0

INTRODUCCIÓN A PEOPLE ANALYTICS

0.1 DEFINICIONES Y CONCEPTOS

"In God we trust, all others must bring data"

"A Dios le creemos, los demás que vengan con datos"

Edward Demming

Antes de profundizar en el mundo de *People Analytics*, es conveniente que definamos adecuadamente el término, así como el ámbito y el alcance al que nos referiremos.

La definición de *People Analytics* entraña cierta dificultad porque se trata más de una filosofía que de un producto o metodología concreta. En general, suele estar asociado con el concepto de *"Decisiones Data Driven"* (decisiones derivadas de los datos) recientemente popularizado en la mayoría de las organizaciones empresariales y en el ámbito de los Recursos Humanos (RRHH). Las culturas *"Data Driven"* basan todas sus decisiones en datos intentando alcanzar la mayor objetividad posible.

Una aproximación al concepto de *People Analytics* puede ser la siguiente: *"People Analytics es un enfoque o cultura de gestión de los recursos humanos basado en datos y herramientas analíticas"*. Este enfoque utiliza modelos y análisis cuantitativo para comprender, medir y optimizar el rendimiento de los recursos humanos y la gestión del talento en las organizaciones. Con este propósito, recoge y utiliza datos relacionados con los empleados, información demográfica, desempeño

laboral, comportamientos, habilidades y otros indicadores y variables relevantes para la actividad empresarial.

El objetivo principal de *People Analytics* es proporcionar a los líderes de las organizaciones información útil y lo más objetiva posible para tomar decisiones más informadas y estratégicas en áreas como la contratación, la retención de talento, el desarrollo de habilidades, la gestión del rendimiento y la planificación de la sucesión. El objetivo final es mejorar la eficiencia, la productividad y la toma de decisiones relacionadas con el capital humano, contribuyendo así al cumplimiento de los objetivos empresariales. Antes de profundizar en algunas de las posibles aplicaciones, vamos a repasar brevemente la historia, quién ha utilizado este término hasta la fecha y cómo lo ha hecho.

En el mundo actual, tan analítico y tecnificado, en el que las culturas "*Data Driven*" se han generalizado, no sorprende la utilización de datos analíticos en todos los ámbitos de decisión. Sin embargo, en el campo de los recursos humanos estas dinámicas son relativamente recientes. En los últimos 20 o 30 años, las prácticas y enfoques modernos de gestión de personas se han implementado en una amplia gama de organizaciones, desde grandes corporaciones hasta pequeñas y medianas empresas.

Hasta finales de los años 80, la gestión de los recursos humanos era, esencialmente, una función administrativa centrada en la gestión de las nóminas y en las relaciones laborales. A partir de los años 90, la gestión de los recursos humanos empezó a convertirse en un elemento estratégico en el ámbito empresarial, evolucionando hacia un enfoque más centrado en las personas y en la contribución de los empleados al éxito de la organización. Desde entonces, se empezaron a introducir conceptos como la gestión del talento, el desarrollo del liderazgo o la planificación de la sucesión.

Con los primeros años del siglo XXI llegó la automatización de los procesos administrativos. Al liberarse de parte de estos procesos, los equipos de recursos humanos pudieron centrarse en actividades estratégicas con mayor capacidad de creación de valor. De este modo, empresas de todos los tamaños comenzaron a adoptar este tipo de tecnologías para mejorar la gestión de los recursos humanos.

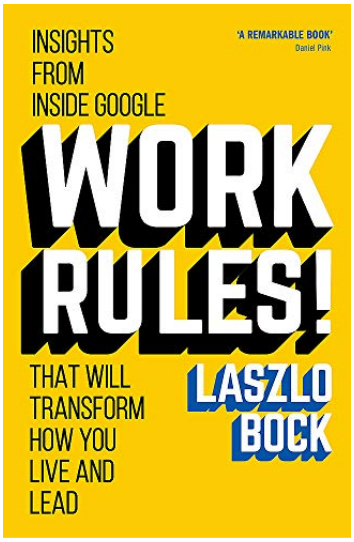
En los últimos 10 o 15 años, los profesionales de recursos humanos han continuado evolucionando, con un enfoque cada vez más centrado en la experiencia del empleado, la diversidad e inclusión y la gestión del desempeño. Además, se comenzó a dotar de una mayor importancia a la capacidad de crear culturas organizativas sólidas, promover la colaboración y la innovación, y adaptarse a las demandas cambiantes del mercado laboral. Todos estos campos debían ser integrados en los procesos de decisión de las empresas y estos procesos, cada vez con mayor

frecuencia, se basaban en datos objetivos y en análisis cuantitativos. Con todo ello se popularizó el uso de la analítica al ámbito de los recursos humanos.

Dicho esto, hemos hablado, principalmente, de la definición de *People Analytics* en el contexto de la gestión de recursos humanos, no del término en sí. Específicamente, el origen del término se encuentra en los equipos de recursos humanos de *Google*, que se denominaban a sí mismos como *"People Operations"*. Aunque antes de *Google* ya se hablaba de *HR Analytics* o *Talent Analytics*, las primeras referencias más o menos consistentes se remontan a la década de 1990, cuando empresas pioneras como *General Electric*, *Motorola* o *American Express* comenzaron a utilizar herramientas de análisis de datos para mejorar los procesos de selección de candidatos, de evaluación del desempeño y de la gestión del talento. La introducción de herramientas basadas en datos objetivos permitió a estas corporaciones mejorar la gestión de sus profesionales.

En un principio, *Google* creó el departamento de *People Operations*, definiendo métodos de trabajo y procesos a partir de datos y algoritmos. Después, unificó los departamentos de *People Operations* y *HR Analytics*, y de esta unión nace el concepto de *People Analytics*.

Para quien esté interesado en cómo implementó *Google* la analítica en la gestión del talento, le recomendamos la lectura del libro *"Work Rules!"* (2015) de *Laszlo Bock*, máximo responsable de *People Operations* de la compañía entre 2006 y 2016. En este libro, *Bock* describe aspectos relacionados con los principios y la cultura de *Google* y varias situaciones en las que, a través de análisis de datos, la compañía reformuló sus procedimientos.



Uno de los ejemplos más llamativos del mundo *Google*, procede de su plan de incentivos. Desde sus inicios, *Google* permitía que sus ingenieros dedicaran parte de las jornadas laborales a trabajar en proyectos personales. Al cabo del año, aquellos proyectos considerados de mayor calidad y utilidad eran premiados con un millón de dólares.

Fruto de esta iniciativa surgieron proyectos verdaderamente prometedores. Sin embargo, con el paso del tiempo la compañía descubrió que, sistemáticamente, los proyectos eran abandonados con independencia de su potencial. La razón de este

abandono se debía a que los ingenieros más brillantes (aquellos que recurrentemente presentaban los mejores proyectos) sabían que no podrían ganar el premio un segundo año con el mismo proyecto. En consecuencia, abandonaban el proyecto y empezaban a pensar en una nueva iniciativa que tuviera posibilidades de ganar. Un millón de euros era un estímulo tan relevante que desvirtuaba el propósito de la iniciativa. El programa perseguía que profesionales brillantes pudieran dedicar su atención a proyectos que les apasionaban, y el premio debía ser un estímulo adicional. Sin embargo, por su propia relevancia, el premio se convirtió en el fin último del programa. Para no alargar la historia, *Google* mantuvo la iniciativa, pero cambió la recompensa. Sustituyo el premio en metálico por otros reconocimientos en especie, igualmente singulares, pero que no eran el fin último del concurso.

Aunque este es solo un ejemplo entre muchos otros, podemos concluir diciendo que, desde su génesis en *Google*, *People Analytics* aborda una amplia gama de temas relacionados con los empleados y los equipos profesionales utilizando datos y análisis cuantitativo. Pese a que en los siguientes apartados lo desarrollaremos con una mayor profundidad, podemos adelantar que los temas que serán abordados incluyen la selección y contratación de personal, la medición del *engagement* (compromiso) y la satisfacción de los empleados, la formación y el desarrollo profesional, la retención y la productividad… En cuanto al análisis, dependiendo del ámbito de actuación, el abanico de herramientas disponibles es muy amplio. Desde modelos sencillos que se nutren de hojas de cálculo hasta técnicas del *Aprendizaje Máquina* (*Machine Learning*) que permiten identificar comportamientos y tendencias de carácter complejo. Como resulta evidente, lo fundamental es utilizar datos que nos permitan objetivar la decisión. Es decir, el elemento diferencial en *People Analytics* es el uso de datos cuantitativos y objetivos para la elaboración de modelos que facilitan la toma de decisiones en los ámbitos relativos a la gestión de personas. Sin embargo, tal y como ha sido definido en un principio, *People Analytics* parece un aspecto más de la gestión empresarial en la actualidad en el que la toma de decisiones se basa en los datos. Aun así, presenta diversos matices que comentaremos en los siguientes apartados.

0.2 ALGUNAS PECULIARIDADES DE *PEOPLE ANALYTICS*

Las culturas "*Data Driven*" (*People Analytics* es un caso particular) se apoyan en el uso de datos; uno de los principales problemas, sino el principal, al que se enfrentan los equipos de *People Analytics* es precisamente el de la gestión de los datos.

Las organizaciones con culturas "*Data Driven*" tienden a cuidar y desarrollar sus datos tanto con tecnología como con modelos de gobernanza del dato. Estos

últimos resultan fundamentales en el desarrollo de los llamados *"Data Lakes" o "lagos de datos"* (sistemas de almacenamiento que permiten guardar grandes cantidades de datos en su formato original, sin necesidad de estructurarlos previamente). Sin embargo, este cuidado con los datos tiende a limitarse a las variables propias de la producción, las ventas y las finanzas. En otros territorios, menos "críticos" para los modelos de gestión tradicionales, no suele existir esa preocupación por el cuidado de los datos, lo que se traduce, generalmente, en datos de mala calidad y difícil explotación y en interpretaciones y conclusiones erróneas.

