el peso que tiene cada variable para el equipo de reclutamiento de la empresa, pero antes de analizar esos resultados, presentamos los algoritmos que hemos utilizado.

En primer lugar, importamos las librerías necesarias para el ejercicio de programación a realizar.

Figura 1: librerías importadas para la programación de algoritmos en R-Studio



Después de leer los datos utilizamos el comando *summary (x)* para entender la distribución de nuestros datos, esto es, donde se sitúan los cuartiles, la media y los extremos de cada categoría.

Figura 2: resumen de distribución de las variables

```
      Age            Gender      EducationLevel  ExperienceYears  PreviousCompanies
 Min.   :20.00   Min.   :0.000   Min.   :1.000   Min.   : 0.000   Min.   :1.000
 1st Qu.:27.00   1st Qu.:0.000   1st Qu.:2.000   1st Qu.: 4.000   1st Qu.:2.000
 Median :35.00   Median :0.000   Median :2.000   Median : 8.000   Median :3.000
 Mean   :35.15   Mean   :0.492   Mean   :2.188   Mean   : 7.694   Mean   :3.002
 3rd Qu.:43.00   3rd Qu.:1.000   3rd Qu.:3.000   3rd Qu.:12.000   3rd Qu.:4.000
 Max.   :50.00   Max.   :1.000   Max.   :4.000   Max.   :15.000   Max.   :5.000
 DistanceFromCompany InterviewScore     SkillScore     PersonalityScore RecruitmentStrategy
 Min.   : 1.031      Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :1.000
 1st Qu.:12.839      1st Qu.: 25.00   1st Qu.: 25.75   1st Qu.: 23.00   1st Qu.:1.000
 Median :25.502      Median : 52.00   Median : 53.00   Median : 49.00   Median :2.000
 Mean   :25.505      Mean   : 50.56   Mean   : 51.12   Mean   : 49.39   Mean   :1.893
 3rd Qu.:37.738      3rd Qu.: 75.00   3rd Qu.: 76.00   3rd Qu.: 76.00   3rd Qu.:2.000
 Max.   :50.992      Max.   :100.00   Max.   :100.00   Max.   :100.00   Max.   :3.000
 HiringDecision
 Min.   :0.00
 1st Qu.:0.00
 Median :0.00
 Mean   :0.31
 3rd Qu.:1.00
 Max.   :1.00
```

Posteriormente, preparamos los datos para el análisis, convirtiendo las variables categóricas en *dummies*, limpiando los nombres de cada variable de espacios o caracteres especiales que pudieran complicar el análisis, y normalizamos los datos para que puedan ser utilizados en algoritmos como ANN (*Artificial Neural Network*) o KNN (*K-Nearest Neighbors*).

Figura 3: preparación de los datos

```
```{r}

recruitment_mm <- as.data.frame(model.matrix(~., -1, data = recruitment))
recruitment_mm <- clean_names(recruitment_mm)

recruitment_mm$intercept <- NULL


normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}


recruitment_norm <- as.data.frame(lapply(recruitment_mm, normalize))

str(recruitment_norm)

```
```

Por último, antes de pasar a presentar los modelos de predicción utilizados, dividimos los datos en *train* y *test*, ya que una porción de los datos será utilizada para entrenar los modelos y otra para probarlos. Para ello, establecemos una semilla de manera que al dividir los datos aleatoriamente estos siempre sean los mismos, y establecemos una ratio de 0.5 de manera que la mitad de los datos serán de entrenamiento y la otra mitad de comprobación.

Figura 4: división en train y test

```r
set.seed(12345)


ratio <- 0.5


train_rows <- sample(1:nrow(recruitment_norm), ratio*nrow(recruitment_norm))
train_data <- recruitment_norm[train_rows, ]
test_data <- recruitment_norm[-train_rows, ]


train_data_predictors <- train_data[,-11]
test_data_predictors <- test_data[,-11]

train_data_labels <- train_data[,11]
test_data_labels <- test_data[,11]
```

Ejecutados estos pasos preliminares, pasamos a la construcción de los modelos de aprendizaje. En primer lugar, se utiliza un modelo de regresión logística que en base a una combinación lineal de todas las variables calcula una probabilidad que en caso de ser mayor del 50% clasificará la predicción como “contratado” y en caso de ser inferior como “no contratado”.

Figura 5: modelo de regresión logística

```r
set.seed(12345)


lr_model <- glm(hiring_decision ~., family = "binomial", data = train_data)
summary(lr_model)

lr_predict <- predict(lr_model, test_data, type = "response")
lr_bin <- ifelse(lr_predict >= 0.5, 1, 0)

confusionMatrix(as.factor(lr_bin), as.factor(test_data_labels), positive = "1")
```

En segundo lugar, construimos un árbol de decisión de tipo C5.0 que construye un árbol en el que los datos se van clasificando en “ramas” a distintos niveles, en función de las variables más relevantes. Al final genera un diagrama similar a un árbol invertido.

