ManagementSolutions
Making things happen

Auto Machine Learning, hacia la automatización de los modelos

I+D

www.managementsolutions.com

Diseño y Maquetación
Dpto. Marketing y Comunicación
Management Solutions - España

Fotografías
Archivo fotográfico de Management Solutions
iStock

Índice

Introducción

"Media docena de monos, provistos de máquinas de escribir, producirán en unas cuantas eternidades todos los libros que contiene el British Museum"

– Jorge Luis Borges[1]

Un modelo matemático es, en cierto modo, una simplificación de la realidad que aprovecha la información disponible para facilitar la toma de decisiones. Esta simplificación permite evaluar hipótesis sobre el comportamiento tanto de variables como de sistemas a través de su representación resumida bajo un conjunto de postulados, habitualmente basada en datos y aplicando criterios de inferencia. Tiene como principal fin explicar, analizar o predecir el comportamiento de una variable.

La revolución de las técnicas de modelización, junto con el incremento de la potencia de computación y la mayor accesibilidad y aumento de la capacidad de almacenamiento de datos, ha cambiado de manera radical la manera de construir modelos en los últimos años. Esta revolución ha sido un factor clave que ha estimulado no solo el uso de estas nuevas técnicas en los procesos de toma de decisiones donde clásicamente se usaban enfoques tradicionales, sino también en ámbitos donde no era tan habitual el uso de modelos. Por último, en algunos sectores, como por ejemplo en el sector financiero, el uso de modelos ha sido impulsado además por la regulación. Normas como las NIIFs 9 y 13 o Basilea II han promovido el uso de modelos internos con el objetivo de añadir sensibilidad y mejorar la sofisticación del cálculo del deterioro contable o de la determinación de los riesgos financieros.

Aunque pudiera parecer lo contrario, las técnicas de modelización más comúnmente aplicadas en el ámbito empresarial no tienen un origen reciente. En concreto, las regresiones lineales y logísticas datan del siglo XIX. No obstante, de un tiempo a esta parte se ha producido un significativo desarrollo de nuevos algoritmos, que tiene como objetivo sofisticar la forma en la que se encuentran patrones en los datos, pero también introduce nuevos retos como la mejora de las técnicas de interpretabilidad. La aplicación de estos nuevos modelos matemáticos a la computación es una disciplina científica conocida como aprendizaje automático o *machine learning*, ya que permite que los sistemas puedan aprender y encontrar patrones sin ser explícitamente programados para ello.

Existen múltiples definiciones del aprendizaje automático. Entre ellas, dos de las más ilustrativas son las de Samuel y Mitchell. Para Arthur Samuel[2], el aprendizaje automático es "el campo de estudio que da a las computadoras la habilidad de aprender sin ser programadas explícitamente", mientras que para Tom Mitchell[3] se define como "un programa que aprende de la experiencia E con respecto a alguna clase de tareas T y en función de una medida de rendimiento P, si este rendimiento en las tareas en T, según la medida de P, mejora con la experiencia E". Estas dos definiciones se relacionan habitualmente con el aprendizaje no supervisado y el aprendizaje supervisado respectivamente[4].

Como consecuencia de todo ello, el apetito para comprender adecuadamente y extraer conclusiones de los datos se ha incrementado drásticamente. Pero, de manera paralela, la implantación de estos métodos ha requerido modificar múltiples aspectos en las organizaciones[5], y es, a su vez, foco de posibles riesgos derivados de errores en su desarrollo o implementación, o de su uso inadecuado.

La modelización avanzada permite mejorar los procesos comerciales y operativos, o incluso facilita la aparición de nuevos modelos de negocio. Un ejemplo puede encontrarse en el sector financiero, donde las nuevas metodologías, en el contexto de la digitalización, están modificando la propuesta de valor actual, pero también añadiendo nuevos servicios. Según una encuesta realizada por el Bank of England y la Financial Conduct Authority sobre casi 300 empresas del sector financiero y asegurador, dos tercios de los participantes utilizan *machine learning* en sus procesos[6]. Las técnicas de *machine learning* se utilizan con frecuencia en tareas típicas de control, como la prevención del blanqueo de capitales (AML, por sus siglas en inglés), el análisis de las amenazas relacionadas con ciberseguridad o la detección de fraude, así como en los procesos de negocio, tales como la clasificación de clientes, los sistemas de recomendación o la atención a clientes a través del uso de *chatbots*. También son utilizadas en la gestión del riesgo de crédito, en *pricing*, en la ejecución de operaciones o en la suscripción de seguros.

[1] Jorge Luis Borges, "La biblioteca total" (1939). Escritor, poeta, ensayista y traductor argentino, autor, entre otros, de Ficciones y El Aleph.
[2] Samuel, 1959.
[3] Mitchell, 1997.
[4] Management Solutions, 2018.
[5] Ibídem.
[6] Bank of England, 2019.

En otros sectores se puede observar un grado de desarrollo similar. El uso de modelos de *machine learning* es habitual en sectores como el manufacturero, el transporte, la medicina, la justicia o los sectores de *retail* y gran consumo. Esto ha hecho que la inversión en empresas dedicadas a la inteligencia artificial aumentase de 1.300 millones de dólares en 2010 a 40.400 millones en 2018 a nivel global[7] (véase figura 1). El retorno esperado justifica esta inversión: el 63% de las empresas que han adoptado el uso de modelos de *machine learning* en sus unidades de negocio informan de un aumento de los ingresos, siendo de más del 6% para aproximadamente la mitad de ellas. Asimismo, el 44% de las empresas reportan un ahorro de costes, siendo superior al 10% para aproximadamente la mitad de ellas[8].

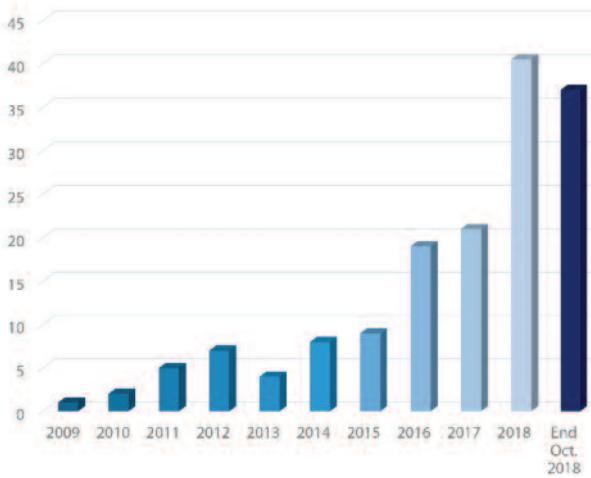
De los distintos cambios que se han registrado en las organizaciones para adaptarse a este nuevo paradigma, la captación y retención del talento está siendo uno de los elementos centrales. En primera instancia, se ha requerido un aumento de los equipos especialistas en *machine learning*. La demanda de profesionales en este ámbito ha aumentado un 728% entre 2010 y 2019 en Estados Unidos[10] (véase figura 2), registrándose también un cambio cualitativo en la demanda de habilidades y conocimientos de los *data scientists*.

Pero esta demanda no es genérica: con la intención de explotar la cada vez mayor cantidad de datos disponibles mediante herramientas cada vez más sofisticadas, los requisitos se han vuelto más específicos (entre los que se incluye el conocimiento de diferentes lenguajes de programación, como Python, R, Scala o Ruby, capacidad para el tratamiento de bases de datos en arquitecturas *big data,* conocimientos en *cloud computing,* conocimientos matemáticos y estadísticos avanzados, encontrarse en posesión de posgrados especializados, etc.), existiendo una gran diversidad de posiciones, con requisitos muy concretos y, por tanto, difícil de cubrir. Además, el gran aumento registrado en el volumen de generación de datos por parte de las empresas hace que, incluso con una oferta estable de *data scientists,* la solución actual de contratación de recursos no sea escalable.

Esta inclusión de modelos de *machine learning* no solo hace necesario el establecimiento de equipos especialistas, sino también el uso de nuevos procedimientos de desarrollo, y la revisión de los métodos de validación, revisión y evaluación de los modelos dentro de los ámbitos de validación y auditoría, así como un importante cambio cultural en el resto de áreas para conseguir una implementación efectiva. La inclusión de estos nuevos procesos genera una reacción en cadena que afecta a todo el ciclo de vida de los modelos, destacando entre ellos la

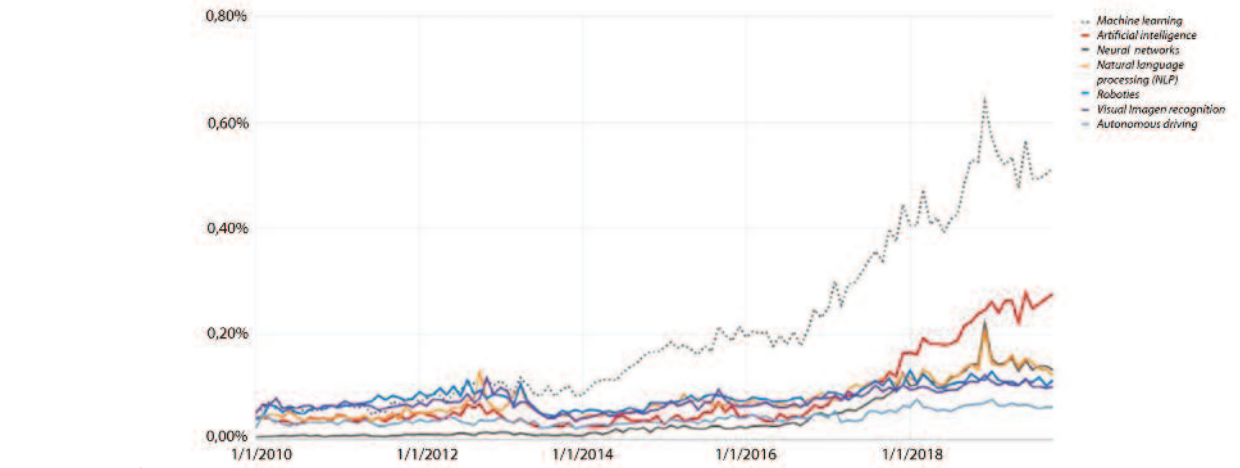
[7] Incluye únicamente inversiones por importe superior a 400.000 dólares, Stanford University, 2019.
[8] Statista, 2019.
[9] Stanford University, 2019.
[10] Ibídem.

Figura 1: inversión anual (miles de millones de dólares) en empresas de IA[9].



Fuente: Stanford University, 2019.

Figura 2: aumento de la demanda de perfiles con conocimiento en machine learning e inteligencia artificial.



Fuente: Burning Glass, 2019.

identificación y gestión del riesgo de modelo, así como su gobierno[11]. Muchos de estos modelos requieren, adicionalmente, la aprobación de organismos supervisores, como ocurre en la industria financiera (por ejemplo, en los modelos de capital o provisiones), o en la industria farmacéutica, lo que añade retos adicionales a los ya existentes, como son la necesidad de garantizar la interpretabilidad de los modelos utilizados, así como desarrollar el resto de elementos de confianza de los modelos.

Otro aspecto destacable sobre la inversión en métodos de *machine learning* es que esta tiene un desarrollo desigual: la obligación de superar procesos de validación, auditoría y aprobación, según la regulación establecida, o el requisito de mantener unos estándares específicos de documentación, está generando diferencias en la implantación de modelos internos de las empresas. De acuerdo con el informe sobre *big data* y *analytics* de la EBA[12], las entidades financieras están adoptando programas de transformación digital o impulsando el uso de técnicas de *machine learning* en aspectos como la mitigación de los riesgos (incluyendo *scorings* automáticos, gestión de riesgo operacional o fraude) y en procesos de *Know Your Customer*. No obstante, "aunque la aplicación de *machine learning* puede suponer una oportunidad para optimizar el capital, desde la perspectiva de un marco prudencial es prematuro considerar apropiado el uso de técnicas de *machine learning* para determinar los requerimientos de capital"[13].

[11] Management Solutions, 2014.
[12] European Banking Authority, 2020.
[13] Ibídem.