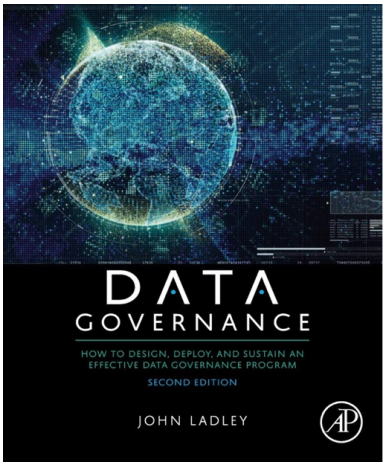
Aunque en el ámbito de los recursos humanos encontramos datos que sí son críticos (y que suelen estar relacionados con el pago de los salarios) también existen otros "menos críticos" pero, al mismo tiempo, muy relevantes para gestionar adecuadamente los equipos y los profesionales. Estos suelen referirse a información relacionada con la evaluación de los profesionales, (sus ámbitos de responsabilidad, su formación, sus habilidades y capacidades, su carrera interna) o con los procesos de contratación. En general, estos conceptos son más ambiguos e imprecisos y, por lo tanto, difíciles de medir. Por ello, frecuentemente nos encontramos con ciertos errores y ambigüedades a las que raramente se le presta atención.

Además, toda esta información que existe en las organizaciones, habitualmente, no se encuentra estructurada, ni se almacena en un repositorio accesible a los responsables. Esto implica cierta probabilidad de encontrar inconsistencias en los datos o el desconocimiento sobre su ubicación y características y, en consecuencia, dificulta el acceso y el procesamiento de la información necesaria para la toma de decisiones.

En este punto es crítico pensar en determinados errores: el mismo nombre con distintas ortografías, ausencia de un identificador del empleado que permita hacer un seguimiento a lo largo de varios años (DNI, número identificador del empleado, la unión de nombre y apellidos, la unión de apellidos y nombre…). Estos errores son habituales tanto en grandes corporaciones internacionales como en pequeñas empresas locales. Además, se trata de errores que no se solucionan con grandes inversiones en *data lakes*, o con grandes bases de datos corporativas. La solución adecuada pasa por desarrollar internamente una cultura que dote a cada dato (o característica o variable) de un significado preciso y del valor e importancia que merece y, en consecuencia, defina un protocolo para el procesamiento de los datos (recogida, almacenamiento, extracción e interpretabilidad). Si nuestra organización adolece de alguno de estos defectos, deberíamos empezar por resolverlos.

Por lo tanto, un paso previo al desarrollo de modelos de *People Analytics* consiste en verificar la calidad y consistencia de los datos en relación con aquellas situaciones similares a las descritas en el párrafo anterior. El resultado debe dar lugar a modelos de datos compuestos por información estructurada, gobernada por

los mismos sistemas y criterios que se aplican a la contabilidad o a las ventas, y accesibles a los profesionales que deben utilizarlos.



El campo de la ciencia de datos que se ocupa de la calidad de los datos se conoce como "*Data Governance*" y, aunque escapa al alcance de este libro, es fundamental para el desarrollo de modelos de *People Analytics*. Para aquellos lectores interesados, sugerimos la lectura del libro "*Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program"*, publicado en el año 2012 y escrito por *John Ladley*.

Una vez que es posible garantizar una mínima calidad en los datos, llega el momento de pensar en conceptos de ciencia de datos y estadística avanzada. En definitiva, en el desarrollo y la aplicación de modelos de *People Analytics*.

Además de la problemática descrita, los modelos de *People Analytics* tienen otras dos peculiaridades respecto de la utilización de modelos de ciencia de datos en otros ámbitos de la empresa:

- La primera es **la cantidad de información**. Habitualmente, los volúmenes de información son relativamente pequeños. Si bien es verdad que existen grandes organizaciones con decenas de miles de empleados donde es posible manipular grandes volúmenes de información, también es cierto que existen muchas más, con menores volúmenes de profesionales e información, en las que las soluciones y los modelos deben tener en cuenta esta realidad. Esta limitación es especialmente relevante cuando aplicamos modelos estadísticos o herramientas del *Aprendizaje Máquina*, pues suelen proporcionar mejores resultados cuantos más datos se emplean para su ajuste.

- La segunda es **la naturaleza de la información**. En el ámbito empresarial, y en términos generales, se trabaja con datos cuantitativos y objetivos. Las cantidades suelen responder a magnitudes reales que se miden con cierta precisión y cuyo significado no admite interpretación (cada concepto está definido de manera detallada). Sin embargo, en el ámbito de los recursos humanos, además de datos como el coste mensual de la

nómina o el número de empleados dados de alta en la *Seguridad Social*, encontramos otros que son:

- Necesariamente *subjetivos*. Por ejemplo, aquellas variables que dependen de percepciones o sentimientos, como ocurre en procesos de análisis de la satisfacción y del compromiso del empleado.
- *Imprecisos*. Una evaluación de la personalidad se concreta en un perfil de individuos. Sin embargo, si repetimos la prueba sobre el mismo individuo, probablemente nos encontremos con ciertas diferencias debido a la subjetividad inherente al proceso de evaluación. Esta situación conlleva a imprecisión en las métricas resultantes.
- Relativamente *ambiguos*. El significado del concepto "compromiso profesional" no es exactamente el mismo para cada uno de los directivos de la organización, no digamos de la diferente percepción entre directivos y empleados. Aunque todos tenemos una noción de determinados conceptos, para cada uno de los individuos esa noción tiene matices que dificultan un tratamiento homogéneo.

La consecuencia inmediata de emplear este tipo de datos es que debemos aprender a trabajar en entornos ambiguos e inciertos (en definitiva, entornos que involucran subjetividad) donde, frecuentemente, las tendencias son más relevantes que las métricas absolutas. Es decir, hemos de prestar mayor atención a la distribución global de los datos, y relegar las relaciones e interpretaciones de carácter local a un segundo plano. De este modo, las conclusiones que extraigamos de los datos serán más robustas a estos factores (subjetividad, imprecisión y ambigüedad).

Estas dos peculiaridades deben considerarse como posible causa al plantear soluciones o modelos que proporcionan resultados erróneos, pese a haber utilizado una aproximación correcta para su desarrollo.

0.3 DATOS EN *PEOPLE ANALYTICS*

Mas allá de las peculiaridades descritas al final del apartado anterior, los datos deben responder a unos criterios de calidad similares a los que nos podemos encontrar en un problema típico de análisis de datos. Aunque el significado de estos criterios se detalla en el próximo capítulo (*Capítulo 1: Introducción a la Analítica*), a continuación, se proporciona una breve introducción a los mismos desde una perspectiva enfocada en la gestión empresarial:

- **Claridad y precisión:** los datos deben de estar documentados y conocerse el alcance y significado de las diferentes magnitudes que se consideren. Por ejemplo, en relación con el compromiso, debemos tener claro de qué estamos hablando específicamente y establecer un consenso general. De lo contrario, no será posible interpretar correctamente las conclusiones de nuestros modelos ni compartirlas con la organización.

- **Consistencia**: es necesario que los datos sean coherentes entre sí (mismo origen, sistema de evaluación, escala de medición...) De lo contrario, de llegar a alguna conclusión no sabremos evaluar sus implicaciones para la organización. Métricas como el absentismo o la rotación deben de estar perfectamente definidas y deben de manejarse siempre en las mismas unidades. Este aspecto es especialmente importante cuando los datos proceden de la evaluación de diferentes profesionales que, necesariamente, van a introducir sesgos y ruido en sus evaluaciones.

- **Tiempo real**: otro punto clave, pues es necesario que los datos sean representativos de la realidad actual. Los cambios en usos y costumbres pueden inutilizar los datos históricos. Si la obsolescencia de tecnologías y productos puede restar valor a las métricas de hace 4 o 5 años, el paso del tiempo puede tener un impacto aun mayor en las métricas que describen los usos y costumbres de los profesionales. No es necesario explicar que el período en el que la mayoría de los profesionales fueron obligados a trabajar en remoto como consecuencia del *COVID* no puede ser representativo de lo ocurrido el año anterior o dos años después. Sin embargo, frecuentemente se olvida y se acaban utilizando datos de momentos que no son representativos de la realidad que se pretende modelar (restando validez a los resultados finales).

- **Accesibilidad**: la accesibilidad se refiere a la medida en que los datos están disponibles para su extracción y utilización por los profesionales. Si los datos no están disponibles o no proceden de las mismas fuentes, se pueden generar problemas en la actualización de los modelos, dando lugar a resultados volátiles. Además, si los datos son difíciles de obtener o el proceso de obtención es costoso, nos encontraremos con pocas iniciativas para incorporar modelos de *People Analytics* en la tomas de aquellas decisiones que afectan a los profesionales.

0.4 SOBRE LA APLICACIÓN DE PEOPLE ANALYTICS

Aunque *People Analytics* se centra en la aplicación de modelos analíticos en el ámbito de los recursos humanos, y las posibilidades en este ámbito son muy amplias, en este libro nos vamos a centrar en la utilización de ciertas herramientas que van más allá de los modelos cuantitativos clásicos. Concretamente, en herramientas del ***Aprendizaje Máquina*** o ***"Machine Learning" (ML)***; rama de la *Inteligencia Artificial* (*IA*) enfocada en el diseño y desarrollo de algoritmos capaces de aprender de los datos y mejorar su desempeño de manera autónoma en una amplia variedad de problemas de carácter cuantitativo, sin necesidad de intervención directa por parte del humano. Muchas son las aplicaciones de los modelos de *Inteligencia Artificial* y, específicamente, del *Aprendizaje Máquina*, en el ámbito empresarial (incluida la gestión de los recursos humanos), de ahí su creciente popularidad en los últimos años.

Resulta relativamente sencillo pensar en problemas estándar que podrían ser tratados con esta tecnología: previsión de demanda (modelos de regresión, modelización de series temporales), segmentación y clasificación de clientes (técnicas de agrupamiento o *clustering*), gestión de stocks (*clustering* y modelos de clasificación), mantenimiento predictivo (modelos de regresión, redes neuronales)… Todos ellos tienen un denominador común; responden a problemas específicos de la actividad empresarial. Es decir, cada posible modelo responde a una problemática que ya existía en las organizaciones desde antes de que se utilizara la *Inteligencia Artificial*.

En Recursos Humanos existen múltiples problemas desde hace años (históricamente resueltos con cierta dificultad) a los que la *Inteligencia Artificial* les da la oportunidad de una sofisticación que hasta ahora no era imaginable. Esta combinación de tecnología y problemas de gestión de personas es la base de *People Analytics*. Como ocurre en otros ámbitos de la empresa, *People Analytics* no sólo puede dar respuesta a los problemas tradicionales, sino que permite desarrollar utilidades que hasta hace poco nadie se planteaba (por ejemplo, *chatbots* para resolver dudas internas o como asesores de planes de carrera, aconsejando a cada profesional sobre las mejores oportunidades de desarrollo). La siguiente figura enumera las tipologías de métricas y e indicadores clave del rendimiento o *KPI's* (*Key Performance Indicators*) más comunes en contextos de *People Analytics*.

