Facultad de Ciencias Económicas y Empresariales

PROBLEMAS ÉTICOS EN TORNO A LA TOMA DE DECISIONES DE RECURSOS HUMANOS CON INTELIGENCIA ARTIFICIAL

Autor: Carlota Doussinague Gutiérrez

Director: Raúl González Fabre

MADRID | Junio 2023

ÍNDICE

Índice de figuras

Resumen

La inteligencia artificial (IA) ha sido uno de los temas más relevantes en la última década, debido a su capacidad para transformar los procesos de toma de decisiones en diversas áreas de la empresa, incluyendo los recursos humanos. Los sistemas de IA se basan en la recolección de grandes cantidades de datos y su procesamiento mediante algoritmos y modelos matemáticos, con el fin de detectar patrones y realizar predicciones con alta precisión. Esto ha llevado a una gran cantidad de aplicaciones de la IA en diferentes sectores, incluyendo la salud, la educación, el comercio y la industria, entre otros.

En este trabajo, se aborda el tema de la aplicación de la IA en los recursos humanos para la toma de decisiones, y las implicaciones éticas que eso conlleva. La gestión de los recursos humanos es un área crítica para el éxito de cualquier organización, y al tratar con datos sensibles de los empleados, las organizaciones se han de asegurar de que las decisiones tomas al respecto sean equitativas, justas y transparentes. Estas asunciones entran en conflicto cuando la empresa no es capaz de explicar el proceso desarrollado por la IA que ha intervenido. En el trabajo se abordará esta problemática y se propondrán soluciones que puede adoptar toda empresa, además de proporcionar un ejemplo de estas prácticas.

Palabras clave

Inteligencia artificial, algoritmo, *machine learning*, ética, IA responsable

Abstract

Artificial intelligence (AI) has been one of the most relevant topics in the last decade, due to its ability to transform, automate and improve decision-making processes in various areas of the company, including human resources. AI systems are based on collecting large amounts of data and processing them using algorithms and mathematical models, in order to detect patterns and make highly accurate predictions. This has led to a large number of AI applications in different sectors, including health, education, commerce and industry, among others.

In this thesis, issues associated with the application of AI in human resources for decision-making are addressed, and the ethical implications that this entails. Human resource management is a critical area for the success of any organization, and when dealing with sensitive employee data, organizations need to ensure that the decisions they make in this regard are equitable, fair and transparent. These assumptions come into conflict when the company is not able to explain the process developed by the AI that has intervened. The thesis addresses these problems and offers solutions that can be adopted by any company, in addition to providing an example of these practices.

Key words

Artificial intelligence, algorithm, machine learning, ethics, responsible AI

AUC	*Area under the curve*
BCG	Boston Consulting Group
CV	*Curriculum vitae*
GBM	*Gradient Boosting Machine*
IA	Inteligencia artificial
IBM	International Business Machines
ML	*Machine learning*
MLOps	*Machine learning operations*
RAI	*Respinsible artificial intelligence*
ROC	*Receiver operating characteristic*
RRHH	Recursos humanos

1. INTRODUCCIÓN

1.1 Contextualización del tema

La inteligencia artificial (IA) se refiere a la capacidad de las máquinas para realizar tareas que normalmente requieren inteligencia humana, como el aprendizaje, el razonamiento y la percepción. La IA se basa en la combinación de técnicas de estadística, informática y matemáticas para procesar grandes cantidades de datos y proporcionar resultados precisos y fiables. En la última década, la IA ha experimentado un crecimiento significativo en su aplicación en diferentes áreas, incluyendo los recursos humanos.

Las organizaciones buscan cada vez más soluciones que les permitan mejorar su eficiencia de manera transversal, lo que implica buscar la optimización en la gestión de los recursos humanos y, en consecuencia, lograr el mejor desempeño de los trabajadores. Por ello, la IA se ha convertido en una herramienta clave en esta área de la empresa y se utiliza para automatizar procesos, como la selección de candidatos, la evaluación del desempeño de los empleados y la planificación de la carrera profesional de los mismos, entre otros. Estos procesos pueden ser mejorados significativamente mediante el uso de algoritmos de aprendizaje automático, que pueden procesar grandes cantidades de datos y detectar patrones que resultan difíciles de encontrar para los seres humanos.

La IA es capaz de proporcionar información valiosa para la toma de decisiones en recursos humanos. La información de los empleados, como su desempeño, historial de empleo y calificaciones, se puede procesar y analizar para proporcionar una comprensión más profunda de las necesidades y desafíos de la organización. De esta manera, se pueden tomar decisiones más informadas y precisas en la gestión de los recursos humanos, lo que a su vez contribuye a mejorar la productividad y la eficiencia de la organización, y se traduce en un impacto significativo en los resultados empresariales.

Sin embargo, la automatización de la toma de decisiones en el ámbito de los recursos humanos mediante la IA ha suscitado preocupaciones éticas y sociales significativas. Esto se debe a que los algoritmos de aprendizaje automático utilizados para la selección, gestión y promoción de los empleados pueden generar resultados injustos, discriminatorios y

perjudiciales para las personas involucradas. Además, se han levantado problemáticas en torno a la transparencia y trazabilidad de dichas decisiones tomadas por máquinas.

1.2 Objetivos del TFG

El objetivo principal de este TFG es analizar los problemas éticos que pueden surgir en la aplicación de sistemas de inteligencia artificial en la toma de decisiones de recursos humanos y proponer medidas para abordarlos. Para lograr este objetivo, se plantean los siguientes objetivos secundarios:

- Identificar las áreas de aplicación de la inteligencia artificial en los recursos humanos y el funcionamiento de los sistemas de inteligencia artificial en la toma de decisiones.
- Revisar críticamente la literatura existente sobre la aplicación de la inteligencia artificial en la toma de decisiones de recursos humanos y analizar los problemas éticos que se han identificado.
- Diseñar un estudio que permita analizar posibles problemas éticos que puedan surgir a partir de un algoritmo capaz de tomar decisiones de recursos humanos en base a un *dataset*.
- Realizar un análisis detallado de los resultados obtenidos del estudio y proponer medidas para abordar los problemas éticos identificados.

Así pues, el objetivo principal de este TFG está estrechamente relacionado con el factor ético en la toma de decisiones de recursos humanos a través de sistemas de inteligencia artificial. Los objetivos secundarios se enfocan en aspectos específicos que ayudarán a cumplir con el objetivo principal, como la revisión de la literatura existente y el diseño y análisis empírico con datos ficticios que puedan extrapolarse a *datasets* reales. Además, se utilizarán herramientas estadísticas y de análisis de datos, y se recopilarán y analizarán datos de diferentes fuentes para llevar a cabo la investigación.

1.3 Metodología

En primera instancia se hará una revisión de la literatura con el fin de contextualizar el trabajo de investigación y asentar una base clara para aterrizar correctamente lo que se va a analizar más adelante en la parte práctica del trabajo, es decir, qué temas se deben tratar, qué problemáticas se deben abordar y cuáles son los posibles usos que se le puede dar.

Con el fin de analizar los problemas éticos asociados a la aplicación de algoritmos de inteligencia artificial en la toma de decisiones de recursos humanos, se llevará a cabo un análisis empírico utilizando un *dataset* con datos ficticios y será aplicable a datos reales. El estudio consistirá en aplicar un algoritmo de aprendizaje automático al *dataset* y analizar los resultados obtenidos. El objetivo del estudio será identificar posibles problemas éticos que puedan surgir a partir de un algoritmo capaz de tomar decisiones de recursos humanos en base a ese *dataset*.

Para llevar a cabo este estudio, se utiliza un *dataset* que contine información relevante para la toma de decisiones de recursos humanos.

Una vez seleccionado el *dataset*, se aplica un algoritmo de aprendizaje automático para analizar los datos y obtener resultados. Es importante la elección del algoritmo y que este sea adecuado para la toma de decisiones de recursos humanos, puesto que afectará directamente a los resultados obtenidos.

Finalmente, se analizarán los resultados obtenidos para identificar posibles problemas éticos que puedan surgir a partir de un algoritmo capaz de tomar decisiones de recursos humanos en base a ese *dataset*. Se deberán tener en cuenta aspectos como la equidad y la justicia en el empleo, la privacidad de los datos y la transparencia.

1.4 Desarrollo

El trabajo se estructura de la siguiente manera:

En el capítulo 2, '**Error! Reference source not found.**', se hace una revisión de la literatura existente acerca de los conceptos básicos de inteligencia artificial, las áreas de aplicación de la IA en recursos humanos, y el funcionamiento de los sistemas de IA en dicho departamento.

En el capítulo 3, 'DESAFÍOS QUE PROPONE LA INSERCIÓN DE LA IA EN RRHH', se abordan los principales desafíos que propone la inserción de sistemas de IA en el proceso de toma de decisiones de recursos humanos de la empresa. Se estudian tanto problemáticas éticas como cuestiones legales.

En el capítulo 4, 'IA RESPONSABLE', se profundiza en lo que es la RAI (Responsible Artificial Intelligence), cómo adelantarse a la regulación que todavía no se ha publicado y qué herramientas existen para contribuir con el uso responsable de la IA. Además, se desarrolla en profundidad uno de los mayores problemas a los que se enfrentan las empresas que incorporan sistemas de IA en sus procesos, la transparencia de los modelos.

En el capítulo 5, 'PRÁCTICA', se proporciona una demostración práctica de la herramienta Azure ML que permite la incorporación de prácticas responsables durante el desarrollo del modelo. El estudio se centra en datos de RRHH y se presentarán los resultados obtenidos del mismo.

2. MARCO TEÓRICO

2.1 Conceptos básicos de IA

El presente apartado pretende desarrollar conceptos básicos sobre IA, qué es y qué elementos intervienen en su desarrollo. Esta explicación será breve y no se entrará en detalle, pero pretende orientar al lector y ayudar a su comprensión transversal del estudio.

La inteligencia artificial se refiere a la capacidad de las máquinas para realizar tareas que requieren inteligencia humana, como el aprendizaje, la toma de decisiones, el reconocimiento de voz y demás. La clave para el funcionamiento de la IA es la capacidad de las máquinas para aprender de los datos. Para ello, se utilizan algoritmos de aprendizaje automático (o *machine learning*, ML), que serán explicados a continuación. Estos algoritmos se entrenan con grandes cantidades de datos, que a menudo son recopilados y etiquetados por seres humanos. (Kontsevoi, 2021)

En su libro *Introduction to Machine Learning*, Alpaydin (2010) habla sobre cómo el ML se basa en la idea de que estos modelos, de aprendizaje automático, son capaces de identificar patrones en los datos y usarlos para hacer predicciones o tomar decisiones, sin la necesidad de que se les haya dado una explicación detallada de cómo hacerlo. En lugar de ser programadas explícitamente, las máquinas son entrenadas en un proceso iterativo, utilizando datos de entrenamiento para ajustar los parámetros del modelo y mejorar su capacidad para hacer predicciones precisas. Existen tres tipos principales de ML: el aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje por refuerzo.

El aprendizaje supervisado se utiliza cuando se tienen datos etiquetados y se quiere predecir una variable de salida en función de las entradas. Por ejemplo, tras entrenar al modelo con un historial de precios de casas con una serie de características definidas, el algoritmo debería ser capaz de predecir el precio de las futuras casas con características diferentes y "nuevas" para el modelo. (Alpaydin, 2010)

El aprendizaje no supervisado se utiliza cuando no se tienen etiquetas y se quieren descubrir patrones o estructuras ocultas en los datos. Se utilizan algoritmos de *clustering*, por ejemplo,

para agrupar los datos en grupos con características similares; si, por ejemplo, alimentas al algoritmo de *clustering* con un historial de compras sin haber etiquetado cada transacción, el algoritmo debería ser capaz de agrupar las transacciones en función del tipo de producto, la hora del día, el método de pago y demás, y sacaría patrones como los productos que más se compran, conjuntos de productos que suelen comprarse juntos, o qué tipo de consumidor suele comprar a qué hora, entre otros. (Alpaydin, 2010)

Por último, el aprendizaje por refuerzo se utiliza cuando el modelo aprende a través de la retroalimentación y la recompensa, de manera que busca maximizar una función de recompensa mientras interactúa con su entorno. Este tipo de algoritmos se entrenan con la prueba y el error de manera iterativa, por ejemplo, con partidas de ajedrez, la resolución de puzles o escapando de un laberinto, de manera que el algoritmo aprende y mejora después de cada intento. (Alpaydin, 2010)

Además, es importante tener en cuenta la calidad de los conjuntos de datos (*datasets*) utilizados para entrenar y evaluar los algoritmos de IA, ya que es crucial para su éxito. Estos *datasets* han de ser lo suficientemente abundantes, precisos, representativos y relevantes para la tarea que se está realizando, y han de estar adaptados a la población a la que se aplicará el modelo de IA en cuestión.