Existen además riesgos operacionales difíciles de detectar, como pueden ser los de carácter humano durante el proceso de implementación de un modelo o los relacionados con la seguridad del almacenamiento de datos, que deben ser gestionados convenientemente para garantizar el uso de estos sistemas en un entorno adecuado. Esto adquiere una relevancia significativa para las empresas que operan en entornos considerados de alto riesgo. Un ejemplo de ello es el marco que ha establecido la Comisión Europea en estos casos y que engloba distintos aspectos del proceso de modelización[14]. Por último, y también debido tanto a criterios regulatorios como de gestión, los modelos deben funcionar de forma fiable y ser utilizados de forma ética, de modo que el usuario pueda confiar en ellos para su uso en los procesos de toma de decisiones. En esta línea, resulta de especial interés la propuesta de la EBA basada en siete pilares de confianza[15]: ética, interpretabilidad, eliminación de discriminaciones, trazabilidad, protección y calidad de los datos, seguridad y protección al consumidor. Estas cuestiones se han identificado como elementos clave también desde los ámbitos universitario y empresarial[16].

En este contexto las tareas de desarrollo de modelos demandan dedicaciones muy desiguales: las tareas previas y complementarias al análisis requieren de una gran cantidad de tiempo y recursos dirigidas a la preparación, limpieza y tratamiento general de los datos; el 60% del tiempo de un *data scientist* se dedica a la limpieza de datos y a organizar la información, mientras que un 9% y un 4% se focaliza en tareas de *knowledge discovery* y el refinamiento de algoritmos, respectivamente[17]. Todo ello impulsa la necesidad de cambiar la forma de abordar el desarrollo, la validación y la implementación de modelos, de forma que se aprovechen las ventajas de las nuevas técnicas, pero resolviendo las dificultades asociadas a su utilización, así como mitigando sus posibles riesgos.

Derivado de los motivos comentados anteriormente, existe una tendencia clara hacia la automatización de los procesos relacionados con la aplicación de técnicas de *advanced analytics*, que se ha denominado, de manera general, *machine learning* automatizado (AutoML o *automated machine learning,* indistintamente), cuyo objetivo es no solo automatizar aquellas tareas donde los procesos heurísticos son limitados y fácilmente automatizables, sino también permitir la generación de procesos de búsqueda de patrones y de algoritmos más automática, ordenada y trazable. De acuerdo con Gartner[18], más del 50% de las tareas *data science* estarán automatizadas en el año 2025.

Pero además, esta tendencia a la automatización ofrece una serie de oportunidades, como las que brinda la arquitectura de sistemas utilizados en la automatización en términos de diseño de los *workflows*, de inventario de modelos o de validación por componentes. Los sistemas de AutoML integran diversas herramientas para desarrollar modelos, reduciendo además el coste, el tiempo de desarrollo y los errores en la implementación de dichos sistemas.

[14] European Comission, 2020.
[15] European Banking Authority, 2020.
[16] Por ejemplo, la cátedra iDanae, fruto de la colaboración entre la Universidad Politécnica de Madrid y Management Solutions, ha publicado newsletters sobre interpretabilidad (Cátedra iDanae, 3T-2019) y ética en la inteligencia artificial (Cátedra iDanae, 4T-2019).
[17] Según una encuesta realizada por la plataforma en Inteligencia Artificial CrowdFlower (CrowdFlower, 2017).
[18] Gartner, 2019.

Tanto los sistemas como los métodos de AutoML persiguen, entre otras cosas:

- Reducir el tiempo dedicado por los *data scientists* en el desarrollo de modelos a través de técnicas de *machine learning*, e incluso permitir el desarrollo de algoritmos de *machine learning* por parte de equipos no especializados en *data science*.
- Mejorar el desempeño de los modelos desarrollados, así como la trazabilidad y comparabilidad de los modelos obtenidos frente a las técnicas de búsqueda de hiperparámetros manual.
- Permitir cuestionarse los modelos desarrollados mediante otros enfoques.
- Reaprovechar la inversión realizada tanto en tiempo como en recursos para el desarrollo de códigos y mejorar y refinar los componentes incluidos en los sistemas de forma eficiente y con mayor trazabilidad.
- Simplificar la validación de los modelos y facilitar su planificación.

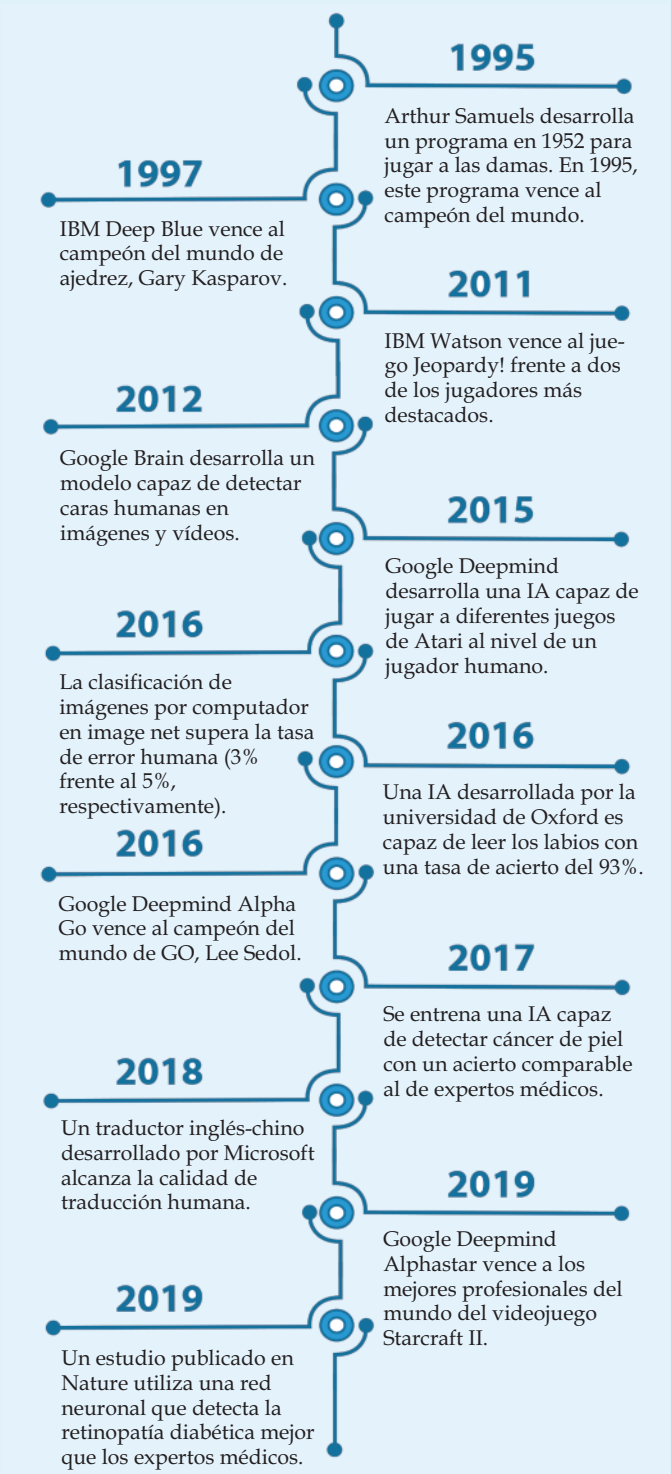
En este contexto, el presente documento pretende describir los principales elementos sobre los sistemas de AutoML. Para ello, se ha estructurado en tres apartados, que se corresponden a su vez con tres objetivos:

- En el primer apartado se analiza la evolución en la automatización de los procesos de *machine learning*, así como los motivos que subyacen en el desarrollo de sistemas de AutoML, tanto a través de la componentización, como de la automatización de los mismos.

- En el segundo bloque se aporta una visión descriptiva sobre los principales marcos de AutoML, y se explica qué enfoques se están siguiendo, tanto en el ámbito académico como en experiencias prácticas dirigidas a automatizar los procesos de modelización a través de técnicas de *machine learning*.

- Por último, el tercer apartado tiene como objetivo ilustrar los resultados del desarrollo de sistemas de AutoML, presentando como caso de estudio una competición organizada por Management Solutions a inicios de 2020 dirigida a los profesionales de la firma y cuyo objetivo fue el diseño de un modelo de Automated Machine Learning.

Principales hitos en el desarrollo de ML



Reporte sobre *big data* y Advanced Analytics de la EBA

La Autoridad Bancaria Europea (EBA) ha publicado un informe sobre *big data* y *advanced analytics,* con el objetivo de dar a conocer su uso en el sector financiero europeo, así como aportar su entendimiento sobre (i) la identificación de cuatro pilares clave para su desarrollo, implementación y adopción, (ii) los principales elementos de confianza sobre los que se ha de basar un marco de *big data* y *advanced analytics,* y (iii) señalar principales observaciones, oportunidades y riesgos que se derivan de la aplicación de estas soluciones.

I. Pilares clave de un marco de *big data* y *advanced analytics*

II. Elementos de confianza

III. Principales observaciones, oportunidades y riesgos en la utilización

Pilares clave de un marco de *big data* y *advanced analytics*

GESTIÓN DE DATOS

- Tipos de datos y fuentes: datos estructurados, desestructurados y semi-estructurados.
- Seguridad y protección de los datos: protege la confidencialidad, integridad y disponibilidad de los datos.
- Calidad de los datos: asegura la calidad de los datos durante todo el ciclo de *Big Data* y *advanced analytics* (BD&AA).

METODOLOGÍA DE ANÁLISIS

- Preparación de los datos: incluye la ingeniería de datos y el análisis de importancia.
- Analítica: entrenamiento de modelos calibrado, validación y selección, así como prueba y lanzamiento.
- Operaciones: control y actualización de modelos.

INFRAESTRUCTURA TECNOLÓGICA

- Infraestructura: enlaza recursos para transmitir *big data.*
- Plataforma de los datos: gestiona todos los datos utilizados por un sistema de análisis avanzado.
- Procesamiento: proporciona el software necesario para sostener la implementación de aplicaciones de analítica avanzada.

ORGANIZACIÓN Y GOBIERNO

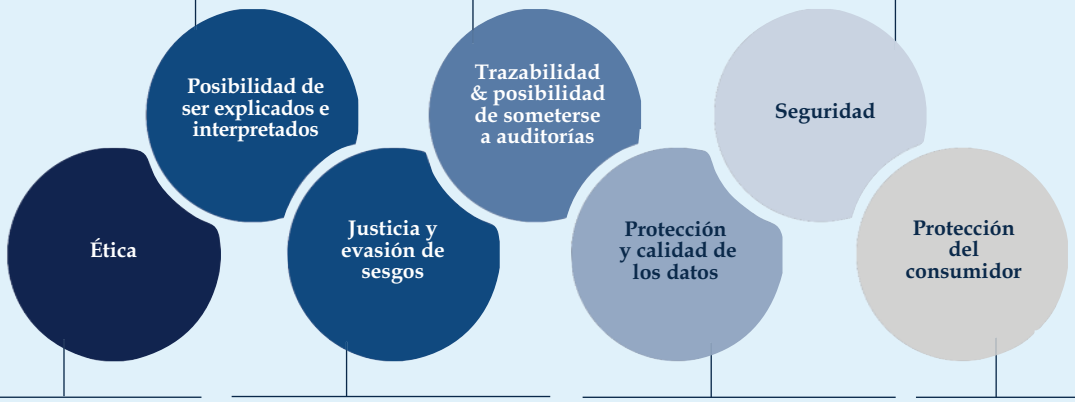
- Estructura y medidas internas de gobierno: estructura de gobierno, estrategia y gestión del riesgo, transparencia, desarrollo externo y subcontratación.
- Habilidades y conocimiento: nivel de comprensión del cuerpo de dirección y la alta dirección, segunda y tercera líneas, desarrolladores de sistemas avanzados de análisis.

Elementos de confianza

Una parte fundamental para construir modelos confiables es que sean explicables e interpretables, transparentes, comprendidos correctamente y que aporten justificaciones claras.

El uso de soluciones que puedan ser trazables ayuda a seguir todos los pasos, criterios y decisiones del proceso, lo que permite la repetición de procesos que resultan en las decisiones tomadas por el modelo y ayuda a la auditoría del sistema.

Es importante mantener una vigilancia técnica de los últimos ataques de seguridad y las técnicas de defensa relacionadas, para asegurar que el gobierno, la supervisión y la infraestructura técnica se encuentran correctamente instaladas para una efectiva gestión del riesgo ICT *(Information and communications technology).*



El desarrollo, lanzamiento y uso de cualquier solución de inteligencia artificial debe adherirse a unos principios éticos fundamentales.