Figura 6: árbol de decisión predictivo

```r
set.seed(12345)


tree_model <- C5.0(as.factor(hiring_decision) ~. ,data = train_data)


tree_predict <- predict(tree_model, test_data)

confusionMatrix(as.factor(tree_predict), as.factor(test_data_labels), positive = "1")
```

Construimos también un algoritmo de predicción KNN que predice la variable objetivo en función de los 21 vecinos más cercanos en un plano de similitud en función de las variables dependientes.

Figura 7: modelo de predicción KNN

```r
```{r}
set.seed(12345)


knn_predict <- knn(train_data_predictors, test_data_predictors, train_data_labels, k = 21)


knn_matrix <- as.data.frame(as.table(confusionMatrix(as.factor(knn_predict),
as.factor(test_data_labels), positive = "1")))

confusionMatrix(as.factor(knn_predict), as.factor(test_data_labels), positive = "1")

```
```

Hacemos uso también del modelo ANN, un modelo basando en el sistema de procesamiento del cerebro humano para detectar relaciones complejas no lineales entre los datos. En nuestro caso se incluye una sola capa oculta de 3 neuronas, estableciendo el criterio de convergencia en 0.01, con un número de iteraciones no muy elevado para agilizar la construcción del diagrama, utilizando el algoritmo “slr” de mínimos cuadrados, y estableciendo el umbral de clasificación en 0.3.

Figura 8: modelo predictivo ANN

```r
```{r}
set.seed(12345)


ann_model <- neuralnet(hiring_decision ~ ., data = train_data, hidden = c(3), threshold = 0.01, stepmax
= 1e+05, algorithm = "slr")


plot(ann_model)


ann_predict <- predict(ann_model, test_data, type = "response")
test_ann_predict <- ifelse(ann_predict >= .3, 1, 0)


ann_matrix <- as.data.frame(as.table(confusionMatrix(as.factor(test_ann_predict),
as.factor(test_data$hiring_decision), positive = "1")))

confusionMatrix(as.factor(test_ann_predict), as.factor(test_data$hiring_decision), positive = "1")

```
```

También analizaremos los resultados de un modelo predictivo *random forest* que construye un número dado de árboles de decisión (en nuestro caso 200) utilizando subconjuntos de datos distintos.

Figura 9: modelo random forest

```r
set.seed(12345)

rf_model <- randomForest(hiring_decision  ~ ., data = train_data, ntree = 200)

plot(rf_model)

rf_predict <- predict(rf_model, test_data, type = "response")


confusionMatrix(as.factor(rf_bin_predict), as.factor(test_data$hiring_decision), positive = "1")
```

Para terminar, hemos utilizado modelos de *support vector machine* que predicen la variable objetivo clasificando los datos en función del hiperplano que los separe de manera más amplia. En estos modelos se utilizan *kernels* o funciones para medir la similitud de los datos, en nuestro caso hemos utilizado las funciones *Laplace, Vanilla y RBF.*

Figura 10: modelos de support vector machine con distintos kernels

```r
# Laplace Model
svm_laplace_model <- ksvm(hiring_decision ~ ., data = train_data, kernel = "laplacedot")
svm_laplace_predict <- predict(svm_laplace_model, test_data)
confusionMatrix(as.factor(svm_laplace_bin), as.factor(test_data_labels))
```

```r
# Vanilla Model
svm_vanilla_model <- ksvm(hiring_decision ~ ., data = train_data, kernel = "vanilladot")
svm_vanilla_predict <- predict(svm_vanilla_model, test_data)
confusionMatrix(as.factor(svm_vanilla_bin), as.factor(test_data_labels))
```

```r
# RBF Model
svm_rbf_model <- ksvm(hiring_decision ~ ., data = train_data, kernel = "rbfdot")
svm_rbf_predict <- predict(svm_rbf_model, test_data)
confusionMatrix(as.factor(svm_rbf_bin), as.factor(test_data_labels))
```

4.3. Resultados y conclusiones

Presentados los modelos predictivos de *machine learning* que hemos utilizado para entender mejor como afectan las variables dependientes de nuestro conjunto de datos en la clasificación de un candidato como “contratado” o “no contratado”, pasamos a analizar los resultados obtenidos.

El anexo II contiene los resultados del modelo de regresión logística, donde el modelo ha alcanzado una precisión de más del 85% y un valor kappa de más del 60%. El valor kappa mide la precisión en la predicción de la variable objetivo excluyendo el factor del azar. El anexo I contiene la tabla elaborada por Landis y Koch sobre la índice kappa, por la que un acuerdo entre las predicciones del modelo y los valores reales de más del 60% se considera

"sustancial". Se trata de un modelo considerablemente preciso para predecir la decisión de contratación en procesos de acuerdo con las variables estudiadas.

Otro elemento muy interesante dentro de los resultados del modelo de regresión es el nivel de significancia estadística de las variables a la hora de predecir la variable objetivo. En función del *p-value* de cada variable estas influyen en mayor o menor medida en la decisión de contratación. Como podemos ver en nuestro modelo son muy significativas, por tener un *p-value* cercano a cero, el nivel de educación, los años de experiencia, el resultado de la entrevista, las habilidades técnicas y personales y la estrategia de contratación. Nuestro conjunto de datos es por lo tanto un buen ejemplo de una empresa en la que se hace un uso ético de la IA y los algoritmos utilizados en el proceso de contratación. Variables como el género, la edad o la distancia hasta el trabajo no son significativas a la hora de contratar a un candidato o no en nuestro modelo, comprobándose que la empresa que presenta este conjunto de datos no hace uso de modelos que pudieran ser clasificados como discriminatorios.