En cualquier caso, el objetivo de toda IA es que el algoritmo sea capaz de generalizar a nuevos datos, es decir, que consiga hacer predicciones precisas en datos que no ha visto antes. En este trabajo de investigación se desarrollará un modelo de ML y se expondrá su funcionamiento, y el proceso que el algoritmo lleva a cabo para realizar predicciones basadas en los datos que se le proporciona.

2.2 Aplicaciones de IA en RRHH

La introducción de sistemas de inteligencia artificial en los distintos departamentos de la empresa es un tema que está a la orden del día, y el departamento de recursos humanos no es una excepción. Se trata de una tecnología disruptiva capaz de transformar radicalmente los procesos, actividades, productos y servicios que desarrolla la empresa en su día a día.

En este epígrafe se expondrán las principales funciones del departamento de RRHH de la empresa, y cómo la IA puede ser utilizada en los distintos procesos que se llevan a cabo dentro de éste.

La función de recursos humanos dentro del ámbito empresarial es aquella que opera y trata con los activos más importantes y valiosos de la organización – las personas – por lo que desarrolla un papel vital para el crecimiento de la empresa.

Las principales tareas que desempeña este departamento son (Mayhew, 2019):

- El reclutamiento de los empleados, publicando anuncios de puestos de trabajo disponibles, realizando el seguimiento de los candidatos y desarrollando el proceso de selección de éstos.
- Realizar los entrenamientos y contribuir con el desarrollo profesional de los trabajadores, centrándose en el área de conocimiento de cada uno de ellos para permitirles evolucionar en su trayectoria profesional dentro de la empresa.
- Asegurarse de la salud y bienestar de los empleados.
- Gestionar las compensaciones y beneficios de los empleados en sus diversas funciones, con el fin de asegurar un sistema de compensación justo, basado en las capacidades y habilidades de cada uno y de las reglas del mercado. Además, tratan temas relacionados con seguros, pensiones y demás beneficios de los empleados.
- Realizar seguimientos del desempeño de las tareas de los empleados en el entorno laboral.
- Asegurarse del cumplimiento de la ley en materia laboral por parte de la organización hacia las condiciones laborales de su plantilla.

La selección de personal es una de las áreas de RRHH que más rápido ha adoptado el uso de la IA. Según Dattner, Chamorro-Premuzic, Buchband y Schettler (2019) , el uso de la IA a la hora de realizar procesos de selección permite a las empresas mejorar su capacidad de identificar a los mejores candidatos de manera más rápida y barata. Esto supone un poder sin precedentes a la hora de tomar decisiones, en cuanto a capital humano se refiere, basándose

meramente en datos. Lo mismo sucede con el poder proporcionar *feedback* personalizado a cada candidato, basándose en datos objetivos observados por la máquina a nivel personalidad, gesticulación y tono de voz, entre otros. La IA se utiliza para procesar grandes cantidades de currículums (CVs), buscar patrones para identificar a los candidatos más adecuados para el puesto, y realizar entrevistas en línea siendo capaz de analizar el lenguaje corporal e incluso las emociones de los candidatos.

Los procesos de formación y desarrollo también se están viendo afectados por la introducción de sistemas de IA. Según Kim-Schmid y Raveendhran (2022), una de las mayores causas por las que 4,2 millones de personas abandonaron su puesto de trabajo en agosto de 2022 es la falta de oportunidades de formación y desarrollo para las personas dentro de las organizaciones. La IA resulta interesante para abordar este problema, ya que es capaz de ayudar a las empresas a identificar áreas en las que los empleados necesitan capacitación, y de proporcionar contenido personalizado para satisfacer sus necesidades individuales. La IA también puede ayudar a las empresas a evaluar el impacto que dichas formación y desarrollo tienen en el rendimiento de los empleados, mediante el análisis de datos en tiempo real y el seguimiento de las mejoras en el rendimiento y la productividad. Como resultado, las empresas pueden tomar decisiones más informadas sobre la inversión en la formación y el desarrollo de los empleados, y optimizar sus estrategias de gestión del talento para impulsar el éxito del negocio.

Acompañando a lo anterior, las empresas también pueden aprovechar la IA para la evaluación del rendimiento de sus empleados, ya que contribuiría a identificar áreas en las que los empleados necesitan mejorar y a proporcionar retroalimentación personalizada. Este *feedback* podría ser emitido en tiempo real a los empleados, de manera que verían cómo pueden optimizar su trabajo diario y se les daría la oportunidad de ajustar su comportamiento en consecuencia y, como resultado, mejorar su desempeño en general. Además, se podría llegar a predecir el éxito futuro de cada empleado midiendo su potencial e identificar a aquellos empleados que pueden estar en riesgo de abandonar la empresa, de manera que se darían soluciones adecuadas para intentar retener a estas personas. (Kim-Schmid & Raveendhran, 2022)

En definitiva, se identificarían las habilidades y competencias clave de los empleados, se analizarían sus puntos fuertes y débiles, y se proporcionarían recomendaciones personalizadas de capacitación y desarrollo, con el fin de mejorar el rendimiento y el potencial de estos.

La IA también se puede utilizar para monitorear la salud y bienestar de los empleados de una organización. La empresa pretende abogar por la tranquilidad, comodidad y sentido de pertenencia que los empleados sienten hacia la firma. Sin embargo, se tratan de datos altamente intrusivos, y la empresa ha de ser transparente con su recolección y uso. La IA podría ofrecer recomendaciones personalizadas con el fin de mejorar la calidad de vida de los empleados y, en consecuencia, generar un impacto positivo en la retención de talento dentro de la empresa.

A la hora de tomar decisiones en cuanto a la compensación de los empleados, la IA es útil para tener en cuenta factores que un mánager no consideraría al no estar respaldado por un soporte estadístico. Además del rendimiento, existen muchos otros factores que están directamente relacionados con la compensación que cada empleado debería tener, de manera que la IA contribuye a evitar pagar de menos o de más. Por ejemplo, es importante considerar el precio de mercado que tienen las habilidades de cada uno, cuánta demanda hay en el mercado para dichas habilidades o si un buen rendimiento debería ser compensado a través de la retribución fija o variable. Todo ello contribuye a la fidelización de los empleados, de manera que si se ven justamente recompensados por su trabajo disminuye su intención de abandonar la empresa.

La gestión del desempeño es otra área de aplicación común de la IA en los recursos humanos. Los sistemas de IA pueden analizar datos de desempeño y proporcionar comentarios y sugerencias personalizadas para mejorar el rendimiento de los empleados. Según, el uso de la IA en la gestión del desempeño puede mejorar la eficiencia y la precisión de la evaluación del desempeño, así como proporcionar una retroalimentación más personalizada y significativa para los empleados (Hawksworth, Berriman, & Goel, 2020).

En definitiva, la IA en RRHH es realmente candente en la actualidad. El departamento cuenta con una multitud de procesos repetitivos y basados en datos que la empresa puede recopilar y en consecuencia automatizar el análisis de estos y de los procesos para los que son relevantes. Esto no solo favorece a la productividad de la empresa, sino que también supone una mejora en la calidad de vida y condiciones laborales de los empleados. Si bien parece tener impactos significativamente positivos en muchos aspectos, se han de tener en cuenta las limitaciones, riesgos y desafíos que propone esta tecnología. Estos serán presentados en el capítulo 'DESAFÍOS QUE PROPONE LA INSERCIÓN DE LA IA EN RRHH'.

2.3 Toma de decisiones: Personas vs Máquinas y su colaboración para RRHH

En este epígrafe se explorará cómo difiere la forma en la que las personas toman decisiones con respecto a una máquina o una IA y las implicaciones que esto conlleva a la hora de tomar decisiones empresariales. Además, se profundizará en la importancia de la interlocución entre ambas partes en la toma de decisiones de RRHH.

El proceso de toma de decisiones de las personas en comparación con el desarrollado por una IA difiere en muchos aspectos. Aunque el enfoque hacia la resolución de problemas y el empleo de habilidades cognitivas es común, la forma y los recursos utilizados no son comparables.

Como se ha explicado anteriormente, la IA funciona de forma automática una vez se le han proporcionado los datos con los que va a trabajar. De manera que el proceso que lleva a cabo para generar una solución consiste en: recopilar los datos, procesarlos (limpieza, normalización, tratamiento de valores atípicos, etc.), condensarlos (extraer las características más relevantes o generar variables interesantes a partir de las que se tienen), construir el modelo (entra en juego el ML, entrenamiento y demás), evaluar el rendimiento del mismo y, por último, tomar decisiones en base a los resultados, es decir, proporcionarle nuevos datos para que genere una respuesta.

La IA se basa en principios de inteligencia computacional para la toma de decisiones, alimentándose de algoritmos de ML, y procesos estadísticos capaces de analizar cantidades masivas de datos y generar resultados (Alpaydin, 2010).

Por estas razones, es fundamental comprender cómo se formulan los algoritmos, qué datos se utilizan para su entrenamiento y qué criterios se tienen en cuenta en el proceso de toma de decisiones con estos sistemas de IA.

Como establecen McKendrick y Thurai (2022), la IA está diseñada para tomar decisiones cuando se emplean datos, parámetros y variables que sobrepasan las capacidades de procesamiento humanas. Si bien esta tecnología, por lo general, toma las decisiones correctas en base a las directrices que se le proporcionan, es incapaz de incorporar el factor humano que subyace en toda decisión de la vida real.

En contraposición, la persona se basa en intuición. Está limitada por la cantidad de información que es capaz de procesar; le resulta complejo identificar patrones poco evidentes o predecir de manera acertada, y utiliza la percepción, la razón y el juicio. Por ello las decisiones de una persona pueden estar sesgadas, mayoritariamente, en función de sus experiencias, emociones y prejuicios. Si bien es importante mencionar que esto no difiere del todo con la IA, ya que ésta también puede presentar sesgos, como se explicará en el apartado 'Implicaciones éticas en torno a la toma de decisiones en la empresa'. Además, la velocidad a la que una persona es capaz de procesar información para llegar a una conclusión es considerablemente más lenta que aquella de la IA.

El principal problema con el que nos encontramos respecto a la IA en la toma de decisiones es su falta de capacidad de razonar la respuesta que ha proporcionado, lo cual explica por qué se utiliza el conocido término de "caja negra", haciendo referencia a su opaca naturaleza. Es decir, que no siempre está claro cómo el algoritmo ha llegado a una determinada decisión. Esta percepción de caja negra es contraria a la que se tiene con las personas, ya que estas son transparentes y su lógica puede ser entendida y evaluada a través del razonamiento y la introspección. (McKendrick & Thurai, 2022)

Aplicando esta lógica a la introducción de IAs en procesos empresariales, se puede inferir que estos sistemas tienen el potencial de impactar negativamente en decisiones empresariales. McKendrick y Thurai (2022) hacen hincapié en la importancia de la convivencia entre humano y máquina a la hora de tomar decisiones en la empresa, y más aún

en aquellas que afectan a la vida de las personas, como son las tomadas en RRHH. Los autores definen tres niveles de decisión en los que las máquinas están involucradas:

- Soporte en decisiones de alto nivel, en el que las decisiones se toman principalmente por las personas.
- Soporte aumentado de las máquinas, en el que una IA genera soluciones, propone recomendaciones y hace análisis de datos que más tarde serán validados y estudiados por las personas.
- Configuraciones altamente automatizadas, en las que intervienen principalmente máquinas. Sin embargo, según los autores, estos procesos también necesitarían un *human–in–the–loop* para alguna excepción.

En definitiva, todos los escenarios en los que una IA se ve involucrada requieren de intervención humana en algún momento del proceso.

En materia de RRHH esta cuestión cobra especial importancia debido a la naturaleza de las decisiones que toma el departamento, que afectan a la vida de los empleados.

Como ejemplo se encuentra el conocido caso de Amazon en 2015, en el que un equipo de la empresa desarrolló un algoritmo de ML que parecía una cómoda solución al tedioso proceso de reclutamiento. Se trataba de alimentar al algoritmo con los CVs de todos los candidatos para el puesto y que este recomendara los cinco mejores. El equipo lo desarrolló con la intención de contratar a esos cinco perfiles sin llevar a cabo otro proceso adicional. Unos meses después llegaron a la conclusión de que el algoritmo estaba discriminando en cuestión de género y únicamente recomendaba hombres. Esto se debe a que los datos de entrenamiento constituían 10 años de perfiles contratados por la empresa, de los cuales, la gran mayoría, eran perfiles masculinos. Una vez identificado el problema, y pese a tener la capacidad de erradicar el sesgo del algoritmo, nadie garantizaba que éste no fuese a encontrar otro atributo no técnico para seleccionar. Además, el algoritmo era capaz de diferenciar a mujeres de hombres aun contando con CVs anónimos, debido a logros como “capitana del equipo femenino de ajedrez”. De manera que, en la actualidad, se utiliza la herramienta en el proceso

de selección, pero no como único trámite. (Kim-Schmid & Raveendhran, 2022) (McKendrick & Thurai, 2022)

Este ejemplo refuerza la idea de que se necesita capital humano para la supervisión de los algoritmos o que colabore e intervenga en el proceso de toma de decisiones. Pese a tener controles sobre la tecnología, ésta sigue teniendo problemas de transparencia y falta de capacidad de razonamiento subjetivo, propio de las personas humanas.