La justicia e imparcialidad requiere que el modelo asegure la protección de determinados grupos contra la discriminación, y para garantizarlo el modelo debe estar libre de sesgos.

Los datos deberían ser correctamente protegidos con un sistema de BD&AA confiable, y su calidad debe tenerse en cuenta durante todo el proceso.

Un sistema de BD&AA confiable debe respetar los derechos de los consumidores y proteger sus intereses.

Principales observaciones, oportunidades y riesgos en la utilización

Observaciones clave

- Las instituciones están en distintas etapas del desarrollo de BD&AA. Algunos de los usos más comunes son la detección del fraude, el CRM y la automatización de procesos.
- Existe una mayor dependencia en datos internos, en vez de en datos externos o redes sociales. Se incorporan soluciones de código abierto. Hay un uso limitado de algoritmos complejos.
- Hay diferentes niveles de integración y gobiernos de la analítica avanzada en las instituciones.
- Se observa una dependencia creciente en compañías tecnológicas para la provisión de servicios de infraestructura y computación en la nube.

Oportunidades clave

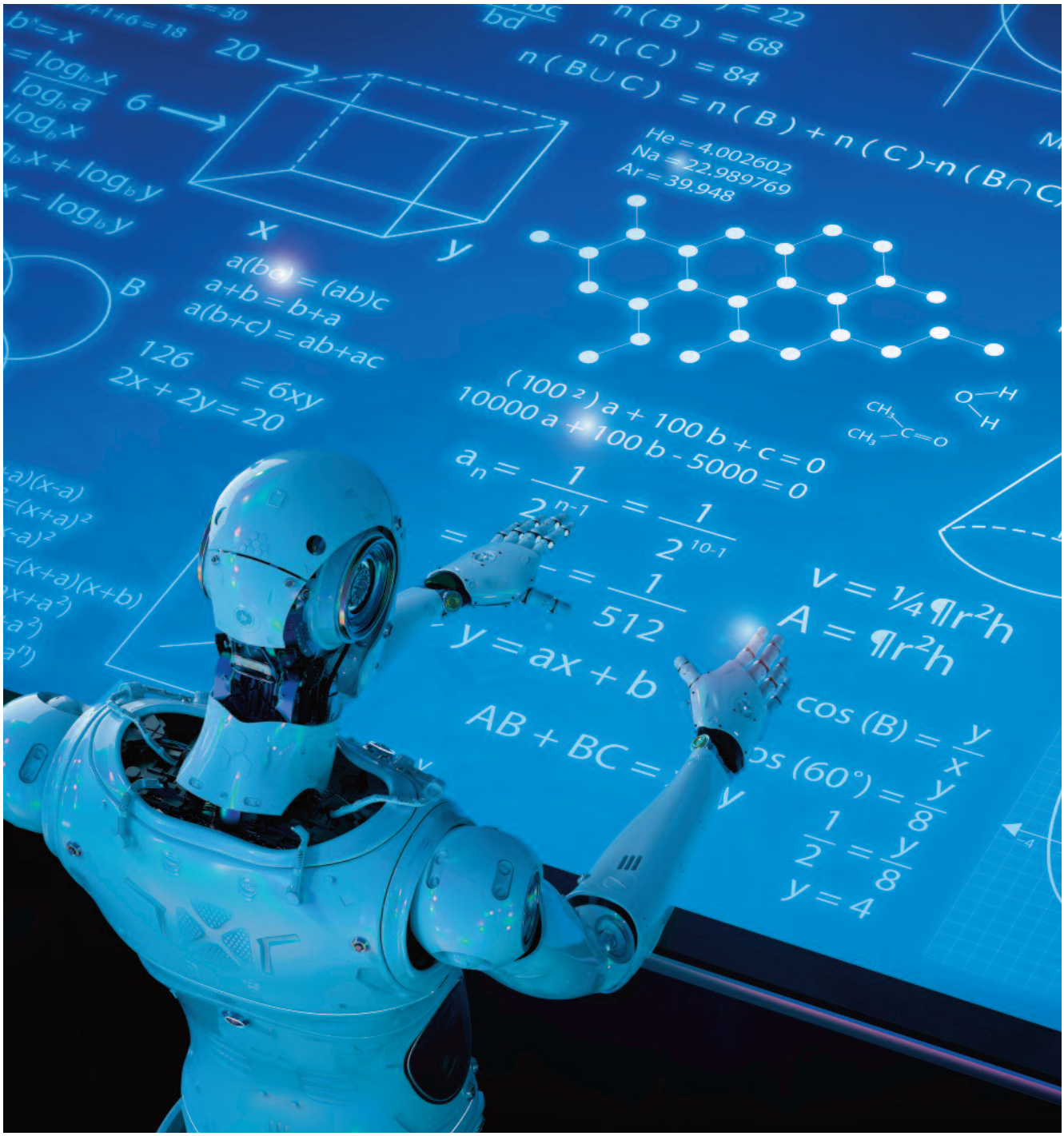
- Los clientes de los servicios financieros de los sectores de comercio minorista y de ocio esperan un servicio más personalizado. Hay confianza en el sector financiero respecto al cumplimento del RGPD.
- El aumento de la satisfacción del consumidor y el uso de información para mejorar la oferta, reducir la pérdida de clientes, optimizar procesos, y ayudar en la mitigación del riesgo y la detección del fraude.
- Hay muchos usos y oportunidades posibles que surgen del uso de modelos interpretables.

Riesgos clave y guía propuesta

- El resultado de los modelos puede ser complejo, no determinista y correcto de acuerdo con una medida de probabilidad, lo que podría dañar a la institución o a sus clientes. Debe asegurarse que los resultados de estos sistemas no violan los estándares éticos de las instituciones. Además, un ser humano debe involucrarse en el ciclo de decisión, por lo que es necesario realizar entrenamiento del personal.
- La implementación de un marco metodológico y de gobierno de BD&AA podría promover su uso responsable, lo que debería incluir la documentación adecuada, una justificación suficiente y otras técnicas explicativas y de seguimiento, incluyendo el uso de soluciones trazables. La explicación debería basarse en una aproximación basada en el riesgo.
- Existe la necesidad de que los modelos sean precisos y de realizar controles regulares.
- El uso de soluciones de *machine learning* podría dar lugar a riesgos ICT: seguridad de los datos, seguridad del modelo, calidad de los datos, gestión del cambio y continuidad y resistencia del negocio.
- Como consecuencia de la dependencia en marcos de código abierto, o en herramientas y sistemas desarrollados por terceras partes, tanto sus potenciales riesgos (falta de control y conocimiento de la tercera parte, alta dependencia de un proveedor, riesgo de concentración mantenimiento de un modelo, etc.) como la responsabilidad que debe siempre mantenerse en la entidad, deben evaluarse.
- Por último, la importancia de la calidad de los datos, la protección y la seguridad es enfatizada, tanto para propósitos regularorios, (incluyendo el cumplimiento de RGPD) como para asegurar la idoneidad del modelo.

Resumen ejecutivo

"Bastaría, en rigor, con un solo mono inmortal"
Jorge Luis Borges



El contexto de la automatización de los modelos de machine learning

1. La incorporación de las técnicas de *big data* y *advanced analytics* en la economía está cambiando la manera en que se usa la información. Partiendo de la combinación de distintos conocimientos relativos a la explotación de datos y de negocio, la capacidad de análisis se ha incrementado radicalmente, aunque, al mismo tiempo, es foco de posibles riesgos derivados de errores en su desarrollo o implementación, su uso inadecuado o el exceso de confianza en su aplicación.

2. Para poder aprovechar el potencial de estas nuevas técnicas, las entidades están transformando su forma de trabajar. Estos cambios afectan directamente al desarrollo y validación de los modelos, pero también a otros procesos como son los relativos a las estructuras tecnológicas, la selección, formación y retención de perfiles especialistas o, de manera más amplia, a la cultura de trabajo.

3. Existen también riesgos operacionales difíciles de detectar y, en algunos casos, regulación relativa al uso, calidad y procesos relacionados con datos, así como con la aplicación o la interpretabilidad de modelos que generan la necesidad de contar con procesos trazables, y validaciones y revisiones recurrentes de los modelos.

4. En este contexto, se observa una tendencia clara hacia la automatización de los procesos relacionados con la aplicación de técnicas de *advanced analytics*, cuyo objetivo es no solo automatizar aquellas tareas donde los procesos heurísticos son limitados y fácilmente automatizables, sino también permitir la generación más automática, ordenada y trazable de modelos.

5. Con todo ello, se favorece la reducción del tiempo dedicado a tareas complementarias y repetitivas; el acceso a estas técnicas por parte de equipos no especialistas; el desempeño, la trazabilidad y la comparabilidad de los modelos; el reaprovechamiento de los desarrollos de códigos en proyectos específicos, la mejora y el refinamiento de las técnicas; e incluso la mejora de los procesos de validación, incluyendo la generación de modelos *challenger*.

Hacia la automatización de la modelización

6. En el proceso de automatización del flujo de trabajo (*workflow*) de modelización existen múltiples retos. Entre ellos, la necesidad de contar con procesos que garanticen la adecuación de la carga de datos y aquellos relativos al desarrollo, validación, implementación y seguimiento de los modelos. Se debe garantizar también la trazabilidad de los procesos de construcción, así como su interpretabilidad y un adecuado gobierno que permita su integración en la gestión, además de cumplir con la regulación existente.

7. Con relación al proceso de tratamiento de datos, el 60% del tiempo de un *data scientist* se dedica a limpieza y organización de la información, existiendo mucho recorrido en términos de automatización de estos procesos.

8. Sobre el *workflow* de modelización, es posible automatizar este proceso mediante dos opciones, que usualmente se combinan: la componentización de los distintos procesos en elementos segregados, y la ejecución automática de estos componentes, sistematizándolos a través de reglas preestablecidas y técnicas estadísticas.

9. La componentización se basa en la separación de las tareas de modelización en distintas piezas, y su programación y desarrollo de forma independiente. Cada uno de estos componentes recibe un input determinado y ejecuta una tarea específica.

10. Las ventajas de la componentización son la estandarización de los procesos, el aumento de la calidad y de la eficiencia, la especialización en el desarrollo, la mejora de la usabilidad, y el impulso de la escalabilidad.

11. Por otra parte, la automatización del proceso de construcción de modelos se fundamenta en el uso de criterios automáticos para seleccionar sus atributos, de forma que el procedimiento puede replicarse y auditarse. También garantiza que la selección final se ha hecho mediante un proceso que garantice alcanzar un poder predictivo óptimo dadas unas restricciones.

12. Las ventajas de la automatización en la búsqueda son la optimización del proceso de selección de hiperparámetros, la generalización de los problemas de modelización, la adaptación de los espacios de búsqueda de parámetros a cada problema, y la posibilidad de experimentación fuera de los rangos habituales.

Marcos de automatización de procesos de machine learning

13. En la práctica, la forma de automatizar estos procesos se ha basado en (i) sistematizar la mayor parte de los aspectos relativos al análisis y tratamiento previo de los datos, incluyendo la transformación de variables y su preselección, (ii) generar un espacio de búsqueda de posibles modelos y parámetros, así como un proceso de desarrollo y selección de modelos que evite tanto el *overfitting*[19] como el *underfitting*[20], y (iii) automatizar la aplicación de técnicas de interpretabilidad.

14. Existe una gran variedad de opciones para poner en producción estos sistemas, recogidos en enfoques basados en modelos (*model-based schemes*) y en procesos basados en datos (*data-driven approaches*).

15. Los procesos de generación y evaluación de modelos se basan fundamentalmente en dos componentes: el optimizador y el evaluador.

16. El optimizador genera y actualiza combinaciones de parámetros dentro de la frontera de posibilidades definida en función del modelo y de los datos utilizados. Posteriormente el evaluador se encarga de medir el desempeño de las opciones propuestas por el optimizador, pudiendo influir en la estrategia de búsqueda en función de los resultados.

17. El optimizador utiliza diversas técnicas para encontrar la mejor configuración. Estas técnicas se pueden clasificar en simples (como *grid search, random search,* algoritmos evolutivos u optimización bayesiana) o basados en experiencia (como *meta-learning* o *transfer learning*).