Métricas y KPI's utilizados en ámbitos de People Analytics

1. Productividad : EBIT / ingresos / rotación / beneficio por empleado; ROI de capital humano, o la relación de ingresos o ingresos a capital humano.
2. Costes : costos totales de la fuerza laboral, costes directos/costes indirectos, ciclo de vida del empleado, costes de contratación.
3. Contratación, movilidad y rotación : tiempo medio para cubrir vacantes; tiempo medio para cubrir posiciones críticas; % de posiciones cubiertas internamente; % de puestos críticos de negocios cubiertos internamente; tasa de rotación, tasa de movilidad interna.
4. Ética : número y tipo de reclamaciones presentadas; número y tipo de acciones disciplinarias concluidas; % de empleados que han completado la capacitación sobre cumplimiento y ética.
5. Diversidad de la fuerza laboral : Indicadores por áreas : departamento/ Staff/ operaciones y edad, género, discapacidad y diversidad del equipo de liderazgo.
6. Liderazgo : Evaluación del liderazgo mediante encuestas a los empleados.
7. Seguridad, salud y bienestar : tiempo perdido por lesiones; número de accidentes laborales, índices de salud/wellbeing: estrés, actividad, … .
8. Habilidades y capacidades : costos totales de desarrollo y capacitación .

En este texto nos vamos a centrar en aplicar *People Analytics* sobre aquellas preguntas de negocio típicas a las que deberá hacer frente el profesional de recursos humanos en algún punto de su carrera. Algunos de los puntos a tratar serán:

1. **Selección y contratación del personal:** *People Analytics* permite identificar los rasgos, habilidades y características que son importantes para el éxito en una posición determinada y, a partir de estos, realizar una selección más adecuada de los futuros empleados. Existe un consenso generalizado sobre cómo la contratación errónea tiene consecuencias terribles, tanto para la empresa como para el propio profesional contratado.

2. **Gestión de la rotación y retención del talento:** *People Analytics* permite identificar los factores que, en mayor medida, explican la rotación del personal. A partir de este conocimiento las organizaciones pueden concentrar sus esfuerzos en aquellas políticas más eficaces para retener

a los empleados clave. Ligado a esta gestión, existe el concepto del ciclo de vida del empleado. Este ciclo de vida es otra perspectiva de la gestión de la rotación, y cobra una especial relevancia en aquellos modelos de negocio en los que ciertos ratios de rotación son esenciales para explicar la propuesta de valor (modelos de gestión "*up or out",* cadenas de *retail* que buscan un determinado perfil de dependiente…).

3. **Gestión y desarrollo del talento:** *People Analytics* tiene en cuenta el perfil y el *performance* (rendimiento) de los empleados para identificar las áreas en las que se necesita una mayor capacitación o desarrollo y, en consecuencia, anticipar futuras necesidades. Esto ayuda a las organizaciones a adelantarse a ciertas problemáticas, creando planes de capacitación y de desarrollo personalizados y efectivos (tanto para el empleado como para la organización).

4. **Evaluación y gestión del rendimiento y el compromiso**: *People Analytics* permite medir y evaluar el rendimiento de los empleados para tomar mejores decisiones sobre el reconocimiento de los profesionales y las oportunidades de desarrollo optimas. Un aspecto cada vez más importante de este rendimiento es el llamado *Engagement* (sinónimo de compromiso o motivación) de los profesionales en relación con la organización. En este caso, el consenso también es general respecto de cómo aquellas organizaciones que disfrutan de mejores métricas en este ámbito construyen más valor para sus accionistas. En este sentido, es fundamental identificar qué factores contribuyen a potenciar dicho compromiso y cuáles tienden a erosionarlo.

5. **Gestión de la comunicación interna:** *People Analytics* facilita herramientas de procesamiento del lenguaje natural o *NLP* (*Natural Language Processing*) que permiten optimizar la comunicación y las conversaciones entre la organización y sus profesionales (recogiendo y procesando de manera objetiva las opiniones de los trabajadores). Este ámbito tiene una creciente relevancia porque gran parte de la "percepción" del empleado sobre su entorno laboral (estrategia, decisiones, información interna…) se articula a través de canales de la comunicación y esta percepción, en gran medida, explica su compromiso y motivación. Por ejemplo, es común encontrar situaciones en las que la organización lanza un mensaje y los profesionales entienden algo diferente de lo que pretendía el emisor, dando lugar a frustración, desalineamiento y rotación no deseada.

6. **Gestión de la diversidad**: *People Analytics* permite identificar y mitigar de forma objetiva desequilibrios o sesgos en la distribución de los

diferentes grupos en la organización, proporcionando información que facilita el desarrollo de estrategias orientadas a promover la diversidad y a fomentar una cultura inclusiva. Por ejemplo, mediante un análisis de los datos se puede determinar si existe una brecha salarial de género (o en relación con otras características) en la organización y establecer si esta diferencia se debe a comportamientos o sesgos específicos, o a factores no observables.

A lo largo de los próximos capítulos abordaremos los principales desafíos de cada una de estas categorías utilizando herramientas y modelos de la *Inteligencia Artificial* y ejemplificaremos su desarrollo y tratamiento mediante múltiples casos prácticos. Ahora bien, dado que la cultura *People Analytics* se fundamenta en la aplicación de la *Inteligencia Artificial* en el ámbito de los recursos humanos, decidimos reservar el siguiente capítulo como introducción a la ciencia de datos. En concreto, exploraremos las estrategias y metodologías clave utilizadas para abordar problemas típicos de análisis de datos, comentaremos algunas herramientas populares y explicaremos los conceptos fundamentales que subyacen. De este modo dispondremos de una sólida base de conocimiento para comprender el funcionamiento de las herramientas de *People Analytics* que serán introducidas y aplicadas en los capítulos siguientes, donde abordaremos diversas problemáticas en los seis ámbitos de la gestión de recursos humanos que fueron enumerados anteriormente.

La ciencia de datos es un campo interdisciplinario que utiliza métodos científicos, sistemas matemáticos y algoritmos para extraer conocimiento e información de los datos (estructurados y no estructurados) y completar tareas de forma dinámica (adaptándose al entorno). Combina aspectos de la estadística, la informática, la *Inteligencia Artificial* y el análisis de datos para interpretar realidades complejas y solucionar problemas prácticos (modelización, generación). En el siguiente capítulo introduciremos algunos de estos conceptos, no sin antes comenzar por la importancia del dato como entidad de referencia.

1

INTRODUCCIÓN A LA CIENCIA DE DATOS

1.1 DATOS, DATOS Y MÁS DATOS…

Que vivimos en el siglo de los datos y la información no le pasa desapercibido a nadie. Todos llevamos un dispositivo que genera datos de todo tipo, en cualquier momento y en cualquier lugar. Este dispositivo nos permite cambiar los datos que generamos (modificando vídeos o imágenes, por ejemplo) y, además, nos proporciona salida al mundo exterior a través de navegadores web y aplicaciones de redes sociales donde seguimos generando contenido que, básicamente, es información y datos. Es un dispositivo de uso universal conocido como…*smartphone*. Junto a nuestro teléfono tenemos todo tipo de dispositivos, como son los portátiles y las *tablets*, que están en la misma línea de generación, modificación, envío y recepción de todo tipo de información. Si nos vamos a las estadísticas, tenemos las enormes cifras de cantidad de información que nos podemos encontrar si realizamos una búsqueda de información generada, tráfico en *Internet*, *Redes Sociales*… Todo son números enormes y que, además, se han visto incrementados tras la pandemia. Existe una famosa página web (cuyo nombre lo indica todo) que hace referencia a estas cifras, *https://www.domo.com/data-never-sleeps*. Esta empresa, *Domo*, lleva realizando informes sobre actividad en *Internet* y *Redes Sociales*, presentando datos como:

A partir de abril de 2022, Internet llega al 63 % de la población mundial, lo que representa, aproximadamente, 5.000 millones de personas. De este total, 4.650 millones (más del 93 %) eran usuarios de Redes Sociales. Según Statista, la cantidad total de datos que se prevé que se creará, capturará, copiará y consumirá a nivel mundial en 2022 es de 97 zettabytes, un número proyectado que aumentará a 181 zettabytes para 2025.

En el párrafo anterior aparece un prefijo de cantidad junto con una unidad de medida de información. Es interesante destacar la magnitud de estos prefijos para apreciar la cantidad de datos de que estamos hablando. El *byte* es la unidad de información fundamental, compuesto por 8 *bits*. Cada *bit* representa un valor binario, es decir, un 0 o un 1, correspondiendo a un "si" o "no" en términos lógicos. Con un *byte* se pueden codificar hasta $2^8 = 256$ instrucciones diferentes correspondientes a las posibles combinaciones de esos ocho valores binarios: 00000000, 00000001, 00000010… Y así hasta 256 diferentes combinaciones. Nuestro alfabeto tiene 27 letras por lo que, con un *byte*, ¡lo puedo codificar entero y me sobran términos! Hagamos una analogía sencilla para entender la clase de cantidades que estamos manejando. Supongamos que un *byte* es equivalente a un milímetro. Entonces, los ordenadores de los años 80 (*Spectrum, MSX, Amstrad*, seguro que algún lector los reconoce), que tenían de media 16 *kilobytes* de memoria física, tendrían una longitud de 16 metros. Los actuales discos duros de 1 *terabyt*e donde almacenamos películas, vídeos, libros, etc., tendrían una longitud de 1 millón de kilómetros (ya empiezan a aparecer números interesantes), lo que equivale a 2,6 veces la distancia de la Tierra a la Luna. Un *zettabyte* se correspondería con 10^{15} kilómetros y, evidentemente, es una distancia que entra dentro de lo astronómico (y eso que hemos partido de un milímetro).

Una vez que hemos comprendido la cantidad de información que suponen estos "inocentes" prefijos, vamos a dar algunos datos de lo sucedido en un minuto en las *Redes Sociales* o el *Internet* durante el año 2022:

- Se envían 16 millones de mensajes de *SMS*.
- Los usuarios de *Instagram* comparten 66000 fotos.
- Los usuarios de *Google* realizan 5,9 millones de búsquedas.
- Los usuarios de *Facebook* comparten 1,7 millones de piezas de contenido.
- Los compradores de *Amazon* gastan 443.000 dólares.
- Los usuarios de *Twitter* escriben 347200 *tuits*.
- Los usuarios de *YouTube* suben 500 horas de vídeo.
- Los usuarios de correo electrónico envían 231,4 millones de mensajes.