Como podemos ver en los anexos III a VII, en los resultados obtenidos para el resto de los modelos se alcanza una precisión de entre el 80% y el 90% y un índice Kappa Cohen de entre 0.6 y 0.7. En todos los casos el modelo clasifica con una precisión sustancial la decisión de contratación.

Se trata, en conclusión, de un conjunto de datos que explica las variables tenidas en cuenta por una empresa anónima en la utilización de modelos e IA para automatizar los procesos de contratación. Los modelos predictivos permiten entender como las variables más significativas en este caso son aquellas verdaderamente deben influir en la decisión. Se trata de un conjunto de datos que muestra poco sesgo discriminatorio en atención a variables como la edad o el sexo, siendo un buen ejemplo de utilización de modelos éticos para la agilización de los procesos de contratación.

5. ESTRATEGIAS PARA MITIGAR EL SESGO

5.1. Diseño ético de algoritmos

Dados todos los impactos negativos que pueden tener los sistemas de IA, el diseño ético de algoritmos se ha constituido como un pilar esencial para que estos estén alineados con los valores humanos fundamentales.

Una de las principales metas que tiene el desarrollo ético de algoritmos es luchar contra la opacidad de aquellos algoritmos que funcionan en sistema de "caja negra", lo que dificulta la identificación de sesgos que permitan exigir responsabilidades a los desarrolladores del sistema de aprendizaje o a aquellos que seleccionaron los datos, no existiendo tampoco la posibilidad de corregir el algoritmo (Martin, K., 2021). Para solucionarlo se proponen sistemas que permitan entender cómo interpreta la IA los datos y cómo influye cada variable en sus decisiones (Floridi, L., & Cowls, J., 2019). Se pretende alcanzar un nivel de auditoría efectiva mediante herramientas como los modelos sustitutos o los sistemas de visualización de impacto de cada variable (Tsamados, A. et al, 2022).

El segundo aspecto clave en el diseño ético de algoritmos es la equidad, que busca reducir los sesgos que surgen por prejuicios existentes, y que excluyen a ciertos grupos del acceso a empleos, créditos o servicios públicos esenciales, por ejemplo. Para solucionarlo es esencial el preprocesamiento de los datos iniciales, la introducción de restricciones en el entrenamiento del modelo y la postvalidación de los resultados que ofrezcan los algoritmos (Floridi, L., & Cowls, J., 2019).

Otro factor vital en el diseño ético de algoritmos es la responsabilidad de las decisiones adoptadas por la IA. Ante una situación en la que no se conoce quien es el responsable de la introducción de una determinada variable o categoría de datos, se produce un vacío ético en el que decisiones de un calado muy profundo, no pueden ser atribuidas a un actor especifico. Para aplacar este problema se introducen medidas de gobernanza, evaluación de impacto o *impact assesment*, y de rendición de cuentas que permiten identificar a los responsables de una determinada decisión y al mismo tiempo obligan a las empresas a mitigar riesgos (Tsamados, A. et al, 2022).

El diseño ético de algoritmos requiere también la protección de la privacidad de los individuos, y es que muchos de los modelos de IA más utilizados están entrenados con datos personales que pueden plantear preocupaciones en torno a la privacidad o al uso indebido

de esos datos. Es necesario establecer técnicas de anonimización y normas estrictas de manejo de los datos para garantizar la privacidad (Floridi, L., & Cowls, J., 2019).

Para asegurar el cumplimiento de todos los criterios que hemos establecido para el diseño ético de los algoritmos es esencial que en su desarrollo intervengan no solo profesionales en el campo de la informática o en el desarrollo de software sino también expertos en otras áreas como la ética, el derecho y otras áreas de las ciencias sociales (Tsamados, A. et al, 2022). Esto permite asegurar la eticidad global de la IA y evitar consecuencias no deseadas con la entrada en sociedad del algoritmo.

Es vital también conservar cualquier tipo de documentación en la que se detalle el proceso de desarrollo del algoritmo, así como su entrenamiento para facilitar la trazabilidad de las decisiones humanas adoptadas en el proceso.

Resultan muy interesantes también las iniciativas de co-diseño en las que los usuarios finales también intervienen en el desarrollo y en la configuración del algoritmo.

El diseño ético de algoritmos requiere de múltiples enfoques profesionales, sistemas de gobernanza eficientes, sistemas de asunción de la responsabilidad y de sistemas de evaluación continua de la IA.