18. El evaluador se encarga de comprobar si la configuración proporcionada por el optimizador es óptima. Existen también distintos enfoques de optimización como son (i) el *early stop,* en los que el evaluador deja de evaluar si el desempeño es muy bajo en las primeras iteraciones, (ii) la reutilización, basado en el uso de configuraciones utilizadas en entrenamientos anteriores, o (iii) el uso de modelo subrogados en la evaluación.

19. En este tipo de sistemas, los retos existentes son la inclusión de conocimiento previo en él, como puede ser el conocimiento de negocio o criterios expertos, así como el desarrollo de sistemas que abarquen todo el proceso de construcción de modelos.

20. Una alternativa al uso de un sistema basado en la interacción de un optimizador y un evaluador es la búsqueda de arquitecturas neuronales (NAS) por sus siglas en inglés. Esta técnica, utilizada en modelización del lenguaje o clasificación de imágenes, realiza de manera simultánea las tres tareas requeridas para llevar a cabo la automatización: la determinación del espacio de búsqueda, la estrategia de búsqueda en este espacio y la estimación de los modelos obtenidos en cada estimación.

21. En este enfoque, si bien el rendimiento es elevado, resulta más difícil explicar por qué se llega a determinadas configuraciones y si estas sirven para extender su uso a otros tipos de problemas.

22. Actualmente, y pese a tener aún un gran margen de mejora, los sistemas de AutoML han llegado a un grado de desarrollo que puede competir y batir a los expertos humanos en *machine learning,* configurándose como una herramienta fundamental, que puede modificar el tipo de trabajos desarrollados.

[19] Característica de un modelo que se da cuando este se ha ajustado demasiado a la muestra de entrenamiento, de forma que no consigue resultados satisfactorios sobre muestras diferentes a esta.

[20] Característica de un modelo que se da cuando este no se ha ajustado lo suficiente a la muestra de entrenamiento, de forma que no consigue resultados satisfactorios sobre muestras diferentes a esta.

23. Los retos pendientes, tanto para sistemas basados en la optimización de hiperparámetros como para los métodos NAS se refieren a cuestiones relacionadas con la interpretabilidad, la reproducibilidad, así como el reaprovechamiento en las configuraciones de los ejercicios previos y con objeto de facilitar una mejor interacción con el usuario.

Competiciones de AutoML: una herramienta de exploración de enfoques de AutoML

24. Para profundizar en el entendimiento y puesta en práctica de enfoques de AutoML se han diseñado y ejecutado competiciones que confrontan metodologías. Un ejemplo de estas competiciones se expone en el apartado "Competiciones de AutoML: una herramienta de exploración de enfoques de AutoML".

25. En el caso de la competición sobre AutoML llevada a cabo por Management Solutions, los participantes han llevado a cabo enfoques similares a los comentados anteriormente, utilizando enfoques *grid, random search*, algoritmos genéticos o búsquedas bayesianas para llevar a cabo la generación de modelos. Para evaluar la configuración se han utilizado técnicas de *cross* validation.

26. De dicho ejercicio se han podido extraer algunas conclusiones de utilidad: i) el tratamiento de datos ha resultado bastante homogéneo, mostrando que existen diversas técnicas estandarizadas en el sector, ii) la reducción de la dimensionalidad ha sido realizada por la mayoría de los participantes, debido a la gran reducción de coste computacional que ello supone, iii) los sistemas de AutoML mejoran al realizar optimización de todo el pipeline, la utilización de modelos stacking o la paralelización de las tareas en diversos núcleos.

Reflexiones finales

27. Actualmente la configuración de los modelos de *machine learning* depende de manera significativa de apriorismos y ajustes manuales, lo cual puede llevar a subóptimos, como consecuencia tanto de *overfitting* como *underfitting*, en función del tamaño del *dataset* y las técnicas utilizadas. En algunas técnicas sigue siendo habitual el *overfitting* de modelos, lo que parece indicar que existe recorrido de mejora en la generación de sistemas de AutoML en casos específicos.

28. Aunque los enfoques de AutoML han alcanzado un alto grado de desarrollo, se siguen encontrando limitaciones ligadas a que el pipeline no está totalmente automatizado o a la ausencia o escasez de objetivación en algunas de las decisiones, así como en el espacio de búsqueda.

29. Otro reto es permitir el acceso de perfiles no expertos a los entornos de AutoML, para que puedan interactuar directamente con estos métodos y sistemas, de forma que se pueda incorporar la intuición de negocio, o evaluar directamente la interpretabilidad de los modelos. Por último, y respecto a la interpretabilidad, esta sigue siendo una de las cuestiones abiertas en los sistemas de AutoML.

30. En términos de avances recientes, estos son más habituales en la optimización del *feature engineering,* así como en la selección de modelos, en detrimento del tratamiento o la preparación de datos.

31. Por último, se prevé que los sistemas de AutoML se configuren como una herramienta fundamental, que modificará los trabajos desarrollados por los *data scientists,* de forma que estos se centren en los análisis previos o posteriores del desarrollo de modelos, en la generación de componentes y sistemas de AutoML, así como en la resolución de problemas específicos donde un sistema de AutoML no consiga buenos resultados.

Hacia la automatización de la modelización

"Sin embargo, si se sigue jugando cien años, mil años, cien mil años, con toda probabilidad saldrá una vez, por casualidad, un poema. Y si se juega eternamente tendrán que surgir todos los poemas, todas las historias posibles"

Michael Ende[21]

El desarrollo e implantación en la gestión de modelos de *machine learning* genera un conjunto de beneficios, que son consecuencia tanto de la mejora de los procesos de toma de decisiones como de la automatización de tareas en el desarrollo de modelos. Estos beneficios se materializan, por ejemplo, a través de una predicción de la demanda más acertada, en mejoras en la gestión del stock, en las estrategias de *pricing*, en el aumento de la fidelidad de los clientes, o en mejoras en la eficiencia y la reducción de los costes de producción, entre otros. Esto a su vez implica mejores resultados en el desarrollo de productos o en la prestación de servicios, una distribución más eficiente de los recursos, o un mejor posicionamiento en el mercado, pudiendo generar ventajas competitivas frente a competidores que hagan un menor uso de estas técnicas.

Sin embargo, en el proceso de construcción de modelos que presentan también múltiples retos ligados al desarrollo y la implementación de estos nuevos métodos:

- Por un lado, en numerosas ocasiones los modelos de *machine learning* requieren grandes cantidades de datos para evitar el *overfitting*, lo que supone la necesidad de invertir en la obtención, ingesta, almacenamiento y gestión de fuentes de datos y de arquitecturas tecnológicas, con soluciones *in-house* o *cloud* para garantizar la disponibilidad y la calidad de los datos utilizados.

- Por otro, se ha de invertir en el desarrollo de los modelos, su validación, la implementación en los procesos de gestión, y el seguimiento y mantenimiento de los algoritmos.

- Asimismo, se ha de considerar la trazabilidad de los procesos de construcción, así como garantizar la interpretabilidad de los algoritmos y de los resultados obtenidos, ya que las decisiones basadas en algoritmos deberían estar respaldadas por este conocimiento, aunque sea parcialmente.

- Todo lo anterior exige una gobernanza adecuada, que garantice la consideración tanto de los elementos de gestión y éticos en el uso de modelos como de los requerimientos regulatorios. Estos impactos resultan aún mayores para las entidades que operan en sectores regulados, puesto que existen limitaciones en la implantación y uso de estos modelos para determinados fines.

- Por último, para garantizar que se cumple lo anterior, es necesario contar con perfiles especialistas, bien mediante la contratación directa de *data scientists* y desarrolladores, bien mediante la externalización del proceso con compañías especializadas; así como transformar la estructura organizativa y adaptarla en función de las necesidades de desarrollo de modelos, incluyendo nuevas formas de trabajar (por ejemplo, a través de organizaciones *Agile*[22]).

Estos retos han motivado la aparición y desarrollo de sistemas de AutoML, ya que su uso puede dar respuesta a las dificultades planteadas, permite aprovechar los beneficios de la automatización y contribuye a la democratización de los procesos de modelización, facilitando su uso por usuarios no expertos.

Fundamentos de la automatización

El principio de *no free lunch*[23] sugiere que no existe un modelo cuyo desempeño sea siempre mejor que todos los demás, de forma que en función del conjunto de datos analizados, el tipo de modelo que mejor los predice o explica puede ser distinto. Extendiendo esta idea al ámbito de *machine learning*, este principio se puede interpretar como la no existencia de estimadores o combinaciones de configuraciones, hiperparámetros o arquitecturas de redes que sean siempre

[21] Michael Ende, "La historia interminable" (1979). Escritor alemán del siglo XX, conocido principalmente por sus obras de ficción y para niños.
[22] El detalle de la transformación hacia organizaciones *agile* se describe ampliamente en la publicación "De proyectos Agile, a organizaciones Agile", Management Solutions, 2019.
[23] Wolpert & Macready, 1997.

mejores frente a otras alternativas. Si bien a la hora de tratar problemas específicos existen investigaciones académicas que muestran que la selección de valores para hiperparámetros en rangos reducidos pueden garantizar modelos óptimos[24], esta idea no se puede extrapolar a todos los problemas posibles. Esta situación acarrea la necesidad de encontrar métodos para garantizar que un algoritmo de *machine learning* realiza una búsqueda adecuada de las posibles configuraciones para poder maximizar su desempeño.

Dado un problema que se desea resolver mediante técnicas de *machine learning,* la forma de abordarlo se sustenta en establecer las distintas opciones de parámetros y configuraciones que se podrán elegir a lo largo de todo el proceso. Para ello, en primer lugar, es necesario identificar las características de los datos que han de ser tratados, así como las técnicas utilizadas. Posteriormente se debe establecer el enfoque de modelización que se desea usar, así como las métricas con las que se van a seleccionar los modelos y, finalmente, se incorporan las restricciones que puedan existir basadas en el conocimiento del problema (por ejemplo, el signo de las variables). Se configura de esta manera un *workflow* de modelización que servirá para obtener un grupo de modelos ordenados en función de su desempeño. Sobre este flujo es posible realizar una automatización basada en la componentización, es decir, la separación de los distintos procesos de construcción de modelos en componentes que puedan ser ejecutados de manera modular.

Componentización y optimización del workflow de modelización

En el proceso de modelización existe un amplio conjunto de opciones en cada una de las fases que integran el *workflow* de desarrollo, fruto de la combinación de las distintas técnicas usadas en cada apartado. Si bien tanto la selección de los distintos parámetros, como la configuración de las técnicas (cuáles aplicar, en qué orden, sobre qué parte del *dataset*, etc.) varían dependiendo del problema, se pueden desarrollar técnicas para conseguir la automatización de múltiples tareas. Entre los objetivos principales que persigue esta automatización se encuentran reducir el coste y los posibles errores operacionales derivados del desarrollo end-to-end para cada problema de *machine learning*, así como mejorar la eficiencia del proceso de modelización.

Tres de las causas que generan la necesidad de invertir en automatización son las siguientes:

- **Redundancia en el desarrollo:** algunas tareas y funciones programables para generar el proceso de modelización pueden haber sido desarrolladas en otros procesos anteriores, bien internamente por equipos especialistas en la compañía, bien por la comunidad *data science*.
- **Existencia de errores:** el desarrollo de un nuevo código puede conllevar una mayor probabilidad de contener errores, por lo que es necesario realizar procesos de testeo y pruebas, lo que implica un mayor esfuerzo en tiempo y recursos.
- **Busqueda eficiente de estrategias:** que permitan descartar explícitamente combinaciones de configuraciones y rangos de hiperparámetros que se consideren inadecuados o que puedan conllevar errores en la implementación[25].

[24] Ver, por ejemplo, Segal, 2004.

[25] No obstante, aunque la generalización del problema ayuda a conseguir una resolución eficiente, en determinados casos es necesario establecer mecanismos para que el modelizador pueda probar ciertas configuraciones o definición de hiperparámetros fuera del espacio de búsqueda.