¡Y todo lo anterior durante un minuto! Nunca se había generado tal cantidad de datos (que no de información, ya hablaremos de esta diferencia más adelante) en toda la historia de la humanidad. La primera pregunta que se quiere lanzar es: ¿de verdad una empresa renunciaría a los beneficios que ofrecen las posibilidades de usar estos datos? Y, más concretamente, dado que estos datos reflejan el comportamiento, aptitudes y actitudes de los humanos que los generan, ¿debemos seguir empleando técnicas *clásicas* en recursos humanos? Seguiremos con este debate a lo largo del capítulo, pero ya tenemos una pieza esencial para cambiar la forma de actuar de estos departamentos: *existe una cantidad astronómica de datos que se pueden explotar y analizar*. Aquí toca comentar las tendencias puestas de manifiesto según el famoso

economista y premio *Nobel Herbert Simon*; por una parte, la riqueza de información que actualmente tenemos y, por otra, la poca atención que suele recibir. Nuestra capacidad de procesamiento está muy por detrás de la velocidad de generación de los datos. Además, nuestra capacidad de analizar información de diferentes fuentes al mismo tiempo está destinada a fallar. Ahí están todos los estudios de la *Economía del Comportamiento* que demuestran nuestros sesgos e incapacidades para manejar grandes cantidades de información. Por todo ello se hace necesario plantear sistemas automáticos para procesar la información que actualmente tenemos disponible; de ahí el surgimiento de la analítica de datos.

Una vez comprendida la cantidad, pasemos a los tipos de datos que nos podemos encontrar. Actualmente, se tienen dos grandes grupos:

- *Datos estructurados*. Como su nombre indica, son datos que presentan una cierta estructura (tamaño y atributos fijos). Son fáciles de almacenar y procesar y, hasta hace pocos años, eran los datos mayormente utilizados en aplicaciones derivadas de la analítica. Cualquier base de datos tabular con el mismo número de características o atributos por entidad respondería a esta tipología. Por ejemplo, los datos que se almacenan en hojas de cálculo, donde cada celda tiene un tipo de dato, un tamaño y una estructura (y esta estructura es respetada a lo largo de toda la base de datos).

- *Datos no estructurados*. En este caso la situación es opuesta a la del tipo de dato anterior. Los datos no presentan estructura definida, ni un tamaño fijo, por lo que la forma de almacenarlos y de procesarlos requiere de metodologías más complejas en comparación con el caso anterior. Ejemplos de este tipo de datos serían textos, audio, voz, *logs*, etc. En definitiva, todo lo que no se corresponda con una base de datos tabular con estructura fija entraría en esta tipología.

La siguiente tabla resume las diferencias entre estas dos tipologías de datos:

	Datos estructurados	Datos no estructurados
Organización	Está predefinida	No tiene organización predefinida
Tamaño	Pequeño/mediano	Generalmente grande
Accesibilidad	Fácilmente accesible mediante una base de datos	Más difícil de acceder debido a la falta de organización predefinida
Procesamiento	La aplicación de diferentes algoritmos es sencilla	Más difícil de procesar y analizar debido a su desorganización

Tabla 1.1. Diferencias entre tipos de datos; estructurados y no estructurados.

Debido al crecimiento exponencial de la capacidad de generación de datos y a la accesibilidad a sistemas de almacenamiento y procesamiento, en los últimos años se ha popularizado el término *"Big Data"*. Este concepto hace referencia a los datos que cumplen las famosas *3 V's:*

1. *Volumen*: se refiere al tamaño de los datos considerados, de tal manera que su cantidad supone un desafío importante para los sistemas clásicos de almacenamiento y procesamiento. Aunque no hay una distinción específica sobre el volumen de datos, normalmente, este puede variar desde los *terabytes* (10^12 *bytes*) hasta los *exabytes* (10^18 *bytes*).
2. *Velocidad*: los datos se generan a un ritmo muy rápido. Esta alta velocidad puede observarse en el hecho de que una gran proporción de los datos que se utilizan actualmente pertenecen al pasado reciente.
3. *Variedad*: los datos analizados pueden ser estructurados o no estructurados y pueden obtenerse de numerosas fuentes como registros *web*, dispositivos del *Internet de las Cosas* (*IoT*), de *URLs*, de *tweets* de usuarios, de patrones de búsqueda, etc. Asimismo, los datos pueden estar almacenados según diferentes formatos: valores separados por comas o *comma sepparated values* (*CSV*), tablas, documentos de texto, gráficos... Esta cantidad ingente de información es imposible de procesar por nosotros. Tenemos una forma de procesar la información definida por la evolución y la naturaleza de esta. No podemos tratar de abarcar todos los procesos de generación y procesado de los datos (especialmente en los datos creados por máquinas para ser consumidos por otras máquinas) sino en los que nos son más relevantes y/o recientes.

Otro punto importante en el auge de los datos es el aumento de la potencia de cálculo en los dispositivos electrónicos, así como de la capacidad de almacenamiento. Todas las gráficas que representan estas dos características en función del tiempo siguen una curva de tipo exponencial, desde una perspectiva anual (que es una de las curvas que presenta un mayor crecimiento). Para comprender esta relación se suele hacer la siguiente analogía con el mundo de los coches. Si la industria automovilística hubiera seguido la misma evolución que la potencia de cálculo de los ordenadores comerciales, actualmente tendríamos acceso a un coche de alta gama (*Ferrari*, *Lamborghini*, *Rolls Royce* y el que pueda pensar el lector) ¡por menos de un céntimo de euro! Por otra parte, en cuanto a la capacidad de almacenamiento se ha pasado de almacenar unas cuantas líneas de código que hacían posibles esos videojuegos de los años 80 a almacenar películas con alta resolución.

Continuando con el tema de la capacidad de los equipos, en los últimos años aparece otro elemento clave en la revolución de los datos: los sistemas de

computación en la nube (*Cloud Computing*). *Amazon* lanza en 2006 su servicio de *Cloud Computing*, conocido como *Amazon Web Services (AWS)*. El conocimiento que, en aquel entonces, tenía *Amazon* sobre centros de datos (*data centers*) fue aplicado a la venta de recursos en la nube para sus clientes. Actualmente, según un estudio reciente un 94% de las organizaciones, desde corporaciones multinacionales hasta pequeñas y medianas empresas, utilizan la computación en la nube. La computación en la nube facilita la disponibilidad remota de sistemas de almacenamiento y de procesamiento de datos de modo que, prácticamente cualquiera, pueda acceder y utilizar complejas herramientas de la *Inteligencia Artificial* (*IA*) a un precio que podríamos calificar de irrisorio si tenemos en cuenta la ingeniería que subyace. Actualmente, si una *start-up* quiere llevar a cabo una idea de negocio que requiere del procesamiento de datos, que pueden entrar dentro de lo ya mencionado como *Big Data*, no necesita realizar inversiones millonarias en *hardware,* sino que puede contratar el equipo necesario para llevar a cabo el procesamiento en la nube y desarrollar el modelo necesario para su problemática, pagando por el uso y no por el recurso en sí. Se ahorran pues los servidores y el salario de un equipo de especialistas para su mantenimiento. Actualmente, *Amazon, Microsoft, Google e IBM* ofrecen acceso a sus herramientas de *Cloud Computing*, permitiendo a las organizaciones importar datos, entrenar modelos y desplegar soluciones dentro de sus aplicaciones. Esto reduce enormemente el tiempo (y el precio) que se invierte en pasar de la idea al prototipo y del prototipo al despliegue en producción. Además, este tipo de herramientas suelen ofrecer modelos ya desarrollados (sistemas de reconocimiento de imágenes, sistemas *web* de recomendación, modelos predictivos, buscadores.... Son los llamados *servicios cognitivos*. En los últimos años, estos servicios están siendo transformados radicalmente por la rama de la *Inteligencia Artificial* conocida como *IA Generativa*, capaz de proporcionar soluciones autosuficientes ante determinadas preguntas de negocio o tareas planteadas en lenguaje natural.

1.2 ETAPAS DE UN ANÁLISIS DE DATOS

Antes de comenzar nuestro periplo por el mundo de la analítica definiremos los conceptos de variable y patrón, ambos fundamentales. Consideremos una base de datos estructurada (todas las características han de ser las mismas para los diferentes patrones), como la que puede surgir de un proceso de contratación de personal en el que se recogen unas series de características (expediente académico, formación, idiomas, experiencia…) para cada candidato. En este ejemplo, cada candidato (entidad) sería un patrón y cada característica una variable. La siguiente tabla muestra una situación similar a la expuesta (cada fila o entidad se corresponde con un patrón y cada columna o característica con una variable):

Paciente	Años experiencia	Hijos	Nº masters
Emilio	25	0	2
Hector	10	2	3
Antonio	5	0	1

Tabla 1.2. Base de datos estructurada para un problema de análisis de datos (cada fila se corresponde con un patrón y cada columna con una variable).

Tengamos en cuenta que existen casuísticas diferentes, no necesariamente asociadas con bases de datos estructuradas. Por ejemplo, en un problema de clasificación de imágenes cada imagen se corresponde con un patrón y cada uno de los *pixeles* de la imagen con una variable. En relación con las variables, existen múltiples tipologías. Las más fundamentales, y las que trataremos en los diversos casos prácticos que serán planteados en capítulos posteriores, son las siguientes:

1. *Categóricas*: toman valores de un conjunto discreto de posibles opciones o candidatos (en definitiva, de un conjunto finito de elementos). Por ejemplo, si hablamos de colores tendremos como candidatos: rojo, verde, negro, etc. Si hablamos de intensidad del dolor, esta puede ser codificada en varios niveles: bajo, medio o alto, por ejemplo. En el primer caso se tiene una variable categórica nominal y en el segundo se tiene una variable categórica ordinal (existe una relación de orden y gradación entre los diferentes niveles).

2. *Continuas*: en este caso el valor de la variable puede ser un número real. Es decir, son variables que toman valores de un conjunto continuo de números reales, el cual puede estar definido como un intervalo de la recta real o como una distribución de probabilidad continua. Algunos ejemplos de este tipo de variables son el peso de una persona, el nivel de hemoglobina en sangre, el tiempo de permanencia en una determinada página *web*...