5.2. Transparencia y auditabilidad

Resulta esencial para mitigar los riesgos que conlleva la IA, la implantación de sistemas de comprensión y supervisión algorítmicos (Llamas, J.Z., 2022). El principal problema a nivel de transparencia surge con los algoritmos de “caja negra” que hacen muy complicada la identificación de sesgos, errores y discriminación de grupos (Bitzer, T., 2022) Para ello es esencial documentar todos los pasos en el desarrollo de los algoritmos, explicar la estructura de los modelos y especificar los criterios adoptados en la toma de decisiones (Kossow, N. et al, 2021).

La transparencia implica tanto comprender el algoritmo en la toma de decisiones como evaluar su impacto, de cara a corregir los posibles fallos. Para ello muchos países han introducido registros públicos en los que ciudadanos, investigadores y reguladores pueden acceder al detalle de los modelos utilizados en sectores como la administración, la salud y la seguridad (Llamas, J.Z., 2022). Se han introducido también formatos estándar de la documentación que deberá acompañar al modelo con su explicación.

A nivel de procesos de selección laboral, las empresas utilizan modelos de explicabilidad que permiten a los candidatos entender que criterios han sido evaluados y qué importancia se le daba a cada uno (Valderrama, M. et al, 2023). De entre la información proporcionada a los candidatos resulta de vital importancia especificar: que sistemas de decisión de IA se han utilizado, que datos se han procesado y que ha servido de base para la toma de una decisión final. Esto les permite impugnar las decisiones que consideren injustas sirviéndose de una base objetiva (Kossow, N. et al, 2021).

Otra manera de garantizar la transparencia es la medida adoptada por muchas empresas que permite a los candidatos verificar o corregir la información obtenida de sus perfiles digitales. De esta manera, si cierto dato se hubiera extraído del perfil profesional del candidato de manera errónea, el candidato podrá modificarlo evitando que este afecte a su candidatura. Esto permite a los candidatos intervinientes en el proceso, participar en el mismo y tener mayor control sobre la información que tiene en cuenta la IA (Kossow, N. et al, 2021). En algunos países ya se han establecido normas legales que exigen que las decisiones automatizadas en materia de selección laboral de candidatos deban de estar siempre acompañadas de supervisión humana (Valderrama, M. et al, 2023).

La evolución de los marcos normativos, auditorias periódicas, la publicación clara de los criterios de evaluación y la participación activa del humano son de vital importancia de cara a asegurar la transparencia de los algoritmos en sectores como los procesos de contratación y otros muchos de especial importancia.

5.3. Diversidad

Para asegurar la diversidad en las plantillas, se han utilizado distintas medidas en el uso de la IA para procesos de contratación. En primer lugar, una de las técnicas más utilizadas es la eliminación de datos identificativos de los candidatos, sobre todo en las primeras fases del proceso en las que se filtran muchas candidaturas. Esto se consigue mediante técnicas de anonimización que impiden entender el nombre, género y otros datos demográficos que pudieran ser valorados negativamente por el algoritmo, teniéndose solo en cuenta aquellas variables que guardan relación con las habilidades y la competencia necesaria para la realización de un trabajo concreto (Adytia, B.V., et al, 2024).

Otra técnica utilizada para garantizar la diversidad en los equipos es la inclusión de pruebas prácticas y simulaciones que la IA evalúa de manera objetiva, y que limitan la

importancia de los títulos académicos o la experiencia laboral, favoreciendo a aquellas minorías que por tener menos recursos no tengan credenciales académicas elitistas pero que en cualquier caso han mostrado las capacidades suficientes para desempeñar el trabajo (Vivek, R., 2023).

Las empresas pueden también fomentar la diversidad entrenando al algoritmo para que favorezca la candidatura de aquellos perfiles menos representados en la empresa en ese momento, generando un entorno más inclusivo y creativo (Adytia, B.V., et al, 2024). Esto, sin embargo, no es del todo acertado en mi opinión en tanto que es una manera de discriminación para aquellos candidatos más similares a los empleados de la empresa en el momento. Más allá de este asunto, muchas empresas están implantando equipos especializados en el seguimiento de las decisiones tomadas por los algoritmos de cara a corregir los factores que limiten la diversidad (Albaroudi, E. et al, 2024).

Resulta esencial la intervención de los equipos de recursos humanos en la selección de candidatos. Un sistema hibrido entre IA y empleados de recursos humanos permite realizar evaluaciones de candidaturas de manera más eficiente, pero sin dejar de lado el juicio humano (Adytia, B.V., et al, 2024).

Resulta muy interesante las propuestas de algunos autores que defienden la utilización de la IA para cuestiones menos evidente como pueden ser procesos de selección para personas con discapacidad. Mediante la correlación de datos y la adaptación de las pruebas a sus capacidades, la IA puede extraer conclusiones sobre habilidades que pudieran aportar las personas con discapacidad a la empresa (Albaroudi, E. et al, 2024).

6. CONCLUSIONES

Con todo, la implementación de la IA en los procesos de selección de personal es quizás el paso más significativo en la era digital dentro del campo de los recursos humanos. Por un lado, plantea ventajas como la mejora de la eficiencia, la precisión a la hora de elegir candidatos con más probabilidad de éxito y la reducción de las cargas administrativas, pero por otro lado surgen desafíos éticos de gran calado. Habiendo identificado los riesgos que suponen el sesgo algorítmico, la falta de transparencia o la vulneración de la privacidad, podemos adoptar posiciones concretas sobre lo que constituye el uso ético de la IA.