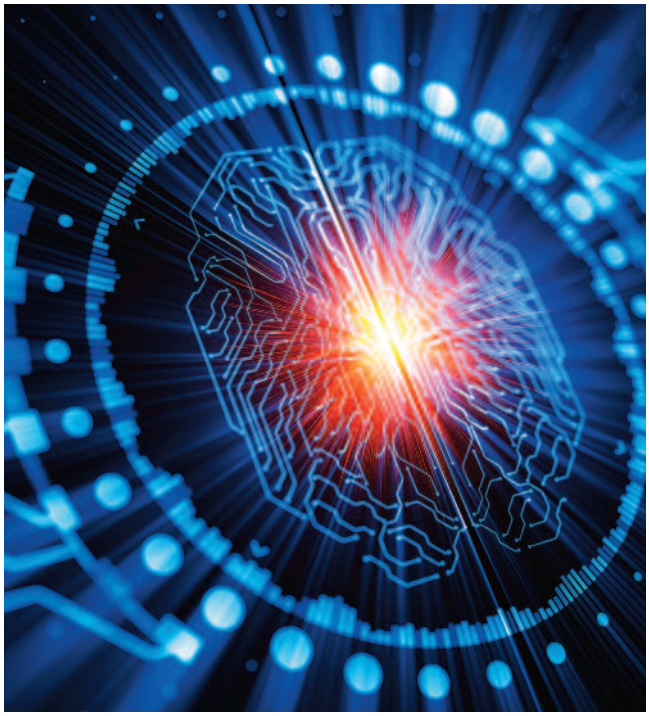
Dos de las formas de incrementar la automatización del proceso de modelización son la componentización de las tareas y la automatización de la búsqueda de configuraciones e hiperparámetros óptimos.

Componentización del workflow

En primer lugar, la segregación de las tareas de modelización en distintas piezas, y su programación y desarrollo de forma independiente, permiten al modelizador el uso automático de cada pieza, en forma de llamadas al código desarrollado, adaptando solo los parámetros y las configuraciones, dentro de las opciones posibles, para resolver una tarea específica. Este tratamiento, que es análogo al aplicado en el desarrollo de librerías en los entornos de programación, o a la programación orientada a objetos, permite aislar la tarea de desarrollo del código, así como del lenguaje de programación específico, de su aplicación posterior. Esto permite contar con un entorno ágil de modelización. Cada uno de estos componentes recibe un *input* determinado (habitualmente, un *dataset* y un conjunto de parámetros), y ejecuta una tarea específica, devolviendo como *output* otro *dataset* con el resultado de la tarea aplicada.

La componentización de los procesos de modelización genera un conjunto de ventajas frente al desarrollo de un *workflow* para cada problema, como las siguientes:

- **Estandarización:** del desarrollo de componentes utilizados en la modelización, lo que reduce la frecuencia de los errores y mejora la comparabilidad.
- **Mejora de la calidad:** en el desarrollo de los componentes, así como en su aplicación.



Elementos de un sistema de *machine learning* automatizado

- **Parámetro:** propiedad interna al modelo que se aprende durante el proceso de aprendizaje, siendo necesario para la realización de predicciones.
- **Hiperparámetro:** parámetro que no puede ser obtenido durante el proceso, y debe ser fijado de forma previa. Los valores que deben tomar los hiperparámetros para resolver un problema específico son desconocidos. El número de árboles en un *Random Forest* o el número de *clusters* en un *K-Means* son ejemplos de hiperparámetros.
- **Configuración:** las posibles combinaciones de valores que pueden tomar los hiperparámetros.
- **Espacio de configuraciones/búsqueda:** conjunto de todas las posibles configuraciones de hiperparámetros sobre el que se busca la configuración óptima para realizar la mejor predicción posible.
- **Arquitectura de redes neuronales:** se refiere conjuntamente tanto al número de capas y neuronas presentes en cada una de estas, así como a la forma en que estas se conectan. En algunos casos también se incluye en el concepto la forma en que la que son entrenadas.
- **Función de coste:** función cuyos mínimos corresponden a las configuraciones óptimas. Buscar la configuración óptima equivale a encontrar los mínimos de la función de coste. Algunas funciones de coste pueden ser el error cuadrático medio o la entropía cruzada, entre otras opciones.

- **Mejora de la eficiencia:** en la aplicación de los componentes, en la revisión por parte de áreas de validación interna y auditoría, así como en los procesos de aprobación.

- **Especialización:** en el desarrollo de cada componente por especialistas en cada materia.

- **Mejora de la usabilidad:** en su aplicación, dado que estos paquetes pueden ser utilizados por distintos tipos de usuarios, incluidos aquellos que no tienen un conocimiento de programación.

- **Escalabilidad:** en el desarrollo, que puede ser basado en un *host* interno o en *cloud*, tanto para cada filial de la empresa como por ámbitos geográficos.

Automatización de la búsqueda de una configuración óptima

Una vez que se han separado los distintos pasos del proceso en componentes, se ha de realizar una selección de los mejores parámetros que configuran el proceso óptimo para llevar a cabo la modelización. Un enfoque que permite aproximarse a dicha selección es la automatización de su búsqueda mediante la definición de distintas estrategias que aborden este problema desde un punto de vista sistemático, ordenado y dejando traza del proceso que genera cada combinación posible, pero que a su vez permita evaluar el impacto de las decisiones tomadas en cada etapa del proceso sobre el desempeño del modelo final.

Sin embargo, durante el proceso de automatización, frecuentemente la evaluación de todo el conjunto de las opciones es muy complejo y habitualmente está condicionada por un tiempo de computación limitado, tanto por el posible número de opciones y combinaciones, por la complejidad del modelo, como por la cantidad de datos analizados. A pesar de lo anterior, la automatización en la búsqueda de configuraciones, aunque limitada, genera un conjunto de ventajas frente a la búsqueda manual, como son:

- **Optimización en la búsqueda:** ya que permite generar un conjunto de combinaciones que se evaluarán y seleccionar aquellas que mejor desempeño generen.

- **Generalización de los problemas:** ya que posibilita generar espacios de búsqueda amplios cuando no exista información previa que permita anticipar cuáles deberían ser los subespacios de búsqueda con mayores probabilidades de generar modelos con un mayor desempeño.

- **Adaptación del espacio de búsqueda:** para aquellos problemas en los que se tiene alguna información sobre cuál debería ser el espacio de búsqueda óptimo, es fácil adaptar este para mejorar los resultados en función de las restricciones computacionales.

- **Experimentación:** dado que permite evaluar el impacto de microdecisiones en cada uno de los componentes incluidos en la automatización sobre el desempeño final. Por ejemplo, permite evaluar el cambio en diversos parámetros del modelo, como puede ser la profundidad máxima de los árboles en un algoritmo *random forest*.

En este sentido, un sistema de AutoML puede definirse como un método que permite la construcción de modelos de *machine learning* sin necesidad de intervención humana y sujeto a unas ciertas limitaciones computacionales[26]. El rol de quien desarrolla un modelo dentro de un sistema de AutoML se centra en la elección de los datos, en seleccionar los criterios de validación de los mismos, así como en elegir las métricas que se van a utilizar para estimar y seleccionar los modelos, en lugar de dedicar su tiempo al tratamiento de datos y la optimización de los hiperparámetros de forma iterativa en función de los resultados del modelo. Todo ello determina el espacio de búsqueda, de forma que se genera un conjunto de opciones que el algoritmo evalúa, sujeto a las condiciones fijadas. Finalmente, como resultado de este proceso se obtiene un conjunto de modelos ordenados en función de su desempeño.

En el siguiente apartado se profundizará en el enfoque de AutoML y su impacto a la hora de resolver los desafíos descritos en esta sección y la transformación que está suponiendo en la forma de modelizar y, en general, de todo el *workflow* de un proceso de *machine learning*.

[26] Yao, y otros, 2018.

Workflow de modelización

La definición de un *workflow* depende tanto del problema que se quiere resolver como del tipo y la calidad de los datos utilizados. Existen diferentes metodologías para el desarrollo de proyectos de *machine learning,* como KDD *(Knowledge Discovery in Databases),* CRISP-DM *(Cross-Reference Industry Standard Process for Data Mining)* o SEMMA *(Sample, Explore, Modify, Model and Assess)*, entre otras. Aunque existen diferencias entre ellas, en todas hay elementos comunes, que son los ejes fundamentales para la construcción de modelos de *machine learning*. A continuación, se resume el proceso de modelización:

Identificación del problema y planificación: en primer lugar, se han de determinar cuáles son los objetivos de negocio y entender el problema que se quiere resolver con un algoritmo de *machine learning,* así como los KPIs que servirán para medir el éxito del proyecto. Con ello, se realiza la planificación del proyecto.

Preparación de los datos: esta fase del proceso hace referencia a un tratamiento previo de los datos, e incluye las fases de recolección (obtención, etiquetado y clasificación de los datos recopilados y mejora de los existentes), limpieza y preparación (tratamiento de los datos para convertirlos en utilizables), análisis (detección de patrones y desarrollo de hipótesis), visualización (representación gráfica para identificar tendencias, *outliers* o patrones), integración (combinación de diferentes *datasets* para tener una visión unificada) y *feature engineering* (conversión de los datos brutos en datos con la forma deseada, generación de nuevas variables y selección de variables a incluir en el modelo).

Desarrollo del modelo: esta fase del proceso hace referencia a la selección del tipo de modelo utilizado, al entrenamiento del mismo, a la evaluación del desempeño y de los criterios estadísticos, así como al ajuste de parámetros, e incluye la elección (valoración de los distintos modelos disponibles para encontrar el que mejor se adapta), el entrenamiento (evaluación de las distintas configuraciones para convertir la información proporcionada en patrones y relaciones), la evaluación (análisis del desempeño del modelo mediante métricas utilizando datos que no se han empleado durante el entrenamiento) y el ajuste de parámetros (revisión de la posibilidad de mejorar la predicción del modelo mediante el reajuste de hiperparámetros).

Evaluación, validación y aprobación: se han de evaluar si se han cubierto los objetivos del negocio establecidos al inicio, y si se cumplen las expectativas iniciales. Asimismo, en función del *governance* definido y de la clasificación de los modelos *(tiering)* establecida en el marco de gestión de riesgo de modelo de la entidad[27], pueden existir fases adicionales en los que equipos de validación y auditoría, independientes de los desarrolladores, realizan una revisión de los distintos aspectos del modelo (datos utilizados, metodología, resultados, documentación, etc.). Esta validación puede incluir técnicas de interpertabilidad[28], con el objetivo de entender las relaciones subyacentes que explican el resultado del modelo. Asimismo, se han de realizar los procesos de aprobación establecidos en los marcos de gobernanza de la entidad. En el caso de modelos regulatorios, es preciso realizar un proceso final de aprobación por parte del supervisor.

Despliegue e integración en la gestión: por último, una vez se han superado las fases anteriores, el modelo se integra en la gestión, mediante la implantación en las arquitecturas tecnológicas, la puesta en producción, y los procesos de seguimiento y monitorización periódica de los resultados.

[27] Puede encontrarse un detalle de estos marcos de gestión en la publicación "Model Risk Management: Aspectos cuantitativos y cualitativos de la gestión del riesgo de modelo", Management Solutions, 2014.

[28] Un detalle de estas técnicas puede encontrarse en la publicación Cátedra iDanae "Interpretabilidad de los modelos de Inteligencia Artificial", UPM y Management Solutions, 2019.

Marcos de automatización de procesos de machine learning

"Hemos oído que un millón de monos con un millón de teclados podrían producir las obras completas de Shakespeare; ahora, gracias a Internet, sabemos que eso no es cierto"

Robert Wilensky[29]

Una vez se han discutido las razones que motivan la componentización y automatización de los *workflows* y algoritmos de *machine learning,* surge como principal cuestión cuál es el mejor enfoque para llevarlo a cabo. De manera específica, cuando se plantea automatizar el proceso de desarrollo de modelos de *machine learning* es necesario contestar a las siguientes preguntas:

- ¿Cuáles son los pasos previos necesarios para preparar los datos antes del proceso de modelización?
- ¿Cómo deben seleccionarse los algoritmos más adecuados para el conjunto de datos que se desea evaluar?
- ¿Cómo se debe determinar el espacio de búsqueda de hiperparámetros y configuraciones posibles?
- ¿Debe seguirse algún enfoque para reducir el tamaño del espacio de búsqueda?

En los enfoques tradicionales la forma de contestar a estas preguntas ha sido a través de la selección manual de estos criterios por parte de un analista basándose en apriorismos y en configuraciones que en el pasado han funcionado, así como sobre una base de ensayo y error que incluye cierto componente de aleatoriedad.