Existen otras tipologías como pueden ser las variables de tipo fecha (codifican instantes temporales), variables indexadoras (almacenan una referencia única para cada patrón) o variables textuales (almacenan texto). Además, también es común hablar de variables discretas, aunque no es la terminología que utilizaremos durante este libro. Estas variables toman valores de un conjunto numerable de elementos (es decir, un conjunto que puede ser contado, ya sea finito o infinito), como puede ser el conjunto de los números enteros. De hecho, las variables categóricas son un caso especial de las variables discretas donde el conjunto de candidatos es finito.

Una vez hemos definido ciertos conceptos fundamentales, procedemos a introducir el procedimiento clásico de análisis de los datos. Para ello supongamos que disponemos de datos similares a los anteriores en cuanto a estructura, ¿cuál sería el siguiente paso? Pues plantear la pregunta de negocio que queremos contestar a partir de los datos. Estos deben tener la suficiente calidad y cantidad como para resolver dicha pregunta. En este libro se formularán muchas y diferentes preguntas de negocio para resolver a partir de datos. La formulación de las problemáticas a tratar es un paso de importancia crucial, pues determinará las herramientas adecuadas para su tratamiento, y debe ser realizado por un especialista en el entorno en cuestión. En la mayoría de los casos, la solución desde una perspectiva de *People Analytivs* consiste en aplicar uno de los dos siguientes enfoques: modelización (regresión o clasificación) o extracción del conocimiento (KDD; *Knowledge Discovery in Databases* o *Descubrimiento del Conocimiento en Bases de Datos)*. En función de la pregunta se requerirá la aplicación de uno, de otro o de ambos. Llegado a este punto, supongamos que disponemos de determinadas preguntas de negocio, datos que contienen información útil para resolverlas (al menos parcialmente) y ciertas herramientas o algoritmos que nos permitirán desarrollar las respuestas desde un enfoque basado en la analítica. Supongamos además que el problema es de extracción del conocimiento. Este tipo de aproximaciones se tratan mediante técnicas del área de la ciencia de datos conocida como *KDD* y son equivalentes en cierto sentido a las diferentes fases o etapas que conforman el flujo típico de un análisis de datos hasta la etapa de modelado (figura 1.1). Sin embargo, existen problemas cuya resolución requiere de modelizar una o varias variables. En este sentido, faltan por definir dos elementos críticos: la métrica de evaluación y la estrategia de validación. La primera es un indicativo del rendimiento del modelo que no depende de su funcionamiento sino de la calidad en las predicciones (o equivalentemente, de la magnitud de los errores). Por ejemplo, si tratamos de predecir el absentismo laboral, ¿con qué grado de acierto estamos satisfechos? ¿cómo definimos matemáticamente, o al menos como pseudocódigo, el grado de acierto? Evidentemente, la respuesta a estas preguntas dependerá de la naturaleza de los datos (variable categórica o continua) y del entorno de aplicación (clases desbalanceadas, escala de medición). Por otro lado, la estrategia de validación define cómo interactúan los datos con la métrica de evaluación para producir un indicativo del rendimiento del modelo, y cómo interpretar ese indicativo (en esencia, que cantidad de los datos utilizar para ajustar el modelo y que cantidad utilizar para evaluar su rendimiento). Sobre todos estos conceptos, hablaremos con un mayor detalle en próximos apartados.

Una vez definidos estos elementos podemos plantear la resolución del problema propuesto mediante un análisis de los datos. Para ello se plantean una serie de etapas que quedan reflejadas en la figura 1.1 y cuyo significado pasamos a comentar a continuación:

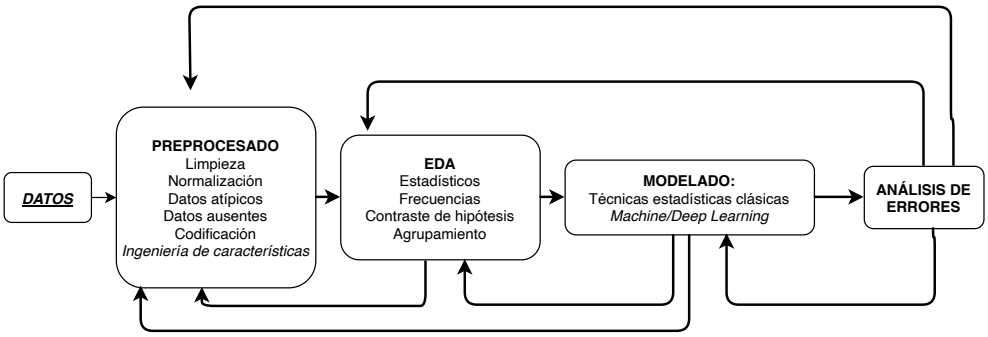


Figura 1.1. Esquema de las diferentes etapas de un análisis de datos (EDA es el acrónimo de Exploratory Data Analysis o Análisis Exploratorio de Datos).

Una de las cuestiones más importantes del análisis de datos es la realimentación de las etapas; no es un proceso secuencial, sino que, según lo obtenido en cada etapa, resulta necesario volver a la(s) etapa(s) anteriores. Incluso en determinados problemas (KDD) puede ser innecesario realizar determinadas etapas (modelado). A continuación, procedemos a explicar con detalle el significado de cada una de estas etapas y su efecto en los resultados.

1.2.1 Preprocesado de los datos

La primera fase de un análisis de datos, el preprocesado, es la más importante y supone el 80% del tiempo invertido en la resolución de cualquier problema basado en datos. Su importancia radica en varios aspectos clave. En primer lugar, permite garantizar la calidad de los datos que se utilizan en el análisis. Los conjuntos de datos suelen contener errores, valores atípicos, valores faltantes y ruido. El preprocesado resuelve estos problemas, lo que resulta en datos más limpios y confiables. En segundo lugar, es necesario para adaptar los datos a los algoritmos de aprendizaje máquina que se utilizarán en el proyecto. Esto implica la transformación de los datos en un formato adecuado para el modelo, lo que, a menudo, incluye la codificación de variables, la normalización de datos numéricos y la selección de características relevantes. Sin esta preparación, los modelos pueden no funcionar correctamente. Además, el preprocesado puede contribuir significativamente a la eficiencia computacional. Reducir la dimensionalidad de los datos, eliminar características irrelevantes y tratar los valores faltantes permite acelerar el proceso de entrenamiento del modelo, algo fundamental en proyectos de gran envergadura o en tiempo real. Otra razón adicional de especial importancia para realizar el preprocesado es la interpretabilidad de los resultados. Cuando se trata de explicar los resultados de un modelo, es esencial que los datos estén en un formato comprensible. Esto implica, en algunas ocasiones, la transformación de los datos de manera que las decisiones

del modelo sean interpretables y significativas. Resumiendo, el preprocesado de los datos es fundamental en un proyecto cualquiera de ciencia de datos porque garantiza la calidad de los datos, prepara los datos para su modelado, mejora la eficiencia computacional y contribuye a la interpretabilidad de los resultados. Ignorar esta etapa puede conducir a resultados inexactos y decisiones erróneas, lo que subraya su importancia en todo el ciclo de vida de un proyecto basado en datos.

En esta etapa se suelen plantear las siguientes fases:

1. *Limpieza de datos*. La tabla 1.2 representa un conjunto de datos "ideal"; no existen datos faltantes (todos los campos de la tabla se conocen) y, además, los valores no parecen estar afectados por ruido ni aparecen datos atípicos (*outliers*). En este punto es importante distinguir entre dato atípico y dato incorrecto. Supongamos que una empresa recopila datos sobre los salarios de sus empleados. En una tabla de datos, se encuentra un registro que indica que un empleado junior con poca experiencia gana un millón de euros al año. Este valor es, claramente, un dato erróneo, pues es extremadamente inusual y poco realista para un empleado de ese nivel. En este caso se trata de un dato incorrecto, posiblemente debido a un error de entrada, y debe corregirse. En el mismo conjunto de datos de salarios de la empresa, se encuentra un registro que muestra que un empleado junior gana 100.000 euros al año. Si la mayoría de los empleados en un puesto similar ganan alrededor de 40.000 euros al año, el salario de 100.000 euros podría considerarse un dato atípico. Aunque no es incorrecto, es inusual en comparación con el patrón general de salarios para ese puesto. Los datos atípicos pueden proporcionar información valiosa, cuya utilidad puede ser la de identificar a los empleados excepcionales o aquellos casos que requieren de una revisión más profunda. La diferencia clave entre ambas connotaciones está en que un dato erróneo es incorrecto y no tiene sentido en el contexto, mientras que un dato atípico es válido, pero inusual en comparación con el conjunto de datos general, y, a menudo, puede ser informativo en lugar de un error.

2. *Normalización.* Algunos modelos de aprendizaje máquina son robustos a una diferencia de rangos (escalas de medición) entre las variables de entrada. Los árboles de decisión y sus modelos derivados (por ejemplo, *Random Forest*) son robustos a este hecho. Sin embargo, en la mayoría de los modelos tendremos problemas en su ajuste. Además, en algunos modelos, esta diferencia de rango provocará que el algoritmo dote de una mayor importancia a ciertas variables frente a otras, no por su papel en el problema a resolver sino por el rango de sus valores. Es el caso de aquellos modelos cuyo funcionamiento depende de la escala de las variables (por ejemplo, modelos que calculan distancias entre patrones

en el espacio de características original). En *People Analytics* es común trabajar con múltiples variables como salarios, años de experiencia, calificaciones, evaluaciones de desempeño... Estas variables pueden tener unidades y escalas diferentes. En definitiva, la normalización garantiza que todas las variables se representen en una escala similar, lo que facilita la comparación, el análisis equitativo y la estabilidad de determinados algoritmos predictivos. Por ejemplo, sin normalización, el salario medido en euros podría dominar la importancia de otras variables, como la calificación del desempeño si se mide en una escala de 1 al 5.

3. *Detección de outliers.* La presencia de datos atípicos (*outliers)* puede impactar negativamente en la estabilidad y calidad de un determinado modelo matemático o algoritmo. Por ejemplo, el resultado de un modelo de regresión lineal (apartado 3.4) se puede ver seriamente afectado por la presencia de un solo valor atípico en los datos. Esto se debe a que este tipo de modelos tratan de ajustar la media de los valores, lo que hace que la presencia de este tipo de patrones modifique en gran manera el ajuste (la media es sensible a los valores extremos), conduciendo a errores en las interpretaciones o durante la evaluación de la capacidad de generalización del modelo en datos que no se han observado previamente. La figura 1.2 representa una situación de este estilo: en esta figura se consideraba añadido un punto atípico (marcado como un punto circular), el cual modifica en gran manera el modelo obtenido anteriormente.