3. DESAFÍOS QUE PROPONE LA INSERCIÓN DE LA IA EN RRHH

3.1 Implicaciones éticas en torno a la toma de decisiones en la empresa

En línea con lo discutido en el apartado anterior, 'Toma de decisiones: Personas vs Máquinas y su colaboración para RRHH', la IA presenta una serie de desafíos éticos y morales que han de ser abordados por las organizaciones que incorporen estos sistemas en sus procesos. En el presente epígrafe se exponen las implicaciones éticas que tiene toda decisión empresarial que se toma con la contribución de una IA.

La ética en el lugar de trabajo es un tema complejo que requiere habilidades y reflexión moral para tomar decisiones éticas. En su libro *Moral Reasoning at Work: Rethinking Ethics in Organizations*, Kvalnes (2019) aborda esta cuestión y sostiene que la ética no es algo que se pueda enseñar simplemente como una lista de reglas o principios, sino que es una habilidad que debe ser desarrollada a través de la práctica y la reflexión.

Según Kvalnes (2019), la moralidad en el trabajo no es un tema aislado de la moralidad en general, sino que se relaciona directamente con la identidad personal de los individuos y se ve afectada por el entorno laboral y la cultura organizacional. En este sentido, las organizaciones tienen una responsabilidad en la promoción de una cultura ética y en la formación de líderes que sean capaces de tomar decisiones éticas.

Sin embargo, en la actualidad, con el avance de la tecnología y la inteligencia artificial, surge la pregunta de cómo las máquinas tomarían decisiones éticamente aceptables. En contraposición a las habilidades humanas, la inteligencia artificial se basa en la aplicación de algoritmos y reglas preestablecidas que le permiten tomar decisiones rápidas y precisas en función de los datos proporcionados (Russell & Norvig, 2010).

Por lo tanto, mientras que los seres humanos se basan en la reflexión moral y la empatía para tomar decisiones éticas en situaciones complejas, las máquinas sólo pueden seguir reglas predefinidas y patrones de datos, por lo que no son capaces de tomar decisiones éticas como tal, sino que están condicionadas por las normas establecidas por su desarrollador. De manera

que, el sistema de IA es programado con una serie de criterios, pero no es capaz de discernir o identificar su significado ético.

Además, las decisiones tomadas por una IA pueden estar sesgadas por los datos que se le han proporcionado o por la forma en que se han programado los algoritmos.

En este sentido, Kvalnes (2019) defiende que la ética no es algo que se pueda resolver mediante la aplicación de principios abstractos, sino que se debe abordar en función de las situaciones particulares y complejas que se presentan en el lugar de trabajo. Esto implica que los líderes y trabajadores deben ser capaces de discernir las características éticas relevantes de cada situación y de aplicar los principios éticos adecuados a cada caso particular. Por estas razones se podría concluir que los seres humanos tienen una ventaja sobre las máquinas en la toma de decisiones éticamente aceptables.

Sin embargo, estos algoritmos "éticos" aún presentan desafíos y limitaciones. La definición de un marco de principios y valores que se adapten tanto a los principios éticos universales como a las consideraciones culturales y sociales específicas es una tarea complicada. Como advierten Floridi y otros (2018), muchas organizaciones han establecido qué valores y principios deberían guiar las decisiones tomadas por IAs. Éstos se resumen en los principios utilizados en bioética: beneficencia, no maleficencia, autonomía y justicia, a los que los autores añaden explicabilidad. Los autores resaltan la importancia de que estos principios sean incluidos en el desarrollo en la IA y que se tome un enfoque con múltiples *stakeholders*, desarrolladores, usuarios y legisladores trabajando conjuntamente. Además, inciden en que su estudio se basa en un enfoque occidental y europeo, y que requiere de también de enfoques adicionales, multiculturales, de manera que los planes de acción que proponen sean dinámicos y estén abiertos al cabio y la evolución.

Además, estos algoritmos podrían perpetuar sesgos existentes en los datos utilizados para su entrenamiento, lo que podría generar discriminación o injusticia en la toma de decisiones (Jobin, Ienca, & Vayena, 2019).

Por otro lado, incluso en caso de desarrollar algoritmos “éticos” libres de sesgos, aún queda la cuestión de cómo garantizar la responsabilidad y la transparencia en la toma de decisiones de la IA. En este sentido, los investigadores están trabajando en el desarrollo de marcos éticos y legales para la IA, que incluyen la definición de responsabilidades y mecanismos de rendición de cuentas (Floridi, y otros, 2018). Sin embargo, estos marcos todavía están en una etapa incipiente y queda mucho por hacer para garantizar que la IA tome decisiones éticamente aceptables de manera responsable y transparente.

En conclusión, la toma de decisiones éticas en el lugar de trabajo es una cuestión compleja que requiere habilidades y reflexión moral. Si bien la inteligencia artificial puede ser programada para tomar decisiones en base a parámetros éticos, todavía existen desafíos y limitaciones que deben abordarse. Los seres humanos siguen teniendo una ventaja sobre las máquinas en la ética de la toma de decisiones, pero la IA puede ser una herramienta valiosa si se desarrolla y se utiliza de manera responsable y ética.

3.2 Desafíos que plantea la IA en RRHH

Las diferentes aplicaciones que se le pueden dar a las herramientas de IA en cuestión de recursos humanos estudiadas en el epígrafe ‘**Error! Reference source not found.**’ llevan a concluir que se trata de algo increíblemente útil para las empresas de hoy en día. El atractivo para los empresarios es indudable: la idea de que la automatización de tareas centradas en la toma de decisiones basadas en datos se traduce en mejora de la toma de decisiones, incrementos en el rendimiento y reducción de costes, es una que nadie podría negar. Sin embargo, esta revolucionaria tecnología supone grandes desafíos, de naturalezas distintas, para la empresa, desde cuestiones éticas a legales a problemas de limitación o de confianza en las herramientas. Los desafíos que se van a exponer a continuación son algunas de las causas que limitan el crecimiento de esta tecnología en el sector.

En primer lugar, se ha de tener en cuenta la limitación en información que se posee sobre los algoritmos y las máquinas que hacen posible la automatización de tareas. Estos algoritmos en ocasiones pueden suponer problemas de transparencia, y se denominan, como se ha mencionado anteriormente en este trabajo, como “cajas negras” en las cuales no se sabe con

exactitud qué procedimientos se están llevando a cabo o cómo se están tratando los datos, de manera que la persona responsable de dicha función se ve fuera de control y termina por no fiarse de la inteligencia artificial. Esto es lo que se conoce como aversión al algoritmo (*algorithm aversion*) y, pese a que la IA sea capaz de eliminar el sesgo en las conclusiones a las que llega, los empleados terminan por fiarse más de sus propias decisiones que de las de la máquina. (Kim-Schmid & Raveendhran, 2022)

Todo ello supone un problema para las empresas. Sin embargo, existen maneras de abordar este problema y conseguir que los empleados se sientan más cómodos utilizando esta tecnología. Por ejemplo, proveyendo formación a los empleados sobre cómo interactuar con estas herramientas, ya que algunos conocimientos en estadística pueden ayudar a que los empleados sean capaces de interpretar mejor las recomendaciones algorítmicas que se les han dado y entender la motivación del algoritmo para llegar a ellas, de manera que confíen más en la solución propuesta por la IA. Además, la aversión algorítmica también se ve reducida en el momento en que la persona tiene más poder de decisión sobre el resultado final, de manera que una persona está más cómoda delegando decisiones objetivas en el algoritmo. En este sentido, se necesita llevar a cabo una selección considerada de cómo es el reparto de tareas entre persona y máquina para llegar a soluciones cocreadas por ambas partes. (Kim-Schmid & Raveendhran, 2022)

En segundo lugar, se ha de tener en cuenta, como se ha explicado en el apartado ‘Conceptos básicos de IA’, que los algoritmos requieren de un entrenamiento con datos pasados. Es decir, las recomendaciones que dan se respaldan de alguna manera en decisiones que se han tomado a lo largo del periodo histórico de entrenamiento por humanos. De manera que, si esos datos están sesgados, las recomendaciones que saldrán como *output* del algoritmo también lo estarán (Kim-Schmid & Raveendhran, 2022).

En cualquier caso, a la hora de reducir el sesgo de un algoritmo, es necesario que cada organización tenga bien definido qué considera justo en materia de *output* algorítmicos y que tenga una serie de estándares que consideren qué nivel de transparencia han de tener sus herramientas de IA. Además, se ha demostrado cómo involucrar a distintos equipos de

ingenieros de perfiles diversos para la creación de un único algoritmo reduce el sesgo algorítmico de éstos (Kim-Schmid & Raveendhran, 2022). Esto se debe a que distintas perspectivas pueden identificar diferentes riesgos, características a tener en cuenta para abordar diversas necesidades, incrementa la transparencia y confiabilidad del algoritmo, y se maximiza la utilidad del mismo para poder aplicarlo a diferentes partes de la organización.

En tercer lugar, la introducción de la IA en recursos humanos plantea preocupaciones sobre la privacidad de datos de los empleados. Las empresas tienen la capacidad de monitorear la actividad de sus empleados en tiempo real, práctica que puede resultar altamente intrusiva si se implementa de manera equivocada. El jugar con la privacidad de los empleados puede resultar en problemas de salud mental, estrés, y disminución del sentido de pertenencia a la empresa, entre otros. El uso de estas herramientas de monitoreo se ha incrementado a raíz de la pandemia del Covid-19 y el aumento del número de gente que trabaja desde casa, con el fin de asegurar la productividad y un comportamiento apropiado durante las horas de trabajo. (Kim-Schmid & Raveendhran, 2022)

Algunos ejemplos de empresas que utilizan estas aplicaciones son (Kim-Schmid & Raveendhran, 2022):

- Uber, en la que los conductores tienen que hacerse un *selfie* cuando empiezan a trabajar, con el fin de que no mientan sobre su identidad; en caso de incumplir esta política se les prohibiría el acceso. Además, el desempeño de los conductores se mide con un sistema de puntuación de cinco estrellas por parte de los clientes, de manera que un conductor con puntuaciones bajas consecutivas sería expulsado de la aplicación (estos casos son estudiados además por una persona, la decisión no se basa meramente en la recomendación del algoritmo).
- Ford utilizó durante un tiempo tecnologías de vigilancia desarrolladas con IA con el fin de realizar un seguimiento de sus empleados para asegurar que mantenían la distancia de seguridad entre ellos (durante la pandemia) a través de pulseras con localizadores.

- Amazon tiene la idea de utilizar IA para monitorear los movimientos de sus empleados para conocer qué músculos utilizan durante su horario laboral, con el fin de reducir el número de movimientos repetitivos que realizan en su día a día y evitar el riesgo de cualquier enfermedad o lesión muscular.

Para mitigar este riesgo, es crucial que la empresa sea transparente con sus empleados acerca del fin con el que recopilan esta información. Con el paso de los años, el grado de información que un empleado comparte con comodidad es cada vez mayor en aquellas empresas donde se ha discutido el uso que se le iba a dar a las herramientas de vigilancia. Por ello, en estas empresas se ha incrementado el uso de estas tecnologías de un 30% a un 50% en la última década. Además, se ha descubierto que los empleados acogen mejor estas tecnologías cuando su fin es informativo y no evaluativo, es decir, no existe involucración humana para evaluar las observaciones de la IA. De esta manera, el empleado recibe *feedback* de la máquina sobre su desempeño con el fin de sentirse motivado a mejorar, pero dicho *feedback* no se utilizará para la evaluación del desempeño de la persona en cuestión por lo que no se genera un ambiente de inseguridad o supervigilancia. (Kim-Schmid & Raveendhran, 2022)

Es importante mencionar los potenciales riesgos legales implicados ya que los empresarios pueden tener problemas legales asociados a actuaciones discriminatorias no intencionales llevadas a cabo como consecuencia de recomendaciones proporcionadas por sistemas de IA. En cualquier caso, la regulación en torno a los derechos legales de empleados y empleadores en cuestión de IA está en un momento muy temprano y sigue en constante evolución.

Como mitigantes para este riesgo, se cuenta con que las organizaciones entiendan y estén actualizadas en las regulaciones que les afecten, sobre todo aquellas empresas que operen internacionalmente ya que tendrán que adaptarse a las condiciones de cada país. Aun estando en un momento preliminar, la regulación vigente en Estados Unidos (la Ley de Iniciativa Nacional de IA de 2020 y el proyecto de ley de Responsabilidad Algorítmica de 2022) aboga por la responsabilidad, la transparencia y la justicia de los sistemas de IA.