En la práctica, tanto para el desarrollo manual como automático, el problema de selección de parámetros obedece a los siguientes retos:

- Un planteamiento maximalista basado en la revisión exhaustiva de todas las combinaciones posibles es inabarcable tanto en tiempo como en requisitos de recursos computacionales. Incluso para *datasets* relativamente reducidos o incorporando restricciones en la búsqueda fruto de la experiencia, esta tarea sigue siendo inabordable, lo que conlleva la obligación de renunciar a la optimización en algunas partes del proceso.
- Las configuraciones utilizadas dependen significativamente de los apriorismos de los analistas y de ajustes manuales, lo que hace necesario programar explícitamente una gran cantidad de código. Por tanto, la elección y el desempeño de muchos de los métodos de *machine learning* utilizados depende de gran cantidad de decisiones sobre su diseño tomadas de forma manual o basadas en hipótesis previas.
- Si se trata de generar una función de evaluación que permita conocer la relación entre los cambios en los hiperparámetros y el desempeño del modelo, la generación de esta puede ser muy costosa, y a veces esta relación no es clara o no permite realizar inferencia sobre los resultados obtenidos.
- Esta restricción no solo no tiene una interpretación global, dado que no se puede inferir adecuadamente cuál es el impacto en la función de pérdida sobre los cambios en los hiperparámetros ni siquiera localmente.
- No se puede optimizar directamente cuando los datasets tienen un tamaño grande, ya que los tiempos de ejecución son prologandos.

Por tanto, aunque existen incentivos para llevar a cabo procedimientos de búsqueda sistemáticos y automáticos, la configuración de estos sistemas pasa por resolver cómo evaluar las posibles configuraciones dadas las restricciones existentes.

Teniendo en cuenta lo anterior, una visión que subyace habitualmente en el desarrollo de los componentes de un sistema de AutoML se basa en:

1. Automatizar la mayor parte de los aspectos relativos al análisis y tratamiento previo de los datos, mediante la generación de sistemas que permitan tratar los datos y transformar las variables utilizando las técnicas más habituales en el tratamiento manual.

[29] Robert Wilensky durante un discurso en 1996. Profesor en la Escuela de la Información de la Universidad de California en Berkeley, su principal ámbito de investigación fue la inteligencia artificial.

2. Generar un espacio de búsqueda de posibles modelos y parámetros donde se configuran un conjunto de opciones para su generación y, mediante un criterio que recorre ese espacio, se puedan obtener, comparar y seleccionar los mejores modelos.

3. Por último, automatizar técnicas de interpretabilidad, aunque de forma separada al modelo de optimización anterior, de manera que se generen reportes que sean más entendibles por los distintos usuarios.

En definitiva, el objetivo es obtener un sistema que, de forma automática, permita encontrar patrones en los datos y seleccionar una forma en la que estos den respuesta a una pregunta del usuario, siendo capaz de explicar adecuadamente los resultados. Con ello, se sustituyen las tareas con mayor complejidad y menos relacionadas con el negocio, se permite el acceso a perfiles expertos en el negocio con una formación menos profunda en ámbitos de *data science*, llevando a cabo todos los procesos de manera eficiente y robusta y teniendo en cuenta restricciones computaciones y de tiempos de ejecución. Por tanto, idealmente, el sistema de AutoML debería permitir automatizar:

- El proceso de tratamiento de datos cuando estos tienen *missings, outliers*, están mal categorizados o contienen errores.

- La posibilidad de combinar, reducir, transformar, crear o eliminar variables con base en criterios estadísticos.

- El proceso de selección de variables.

- La selección de un modelo tratando de evitar tanto el *overfitting* (un ajuste excesivo sobre unos datos de entrenamiento, desvirtuando la predicción para datos desconocidos) como el *underfitting* (el concepto contrario al *overfitting*: cuando un modelo no se ajusta lo suficiente a unos datos como para predecir correctamente).

- La explicación de los patrones identificados en los datos al usuario, de manera que un humano pueda entenderlo.

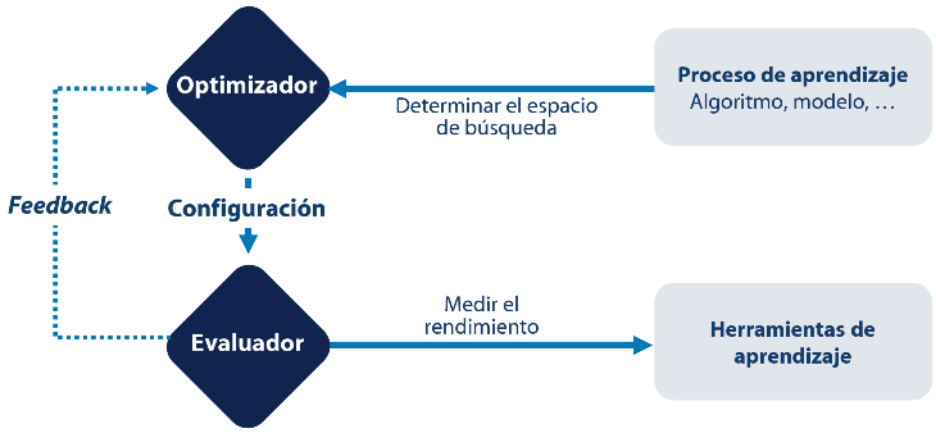
El objetivo del sistema es llevar a cabo todos estos procesos de manera eficiente y robusta, teniendo en cuenta restricciones computacionales y de tiempos de ejecución. Actualmente existen muchas soluciones propuestas, que incluyen marcos que permiten su uso de manera centralizada, distribuida o en *cloud*. Si bien, el grado de desarrollo de estos enfoques puede competir y batir a los expertos humanos en *machine learning*, aún existen muchas cuestiones que deben resolverse para que puedan aplicarse correctamente.

Un *framework* general que engloba todas las posibles partes por automatizar se indica en la figura 3[30]. Este *framework* se basa en la interacción de dos componentes fundamentales: un optimizador, que trabaja en un espacio de búsqueda definido, y un evaluador.

Por un lado, el optimizador genera y actualiza las configuraciones utilizando un espacio de búsqueda determinado en función del modelo elegido y el tratamiento de los datos que se haya realizado previamente. Posteriormente, el evaluador se encarga de medir el desempeño de las configuraciones propuestas por el optimizador. En función del enfoque seleccionado, el evaluador podría afectar a la estrategia de búsqueda del optimizador.

[30] Yao, y otros, 2018.

Figura 3: framework general para un sistema de AutoML.



Fuente: Yao, y otros, 2018.

De forma general, los componentes que se automatizan dentro del flujo son algunas fases del tratamiento de datos (preprocesamiento, *feature engineering,* tratamiento de *missings*, escalado, etc.), el modelado (selección del algoritmo, optimización de los hiperparámetros, etc.) y finalmente la evaluación de los resultados. Es frecuente que algunas piezas del tratamiento de datos se dejen fuera del proceso de automatización, dado que dependen en mayor medida del conocimiento del negocio. Del mismo modo, la interpretabilidad del modelo no se evalúa de forma automática, si bien sí que se suelen incluir herramientas que ayuden a la comprensión de los resultados.

Aunque existe una gran variedad de opciones, es necesario generar un sistema de AutoML en el que el método, descrito como marco teórico, se convierta en un conjunto de tareas (en forma de programas) que se relacionan entre sí (mediante la separación de las tareas en componentes, o a través de un diseño *end-to-end)*. De esta forma, este sistema consta de un flujo de trabajo que automatiza el diseño del *workflow* de modelización.

En general, en el mercado existen enfoques mixtos, en los que algunas fases están separadas (como son la preparación de los datos y el tratamiento de variables, así como la explicabilidad posterior a la selección del modelo), mientras que los componentes de *feature engineering,* selección del algoritmo y evaluación del modelo están incluidas en un modelo de optimización.

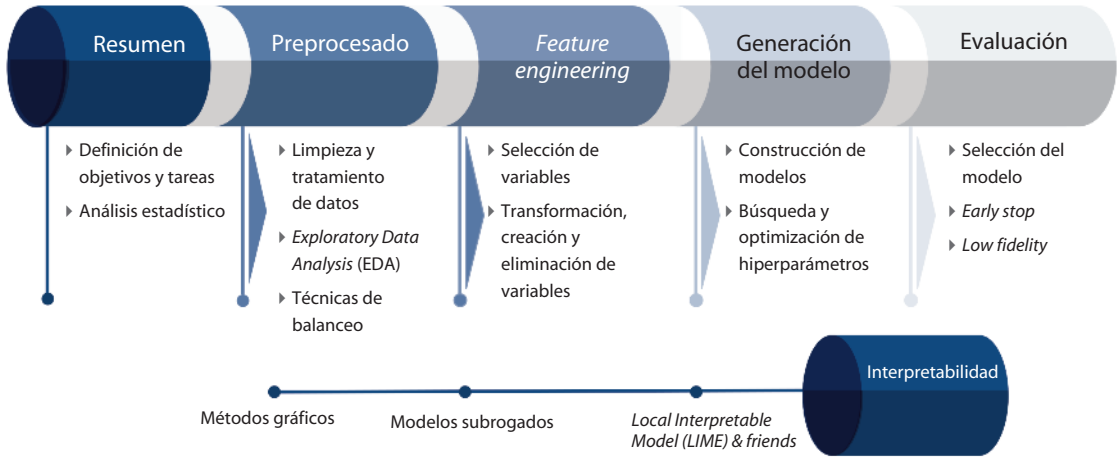
A su vez, los enfoques de modelización en el diseño de dicho flujo se clasifican en enfoques que ponen énfasis en el proceso de modelización *(model-based schemes)* y aquellos que lo hacen sobre los datos *(data-driven approaches)*. En el primero, la modelización requiere un conocimiento a priori tanto de negocio como del componente matemático-estadístico que sustenta el modelo. En el caso del segundo enfoque, la alternativa consiste en procesar la información de los datos directamente, sin realizar particiones por componentes en el proceso de modelización.

Componentes de un sistema de AutoML

En línea con lo anterior, un sistema de AutoML se puede separar en diversos componentes, de forma que, y como puede verse en la figura 4, la anatomía básica contiene los siguientes módulos:

- **Resumen:** fase exploratoria del conjunto de datos que definirá el grueso del conjunto de opciones que tendrá que afrontar el proceso de AutoML.
- **Preprocesado**: etapa de limpieza y transformación de los datos brutos antes del procesamiento y análisis.
- ***Feature Engineering:*** proceso en el que se utiliza el conocimiento que aportan los datos para generar variables que permitan que los algoritmos de *machine learning* realicen un mejor desempeño.
- **Generación del modelo:** proceso de búsqueda de hiperparámetros y optimización del modelo
- **Evaluación del modelo:** conjunto de métricas que permiten evaluar la precisión de los modelos obtenidos.
- **Interpretabilidad:** combinación de técnicas o modelos que permiten interpretar el resultado obtenido.

Figura 4: diferentes componentes de un sistema de AutoML[31].



Fuente: He, Zhao, & Chu, 2019.

Optimización de hiperparámetros

El optimizador utiliza diversas técnicas para encontrar la mejor configuración de los hiperparámetros, de forma que el desempeño del modelo sea el mejor posible. Desde un punto de vista técnico, la función del optimizador es la de buscar configuraciones óptimas en el espacio de búsqueda, de forma que encuentre el mínimo global o, al menos, un mínimo local, de la función de coste. Se puede distinguir entre técnicas simples (como *grid search*, *random search,* algoritmos evolutivos u optimización bayesiana) o basados en experiencia (como *meta-learning* o *transfer learning*).

El evaluador, por su parte, utiliza diversas técnicas para estimar el desempeño de las configuraciones propuestas por el optimizador, siendo la más simple la de entrenar el modelo.

Cuando esto resulta demasiado costoso computacionalmente, puede ser necesaria la utilización de submuestras o la inclusión de un *early stop*.

Técnicas para el optimizador: de los métodos greedy al meta-learning

Una vez que se ha fijado el espacio de búsqueda, es necesario establecer un optimizador que permita realizar búsquedas de configuraciones en el espacio. Dos de las aproximaciones más comunes son el *Grid Search* y el *Random Search,* en las que no se realizan asunciones sobre el espacio de búsqueda.

[31] He, Zhao, & Chu, 2019.