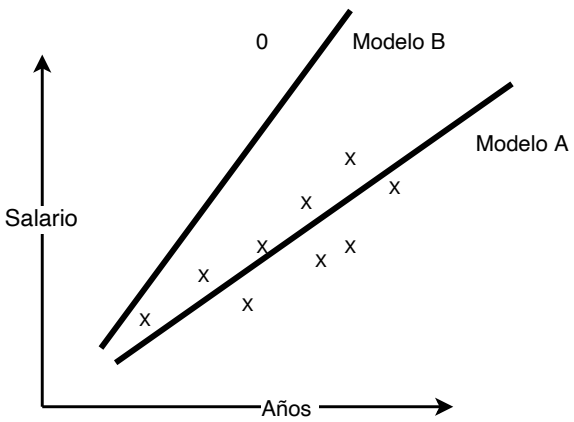


Figura 1.2. Impacto de un outlier en un modelo de regresión lineal. El problema considerado consiste en modelizar una variable continua en función de los años y el salario del empleado. Se entrena el mismo modelo de regresión lineal sobre dos conjuntos de datos diferentes: el formado por las "x" (patrones no atípicos) resultando en el modelo A y el formado por todos los puntos (se añade al anterior conjunto el patrón atípico, marcado como "0"). El efecto del outlier en la recta de regresión es evidente (modifica sustancialmente la tendencia global).

Existen diferentes procedimientos para la caracterización de este tipo de datos, dependiendo de si buscamos un valor atípico dentro de una variable, o bien, un patrón atípico dentro de una base de datos. El primer caso es más sencillo. Como ejemplo introductorio de métodos de detección de outliers, la métrica denominada del z-score y derivada de la estadística (en concreto del concepto de distribución normal o gaussiana) permite identificar outliers bajo ambos enfoques (local o global). Según un enfoque global (la variable contiene outliers o no) este método consiste en calcular la diferencia entre la media y la mediana de la variable en cuestión y, si esta toma un valor muy diferente a cero, entonces podemos considerar que contiene *outliers.* El *z-score* de un patrón mide cuántas desviaciones estándar está el patrón por encima o por debajo de la media. Los *outliers* se identifican cuando los valores de este índice son significativamente altos o bajos. Por lo general, un *z-score* por encima de 3 o por debajo de -3 se considera un *outlier*. Por otro lado, la búsqueda de patrones atípicos es más complicada. Existen múltiples algoritmos para este fin; entre ellos destacar todos los que derivan de los algoritmos de agrupamiento (*clustering*) de los datos. Si lo particularizamos al contexto de *People Analytics*, se agrupan los datos de comportamiento de los empleados en grupos (*clusters*) basados en similitudes. Posteriormente, se identifican aquellos empleados cuyo comportamiento no se ajusta bien a ningún *cluster*, lo que puede considerarse como un comportamiento atípico. A modo de ejemplo, supongamos que se analiza el uso del tiempo en el lugar de trabajo agrupando a los empleados en función de sus patrones de asistencia y productividad. Si se encuentra un empleado cuyos patrones no se asemejan a ningún grupo en particular, eso podría indicar que su comportamiento es inusual y merece una revisión.

4. *Datos ausentes*. Existen modelos que pueden funcionar con datos ausentes, por ejemplo, los modelos gráficos probabilísticos, pero son los menos comunes. La gran mayoría de modelos necesitan que los patrones de entrada estén completos. En nuestro contexto, la ausencia de datos puede deberse a varias razones, como errores de entrada, falta de recopilación de información en ciertas áreas, o incluso la negativa de los empleados a proporcionar ciertos datos por razones de privacidad. En este sentido, una solución consiste en sustituir de manera inteligente los campos vacíos por valores. Este tipo de procedimientos se conocen como de *imputación de datos faltantes* o *missing data imputation* y el proceso de sustituir datos ausentes por valores se conoce como "imputar". Aquí también se divide la estrategia según se considere un punto de vista enfocado a la variable o al patrón. Además, como suele ocurrir en este tipo de problemas, no existe una receta o procedimiento por excelencia, sino que disponemos

de múltiples metodologías y algoritmos. Desde métodos sencillos hasta algoritmos de *aprendizaje automático*. A continuación, explicaremos el funcionamiento de uno de los más sencillos y utilizados desde una perspectiva de la variable: sustituir el dato ausente por el valor promedio de la variable en cuestión. Si la variable es categórica, el valor ausente de un patrón cualquiera se sustituye por la moda (categoría que aparece con mayor frecuencia) de dicha variable. Si la variable es continua, se sustituye por la media (o la mediana). Esta metodología, aunque efectiva, tiene la desventaja de ignorar la información del resto de variables y patrones y carecer de capacidad de adaptación (la sustitución es la misma para cada patrón). Por otro lado, si se escoge una aproximación a nivel de patrones, una técnica más robusta y muy utilizada consiste en buscar aquellos patrones más parecidos o cercanos a los patrones que presentan valores faltantes y se escogen los valores de las variables que se tienen para ser consideradas en los patrones que no las tienen. La imputación comentada es un método simple y efectivo para abordar valores faltantes en *People Analytics*al utilizar la información de registros/patrones similares. Puede ser particularmente útil cuando se trata de datos de comportamiento de empleados o evaluaciones de desempeño, donde las similitudes en los patrones pueden ser informativas para la imputación. Todo lo comentado depende, evidentemente, del porcentaje de datos faltantes. En definitiva, nos estamos inventando una realidad que no es seguro que tengamos. Esto supone que, en muchas ocasiones, se opta por descartar aquellos patrones que contienen datos faltantes (siempre y cuando esto no afecte demasiado a la calidad de los resultados y a la validez de las conclusiones). En el contexto de *People Analytics,* es fundamental tener en cuenta las implicaciones éticas y de privacidad al abordar los datos ausentes. Asegurarse de que cualquier imputación o manejo de datos respete las regulaciones y políticas de privacidad es esencial. Además, es crucial comunicar de manera transparente si se han realizado imputaciones de datos y cómo se han abordado los valores faltantes. Esto es importante para mantener la confianza y la credibilidad en los resultados.

5. *Ingeniería de características*. Se trata de una de las etapas más importantes en el aprendizaje automático y que, a priori, no es necesaria en el aprendizaje profundo (de ahí la revolución que supuso frente a su antecesor). Dentro del campo de *People Analytics* se podría definir como el proceso de crear, modificar y seleccionar variables a partir de los datos disponibles con el objetivo de mejorar la calidad de la información y la capacidad de los modelos analíticos para captar relaciones y tendencias en el comportamiento de los empleados y la gestión de recursos

humanos. Como en las fases anteriores, existen múltiples y diferentes metodologías y técnicas para abordar este tipo de procedimientos. Desde métodos univariantes que puntúan variables en base a su relevancia para el problema en cuestión hasta algoritmos de optimización que buscan el mejor subconjunto de variables. La utilidad de estas técnicas es múltiple; a) mejora de la precisión, ya que permite diseñar variables que capturan mejor la información relevante en los datos, lo que lleva a modelos de predicción más precisos y robustos y resultados más confiables; b) mejora la interpretación y el significado de los resultados, ya que ayuda a dar sentido a los datos y facilita la interpretación de resultados en términos comprensibles para la toma de decisiones en recursos humanos; c) reducción del ruido, ya que ayuda a eliminar variables irrelevantes o redundantes, reduciendo la complejidad en los modelos. Algunos ejemplos en *People Analytics* serían; a) *creación de indicadores de desempeño combinados*; se pueden combinar múltiples métricas en un solo indicador que refleje el desempeño global de un empleado, facilitando la toma de decisiones objetivas; b) *categorización de datos cualitativos*; se pueden transformar variables cualitativas, como niveles de educación o tipos de contratos, en variables categóricas para incluir en modelos de análisis; c) *segmentación de empleados*; se pueden agrupar a los empleados en segmentos o grupos según características como habilidades, experiencia, preferencias o lo que puede ayudar en la toma de decisiones estratégicas.

1.2.2 Análisis exploratorio de los datos

Como siguiente etapa en el proceso de extracción del conocimiento de los datos, se encuentra el análisis exploratorio de datos. Aquí se realizan varias tareas, siendo una de las primeras la descripción de los datos (análisis descriptivo). Para ello se calculan estadísticos descriptivos (media/moda, dispersión, percentiles, frecuencias, valores máximo y mínimo…). Por ejemplo, la media de los salarios en la organización, como medida de tendencia central, puede ser información de utilidad en la toma de decisiones sobre la estructura de compensación. En el caso de las variables categóricas, su distribución se puede resumir numéricamente mediante el cálculo de las frecuencias de cada categoría. Por ejemplo, conocer la frecuencia de los empleados por departamento ayudaría a comprender la composición de la fuerza en la organización. Si se dispone de variables temporales (aquellas que asocian un instante temporal único a cada observación) se puede limitar el cálculo de los estadísticos descriptivos a diferentes periodos temporales (por ejemplo, meses o años), lo que puede revelar tendencias estacionales o cambios significativos con el paso del tiempo.

Otra parte del análisis exploratorio es la inferencia estadística (establecer resultados sobre una población a partir de una muestra; requieren de ciertas suposiciones probabilísticas). Una de sus aplicaciones son los contrastes de hipótesis para determinar normalidad, independencia y posibles relaciones lineales entre variables continuas. La estadística inferencial se utiliza en *People Analytics* para generalizar a partir de una muestra de datos y hacer inferencias sobre toda la población de empleados o aspectos relacionados con recursos humanos. El objetivo es obtener una comprensión más profunda de las tendencias, relaciones y patrones en los datos, así como tomar decisiones informadas basadas en esas inferencias. Un primer ejemplo de aplicación sería realizar inferencias sobre la población en función de una muestra representativa. Por ejemplo, se puede estimar la tasa de satisfacción de todos los empleados a partir de una muestra y calcular intervalos de confianza. Un segundo ejemplo sería determinar si la rotación de empleados en un departamento específico es significativamente diferente de la rotación en el conjunto de la organización. Se plantea entonces una hipótesis nula (no hay diferencia significativa) y una hipótesis alternativa (hay una diferencia significativa) y, a continuación, se realiza una prueba estadística, como la *prueba t de Student* o una prueba *Chi-cuadrado*, para evaluar si existe una diferencia significativa.