El primer y más grave problema en la actualidad, es en mi opinión, la falta de transparencia en los modelos. Los modelos de IA están presentes en muchas facetas críticas de la rutina del ser humano, siendo una de ellas los procesos contratación. Se trata de aspectos que definen la vida de las personas, y, por lo tanto, las propuestas de la IA deben de ser fundamentadas y comprensibles. Excusar la falta de explicabilidad y de auditabilidad de un modelo bajo la afirmación de que es muy complejo o técnico no es suficiente, las instituciones deben de invertir recursos materiales y de personal para garantizar la claridad de los algoritmos que utilizan. Esto se hace aún más patente si cabe, cuando la IA se utiliza en procesos de contratación en los que los candidatos querrán tener acceso a las variables que han afectado a una decisión relevante en su carrera profesional.

El segundo aspecto clave en el desarrollo ético de la IA es el regulatorio, por ser el medio para establecer límites homogéneos a instituciones y empresas en el uso de IA. La definición de un marco claro y completo, mitiga los riesgos relativos a la privacidad, la transparencia y el sesgo. Esto mejora la seguridad jurídica tanto para usuarios del modelo que podrán tomar decisiones de manera guiada y evitarán sanciones, como para los afectados por su uso que conocerán los limites conforme a los cuales las instituciones pueden utilizar la IA. Tanto a nivel global como nacional, esta normativa es cada vez más extensa y hasta que llegue al punto deseado de desarrollo es tarea de las empresas tomar la iniciativa adoptando voluntariamente estrategias y órganos de gobernanza que aseguren el uso ético de la IA.

En cuanto a la utilización de la IA en los procesos de contratación, la clave en mi opinión es entender que estas herramientas de IA pueden agilizar el proceso de selección pero que aspectos como las emociones, la empatía o la cultura de la empresa son cuestiones tan subjetivas que el modelo nunca podrá entender y en los que es esencial la labor de los

departamentos de contratación. Se trata de una herramienta muy útil pero complementaria, que en ningún caso deberá sustituir al equipo de selección humano de la empresa. La automatización total del proceso afecta no solo a la precisión en la elección de candidatos (que será más baja por no evaluarse adecuadamente los aspectos subjetivos mencionados), sino también a la experiencia del candidato que percibirá la empresa como un lugar menos atractivo en el que trabajar.

Otro aspecto clave es promover y fomentar la diversidad en los procesos de contratación automatizados, en los que los datos de entrenamiento utilizados pueden contener sesgos históricos. Es necesario adoptar un enfoque proactivo en torno a este riesgo, ya que para promover la diversidad no solo es necesario construir un *dataset* representativo de muchos grupos demográficos, sino que también es necesario que el equipo que interpreta los outputs del modelo sea lo suficientemente diverso como para enfocarlo desde distintas perspectivas. Un gran ejemplo de conjunto de datos no sesgado o discriminatorio es el utilizado en el caso práctico que hemos realizado.

La utilización de procesos automatizados que agilicen el proceso es siempre la mejor opción, pero es necesario asegurarse de que, como en el conjunto de datos del caso práctico que hemos planteado, las variables significativas que influyen en la decisión son las que deben serlo, como el nivel de educación o los años de experiencia, no permitiendo el desarrollo de sesgos discriminatorios basados en la edad o el sexo.

El factor humano es esencial en los procesos de contratación y una manera de contribuir a la utilización ética de la IA es educar y formar a directivos y empleados de recursos humanos en la utilización ética de los algoritmos. Si los equipos de recursos humanos que aportan el componente humano en los procesos de contratación con IA no son capaces de entender el modelo, sus variables, las decisiones que toma y la manera en la que funciona, difícilmente podrán complementar su tarea. Es esencial que los equipos técnicos que intervienen en la definición del modelo aseguren su explicabilidad y que se aporte la formación necesaria a los equipos de recursos humanos para entender esos modelos. El componente humano y el tecnológico son ambos cruciales para asegurar el equilibrio entre eficiencia y efectividad y la falta de formación de los empleados que intervengan en los procesos de selección podrían afectar este equilibrio.

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, [Nombre completo del estudiante], estudiante de [nombre del título] de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "[Título del trabajo]", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación [el alumno debe mantener solo aquellas en las que se ha usado ChatGPT o similares y borrar el resto. Si no se ha usado ninguna, borrar todas y escribir “no he usado ninguna”]:

1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Crítico:** Para encontrar contra-argumentos a una tesis específica que pretendo defender.
3. **Referencias:** Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
4. **Metodólogo:** Para descubrir métodos aplicables a problemas específicos de investigación.
5. **Interpretador de código:** Para realizar análisis de datos preliminares.
6. **Estudios multidisciplinares:** Para comprender perspectivas de otras comunidades sobre temas de naturaleza multidisciplinar.
7. **Constructor de plantillas:** Para diseñar formatos específicos para secciones del trabajo.
8. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
9. **Generador previo de diagramas de flujo y contenido:** Para esbozar diagramas iniciales.
10. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.
11. **Generador de datos sintéticos de prueba:** Para la creación de conjuntos de datos ficticios.
12. **Generador de problemas de ejemplo:** Para ilustrar conceptos y técnicas.
13. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
14. **Generador de encuestas:** Para diseñar cuestionarios preliminares.
15. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 06/04/2025

Firma: JAIME BLANCO LEDESMA

7. BIBLIOGRAFÍA

-Rodgers, W., et al (2023). An artificial intelligence algorithmic approach to ethical decision-making in human resource management processes. Hum. Resour. Manag. Rev.)