Una búsqueda *Grid*, o de fuerza bruta, establece una rejilla en el espacio de búsqueda, y evalúa la combinación dada por cada punto de la red. Este tipo de búsqueda, que se aplicó por primera vez con un enfoque de AutoML en 1990[32], no asegura que se alcance una buena configuración (es decir, un mínimo local) y puede resultar muy costoso computacionalmente en el caso de un gran número de hiperparámetros. A partir de este enfoque se han elaborado otros que mejoran el proceso basados en la utilización de una rejilla inicial para explorar todas las regiones del espacio y posteriormente una rejilla más fina en las regiones con mejor comportamiento, pudiendo iterarse el proceso hasta encontrar un mínimo local. Sin embargo, aunque se mejoran los resultados, el coste computacional de este tipo de técnicas sigue siendo elevado.

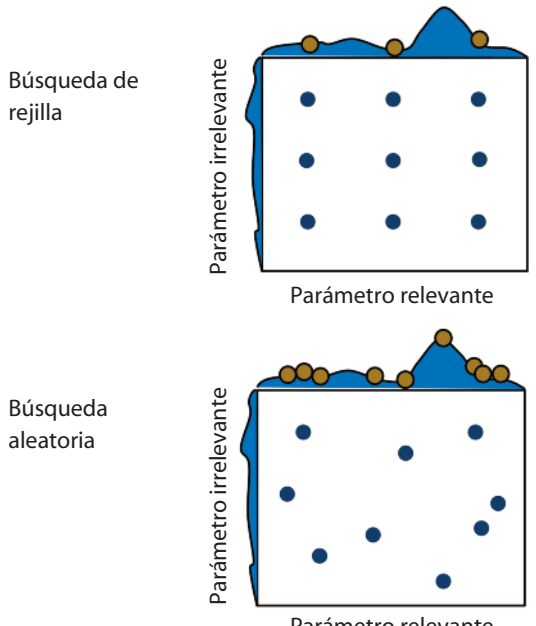
Una de las primeras soluciones que mejora los resultados de una búsqueda *Grid* es *Random Search*, el cual se basa en la selección de un punto en el espacio de búsqueda de forma aleatoria. Esto permite realizar búsquedas en zonas del espacio no equidistribuidas y, por tanto, pudiendo evaluar zonas con mayor desempeño (véase figura 5). Esta técnica sigue siendo costosa a nivel computacional, aunque como solución cumple con la condición de convergencia: cuanto mayor sea el tiempo de búsqueda, más probable será encontrar el set de hiperparámetros óptimo.

Algunos enfoques de algoritmos más elaborados incluyen, por ejemplo, los algoritmos evolutivos (entre los que se encuentran los algoritmos genéticos). Estos algoritmos crean, en una primera fase, una población inicial de configuraciones de forma aleatoria. A continuación, evalúan el desempeño de todos los individuos de la población y se seleccionan los que tienen el mejor desempeño para crear una nueva generación con base en los primeros. Además, es posible añadir mutaciones a las nuevas generaciones, de forma que difieran de la generación anterior. Este tipo de algoritmos permiten optimizar una gran variedad de problemas, pero siguen sin ser muy eficientes en lo que respecta al coste computacional, debido a que sigue siendo necesario evaluar todos los individuos de todas las generaciones.

Tanto los métodos de búsqueda por rejilla o búsqueda aleatoria como los algoritmos evolutivos tienen como riesgo que pueden investigar de forma repetida regiones con muy bajo desempeño del espacio de configuraciones sin que sea posible incluir una condición en la programación del algoritmo que corrija este comportamiento. La optimización bayesiana (utilizada al menos desde 2005[34]) soluciona este problema creando un modelo probabilístico de la función de coste, a través del cual selecciona las posibles mejores configuraciones de los hiperparámetros para evaluarlos y poder estimar la verdadera función de coste. La optimización bayesiana puede actualizar el modelo de forma

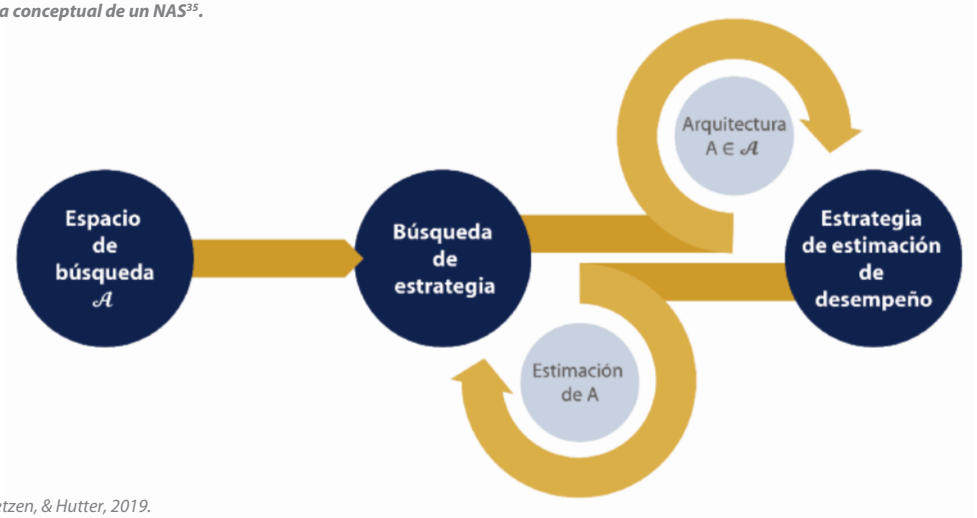
[32] Michie, Spiegelhalter, Taylor, & Campbell, 1994.
[33] Bergstra and Bengio, 2012.
[34] Fröhlich & Zell, 2005.

Figura 5: comparación de búsquedas grid y random[33] .



Fuente: Bergstra and Bengio, 2012.

Figura 6: esquema conceptual de un NAS[35].



Fuente: Elsken, Metzen, & Hutter, 2019.

iterativa haciendo un seguimiento de los resultados de evaluaciones anteriores. Esto permite actualizar el modelo probabilístico en cada cálculo.

Existen casos en los que no pueden aplicarse los procesos anteriores, por ejemplo, al existir escasez de datos. En otros casos, en los que los datasets pueden ser similares a otros estudiados anteriormente, este conocimiento no se aplicará. Con estos objetivos se desarrolló el enfoque *meta-learning,* también conocido como "*learning to learn*", que consiste en diseñar modelos de *machine learning* que sean capaces de imitar el comportamiento humano, aprendiendo nuevos conceptos y habilidades de forma rápida empleando un número reducido de muestras. Es decir, pretende diseñar modelos que puedan adquirir nuevas habilidades y sean capaces de adaptarse a nuevos entornos rápidamente con unos pocos casos.

Técnicas para el evaluador

La manera más simple de evaluar la configuración proporcionada por el optimizador es la evaluación directa sobre los datos de entrenamiento y prueba. Debido al gran número de configuraciones que se espera que el optimizador le proporcione al evaluador en un proceso de AutoML, este método puede resultar muy lento o muy costoso computacionalmente. Por este motivo, existen ciertas aproximaciones para acelerar el proceso de evaluación, aunque ello suele suponer una pérdida en la capacidad predictiva en los modelos obtenidos. Entre estas técnicas se incluyen la evaluación sobre subconjuntos de los datos de entrenamiento; procesos *early stop*, en los que el evaluador deja de evaluar si el desempeño es muy bajo en las primeras iteraciones; la reutilización de parámetros entrenados en modelos anteriores para inicializar el nuevo modelo; o finalmente, la utilización de modelos subrogados para predecir el desempeño, generalmente utilizando la experiencia de pasadas evaluaciones.

Neural Architecture Search (NAS)

Debido al auge de la aplicación de técnicas de *deep learning* en aspectos como el reconocimiento de imágenes, reconocimiento de voz y traducción automática, uno de los ámbitos a los que más interés se ha dedicado es a la configuración de arquitecturas de redes neuronales. De manera análoga a lo comentado anteriormente, estas configuraciones se establecen habitualmente de manera manual por parte de expertos humanos, lo que incurre en los errores anteriormente comentados.

[35] Elsken, Metzen, & Hutter, 2019.

Enfoques para la implementación AutoML

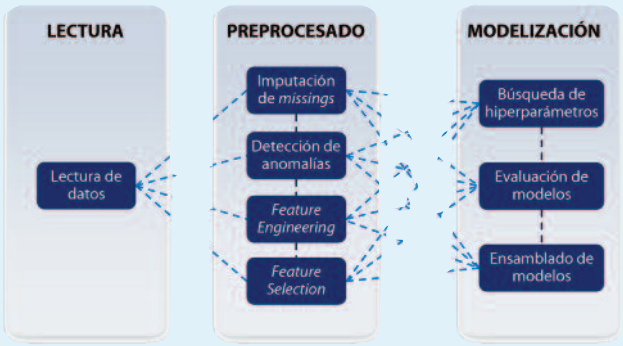
A la hora de implementar un sistema de AutoML en la práctica es necesario tener en cuenta algunas consideraciones, tales como el perfil del usuario que vaya a utilizar el sistema o la profundidad y personalización del análisis requerida. Sin embargo, es posible englobar estas implementaciones en dos enfoques principales:

- Un enfoque consiste en el diseño de un flujo parcial o totalmente editable (figura 7), donde el usuario puede definir el flujo que va a seguir el proceso de tratamiento de datos, así como las técnicas que se aplicarán en cada fase de este proceso. En este caso, el nivel de automatización es menor, ya que solo aplica a la ejecución de los módulos una vez se ha definido su orden. Debido a estas características, este enfoque resulta más adecuado cuando el usuario posee conocimientos técnicos avanzados.
- Un enfoque alternativo es aplicar una automatización *end-to-end,* con un flujo predefinido (figura 8). Los datos siguen un proceso donde el orden de cada componente de AutoML está fijado según el *pipeline* general de construcción de modelos de *machine learning*. De esta forma, el usuario no tiene necesidad de modificar el orden de ejecución de los componentes en el desarrollo. Este puede elegir los tipos de técnicas que aplica en cada componente, pero siempre siguiendo el orden predefinido. Debido a estas características, este enfoque resulta más adecuado cuando el usuario no posee conocimientos técnicos avanzados, lo que es habitual en los perfiles con foco en negocio.

Actualmente, en ninguno de los enfoques se automatiza la generación de nuevas variables a partir de las originales. Los motivos son tanto computacionales (crear transformaciones aleatorias de variables genera un coste computacional muy alto) como de negocio (el conocimiento experto del tipo de problema que se está tratando permite conocer qué transformación es la más apropiada y permite darle un sentido más adecuado a la hora de interpretar el resultado).

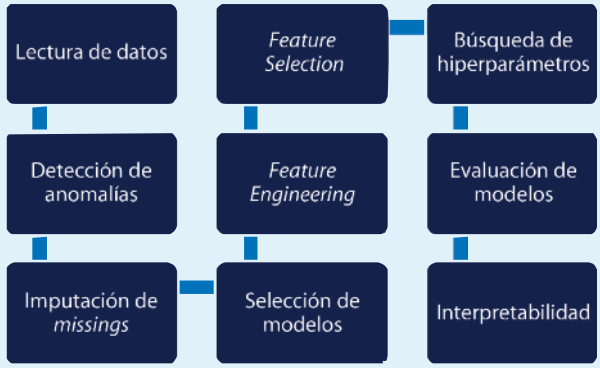
[36] La herramienta de modelización incorpora un módulo, *Model Creator,* que se basa en la automatización *end-to-end,* y un módulo alternativo, *Model Component,* que permite la generación del flujo por parte del usuario.

Figura 7: flujo parcial o totalmente editable.