La siguiente fase es la de visualización de los datos. Este paso, normalmente, no recibe la importancia que tiene. En *People Analytics*, las herramientas de visualización son especialmente útiles, pues ayudan a comunicar resultados, tendencias y relaciones de manera más efectiva a las partes interesadas, incluso a aquellas que no tienen un fondo técnico. En este punto hay que añadir que, al comprender visualmente los datos, los profesionales de recursos humanos pueden tomar decisiones basadas en la evidencia, lo que es esencial para optimizar la gestión del personal. Además, contextualizan los datos de recursos humanos en relación con otros factores, como la economía o la industria, para una mejor toma de decisiones. Sin embargo, su utilización conlleva ciertos riesgos asociados con la subjetividad en las interpretaciones de los diagramas, por lo que estos suelen estar acompañados o apoyados de medidas cuantitativas. Finalmente, cabe destacar el gran número de técnicas de visualización que han aparecido en los últimos años, especialmente en el terreno del análisis multivariante (el estudio de las relaciones entre múltiples variables), lo que se conoce como representaciones multidimensionales. Entre las nuevas técnicas de visualización, destacamos las siguientes:

1. *Diagramas de mariposa*. Estos gráficos permiten comparar dos distribuciones de probabilidad. En *People Analytics* podrían utilizarse para comparar dos grupos de empleados en base a métricas como la satisfacción laboral.

2. *Diagramas de Ternary*. Útiles para representar las relaciones entre tres variables. Por ejemplo, podrían usarse para mostrar cómo se relacionan la experiencia, la edad y el rendimiento de los empleados.

3. *Heatmaps 3D*. Estos gráficos tridimensionales permiten visualizar patrones sobre una matriz de datos. En el contexto de *People Analytics*, podrían utilizarse para analizar las relaciones entre variables como la antigüedad, la rotación y el salario.

4. *Nubes de palabras en 3D*. Las nubes de palabras tridimensionales agregan una dimensión adicional a las nubes de palabras comunes, lo que puede revelar patrones de texto en encuestas de empleados o revisiones de desempeño.

5. *Gráficos de red*. Estos gráficos representan conexiones entre empleados en una organización. Pueden mostrar relaciones profesionales y de colaboración que no son evidentes en otros tipos de visualizaciones.

6. *Gráficos de violín*. Son una combinación de histogramas y gráficos de caja que permiten visualizar la distribución de datos en múltiples dimensiones.

7. *Gráficos de radar*. Útiles para comparar múltiples atributos de empleados. Por ejemplo, podrían utilizarse para comparar las habilidades de diferentes departamentos en una organización.

8. *Gráficos de árbol*. Pueden mostrar la estructura jerárquica de una organización y ayudar a identificar posibles áreas de mejora en la gestión de recursos humanos.

Estos tipos de gráficos menos comunes en *People Analytics* proporcionan una perspectiva más completa y detallada de los datos relacionados con el personal y los recursos humanos. La elección de la visualización depende de los objetivos específicos del análisis y de la naturaleza de los datos en cuestión (¡no todo son diagramas de sectores y de barras!). La siguiente figura muestra diferentes tipos de visualizaciones según lo que se quiere destacar; imprescindible tenerla en cuenta para cualquier presentación de resultados.

Figura 1.3. Técnicas de visualización agrupadas según el tipo de enfoque lo que se pretende destacar. Fuente: https://github.com/Financial-Times/chart-doctor/tree/main/visual-vocabulary

La siguiente etapa en el análisis exploratorio de datos es el agrupamiento de patrones o *clustering* para determinar identificar comportamientos y aspectos similares. Desempeña un papel esencial en el campo de *People Analytics*al ayudar a comprender mejor la diversidad y las similitudes entre los empleados, lo que puede ser valioso para la toma de decisiones estratégicas y la gestión de recursos humanos. Algunos ejemplos de sus aplicaciones son:

- *Segmentación de la fuerza laboral*. Estas técnicas permiten dividir a la fuerza laboral en segmentos o grupos más homogéneos en función de características compartidas. Esto puede ayudar a los profesionales de recursos humanos a comprender mejor a los empleados y a adaptar sus estrategias y políticas según las necesidades de cada grupo.

- *Identificación de diferentes tipos de comportamiento*. El agrupamiento revela tipos de comportamiento y tendencias entre grupos de empleados. Por ejemplo, permitiría identificar grupos de alto rendimiento, empleados en riesgo de rotación o con cierto tipo de habilidades.

- *Personalización de estrategias*. Al comprender las necesidades y características únicas de cada grupo de empleados, las organizaciones pueden personalizar estrategias de capacitación, retención, compensación y desarrollo profesional. Esto conduce a una gestión más efectiva de los recursos humanos.
- *Detección de problemas y oportunidades*. El agrupamiento identifica problemas u oportunidades ocultas en la fuerza laboral. Por ejemplo, puede descubrir un grupo de empleados con altas tasas de rotación y tomar medidas específicas para abordar este problema.
- *Segmentación geográfica*. En organizaciones con múltiples ubicaciones, se puede utilizar para segmentar a los empleados en función de la ubicación, lo que es útil para adaptar las políticas y prácticas de recursos humanos a las necesidades locales.
- *Optimización de recursos humanos*. Al agrupar empleados con características similares, las empresas pueden asignar recursos humanos de manera más eficiente, ya que pueden dirigir esfuerzos y recursos a grupos específicos que requieran de una mayor atención.

El agrupamiento de datos en *People Analytics* es una herramienta poderosa que permite comprender la diversidad y similitudes en la fuerza laboral, lo que a su vez facilita una toma de decisiones más estratégicas y una gestión de recursos humanos más efectiva. Ayuda a las organizaciones a adaptar sus enfoques y políticas a las necesidades de diferentes grupos de empleados, lo que puede tener un impacto significativo en la retención, el desempeño y la satisfacción de los empleados.

1.2.3 Modelización

La siguiente etapa es el desarrollo del modelo. Esto dependerá de la cantidad de datos y del objetivo que se pretenda resolver; ambos factores determinarán si se escoge un modelo de aprendizaje automático o uno de aprendizaje profundo. Aquí hay que tener en cuenta lo que se conoce como el teorema de "*No free lunch*" que refleja que, *a priori*, no existe un mejor modelo para todos los problemas; esto conduce a realizar una búsqueda intensiva para encontrar el mejor modelo que resuelva el problema planteado. En primer lugar, definiremos los tipos de problemas que podremos resolver según el tipo de *aprendizaje* del modelo.

- *Aprendizaje supervisado*. En este tipo de aprendizaje el modelo se ajusta o entrena utilizando datos etiquetados, lo que significa que se le proporciona información de entrada junto con la respuesta correcta (*ground truth*). El modelo aprende a hacer predicciones en base a estas

etiquetas, siendo el objetivo minimizar el error entre las predicciones y las etiquetas verdaderas. Ejemplos en *People Analytics* serían los siguientes:

- *Predicción de la rotación de los empleados*. Se pueden utilizar datos históricos de empleados que se han ido o se han quedado (con etiquetas) para entrenar un modelo supervisado que prediga la probabilidad de que un empleado actual renuncie.
- *Selección de candidatos*. El aprendizaje supervisado se usa para entrenar modelos que evalúan las características de los candidatos (datos de entrada) y determinan si son adecuados o no para el puesto (respuesta correcta).
- *Evaluación de desempeño*. Se pueden utilizar datos de desempeño pasados junto con calificaciones (etiquetas) para entrenar un modelo que prediga el rendimiento futuro de un empleado.

- *Aprendizaje no supervisado*. En este tipo de modelo se buscan patrones, comportamientos o estructuras en los datos sin utilizar etiquetas o respuestas correctas. En *People Analytics* se puede utilizar de las siguientes formas:
  - *Segmentación de la fuerza laboral*. Aquí se utiliza para agrupar a los empleados en función de sus características similares, lo que permite identificar segmentos de la fuerza laboral con necesidades y características comunes.
  - *Detección de patrones ocultos en encuestas de empleados*. Al aplicar técnicas de aprendizaje no supervisado a las respuestas de encuestas de empleados, se pueden descubrir temas y patrones que no estaban predefinidos.

- *Aprendizaje reforzado*. Aquí el agente aprende a través de la interacción con un entorno. Aunque no se utiliza tan comúnmente como los anteriores, se puede aplicar de la siguiente manera en *People Analytics*:
  - *Optimización de turnos de trabajo*. Un sistema de aprendizaje reforzado puede aprender a programar turnos de trabajo para maximizar la satisfacción de los empleados y la eficiencia operativa, basándose en la retroalimentación del personal y la observación del entorno de trabajo.
  - *Recomendaciones de desarrollo profesional*. Se puede ayudar a recomendar rutas de desarrollo profesional para los empleados a medida que interactúan con el sistema, considerando sus preferencias y objetivos a lo largo del tiempo.