-Kot, S., Hussain, H. I., Bilan, S., Haseeb, M., & Mihardjo, L. W. W. (2021). The role of artificial intelligence recruitment and quality to explain the phenomenon of employer reputation. Journal of Business Economics and Management, 22(4), 867–883. https://doi.org/10.3846/jbem.2021.14606

-Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial intelligence: definition and background. In *Mission AI: The new system technology* (pp. 15-41). Cham: Springer International Publishing.

-Davidson S. (2024). The economic institutions of artificial intelligence*. Journal of Institutional Economics.* 20: 1–16.

-Russell S., Norvig P. (2010). *Artificial Intelligence A Modern Approach*. Third Edition, Pearson Education.

-Roberts, J., Baker, M., & Andrew, J. (2024). Artificial intelligence and qualitative research: The promise and perils of large language model (LLM) 'assistance'. *Critical Perspectives on Accounting*, 99, 102722.

-The president of the United States (2023). Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence [Federal Register, 88 (210)

-European Parliament (2024). Regulation (EU) 2024/1689.

-Dam, S. et al (2024). A complete survey on llm-based ai chatbots. *arXiv preprint arXiv:2406.16937*.

-IBM Data and AI Team. (2023). *Types of Artificial Intelligence*. IBM. (disponible en: https://www.ibm.com/think/topics/artificial-intelligence-types; última consulta 06/04/2025)

-Khan, H. (2021). *Types of AI | Different Types of Artificial Intelligence Systems*. ResearchGate.

-Prof. Amita. (2024). *Research Paper on Artificial Intelligence & its Types*. International Journal for Research Trends and Innovation, 202-206.

-Losey, D. et al (2024). *AI—The Good, the Bad, and the Scary*. Virginia Tech Engineering. (disponible en: https://eng.vt.edu/magazine/stories/fall-2023/ai.html; última consulta 06/04/2025)

-Duggal, N. (2025). *20+ Advantages and Disadvantages of AI*. Simplilearn. (disponible en: https://www.simplilearn.com/advantages-and-disadvantages-of-artificial-intelligence-article; última consulta 06/04/2025).

-China, C. R. (2024). *Artificial Intelligence Advantages & Disadvantages*. IBM. (disponible en: https://www.ibm.com/think/insights/artificial-intelligence-advantages-disadvantages; última consulta 06/04/2025).

-Pujari, V., et al (2020). *Advantages And Disadvantages of Artificial Intelligence*. National Seminar on "Trends in Geography, Commerce, IT And Sustainable Development", Aayushi International Interdisciplinary Research Journal

-Cortés, M. (2025). *Advantages and Challenges of AI in Companies*. Esade Business School. (disponible en: https://www.esade.edu/beyond/en/advantages-and-challenges-of-ai-in-companies/; última consulta 06/04/2025).

-Gao, Z., & Qian, Q. (2022). *The Risk and Benefits of Applying Artificial Intelligence in Business Discussions*. University of Leicester. BCP Business & Management, 30. 808-812.

-Agencia Española de Supervisión de Inteligencia Artificial (AESIA). (2023). *Real Decreto 729/2023, de 22 de agosto, por el que se aprueba el Estatuto de la Agencia Española de Supervisión de Inteligencia Artificial.*

-Diario Oficial de la Unión Europea. (2024). *Reglamento (UE) 2024/1689 sobre inteligencia artificial.*

-Parlamento Europeo. (2025). *EU AI Act: First Regulation on Artificial Intelligence*. European Parliament.

-Kusche, I. (2024). *Possible Harms of Artificial Intelligence and the EU AI Act: Fundamental Rights and Risk*. Journal of Risk Research. 1-14.

-Park, S. (2024). *Bridging the Global Divide in AI Regulation: A Proposal for a Contextual, Coherent, and Commensurable Framework*. Washington International Law Journal, 33(2).

-Liu, J. (2024). *Artificial Intelligence and International Law: The Impact of Emerging Technologies on the Global Legal System*. Economics, Law and Policy, 7(2).

-Arvidsson, M., & Noll, G. (2023). *Artificial Intelligence, Decision Making and International Law*. Nordic Journal of International Law, 92(1) 1-8.

-Flores, Á. J. (2024). *Los sesgos algorítmicos en la toma de decisiones automatizadas: retos y oportunidades para el sistema jurídico peruano*. Revista Iberoamericana de Derecho Informático, 15(2), 159-170.

-García-Campos, J. et al (2022). *Tres grandes enigmas de los sesgos cognitivos*. SCIO. Revista de Filosofía, 22, 99-125.