Además, se espera de las empresas que monitoreen de manera activa los proyectos de ley que se publiquen. Esto les permitirá desarrollar prácticas de gestión de riesgos para anticiparse a

dichas legislaciones, es decir, que los sistemas de IA que diseñen cuenten con varios controles durante su proceso de desarrollo. En el siguiente capítulo se expondrán las prácticas que pueden llevar a cabo las empresas para anticiparse a la regulación en materia de IA responsable.

4. IA RESPONSABLE

4.1 Qué es y cómo abordar su implementación

Atendiendo a la problemática que supone la introducción de la IA en los procesos empresariales, las organizaciones están desarrollando un enfoque de esta tecnología que aborda los dilemas éticos en la medida de lo posible. Es la conocida como IA responsable (Responsible AI, RAI). Según Microsoft, actual desarrollador de sistemas de IA y pionero en el campo, "la RAI es un enfoque para desarrollar, evaluar e implementar estos sistemas de manera segura confiable y ética". El objetivo de la RAI es guiar las decisiones tomadas por una IA, de manera proactiva, para llegar a resultados beneficiosos y equitativos. Este objetivo abarca desde la misma finalidad del sistema hasta la manera en que las personas interactúan con esta tecnología. (Sameki & Gayhardt, 2023)

En línea con las implicaciones legales que se han mencionado en el apartado 'Desafíos que plantea la IA en RRHH', los gobiernos están en proceso de desarrollo de la legislación en torno al uso responsable de la IA. En este sentido, las empresas están experimentando una creciente necesidad de contar con leyes y regulaciones que aborden esta cuestión.

Según estudios realizados por BCG con MIT Sloan Management Review (2023), las empresas tardan aproximadamente tres años en obtener progresos significativos en el desarrollo de RAI. Esto, añadido a la presión competitiva impartida por los *peers*, contribuye al aumento de ansiedad e impaciencia existente entre las empresas respecto a la publicación de la regulación.

Se ha observado que, dependiendo del sector, las empresas están preparadas en mayor o menor medida para la llegada de la regulación. Los sectores de telecomunicaciones, tecnología, y aquellos orientados al consumidor están mucho más preparados que el sector energético o sector público y ONGs. La falta de preparación por parte de entidades públicas supone un potencial problema para los gobiernos, ya que puede perjudicar su imagen y la confianza de cara a la población. (Mills, Franke, Gupta, Nopp, & Grebe, 2023)

Por lo general, entran en juego diversos factores que varían radicalmente entre empresas, de manera que es complicado encontrar un enfoque generalizado para la RAI. La dinámica del mercado, las necesidades de las empresas, y su visión respecto a la IA son solo algunos ejemplos. Sin embargo, existen una serie de principios que las empresas han de adoptar para anticiparse y contar con una correcta acogida de la regulación. (Mills, Franke, Gupta, Nopp, & Grebe, 2023)

- Promover el liderazgo de la IA: la empresa ha de contar con un responsable de IA (AI *ethics officer*) encargado de controlar las iniciativas de la empresa. Estos líderes suelen contar con una multitud de conocimientos de distinta índole, tanto técnicos como empresariales y regulatorios. El valor de su trabajo está en su capacidad para integrar esos conocimientos en la estrategia de la empresa de manera transversal y a través de IAs que generen valor. Se rodean de un equipo compuesto por personas de distintos departamentos para facilitar la implementación y la transversalidad de los algoritmos.

- Construcción de un marco ético de IA: la implementación de principios y políticas que estén incrustadas en la RAI es crucial para desarrollar la base de los requerimientos que tengan que cumplir las empresas en el futuro. Independientemente de que una empresa sea local o multinacional, estos principios se han de cumplir y han de tener en cuenta legislaciones ya existentes que puedan aplicar a esto, como leyes de privacidad, por ejemplo. En este sentido, y considerando organizaciones multinacionales, la regulación en general persigue el mismo objetivo, asegurar el uso responsable de la IA. Establecer un marco ético sólido a través de la empresa acompaña a ese objetivo.

- Involucrar a personas en el ámbito de la IA: la regulación requiere de capital humano en relación con el gobierno y responsabilidad. Por lo general, las empresas consideran importante desarrollar IAs que aumenten, pero no reemplacen la capacidad humana. De esta manera, se otorga importancia a cómo se utilizan los sistemas de IA.

- Crear reseñas de IA e integrar herramientas y métodos: uno de los beneficios de la RAI es que prolonga el ciclo de vida de una IA. El objetivo es identificar problemas en el modelo lo antes posible, ya que la capacidad de monitorear los efectos de la IA a lo largo

de su ciclo de vida genera confianza por parte de los gobiernos y de los clientes. Esto se consigue asegurando la correcta implementación de los principios mencionados, y construyendo herramientas de reseñas *end–to–end* de los algoritmos, procesos y resultados.

- Participar en el ecosistema de IA: muchas empresas colaboran de manera conjunta en grupos de trabajo para la consecución de mejores prácticas de RAI. Es una manera ideal tanto de adelantarse a la norma como de mejorar el funcionamiento y rendimiento de los sistemas de IA de la empresa. La colaboración entre empresas y el compartir experiencias y conocimiento, contribuye a mejorar el enfoque para abordar las preocupaciones de la sociedad. Además, los partícipes más inteligentes pueden aprovechar la oportunidad y conseguir una ventaja competitiva.

En definitiva, la responsabilidad y la creación de valor por parte de la IA están indiscutiblemente relacionadas. Sin embargo, existe un gran salto de la aspiración a la implementación. La incertidumbre existente en torno a la regulación que abordará estas cuestiones aumenta la complejidad del diseño e implementación de la RAI. Las empresas han de subirse a la ola y anticiparse a la norma, con el fin de colaborar con su redacción, en lugar de verse sobrecogidos por ella una vez se publique. La manera de desarrollar modelos que se ajusten, en la medida de lo posible, a la futura legislación en un momento tan temprano es centrándose en la confianza y transparencia de estos, además de asegurar su capacidad de crecer y generar valor.

4.2 Qué están haciendo las empresas

En vista de la complejidad que exige la implementación de la RAI en los procesos de las empresas en un momento tan preliminar y sin legislación existente, muchas organizaciones ofrecen facilidades para desarrollar correctamente esta implementación.

Algunas de estas soluciones pertenecen a IBM con *watsonx.gorvernance*, Google con Google Cloud, y Microsoft con Azure Machine Learning. Este trabajo de investigación profundizará en los cimientos y aplicaciones de esta última herramienta, y concluirá con un ejemplo práctico de cómo funciona Azure ML.

Microsoft ha elaborado un estándar de RAI en base a una serie de principios que establece como piedra angular del acercamiento responsable a la IA. Además, la empresa ha incluido en su plataforma de desarrollo de modelos de ML (Azure Machine Learning) herramientas que permiten a los desarrolladores adaptar sus modelos a este estándar. La herramienta permite a las empresas afrontar la llegada de la regulación con mayor tranquilidad y reduce el número de cambios y adaptaciones que tendrán que hacer en cuanto se publiquen estas leyes. (Sameki & Gayhardt, 2023)

Además, la empresa ha hecho públicas sus librerías en materia de RAI. Se trata de una *responsable-ai-toolbox* que permite a los usuarios realizar actividades de exploración y análisis que permiten una mejor comprensión de los sistemas de IA. Esta práctica contribuye al ecosistema que se mencionaba antes de colaboración entre empresas, con el fin de tomar mejores decisiones basadas en datos. Esta *toolbox* incluye herramientas de visualización que desarrollan *dashboards* de IA responsable, análisis de errores, interpretabilidad y justicia.

Los principios mencionados son los siguientes (Sameki & Gayhardt, 2023):

- Equidad e inclusión: los sistemas de IA han de mantener el mismo trato para toda persona en circunstancias similares a otra. En el ámbito de RRHH, por ejemplo, ante igualdad de circunstancias, el sistema de IA no puede proporcionar condiciones de retribución diferentes a dos empleados distintos.

  Para abordar este principio, Azure Machine Learning permite a los responsables del modelo evaluar su equidad, probando distintos grupos definidos sensiblemente en cuestión de género, edad, raza y demás.

- Confiabilidad y seguridad: los sistemas de IA han de ser seguros y coherentes, es decir, deben funcionar en todo momento según cómo fueron diseñados en un principio. Además, su respuesta ante situaciones inusuales o inesperadas ha de ser segura, y deben ser capaces de resistir cualquier manipulación que sea perjudicial para su funcionamiento. Todo ello dependerá de los desarrolladores y las situaciones a las que expongan al modelo durante las etapas de diseño y pruebas.

Azure Machine Learning cuenta con un componente de análisis de errores específico para RAI.

- Privacidad y seguridad: la protección de datos es un tema realmente candente en la actualidad y cada vez la protección y seguridad de información personal es más difícil. En este sentido, para el correcto funcionamiento de la IA, el acceso a los datos es esencial para que sus *output*s sean precisos. De esta manera, los sistemas de IA han de regirse por leyes que "exijan la transparencia sobre la recopilación, el uso y el almacenamiento de datos", y que "obliguen a los consumidores a tener los controles adecuados para elegir cómo se usan sus datos".

  Azure Machine Learning, ofrece la posibilidad de crear una configuración que cumpla con lo exigido por las directrices de la organización. Además, la empresa ha desarrollado dos softwares de código abierto que permiten la incorporación de principios de privacidad y seguridad adicionales (SmartNoise y Counterfit).

- Transparencia: se vincula el concepto con la interpretabilidad, la explicación de la conducta de los sistemas de IA. De esta manera, si el sistema se utiliza para tomar decisiones que afecten significativamente a la vida de las personas, como un algoritmo de contratación, las partes interesadas son capaces de comprender el funcionamiento y motivo de la decisión. Además, en caso de que fuera necesario, sería capaces de identificar problemas de rendimiento, equidad, resultados no deseados y demás.

  En este sentido, Azure Machine Learning cuenta con componentes de interpretabilidad e hipótesis contrafactual, con los que los responsables de los modelos son capaces de generar descripciones sencillas de las predicciones de su modelo.

- Responsabilidad: los sistemas de IA no pueden ser los responsables de las decisiones que se toman con su ayuda. La responsabilidad recae sobre los diseñadores e implementadores de los sistemas. En este sentido, las normas y estándares de la empresa han de asegurar que las personas controlen de manera significativa los sistemas de IA, para que estos no

tengan una decisión final sobre algo que afecte a la vida de las personas ni que sean sumamente autónomos.

Azure Machine Learning incluye operaciones de Machine Learning (MLOps), que facilitan la consecución de IAs responsables. Algunos ejemplos serían: la capacidad de introducir, concentrar e implementar el modelo desde cualquier lugar y poder trazar y seguir a los metadatos asociados a este; avisos y notificaciones sobre el estado o eventos en el ciclo de vida del modelo (finalización, implementación, desfase de datos, etc.), y muchas otras.

Según Boston Consulting Group (BCG) (2023a), escalar el uso de la IA en las organizaciones es crucial para la consecución de una importante ventaja competitiva. Sin embargo, la IA en escala requiere de un cambio de enfoque en la toma de decisiones y operativa de la empresa, y para ello se ha de invertir en capital humano. Se ha visto como las empresas que han conseguido implementar IAs de forma transversal y generar valor de estas inversiones, distribuyen 10% del capital a los algoritmos, 20% a tecnología y el 70% restante en la incorporación de la IA a los procesos de la empresa con la metodología *agile.* De manera que, estas empresas invierten más del doble en capital humano que en la tecnología como tal.

La consultora también establece que la RAI no solo actúa como mitigante del riesgo, sino que contribuye a mejorar el rendimiento de los sistemas de IA. El objetivo es conseguir que los mismos mecanismos que reducen los errores de las IAs sean capaces de acelerar procesos e innovación, promover la diferenciación e incrementar el nivel de confianza del cliente con la empresa.

Por el momento, la implementación de RAI es percibida como amenazadora o compleja por muchas organizaciones. Además, muchas de ellas pretenden esperar a que se formule la regulación y actuar en base a lo que diga (BCG, 2023b).

Sin embargo, como se ha indicado anteriormente, existen herramientas, como la desarrollada por Microsoft, que incorporan los principios de RAI con valiosos beneficios para la empresa, adelantándose a las nuevas normas que están por venir.

4.3 Explicabilidad y transparencia

En este apartado se profundizará en la importancia de la explicabilidad de los sistemas de IA y se expondrá la problemática de la cuestión. Es una de las mayores causas de preocupación de los ejecutivos en cuestión de IA, debido al impacto que puede llegar a tener tanto en los trabajadores como en los clientes.