Como se ha visto, se tienen un gran número de aplicaciones en *People Analytics* de los modelos basados en datos. Veamos ahora uno de los grandes inconvenientes en este tipo de modelos: el sobreajuste. Veamos su explicación mediante un ejemplo sencillo. Supongamos que tenemos un conjunto de datos históricos de rotación de empleados y queremos construir un modelo de aprendizaje automático para predecir qué empleados son propensos a dejar la organización en el futuro. El objetivo es identificar patrones y factores que puedan estar relacionados con la rotación. El modelo puede tener en cuenta variables como la antigüedad, el salario, la satisfacción laboral, la ubicación, etc. Sin embargo, aquí es donde entra el problema del sobreajuste. Si el modelo es muy complejo y se ajusta demasiado a los detalles de los datos que usemos para su ajuste, puede aprender propiedades específicas de esos datos y no será capaz de generalizar, esto es, dar una salida adecuada ante datos que no se le han presentado previamente. En otras palabras, el modelo memoriza los datos usados para su ajuste, pero no aprende las características intrínsecas del problema ni las relaciones relevantes. En nuestro ejemplo esto significa que el modelo puede ser excelente para predecir la rotación de los empleados en función de los datos históricos, pero podría ser ineficaz cuando se trata de nuevos empleados o situaciones que no se parecen exactamente a lo que has visto en el pasado. Como se ha comentado, el problema consiste en que el modelo memoriza los datos que se tienen, pero, ante datos nuevos, no es capaz de dar una salida adecuada, esto es, no es capaz de generalizar. Una estrategia que se podría seguir sería no usar todos los datos para ajustar el modelo dejando una parte de ellos para comprobar la capacidad de generalización del modelo. En esta primera aproximación se divide el conjunto de datos en dos partes; un conjunto de datos para ajustar el modelo, conjunto que se conoce como conjunto de entrenamiento; el resto se conoce como conjunto de *test*. La proporción suele ser de 70%-90% para entrenamiento dependiendo de la cantidad de datos que se tengan. Como regla general para este caso, y lo que viene, el conjunto de *test* nunca se utiliza para nada; se realiza la separación y sólo vuelve a aparecer en el proceso de comprobación del modelo, este conjunto de test definirá el comportamiento del modelo ante datos no observados, definiendo su capacidad de generalización. Una versión mejorada de esta estrategia es dividir el conjunto de entrenamiento en dos, uno que sigue siendo el conjunto de datos para ajustar el modelo (conjunto de entrenamiento) y el nuevo que sería el conjunto de validación (un 80%- 20% suele ser la proporción más típica entre estos dos nuevos conjuntos). Este conjunto, validación, controla el ajuste del conjunto de entrenamiento intentando evitar su sobreajuste; de forma intuitiva daría una estimación del error de test que puede cometer el modelo. Veamos la necesidad de esos conjuntos, entrenamiento y validación. Supongamos que tenemos un conjunto de datos con dos variables: tiempo y una variable que depende de dicho tiempo. Queremos modelizar la variable dependiente mediante un modelo, para lo que dividimos el conjunto de datos en 3

subconjuntos; entrenamiento, validación y *test*. Para entender mejor el proceso de ajuste del modelo, en la siguiente figura se representan los dos primeros conjuntos:

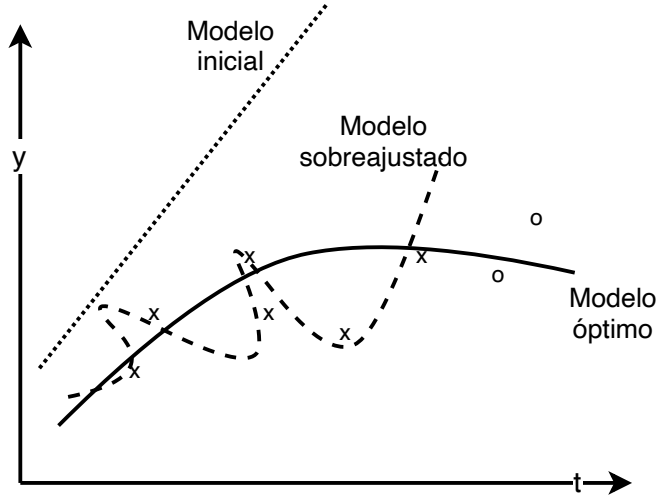


Figura 1.4. Datos de entrenamiento (x) y validación (o). Se muestran 3 modelos, el modelo de partida, el modelo óptimo y el modelo que presenta sobreajuste.

Si se observa la figura anterior podemos inferir el proceso de ajuste (conocido como proceso de aprendizaje) de los modelos. Inicialmente, en la mayoría de los casos, se parte de un modelo aleatorio que, como se observa de la figura 1.4 nada tiene que ver con los datos de entrenamiento y validación. Posteriormente se entra en el proceso de ajuste donde se intenta que el modelo se "acerque" a los datos; esto se consigue mediante lo que se conoce como *función de pérdidas (loss function)* que mide lo bien/mal que se ajusta el modelo a los datos. Conforme el modelo se va ajustando llega un momento que se ajusta de una forma más o menos buena a los datos de entrenamiento y da una respuesta más o menos precisa a los datos de validación (que, recordemos, no usa en el proceso de ajuste). Esta situación se correspondería con el *modelo óptimo* de la figura 1.4. A partir de este momento el modelo empieza a memorizar el conjunto de entrenamiento dando muy bajos valores de la función de pérdidas para este conjunto, pero, sin embargo, dando valores altos de dicha función para el conjunto de validación. De forma intuitiva el modelo se "acerca" mucho a los datos de entrenamiento, pero se "aleja" de los datos de validación. Se tendría el *modelo sobreajustado* de la figura 1.4. Evitar el sobreajuste es clave para llegar a un modelo útil; actualmente todos los modelos son extremadamente flexibles permitiendo ajustarse casi a cualquier tipo de datos por lo que son muy propensos al sobreajuste. Existe otra estrategia cuando se tienen pocos datos y se necesitan todos

para ajustar el modelo; que suele ser lo usual en problemas de *People Analytics* (¡*no siempre se tienen tantos datos como uno querría*!). Un paso más allá de la estrategia definida anteriormente es considerar lo que se conoce como *k-fold cross validation*. En este caso el conjunto de entrenamiento/validación (hay que recordar que siempre hay una primera división entre entrenamiento/ validación y *test*) se divide de acuerdo con el siguiente esquema.

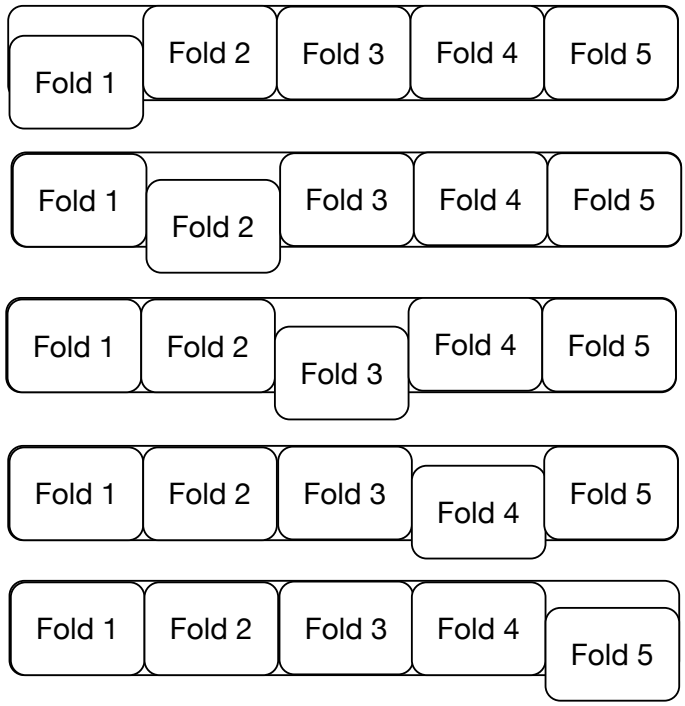


Figura 1.5. División del conjunto de datos según el procedimiento de k-fold cross validation (en este caso 5-fold). El subconjunto que aparece como separado es el que se usa como validación entrenándose el modelo con el resto.

En este caso se plantea dividir el conjunto entrenamiento/validación en un conjunto de subconjuntos (*folds*) de tal manera que un subconjunto se utiliza para validar y el resto para entrenar. Si se divide en *k-folds* significa que obtendremos k valores para el conjunto de entrenamiento para la función que queremos optimizar y k valores para el conjunto de validación. Tras esos cálculos se promedian esos valores obteniendo una medida única para cada uno de los dos procesos, entrenamiento y validación.

Los procedimientos comentados se suelen repetir varias veces para aumentar la fiabilidad de las medidas de funcionamiento obtenidas. En cada una de estas repeticiones la segmentación en los diferentes conjuntos se realiza de forma aleatoria lo que aumenta la robustez de las medidas obtenidas.

Sobre las divisiones de los conjuntos hay que tener en cuenta que el conjunto de *test* no se puede tocar en ningún momento. Si queremos aplicar una transformación a los datos para desarrollar el modelo, esta transformación será la misma para todos los datos (incluidos los de *test*) pero no se usará *en ningún momento*, ni ninguna información que se derive del conjunto de test. Lo comentado se aplica a la división entre entrenamiento/validación; la transformación realizada no puede usar información del conjunto de validación (o del *k-fold* correspondiente). De todas maneras, las implementaciones en los diferentes lenguajes de estos métodos hacen estas divisiones de forma automática y no suele ser necesario el preocuparse de la introducción de información del conjunto de test/validación en el conjunto de entrenamiento (este problema se conoce como fuga de información*, leakage information*).

En cuanto a los tipos de modelos que pueden ser utilizados en cada uno de los capítulos, serán comentados con detalle según se introduzca el problema a resolver. Por adelantar en las siguientes figuras se muestran los principales modelos usados en aprendizaje supervisado y no supervisado (figuras 1.6 y 1.7):

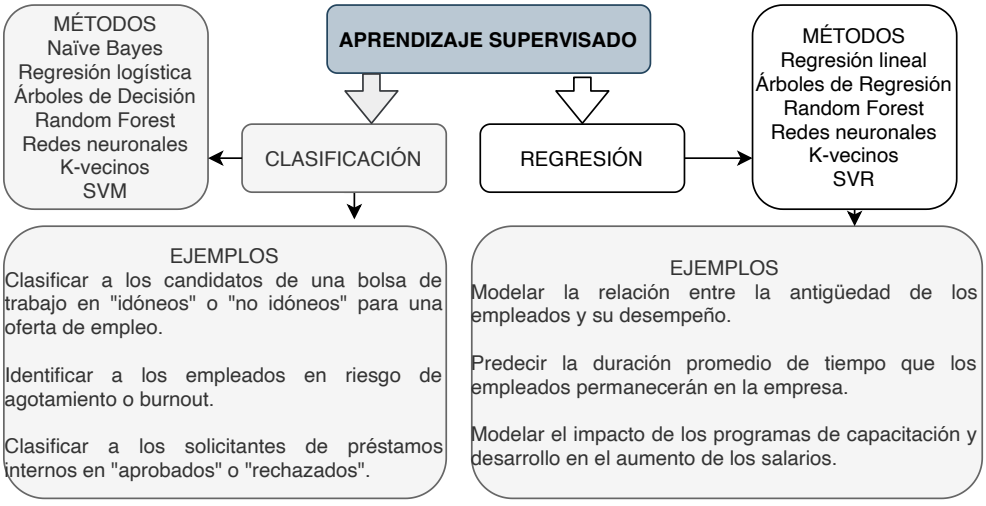


Figura 1.6. Métodos y ejemplos de aplicaciones del aprendizaje supervisado en People Analytics separado por clasificación/regresión.