-Ginès i Fabrellas, A. (2024). *Analítica de personas y discriminación algorítmica en procesos de selección y contratación*. LABOS Revista De Derecho Del Trabajo Y Protección Social, 5, 99-130.

-Moine, M. B. et al (2022). *Sesgos en decisiones de selección de personal que dificultan el acceso de mujeres a puestos de responsabilidad*. XI Congreso de Administración del Centro de la República.

-Cruzado, M. (2024). *Impacto de los sesgos en Recursos Humanos a través de la IA*. Negro sobre blanco.

-Silberg, J., & Manyika, J. (2019). *Notes from the AI frontier: Tackling bias in AI (and in humans)*. McKinsey & Company

-Floridi, L., & Colws, J. (2019). *A Unified Framework of Five Principles for AI in Society.* Harvard Data Science Review.

-Martin, K. (2021). *Designing ethical algorithms: Challenges and best practices*. MIS Quarterly Executive 18(2). 129-142

-Tsamados, A., et al. (2022). *The ethics of algorithms: key problems and solutions.* AI & Soc 37, 215–230

-*Bitzer, T. (2022). Algorithmic Transparency in Action: How and Why Do Companies Disclose Information on Algorithms? Americas Conference on Information Systems.*

-Llamas, J. Z. (2022). *Algorithmic Transparency as a Foundation of Accountability*. SSRN Electronic Journal.

-Kossow, N. et al (2021). *Algorithmic Transparency and Anti-Corruption Measures*. - Transparency International Report.

-Valderrama, M., et al (2023). *State of the Evidence: Algorithmic Transparency*. AI Policy Observatory.

-Vivek, R. (2023). *Enhancing diversity and reducing bias in recruitment through AI: A review of strategies and challenges*. Informatics, Economics and Management, 2(4), 0101–0118.

-Aditya, B. V., et al (2024). *Integrating AI in Recruitment: Pathways to Inclusive and Diverse Talent Management*. International Advanced Research Journal in Science, Engineering and Technology, 11(10), 138–147.

-Albaroudi, E., et al (2024). *A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring*. AI Systems: Theory and Applications, 5, 383–404.

8. ANEXO

Anexo I: tabla de Landis & Koch sobre índice Kappa Cohen

Casi perfecto	>0.8
Sustancial	>0.6
Moderado	>0.4
Regular	>0.2
Ligero	0-0.2
Deficiente	≈0

Anexo II: Resultados del modelo de regresión logística

```
Call:
glm(formula = hiring_decision ~ ., family = "binomial", data = train_data)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -4.7267     0.6195  -7.630 2.35e-14 ***
age                    0.2520     0.3564   0.707   0.4795
gender                -0.2809     0.2158  -1.301   0.1932
education_level        3.1862     0.4238   7.519 5.53e-14 ***
experience_years       1.7513     0.3732   4.693 2.70e-06 ***
previous_companies     0.5913     0.3001   1.970   0.0488 *
distance_from_company -0.5744     0.3728  -1.541   0.1234
interview_score        1.8930     0.3945   4.798 1.60e-06 ***
skill_score            2.6493     0.3946   6.713 1.90e-11 ***
personality_score      2.4447     0.3957   6.178 6.47e-10 ***
recruitment_strategy  -5.3591     0.4394 -12.196  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 949.13  on 749  degrees of freedom
Residual deviance: 552.08  on 739  degrees of freedom
AIC: 574.08

Number of Fisher Scoring iterations: 6

Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 478  58
         1  53 161

               Accuracy : 0.852
                 95% CI : (0.8245, 0.8767)
    No Information Rate : 0.708
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6396

 Mcnemar's Test P-Value : 0.7042

            Sensitivity : 0.7352
            Specificity : 0.9002
         Pos Pred Value : 0.7523
         Neg Pred Value : 0.8918
             Prevalence : 0.2920
         Detection Rate : 0.2147
   Detection Prevalence : 0.2853
      Balanced Accuracy : 0.8177

       'Positive' Class : 1
```