La relevancia del presente epígrafe radica en la percepción de los líderes en tecnología, quienes establecen que aquellos algoritmos que no pueden ser explicados por una persona son considerablemente mejores que aquellos más simples y comprensibles. Esto es lo conocido como la compensación entre precisión y explicabilidad (*accuracy–explainability–tradeoff*), que establece que cuanto mejor pueda entender una persona un algoritmo, menos preciso va a ser. (Candelon, Evgeniou, & Martens, 2023)

Por ello, se hace la distinción entre "cajas blancas" y "cajas negras". Las primeras consisten en modelos simples que utilizan un número limitado de parámetros y normas sencillas ayudándose de un árbol de decisión o una regresión lineal (algoritmos sencillos de ML) para generar resultados. Al tratarse de normas simples y parámetros limitados, los procesos pueden ser comprendidos y explicados fácilmente por una persona.

Sin embargo, en el segundo caso, las "cajas negras" son aquellos modelos que incluyen una multitud de parámetros como *random forests* o redes neuronales complejas para generar un *output* (Candelon, Evgeniou, & Martens, 2023). En este sentido, empresas como OpenAI están intentando abordar este problema. La empresa está empezando a utilizar GPT-4, para intentar explicar sus modelos de lenguaje que incorporan complejas redes neuronales con miles de millones de parámetros. Este enfoque está todavía en desarrollo, y la comprensión sobre cómo funcionan estos algoritmos internamente es todavía limitada, por ejemplo, si cuentan con sesgos heurísticos o si participan en prácticas engañosas. (OpenAI, 2023)

En línea con el dilema entre la precisión y la explicabilidad, Candelon, Evgeniou y Martens (2023) desarrollan los resultados de un estudio llevado a cabo para desmitificar esta afirmación. Explican cómo, en un análisis de unos 100 *datasets* representativos, el porcentaje de algoritmos de caja negra y caja blanca que proporcionaban resultados similares era casi

de un 70%. De manera que se observó que no existía, en su mayoría, una mejora considerable de la precisión de predicción en algoritmos de caja negra.

El artículo continúa estableciendo que estos resultados son consistentes con otros estudios llevados a cabo y con algunos casos de uso que se han probado para explorar el potencial de la explicabilidad de modelos de IA. Algunos ejemplos curiosos en los que se demuestra la falta de necesidad de cajas negras son:

- El uso de COMPAS para el sistema judicial estadounidense. Se trata de una compleja herramienta que predice la probabilidad de futuros arrestos. Se ha demostrado que su precisión es similar a la de un simple modelo predictivo basado en la edad y antecedentes penales.
- Un equipo de investigación elaboró un simple algoritmo capaz de predecir la probabilidad de impago de un préstamo por parte de un cliente. El modelo podía ser comprendido por el cliente medio y su precisión predictiva difería en menos de un 1% de aquella de un modelo de caja negra.

Pese a que los modelos de caja blanca sean una opción viable y justificada en muchos casos, existen muchos otros en los que las cajas negras siguen aportando valor. Sin embargo, por las implicaciones que conlleva, es importante que las empresas se aseguren de seguir los siguientes pasos antes de implementar un modelo de caja negra. (Candelon, Evgeniou, & Martens, 2023)

- Usar algoritmos de cajas blancas por defecto. En caso de querer implementar uno de caja negra, se ha de comparar con el primero para comprobar si realmente existen diferencias significativas en el rendimiento.
- Conocer los datos que se poseen. En primer lugar, la decisión radica en la calidad de los datos: si estos tienen mucho “ruido” (información irrelevante y demás), un algoritmo simple debería ser suficiente. En segundo lugar, se ha de tener en cuenta el tipo de dato ya que en cuanto se requiera de imágenes, audio o vídeo la opción más viable o, en

ocasiones, la única posible es un algoritmo de caja negra; esto pasa con aplicaciones de reconocimiento facial, sistemas de vehículos autónomos o ChatGPT, por ejemplo.

- Conocer a tus usuarios. La transparencia es clave para generar confianza, y en casos de uso específico, la justicia es de extrema importancia para los usuarios. De manera que el poder respaldar y explicar la decisión es vital para la empresa. En materia de RRHH esto es especialmente importante.

- Conocer a la organización. Se ha de tener en cuenta el nivel de preparación de la empresa para la inclusión de sistemas de IA. En una organización en la que los empleados no están acostumbrados a trabajar con esta tecnología, es interesante empezar por la incorporación de algoritmos sencillos que todo el mundo sea capaz de comprender. De esta manera, como se explicaba anteriormente en este trabajo, el empleado confiará más en las recomendaciones o decisiones tomadas por la máquina. Conforme la plantilla se siente más cómoda trabajando con estos modelos y observa resultados positivos, la empresa podrá experimentar con algoritmos más complejos para comprobar si mejora su rendimiento.

- Conocer la regulación. Para abordar los requerimientos de explicabilidad, por el momento, las cajas blancas son la única opción.

- Explicar lo inexplicable. En aquellos casos en los que el algoritmo de caja negra sea indudablemente más preciso, y su uso sea aprobado por la regulación, las empresas han de adaptarse a los requisitos de generación de confianza y seguridad en la medida de lo posible. En ocasiones es posible desarrollar un modelo de caja blanca como proxy para explicar el funcionamiento del modelo complejo. En otros casos, es extremadamente complicado abordar estas explicaciones y la empresa ha de comunicar con sinceridad tanto interna como externamente su desconocimiento, los riesgos asociados, y el trabajo que está desarrollando para mitigarlos.

En definitiva, la compensación entre precisión y transparencia puede ser minimizada y, en ciertos casos, eliminada. Sin embargo, como todo avance tecnológico, se encuentra riesgos

asociados al desarrollo e implementación de algoritmos de caja negra. Estos algoritmos, siguen siendo beneficiosos para la sociedad y han de ser monitoreados y controlados para que los beneficios sigan aumentando mientras que se mitigan los riegos adyacentes.

5. PRÁCTICA

5.1 Descripción y preparación de los datos

En el presente capítulo se realiza una demostración del funcionamiento de Azure Machine Learning. Se resaltan las herramientas y facilidades de RAI que proporciona para incentivar el uso responsable de la IA en cualquier proyecto. En línea con el enfoque de este trabajo de investigación, se utilizará el conjunto de datos de IBM HR Analytics Employee Attrition & Performance (Pavan, 2017). Este *dataset* se ha utilizado en numerosas ocasiones en RRHH para predecir la rotación de una empresa.

Este conjunto de datos contiene información relevante y detallada sobre los empleados: cuestiones demográficas, de satisfacción, desempeño, y otros atributos laborales. Además, incluye si han abandonado o no la empresa. Es importante mencionar que, debido a la naturaleza sensible de los datos de RRHH, el *dataset* utilizado contiene tanto datos reales como ficticios. Por ello, el objetivo del estudio consiste en identificar posibles sesgos en los datos, actuar al respecto si los hubiere, pero, sobre todo, en mostrar el proceso que se ha utilizado para explicar la respuesta del modelo. De esta manera, independientemente de la existencia de sesgos en este caso concreto, el proceso y las técnicas que se utilizan son aplicables a cualquier base de datos real y permiten demostrar el uso de la herramienta de IA responsable que ofrece Microsoft.

El conjunto de datos con el que se va a trabajar consta de 1.470 observaciones y 35 variables. A continuación, se proporciona una breve descripción de las variables utilizadas.

- *Age*: Edad del empleado. Variable numérica.
- *Attrition*: Indica si el empleado ha abandonado la empresa o no. Toma valores 0-1, variable dicotómica.
- *BusinessTravel*: Indica la frecuencia con la que viaja el empleado con la empresa. Toma los valores: *non*_travel (no viaja), *travel_frequently* (viaja frecuentemente) o *travel*_rarely (viaja poco). Variable categórica.
- *DailyRate*: Indica la tarifa diaria de pago de cada empleado. Variable numérica.

- *Department*: Departamento en el que trabaja el empleado. Los empleados pueden pertenecer a Ventas, Investigación y Desarrollo (I+D) o a Recursos Humanos. Variable categórica.
- *DistanceFromHome*: Indica la distancia en Km de la casa del empleado a la oficina. Variable numérica.
- *Education*: Nivel de educación alcanzado. Toma valores del 1 al 5, donde 1 es bajo y 5 es doctorado. Variable categórica.
- *EducationField*: Indica el campo de estudio del empleado. Variable categórica.
- *EmployeeNumber*: Número de identificador del empleado. Variable categórica.
- *EnvironmentSatisfaction*: Nivel de satisfacción del empleado con el ambiente de trabajo. Toma valores del 1 al 4, siendo 1 “bajo” y 4 “alto”. Variable categórica.
- *Gender*: Género del empleado, toma valores femenino o masculino. Variable categórica.
- *HourlyRate*: Tarifa de pago por hora. Variable numérica.
- *JobInvolvement*: Nivel de implicación del empleado en su trabajo. Toma valores entre 1 y 4, siendo 1 el más bajo. Variable categórica.
- *JobLevel*: Nivel jerárquico del puesto de trabajo. Toma valores entre el 1 y el 5, siendo 1 el más bajo. Variable numérica ordinal.
- *JobRole*: Puesto de trabajo o cargo del empleado. Variable categórica.
- *JobSatisfaction*: Nivel de satisfacción del empleado con su trabajo. Toma valores entre 1 y 4, siendo 1 el más bajo. Variable categórica.
- *MonthlyIncome*: Ingreso mensual del empleado. Variable numérica.
- *MonthlyRate*: Tasa mensual del empleado. Variable numérica.
- *NumCompaniesWorked*: Número de empresas en las que ha trabajado el empleado antes de esta. Variable numérica.

- *OverTime*: Indica si el empleado realiza horas extra o no. Toma valores “Si” y “No”. Variable dicotómica.
- *PercentSalaryHike*: Porcentaje de aumento salarial del empleado. Variable numérica.
- *PerformanceRating*: Calificación del desempeño del empleado. Toma valores entre 1 y 5, siendo 1 el más bajo. Variable categórica.
- *RelationshipSatisfaction*: Nivel de satisfacción del empleado con sus relaciones laborales. Toma valores entre 1 y 4, siendo 1 el más bajo. Variable categórica.
- *StandardHours*: Horas estándar de trabajo por empleado.
- *StockOptionLevel*: Nivel de opciones de acciones del empleado. Toma valores de 0 a 3, siendo 0 nulo y 3 alto. Variable categórica.
- *TotalWorkingYears*: Número total de años de experiencia laboral. Variable numérica.
- *TrainingTimesLastYear*: Número de veces que el empleado recibió formación el año anterior. Variable numérica.
- *WorkLifeBalance*: Equilibrio del empleado entre trabajo y vida personal. Toma valores entre 1 y 4, siendo 1 malo y 4 bueno. Variable categórica.
- *YearsAtCompany*: Número de años que el empleado lleva trabajando en la empresa. Variable numérica.
- *YearsInCurrentRole*: Número de años que el empleado lleva en su puesto actual. Variable numérica.
- *YearsSinceLastPromotion*: Número de años desde la última promoción del empleado. Variable numérica.
- *YearsWithCurrManager*: Número de años que el empleado lleva bajo supervisión del mismo mánager. Variable numérica.

- *SalesRating*: Calificación del desempeño del empleado en ventas. Esta variable solo aplica al departamento de ventas. Variable numérica.
- *HireSource*: Fuente de reclutamiento del empleado. Variable categórica.
- *Campus*: Indica a qué campus u oficina pertenece cada empleado. Variable categórica.

5.2 Creación y evaluación de modelos de ML

En este apartado de describirá el proceso de creación y la evaluación de los modelos de aprendizaje automático que se han desarrollado con la ayuda de Azure ML para el *dataset* de IBM. El modelo predice si un empleado abandonará la empresa o no. El objetivo es demostrar cómo las empresas pueden implementar y aprovechar herramientas como Azure ML para garantizar un uso responsable de las IAs y, en consecuencia, tomar decisiones éticamente correctas en cuestión de RRHH.

El proceso llevado a cabo es el siguiente:

1. En primer lugar, se importa el *dataset* a Azure ML, habiendo unificado el formato correctamente y verificado la correcta estructura de los datos. A continuación, se realiza una exploración inicial del *dataset* para la identificación de *missings*, *outliers* y determinar la naturaleza de las variables.
2. Una vez que los datos están preparados para trabajar con ellos, se define la variable objetivo, en este caso "*Attrition*". Es la variable que va a predecir el modelo y determina si un empleado va a abandonar la empresa o no.
3. A continuación, se configura el experimento en Azure ML. Se crea el experimento con el fin de organizar y que quede registrado el flujo de trabajo. Se establecen una serie de parámetros para optimizar la selección del modelo, como métricas de evaluación o el tipo de tarea que se va a realizar, que en este caso sería clasificación. Y previo el entrenamiento de los modelos, se dividen los datos en conjuntos de entrenamiento y prueba, 75% y 15% respectivamente.
4. Además, se han eliminado las variables *StandardHours* y *EmployeeNumber*. La primera porque no aporta valor al modelo, pues todos los trabajadores hacen el mismo

número de horas estándar, 80. La segunda porque al ser el número identificador de cada empleado, no aporta información adicional sobre los empleados y el algoritmo puede interpretar los valores erróneamente en lugar de descartarlos como insignificantes.

5. Con la herramienta de Automated ML de Azure se genera una multitud de modelos de distintos tipos. El algoritmo prueba los datos con distintos modelos hasta que encuentra resultados similares en la últimas 20 – 45 iteraciones.
6. Una vez generados los modelos, se elige el mejor. En este caso se trata de un modelo de clasificación, por lo que se elige el mejor en función del AUC (Area Under the Curve). Esta métrica indica el rendimiento de modelos de clasificación, de manera que cuanto mayor sea el AUC mejor será la capacidad predictiva del modelo (un AUC del 100% indicaría que el modelo predice a la perfección). Se visualiza de manera gráfica con la curva ROC, que indica el ratio de verdaderos positivos y el de falsos positivos.
7. Por último, para examinar la transparencia y explicabilidad del modelo, se utiliza la herramienta *explain the model*, la cual proporciona *insights* sobre las características más relevantes que utiliza el modelo y la relevancia que tienen en el proceso de toma de decisiones. Esta herramienta pretende facilitar la comprensión del razonamiento que hay detrás de las predicciones realizadas, de manera que las decisiones que se toman con este procedimiento son más transparentes.

5.3 Análisis e interpretación del modelo

En el presente apartado se profundiza en el proceso llevado a cabo para llegar al modelo seleccionado. Además, se entrará en el detalle de cómo funciona realmente el algoritmo con la ayuda de las herramientas proporcionadas por Azure ML para la explicabilidad del modelo, y demás cuestiones de RAI.

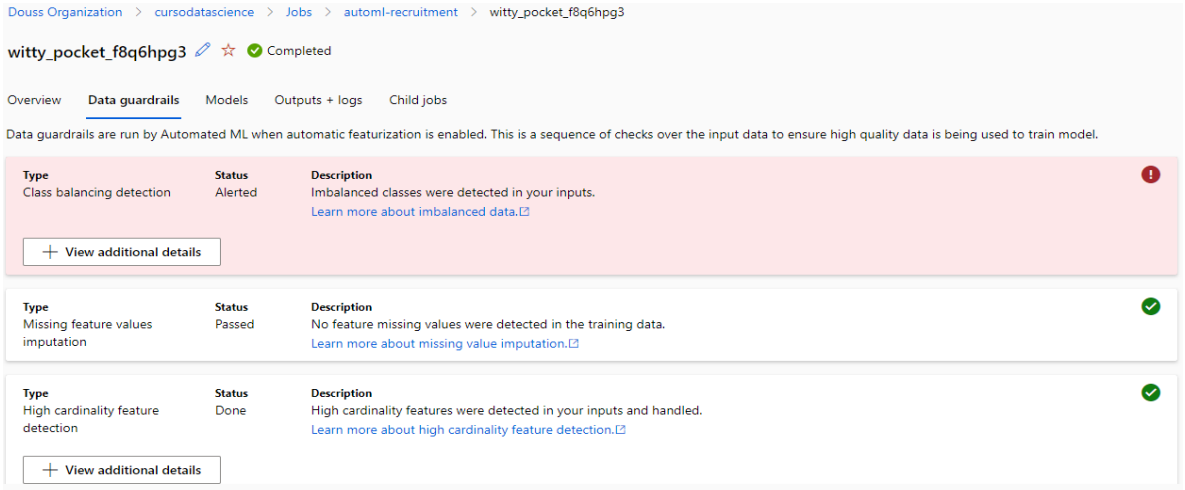
Modelo inicial

En primera instancia, se introdujeron las 33 variables (las originales excluyendo *StandardHours* y *EmployeeNumber*) para tener en cuenta el AUC desde el que se parte, con

el objetivo de aumentarlo lo máximo posible con los tratamientos y filtros de datos que correspondan.

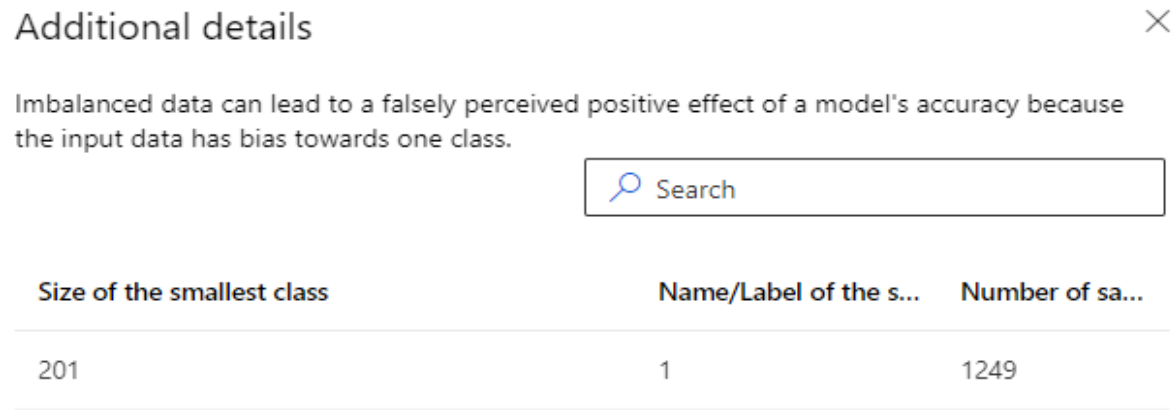
Desde un primer momento se puede intuir que los datos no están correctamente preparados. Azure ML realiza pruebas de calidad de los datos que se van a utilizar, cardinalidad, valores faltantes y balance de clases. En las siguientes figuras se observa el problema de desequilibrio que sufren los datos.

Figura 1 Pruebas en los datos de entrada



Fuente: 'Elaboración propia'

Figura 2 Desequilibrio en la variable objetivo (*Attrition)*



Fuente: 'Elaboración propia'

Este problema se abordará más adelante; por el momento se van a analizar los resultados obtenidos en esta primera ejecución.

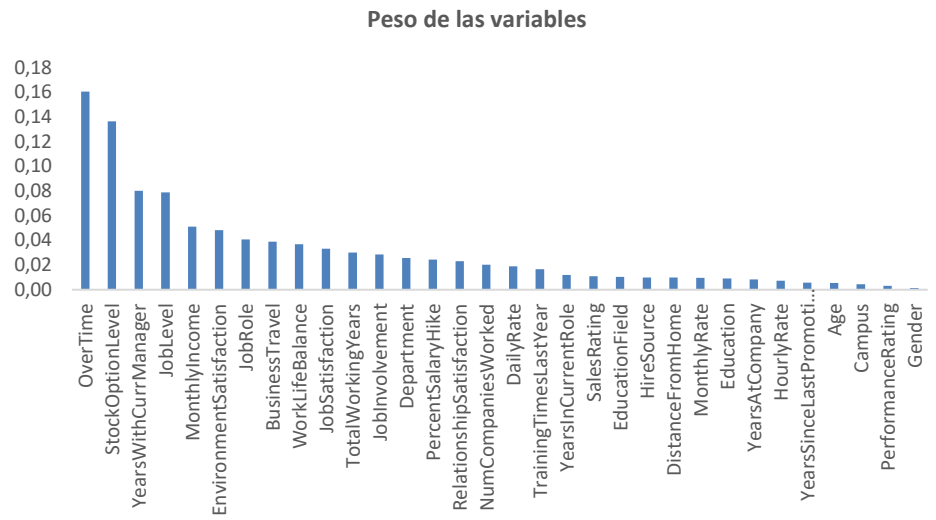
El mejor modelo propuesto es un *VotingEnsemble,* con un AUC ponderado de 81,98%, una precisión del 85,5% y un *F1-score* de 81,42%. El *voting ensemble* es un algoritmo de aprendizaje automático que combina las predicciones proporcionadas por una serie de modelos para tomar la decisión final, de manera que considera las predicciones de varios modelos y las somete a votación para tomar la decisión de manera conjunta. La estrategia de votación puede variar en función de la decisión que se esté tomando: se puede guiar por la mayoría de los modelos que coincidan en la clasificación, o por la media de probabilidades de la clasificación.

El AUC de 81,98% indica que la capacidad predictiva del modelo es razonablemente buena. En cuanto a la precisión, métrica que evalúa la exactitud del modelo para predecir correctamente instancias positivas, en este caso, ha predicho un 85,5% de las instancias positivas correctamente.

Por último, el *F1-score* es una métrica que combina la precisión y la exhaustividad del modelo. La exhaustividad indica la proporción de instancias positivas predichas correctamente respecto a la totalidad de instancias positivas de los datos. El *F1-score* se utiliza para comprobar si existe desequilibrio entre las clases de datos. Por ello, un 81,42% indica que el modelo tiene buen equilibrio entre precisión y exhaustividad, con lo que se logra una correcta combinación de predicciones precisas y la capacidad de capturar correctamente las instancias positivas de los datos.

En cuanto a la explicabilidad de estos resultados, se consideran los datos presentados en la pestaña de “explicaciones” en Azure ML. En primer lugar, se observan los pesos de las variables, con el fin de saber cuánto ha contribuido cada variable en la obtención de los resultados. En este caso, como se muestra en la Figura 3, las tres variables más relevantes son: *Overtime y StockOptionLevel, seguidas de YearsWithCurrManager y JobLevel.*

Figura 3 Importancia de las variables



Fuente: 'Elaboración propia'

Esta gráfica resulta interesante para identificar sesgos en los datos. En el caso de que una de las variables más relevantes fuese la edad, por ejemplo, podríamos concluir que la empresa discrimina a las personas mayores, o si fuese el género se concluiría que la empresa discrimina en este aspecto. En tal caso, se eliminaría la variable en cuestión, ya que no interesa que el algoritmo dé tanta importancia a variables demográficas como las mencionadas, y se volvería a ejecutar el proceso para obtener un nuevo modelo y analizar sus resultados.

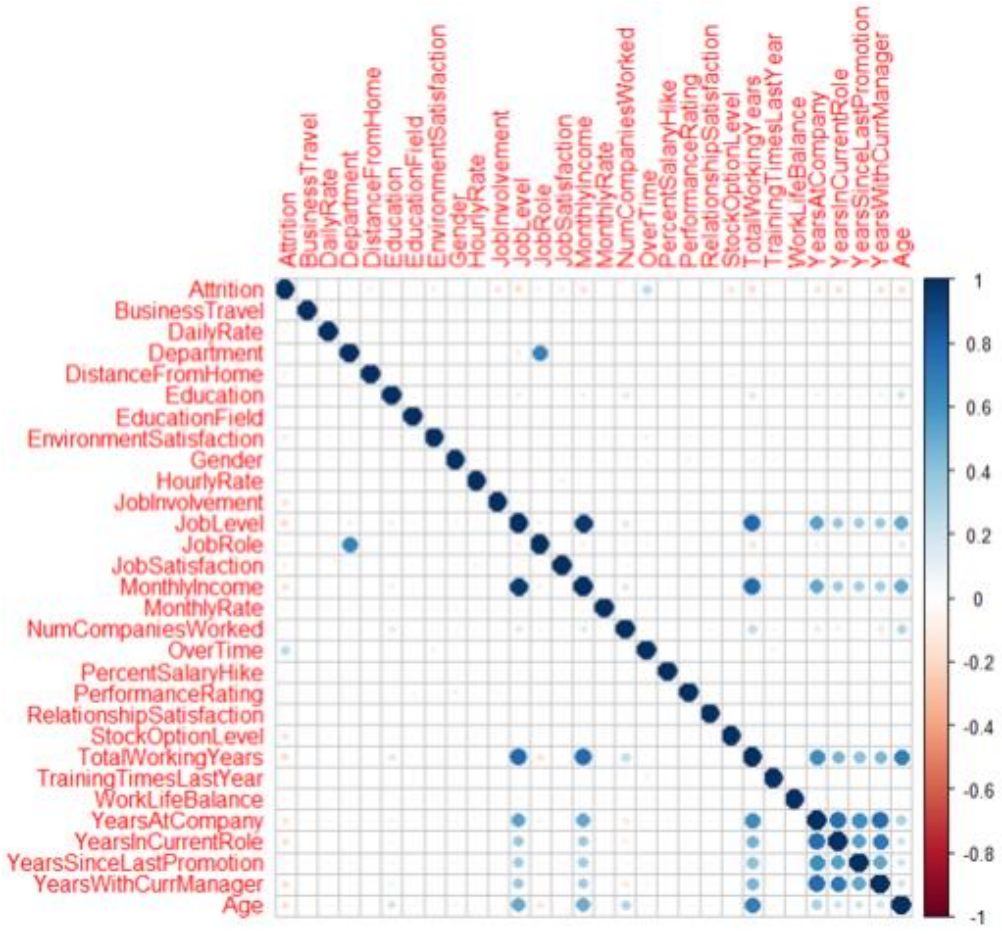
En este caso, no se observan sesgos ni discriminación por parte del modelo en base a datos que deberían ser irrelevantes para la rotación de empleados dentro de la empresa. Lo que sí se observa en la Figura 3, es la existencia de variables que apenas influyen.