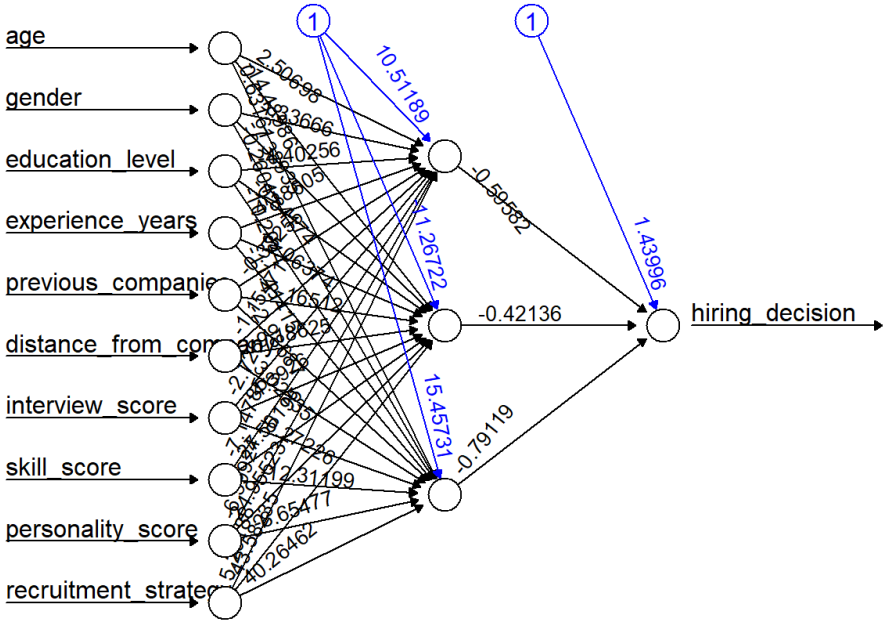
Anexo III: resultados árbol de decisión

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 474  46
         1  57 173

               Accuracy : 0.8627
                 95% CI : (0.8359, 0.8865)
    No Information Rate : 0.708
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6727

 Mcnemar's Test P-Value : 0.3245

            Sensitivity : 0.7900
            Specificity : 0.8927
         Pos Pred Value : 0.7522
         Neg Pred Value : 0.9115
             Prevalence : 0.2920
         Detection Rate : 0.2307
   Detection Prevalence : 0.3067
      Balanced Accuracy : 0.8413

       'Positive' Class : 1
```

Anexo IV: predicciones modelo KNN

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 504 111
         1  27 108

               Accuracy : 0.816
                 95% CI : (0.7864, 0.8431)
    No Information Rate : 0.708
    P-Value [Acc > NIR] : 7.183e-12

                  Kappa : 0.4985

 Mcnemar's Test P-Value : 1.601e-12

            Sensitivity : 0.4932
            Specificity : 0.9492
         Pos Pred Value : 0.8000
         Neg Pred Value : 0.8195
             Prevalence : 0.2920
         Detection Rate : 0.1440
   Detection Prevalence : 0.1800
      Balanced Accuracy : 0.7212

       'Positive' Class : 1
```

Anexo V: resultados modelo de predicción ANN




```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 441  37
         1  90 182

               Accuracy : 0.8307
                 95% CI : (0.8019, 0.8568)
    No Information Rate : 0.708
    P-Value [Acc > NIR] : 5.033e-15

                  Kappa : 0.6176

 Mcnemar's Test P-Value : 3.945e-06

            Sensitivity : 0.8311
            Specificity : 0.8305
         Pos Pred Value : 0.6691
         Neg Pred Value : 0.9226
             Prevalence : 0.2920
         Detection Rate : 0.2427
   Detection Prevalence : 0.3627
      Balanced Accuracy : 0.8308

       'Positive' Class : 1
```

Anexo VI: resultados de predicción modelo random forest

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 429  20
         1 102 199

               Accuracy : 0.8373
                 95% CI : (0.8089, 0.863)
    No Information Rate : 0.708
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6456

 Mcnemar's Test P-Value : 2.244e-13

            Sensitivity : 0.9087
            Specificity : 0.8079
         Pos Pred Value : 0.6611
         Neg Pred Value : 0.9555
             Prevalence : 0.2920
         Detection Rate : 0.2653
   Detection Prevalence : 0.4013
      Balanced Accuracy : 0.8583

       'Positive' Class : 1
```

Anexo VII: resultados de los modelos de support vector machine

Laplace

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 496  72
         1  35 147

               Accuracy : 0.8573
                 95% CI : (0.8302, 0.8816)
    No Information Rate : 0.708
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6369

 Mcnemar's Test P-Value : 0.0005009

            Sensitivity : 0.9341
            Specificity : 0.6712
         Pos Pred Value : 0.8732
         Neg Pred Value : 0.8077
             Prevalence : 0.7080
         Detection Rate : 0.6613
   Detection Prevalence : 0.7573
      Balanced Accuracy : 0.8027

       'Positive' Class : 0
```

Vanilla

```
 Setting default kernel parameters
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 487  60
         1  44 159

               Accuracy : 0.8613
                 95% CI : (0.8345, 0.8853)
    No Information Rate : 0.708
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6573

 Mcnemar's Test P-Value : 0.1413

            Sensitivity : 0.9171
            Specificity : 0.7260
         Pos Pred Value : 0.8903
         Neg Pred Value : 0.7833
             Prevalence : 0.7080
         Detection Rate : 0.6493
   Detection Prevalence : 0.7293
      Balanced Accuracy : 0.8216

       'Positive' Class : 0
```

RBF

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 478  68
         1  53 151

               Accuracy : 0.8387
                 95% CI : (0.8104, 0.8643)
    No Information Rate : 0.708
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6018

 Mcnemar's Test P-Value : 0.2031

            Sensitivity : 0.9002
            Specificity : 0.6895
         Pos Pred Value : 0.8755
         Neg Pred Value : 0.7402
             Prevalence : 0.7080
         Detection Rate : 0.6373
   Detection Prevalence : 0.7280
      Balanced Accuracy : 0.7948

       'Positive' Class : 0
```