Realizando un análisis de correlación, representado en la Figura 4, se observa que existe alta correlación entre las siguientes variables:

- *JobLevel* y *Age*, esta correlación es esperada ya que los empleados con puestos de nivel alto tienden a ser los empleados de mayor edad.

- *MonthlyIncome* y *JobLevel*, correlación esperada ya que cuanto mayor sea el nivel del cargo de un empleado, mayor será su salario.
- *TotalWorkingYears* y *JobLevel* están correlacionadas y es de esperar ya que aquellos que ocupan los puestos más altos suelen haber trabajado durante un mayor periodo de tiempo.
- *YearsWithCurrManager* y *YearsAtCompany* también presentan alta correlación. Esto se puede deber a la baja rotación de equipos de la empresa.
- *YearsAtCompany* y *YearsIn CurrentRole* están correlacionadas, lo que se puede deber a la baja tasa de promoción que hay dentro de la empresa.

Figura 4 Matriz de correlaciones



Fuente: 'Elaboración propia'

Las correlaciones observadas no llegan a tomar valores preocupantes ni parecen tener impactos significativos en el modelo. Por este motivo no se realizarán tratamientos a los datos basados en estos resultados

Modelo elegido

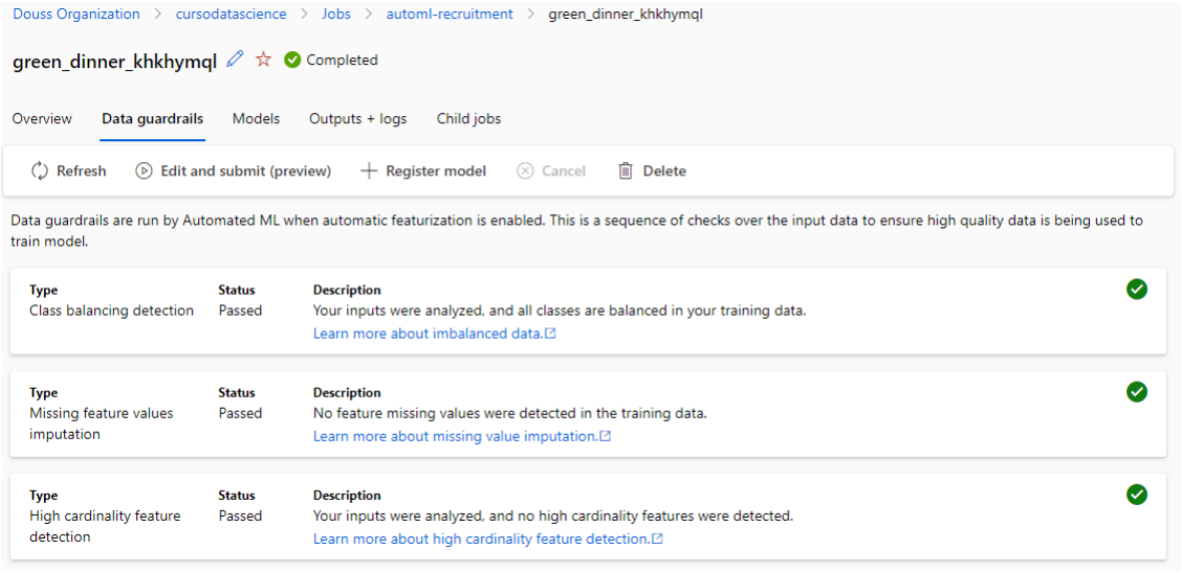
En vista de lo observado en el modelo inicial, se han hecho los siguientes tratamientos a los datos originales, con el fin de obtener un modelo con mejor capacidad predictiva.

- Se han identificado valores faltantes en las variables *SalesRating* (1.024), *HireSource* (618) y *Campus* (1.202). Al tratarse de tres variables categóricas y cantidades significativas de los datos disponibles, se ha procedido a eliminar estas variables.
- El *dataset* presenta un desequilibrio significativo entre las clases de la variable objetivo, *Attrition*, teniendo 1.233 empleados que no han abandonado la empresa y únicamente 237 que sí lo han hecho. Por ello, se ha procedido a realizar tratamientos para equilibrar las clases y mejorar la capacidad predictiva del modelo, con el fin de identificar patrones en la clase minoritaria. Esto se ha conseguido con la función "*ROSE*" de R, con la que se han obtenido 746 observaciones con valor 0 (no han abandonado la empresa) y 724 observaciones con valor 1.

Training set

Tras haber realizado estos tratamientos se ha repetido el proceso en Azure ML. A continuación, se verifica la solución del problema de desequilibrio encontrado en el modelo anterior.

Figura 5 Verificación pruebas



Fuente: 'Elaboración propia'

Como se puede observar en la Figura 5, la prueba del equilibrio de las clases tiene un estatus aceptable.

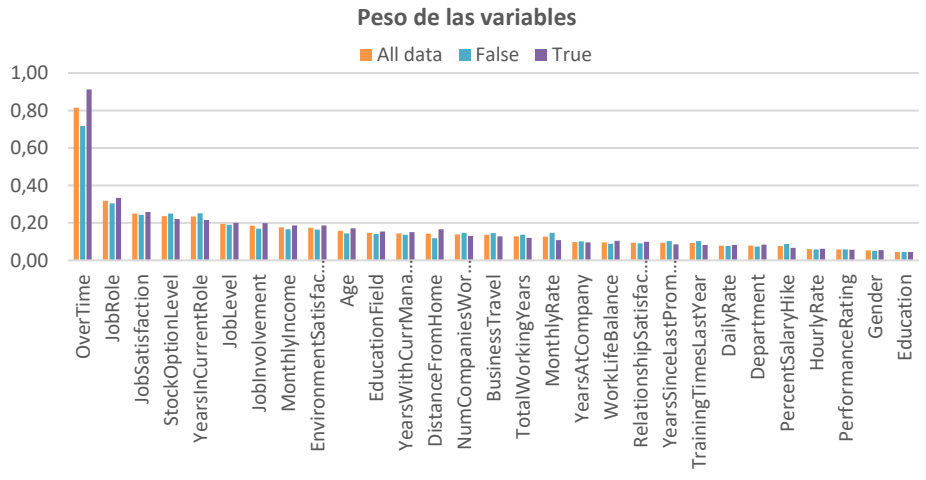
Una vez ejecutado, se observa que el AUC del mejor modelo es de 90,7%, la precisión de 85% y el *F1-score* de 85%, lo que indica que su capacidad predictiva ha mejorado considerablemente respecto a las primera ejecución. El modelo seleccionado es también un *voting ensemble* se pueden obtener los conjuntos de algoritmos (*ensembles*) que se han utilizado para crearlo. A continuación, se detallarán aquellos que más peso tienen en el modelo.

- *MaxAbsScaler, LightGBM:* 46,67% de peso en el *ensemble* final.
    - *MaxAbsScaler:* Algoritmo de preprocesamiento utilizado para escalar las variables de entrada de manera que se mantiene una proporción relativa entre los valores.
    - *LightGBM: Light Gradient Boosting Machine.* Algoritmo de aprendizaje automático que se basa en árboles de decisión y tiene la capacidad de ser entrenado rápidamente.

- *MaxAbsScaler, XGBoostClassifier:* 20% de peso en el *ensemble* final.
    - *XGBoostClassifier: Extreme Gradient Boost Classifier.* Algoritmo de aprendizaje automático que también se basa en árboles de decisión. Trata de combinar múltiples árboles de decisión débiles para obtener decisiones precisas.
- *MaxAbsScaler, ExtremeRandomTrees:* 13,33% de peso en el *ensemble* final.
    - *ExtremeRandomTrees:* Algoritmo de aprendizaje automático parecido al *random forest.* Ambos utilizan conjuntos de árboles de decisión para predecir. Sin embargo, en este caso, las divisiones en los árboles se realizan con mayor aleatoriedad y menor discriminación, mejorando la capacidad predictiva del modelo.

En cuanto a la importancia que el modelo atribuye a cada variable, la distribución cambia significativamente respecto a los valores presentados anteriormente. En esta ocasión, se muestran los resultados del total de los datos, pero también se han creado dos *cohorts* o subconjuntos de datos. Un subconjunto con aquellos empleados que sí han abandonado la empresa y otro con los que no lo han hecho. De esta manera se puede observar la importancia que se le atribuye a cada variable dependiendo de cuál sea el valor de la variable objetivo.

Figura 6 Importancia de las variables en el modelo final



Fuente: 'Elaboración propia'

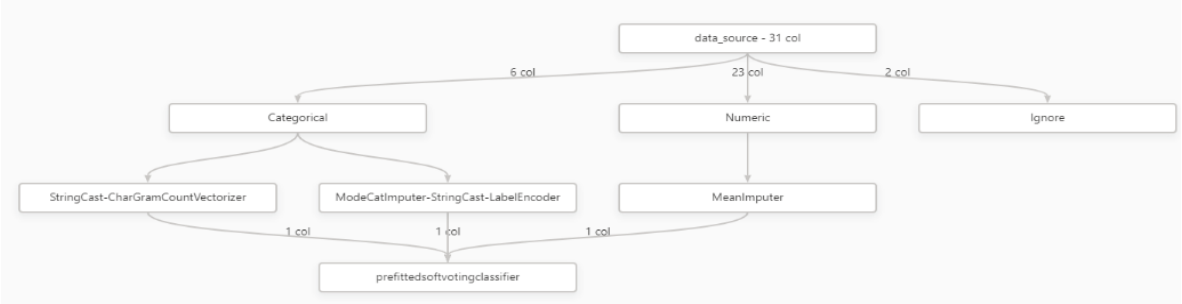
En el conjunto total de datos, las variables con más peso e influencia en la predicción del *Attrition* son *OverTIme* (0,81), seguida de *JobRole* (0,32) y *JobSatisfaction* (0,25).

Comparando las dos cohortes creadas, se observa que algunas variables tienen pesos ligeramente diferentes. Por ejemplo, para los empleados que no han abandonado la empresa, la variable *OverTime* tiene un peso menor (0,72) en comparación con el conjunto total de datos. Ello podría significar que el tiempo extra de trabajo no es tan determinante para quedarse como lo es para abandonar la empresa. Sin embargo, sigue siendo la variable con más importancia. En contraposición, para aquellos empleados que sí abandonan la empresa, la variable *OverTime* tiene un peso de 0,91. Esto sugiere que el tiempo extra de trabajo influye significativamente en tomar la decisión de abandonar la empresa.

En otras variables como *MonthlyIncome, EnvironmentSatisfaction* y *Age*, los pesos son similares para el conjunto total de datos y para ambas cohortes.

Azure ML también permite ver de manera gráfica el proceso que se ha llevado a cabo para el tratamiento de datos que ha realizado internamente. El siguiente diagrama ilustra el preprocesamiento de datos, la ingeniería de características, las técnicas de escalamiento y el algoritmo de aprendizaje automático que la herramienta *Automated ML* ha aplicado para generar este modelo en particular.

Figura 7 Procesamiento del modelo



Fuente: 'Elaboración propia'

Test set

En cuanto a los datos de prueba, se cuenta con una serie de métricas que evalúan el rendimiento del modelo. Los resultados se presentan a continuación.

además del AUC, *F1-score* y la precisión, cuyos valores ya se han especificado antes, obtenemos las siguientes métricas:

- *Average_precision_score:* resume la curva de precisión-sensibilidad y se calcula como la media ponderada de las precisiones obtenidas en cada umbral; se utiliza como peso el aumento en la recuperación desde el umbral anterior. Puede tomar valores entre 0 y 1, cuanto más cercano a 1 mejor. El valor obtenido en este caso es de 0,9 y se considera correcto.
- *Balanced_accuracy:* Media aritmética de la sensibilidad de cada clase. Puede tomar valores entre 0 y 1, cuanto más cercano a 1 mejor. El valor obtenido en este caso es de 0,85 y se considera correcto.
- *Log_loss:* Es la pérdida de entropía cruzada y se define como el logaritmo negativo de la verosimilitud de las etiquetas verdaderas atribuidas a las predicciones. Fomenta que el modelo genere probabilidades más altas para la clase verdadera. Al minimizar esta métrica durante el entrenamiento, el modelo aprende a mejorar sus predicciones y a ajustar sus parámetros para adaptarse mejor a los datos. Puede tomar valores entre 0 e infinito, cuanto más cercano a 0 mejor. El valor obtenido en este caso es de 0,4 y se considera correcto.
- *Matthews_correlation:* Es una medida equilibrada de la precisión que incluso se puede utilizar en datos que presentan desequilibrio entre clases. Toma valores entre -1 y 1, cuanto más cercano a 1 mejor. Un coeficiente de 1 indica que el modelo predice a la perfección, 0 indica predicción aleatoria y -1 indica predicción inversa. El valor obtenido en este caso es de 0,7 y se considera aceptable.
- *Norm_macro_recall:* Es la sensibilidad macro promediada y normalizada, de manera que el rendimiento aleatorio obtiene un valor de 0, significa que el modelo no es capaz de

identificar correctamente instancias positivas, y el rendimiento perfecto obtiene un valor de 1. En este caso, el valor obtenido es de 0,7 y se considera aceptable.

La curva ROC representa la relación entre la tasa de verdaderos positivos (sensibilidad), y la tasa de falsos positivos (1-especificidad). La sensibilidad define a los casos positivos correctamente clasificados y 1-especificidad indica los negativos incorrectamente clasificados como positivos. La curva muestra el rendimiento de un modelo de clasificación binaria en sus distintos umbrales de clasificación. El eje X representa la tasa de falsos positivos, mientras que el eje Y muestra los verdaderos positivos.

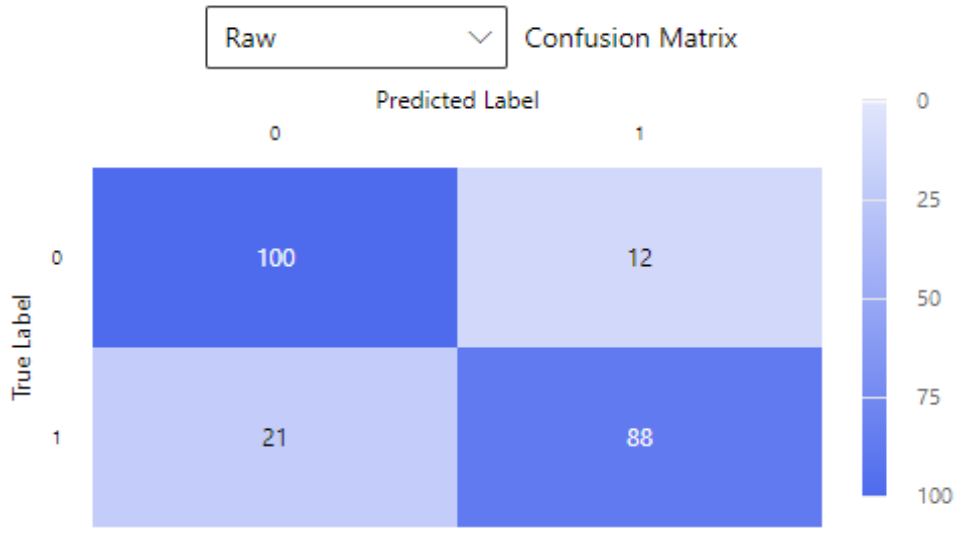
Figura 8 Curva ROC del modelo elegido

Fuente: 'Elaboración propia'

En esta curva se busca el punto que esté más cercano a la esquina superior izquierda, ya que representa la mayor tasa de verdaderos positivos y la menor de falsos positivos. Cuanto más cercano esté el valor a dicha esquina, mejor capacidad predictiva tendrá el modelo. Además, con esta curva también se puede observar el AUC ya que se trata del *Area Under the Curve*, haciendo referencia a la curva ROC, alineado con lo explicado anteriormente acerca del 100% como valor óptimo del AUC. Cuanto mayor AUC, más cerca estará la curva de la esquina superior izquierda.

Azure ML también proporciona la Matriz de Confusión, que permite visualizar el rendimiento de un modelo de clasificación de manera más detallada. La matriz muestra las predicciones del modelo en comparación con los datos reales. Indicando si las predicciones están alineadas o no con los datos reales.

Figura 9 Matriz de Confusión del modelo elegido



Fuente: 'Elaboración propia'

En este caso, el conjunto de prueba consta de 221 observaciones, de las cuales 100 han sido predichas correctamente como negativas y 88 como positivas. En cuanto a los errores cometidos, el modelo ha clasificado erróneamente 12 observaciones como positivas y 21 como negativas.

En definitiva, como se ha podido observar, Azure ML es una potente herramienta capaz de dar explicaciones a decisiones tomadas por complejos algoritmos. El modelo que ha proporcionado es correcto, tiene una capacidad predictiva adecuada y no presenta sesgos u otra característica que pueda clasificarlo como éticamente inaceptable. La capacidad de la herramienta va mucho más allá de lo expuesto en este estudio, pero lo explicado en el presente trabajo ilustra el valor que Azure es capaz de proporcionar. Es importante destacar, una vez más, que la organización ha hecho públicas una serie de librerías y de códigos

capaces de proporcionar esta información para todo aquel que quiera probarlas de manera gratuita. Ello contribuye a que las empresas se vean motivadas a la inclusión de prácticas que garanticen el uso responsable de la IA.

6. CONCLUSIONES

A lo largo del presente trabajo de investigación, se han abordado las distintas funciones del departamento de RRHH en la empresa y el papel que juega la IA en ellas. Se han identificado grandes beneficios que surgen como consecuencia de la introducción de IAs en la toma de decisiones de RRHH. Sin embargo, conforme esta tecnología se implanta en los distintos procesos de toma de decisiones que afectan a las personas, acontecen numerosos desafíos éticos y preocupaciones. Estos están asociados a la privacidad de los datos de los empleados que se utilizan para la toma de decisiones; la equidad a la hora de tomarlas, y la transparencia de la tecnología empleada en los procesos. De manera que los beneficios mencionados están intrínsecamente ligados a determinados riesgos que han de ser abordados y mitigados por las empresas que adopten sistemas de IA en sus procesos. La pregunta sobre si las máquinas son capaces de tomar decisiones éticamente aceptables cobra relevancia en este estudio.

En este sentido, se hace hincapié en que los sistemas de IA son programados en base a una serie de criterios y parámetros, los cuales son ejecutados y cumplidos por la máquina según se han definido. Por ello, el proceso ético y la definición de aquello que es aceptable o no en el marco ético ocurre íntegramente en la persona o equipo de personas que desarrollan la IA. Si bien una vez los sistemas han sido programados e implementados, existen algoritmos conocidos como “cajas negras” que presentan problemas de transparencia.

Se ha evidenciado la colaboración activa entre empresas en la definición de principios éticos aplicables a la implementación de algoritmos en sus procesos. Dicha iniciativa busca anticiparse a la regulación del uso responsable de la IA y facilitar la adaptación a futuras normativas.

En conclusión, la aplicación ética de la IA en la toma de decisiones de RRHH requiere principios como equidad, confiabilidad, privacidad, responsabilidad y explicabilidad. A lo largo del trabajo se atribuye especial importancia a este último, ya que subyace la necesidad de que la empresa se responsabilice de las decisiones tomadas con ayuda de una IA. Para ello, la organización o los desarrolladores de la IA han de ser capaces de explicar el proceso que el algoritmo lleva a cabo internamente para proporcionar un *output* determinado.

Numerosas empresas han lanzado herramientas que facilitan la explicación del proceso interno de los algoritmos. En la demostración práctica llevada a cabo en Azure ML se ha abordado el problema de la explicabilidad de los modelos en cuanto a la predicción de la rotación de empleados dentro de una organización.

En el estudio no se han encontrado sesgos en los datos ni en el modelo elegido por la herramienta, pero sí se ha demostrado la eficacia de ésta en cuanto a explicabilidad del algoritmo. El modelo elegido ha sido un *voting ensemble* o conjunto de algoritmos cuyas predicciones se someten a voto para llegar a una clasificación final. Al tratarse de un cúmulo de algoritmos, la explicación del proceso podría parecer compleja. Sin embargo, Azure ha proporcionado el paso a paso del proceso, las definiciones de lo que se ha empleado, y una multitud de métricas y gráficas que demuestran los resultados.

Por todo ello, se puede concluir con la necesidad existente de que toda empresa que incorpore sistemas de IA en sus procesos de toma de decisiones, sobre todo aquellos de RRHH, se asegure del uso responsable de la IA. Para ello, pueden utilizar recursos ya existentes como Azure ML, y contribuir con el desarrollo y la mejora de estas herramientas. Con el objetivo de promover la responsabilidad y ética en el uso de la IA, así como anticiparse a futuras regulaciones y facilitar su adaptación a las mismas.

Si bien a lo largo del trabajo se han expresado numerosos desafíos éticos que plantea la IA en RRHH, el caso práctico presentado aborda únicamente el problema de la transparencia y explicabilidad en cuanto a la predicción de la rotación de una empresa. En este sentido, se consideran como interesantes líneas de investigación para futuros trabajos tanto la transparencia en otro tipo de decisiones de RRHH, como el estudio de la equidad o privacidad de los datos en las distintas decisiones tomadas por el departamento.

Además, se ha de mencionar que, debido a la sensibilidad de la información, la práctica realizada se ha desarrollado con datos ficticios. Sin embargo, la demostración proporcionada del funcionamiento de Azure ML puede ser aplicada a cualquier base de datos, de distintas empresas y sectores, y sería interesante realizarlo en futuras investigaciones con datos reales de empresas para la toma de decisiones en RRHH.

7. REFERENCIAS

Alpaydin, E. (2010). *Introduction to Machine Learning.* Obtenido de https://kkpatel7.files.wordpress.com/2015/04/alppaydin_machinelearning_2010.pdf

BCG (2023a). *Artificial Intelligence and AI at Scale.* Obtenido de https://www.bcg.com/capabilities/digital-technology-data/artificial-intelligence?utm_source=search&utm_medium=cpc&utm_campaign=digital&utm_description=paid&utm_topic=ai&utm_geo=global&utm_content=ai_in_business_group_topreggeo&gclid=CjwKCAjw1YCkBhAOEiwA5a

BCG (2023b). *Responsible AI.* Obtenido de bcg.com: https://www.bcg.com/beyond-consulting/bcg-gamma/responsible-ai

Candelon, F., Evgeniou, T., & Martens, D. (2023). AI Can Be Both Accurate and Transparent. *Harvard Business Review*. Obtenido de https://hbr.org/2023/05/ai-can-be-both-accurate-and-transparent

Dattner, B., Chamorro-Premuzic, T., Buchband, R., & Schettler, L. (2019). The Legal and Ethical Implications of Using AI in Hiring. *Harvard Business Review*. Obtenido de https://hbr.org/2019/04/the-legal-and-ethical-implications-of-using-ai-in-hiring

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., . . . Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 689-707. Obtenido de https://link.springer.com/article/10.1007/s11023-018-9482-5

Hawksworth, J., Berriman, R., & Goel, S. (2020). Will robots really steal our jobs? *PwC*. Obtenido de https://www.pwc.com/hu/hu/kiadvanyok/assets/pdf/impact_of_automation_on_jobs.pdf

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 389-399. Obtenido de https://www.nature.com/articles/s42256-019-0088-2

Kim-Schmid, J., & Raveendhran, R. (2022). Where AI Can — and Can't — Help Talent Management. *Harvard Business Review*. Obtenido de https://hbr.org/2022/10/where-ai-can-and-cant-help-talent-management

Kontsevoi, B. (2021). What Exactly Is Artificial Intelligence? (Hint: It's All About The *Dataset*s). *Forbes*. Obtenido de https://www.forbes.com/sites/forbestechcouncil/2021/05/04/what-exactly-is-artificial-intelligence-hint-its-all-about-the-*dataset*s/?sh=3e422ed1bc97

Kvalnes, Ø. (2019). *Moral Reasoning at Work: Rethinking Ethics in Organizations.* Noruega: Palgrave Macmillan.

Mayhew, R. (25 de Enero de 2019). *Six Main Functions of a Human Resource Department.* Obtenido de Chron: https://smallbusiness.chron.com/operational-hr-22916.html

McKendrick, J., & Thurai, A. (15 de Septiembre de 2022). AI Isn't Ready to Make Unsupervised Decisions. *Harvard Business Review*. Obtenido de https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions

Mills, S., Franke, M. R., Gupta, A., Nopp, C., & Grebe, M. (11 de Abril de 2023). Obtenido de bcg.com: https://www.bcg.com/publications/2023/how-to-prepare-for-ai-regulation

OpenAI (2023). *Language models can explain neurons in language models.* OpenAI. Obtenido de https://openai.com/research/language-models-can-explain-neurons-in-language-models

Pavan, S. (2017). *IBM HR Analytics Employee Attrition & Performance.* Obtenido de Kaggle: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-*dataset*

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach (3rd ed.).* nEW jERSEY: Prentice Hall.

Sameki, M., & Gayhardt, L. (4 de Abril de 2023). *¿Qué es la inteligencia artificial responsable?* Obtenido de https://learn.microsoft.com/es-es/azure/machine-learning/concept-responsible-ai?view=azureml-api-2