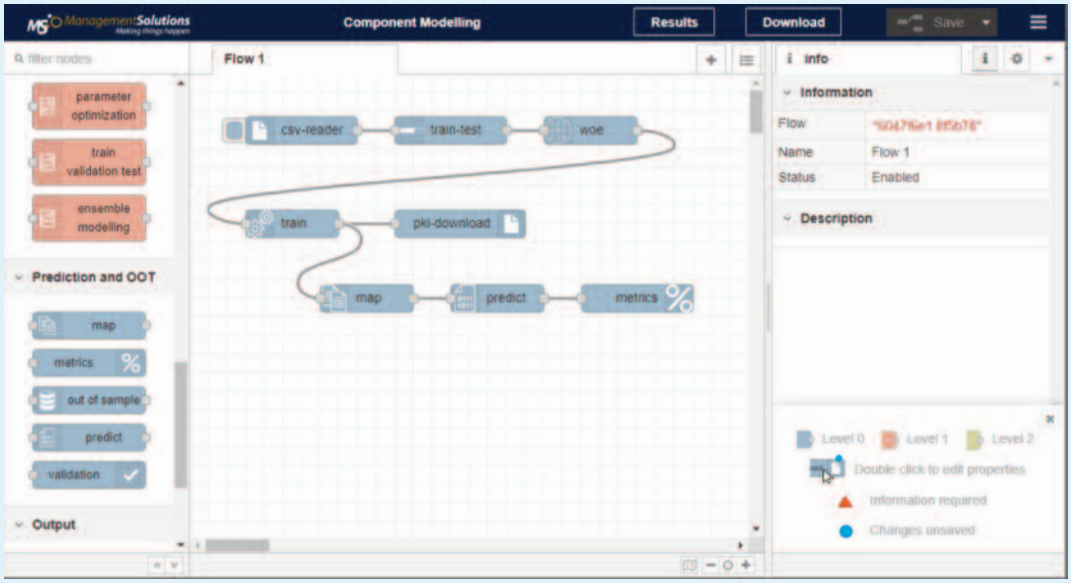
Fuente: Management Solutions.

Figura 8: flujo preestablecido.



Fuente: Management Solutions.

Figura 9: workflow diseñado en la herramienta de modelización por componentes creada por Management Solutions[36].



Fuente: Management Solutions.

Como alternativa, la búsqueda de arquitecturas neuronales (NAS por sus siglas en inglés) se basa en el uso de distintas técnicas para automatizar el diseño de redes neuronales. Los aspectos sobre los que existen parámetros son análogos a los comentados anteriormente: espacio de búsqueda, estrategia de búsqueda y estimación de desempeño. Al usar estas técnicas, todo el proceso se determina de forma simultánea, según lo indicado en la figura 6.

Aunque al aplicar los enfoques NAS los elementos del proceso se determinan de manera simultánea, los procesos previos, como los relacionados con la preparación de datos o el *feature engineering*, suelen ser necesarios y su correcta aplicación redunda en mejoras del poder predictivo.

En el espacio de búsqueda se incluyen diversos elementos, como son el número de capas del algoritmo, el tipo de operación que realiza cada capa, los hiperparámetros asociados a estas operaciones (como el número de filtros o el tamaño del kernel asociado), así como la relación y jerarquía existentes entre las distintas capas en función del algoritmo que se utilice. Si se tiene información a priori sobre posibles arquitecturas que suelen funcionar adecuadamente para una determinada tarea, se puede reducir el tamaño del espacio, simplificando así la búsqueda. En esta misma línea, también se utilizan enfoques para reducir el espacio de búsqueda, como es establecer este a través de bloques de capas en lugar sobre la arquitectura entera.

Respecto a las estrategias de búsqueda, y como se ha comentado anteriormente, existen diversas estrategias entre las que se incluyen la optimización bayesiana, *random search* y los métodos evolutivos o basados en experiencias previas como el *reinforcement learning*. En la práctica, estas técnicas no presentan resultados mejores frente a las estrategias basadas en búsquedas aleatorias. En el ámbito académico, un estudio reciente[37] destaca motivos como el uso de espacios de búsqueda restringido por estos algoritmos, así como los repartos del peso de las distintas capas en la decisión final como los elementos que limitan los resultados.

Por último, en relación con las estrategias para optimizar el desempeño de NAS, existen aproximaciones "de baja fidelidad", en las cuales se utilizan tiempos de entrenamiento más cortos, entrenamiento en subconjuntos de los datos o con menos filtros por capa, con el problema de una subestimación del desempeño. Otras estrategias consisten en la extrapolación de la curva de aprendizaje, en la utilización de modelos subrogados para predecir el desempeño de las nuevas arquitecturas, inicializar la red con pesos obtenidos de redes entrenadas previamente o utilizando teoría de grafos.

Si bien la búsqueda de arquitecturas neuronales ha logrado un nivel de rendimiento que puede competir con la configuración manual, por el momento los motivos que explican por qué las arquitecturas seleccionadas funcionan bien no están claros. Del mismo modo, hace falta más constatación empírica sobre si los motivos que hacen que una configuración funcione pueden generalizarse para diferentes problemas.

Retos actuales de los sistemas de AutoML

Actualmente, y pese a existir aún recorrido de mejora, los sistemas de AutoML han llegado a un grado de desarrollo que les permite competir, e incluso batir, a los expertos humanos en *machine learning*, configurándose como una herramienta fundamental, que modifica el tipo de trabajo desarrollado por

[37] Sciuto, 2019.

los profesionales implicados. De esta forma, los *data scientists* se distribuyen hacia tareas más relacionadas con el análisis previo y posterior al desarrollo de los propios modelos, así como al mantenimiento de estos métodos y sistemas.

Algunos de los retos actuales consisten en la mejora del proceso, así como en incorporar la interpretabilidad y permitir una interacción más sencilla por parte de los expertos:

- Actualmente, la mayor parte de las innovaciones están dirigidas a la selección y optimización de modelos, poniendo menos atención en el tratamiento y la preparación de los datos. Esto se debe a la dificultad de automatizar algunos procesos sin que acarree un gran coste computacional.

- Otra cuestión abierta es cómo tratar elementos con una interpretabilidad muy baja, conocidos como *black boxes,* ya que pueden conllevar problemas tanto legales como éticos y técnicos en su incorporación en las decisiones. En esta línea, algunas de las principales líneas de investigación van dirigidas al *Explainable AI*, la interpretabilidad y la mejora en la trazabilidad y transparencia de los modelos. Este problema también es compartido por la forma tradicional de desarrollar modelos de *machine learning*. Sin embargo, la mayor automatización ofrecida por un sistema de AutoML hace que se enfatice más su uso en este proceso.

- Por otro lado, para que los sistemas de AutoML sean efectivos, deben permitir al usuario interactuar con el sistema, permitiéndole modificar y sobreescribir las decisiones que tome, incorporando el conocimiento de los expertos en negocio respecto a diversos aspectos del proceso, como por ejemplo las predicciones realizadas o con relación a la complejidad e interpretabilidad de los modelos obtenidos.

- Por último, destaca la necesidad de establecer *benchmarks* que sirvan de estándar para poder comparar el desempeño entre las diferentes soluciones propuestas, así como contar con una definición clara de las métricas utilizadas para medir este desempeño.

Augmented machine learning

Uno de los enfoques que está acaparando más atención derivados de la generalización de la aplicación de métodos y sistemas de AutoML es el denominado *Augmented machine learning*. En este enfoque, la automatización de ciertos procesos tiene como objetivo que los sistemas de AutoML permitan lidiar con la complejidad derivada del aumento de posibles arquitecturas, opciones de hiperparámetros y opciones de entrenamiento, pero que siga existiendo un experto que utilice los resultados de la herramienta y evalúe las alternativas y las combinaciones arrojadas de una manera holística. Esto se explica por diversos motivos:

- El primero de ellos es que estos sistemas no pueden incorporar el contexto que sí tiene el usuario sobre los datos, por lo que parece que mejora el proceso el hecho de que el usuario guíe al sistema en la búsqueda de patrones sobre los datos. Este concepto, conocido como *representation engineering,* es, por ejemplo, habitual en ámbitos como la interpretación de búsquedas en internet[38].

- Por otro lado, el análisis de información en silos a través de estas herramientas provoca que se reduzca el valor esperado que puede extraerse mediante las técnicas de *advanced analytics,* por lo que es fundamental el papel de un experto en data science que tome decisiones sobre cuestiones como cuándo deben combinarse las distintas fuentes de datos o en qué casos aplicar técnicas de *transfer learning,* entre otras; dado que no es posible, por el momento, que los sistemas de AutoML aborden el análisis de todas las posibles opciones antes de tomar la decisión.

- Por último, y como se ha comentado en el punto anterior, las cuestiones éticas relativas al objetivo, a los datos utilizados, así como a los posibles sesgos que se generan en el proceso de decisión, requieren que un analista evalúe tanto la pertinencia de utilizar el modelo en un proceso de toma de decisiones como las limitaciones del modelo. En general, el uso como soporte en la toma de decisiones funciona adecuadamente, mientras que la aplicación del juicio automático es imperfecto, si bien está mejorando. En el caso de la predicción del comportamiento humano, sus resultados son dudosos[39].

[38] Abbasi, Kitchens, & Ahmad, 2019.
[39] Narayanan, 2019.

Competiciones de AutoML: una herramienta de exploración de enfoques de AutoML

"¡Ford! -exclamó-. Afuera hay un número infinito de monos que quieren hablarnos de un guion de Hamlet que han elaborado ellos mismos."

Douglas Adams[40]

Como se ha visto en los apartados previos, y a pesar de los avances que se han observado últimamente en esta disciplina, aún sigue sin prevalecer un enfoque sobre el resto de alternativas en cuestiones como el preprocesamiento de datos previo a la modelización, cómo seleccionar los algoritmos o cómo configurarlos adecuadamente. No obstante, sí que empiezan a trazarse algunas tendencias y cánones que debería incorporar cualquier proceso que integre técnicas de *advanced analytics* y, en particular, un sistema de AutoML.

Un enfoque habitual para evaluar los distintos enfoques de AutoML ha sido el desarrollo de competiciones entre *data scientists* con el objetivo de construir sistemas de AutoML. Estas son una buena referencia dado que permiten enfrentar en igualdad de condiciones los distintos enfoques y, por tanto, extraer de sus resultados si existen algunas configuraciones preferibles y bajo qué circunstancias funcionan mejor. En un momento inicial, estas competiciones se basaban en la evaluación de la selección de modelos e hiperparámetros[41]. Posteriormente, este tipo de ejercicios se ha refinado, de forma que se busca que los participantes desarrollen un sistema automático y computacionalmente eficiente capaz de entrenar y evaluar modelos sin ninguna intervención humana[42].

En general, el principal objetivo de estas competiciones es dar respuesta a una serie de cuestiones como son i) conocer el efecto de las limitaciones de tiempo en el diseño de los algoritmos, ii) identificar qué tareas son más difíciles y para qué tipo de participantes, iii) saber si existen ciertas configuraciones que suelen funcionar mejor para determinados tipos de *datasets* o problemas, y iv) evaluar el impacto de la optimización de hiperparámetros y configuraciones sobre el desempeño de los modelos finales.

Un repaso a las competiciones de AutoML

En las distintas competiciones que se han analizado se pueden observar algunos patrones[43]:

- En general, es habitual el uso de enfoques heurísticos o de búsquedas *grid* o uniformes sobre el espacio de búsqueda definido a través de una definición lineal o logarítmica.
- En algunos casos, el método anterior es mejorado a través del uso de métodos de regularización.
- El sobreentrenamiento es controlado mediante la inclusión de condiciones de parada en los métodos de optimización iterativos.
- No se suele optimizar la separación entre la muestra de entrenamiento y validación.

Como resultado, se puede observar que en ningún caso se consigue automatizar todo el proceso y es necesario que exista intervención humana en las tareas más relacionadas con la definición del ejercicio. Aún sigue siendo difícil seleccionar un sistema según el tipo de problema, así como adaptarlo al conjunto de datos existente.

Con relación a las variables, la dificultad está directamente relacionada con la existencia de ciertos atributos en los *datasets* analizados, como son la existencia de datos desbalanceados, la escasez de datos, la existencia de *missings* o la existencia de variables categóricas. En estos casos, la intervención para identificar, tratar y evaluar el impacto del tratamiento en el proceso es mayor.

Sobre el proceso de selección de configuraciones e hiperparámetros, los principales problemas se derivan del uso de técnicas *ad hoc* que empeoran el desempeño del modelo, como son la separación mediante técnicas poco sofisticadas de la muestra en entrenamiento y test, la selección inapropiada de la complejidad del modelo, la selección de hiperparámetros considerando únicamente la muestra de test, no usar todos los recursos computacionales o la definición inadecuada de métricas de *performance*.

[40] Douglas Adams, "Guía del autoestopista galáctico" (1979). Escritor y guionista inglés, conocido especialmente por la saga de nombre homónimo.
[41] Véase, por ejemplo, NIPS 2005.
[42] NIPS 2016, ICML 2016 y PAKDD 2018.
[43] Hutter, Kotthoff, & Vanschoren, 2019.

Desde el punto de vista de las técnicas de búsqueda, se observa un amplio uso de técnicas basadas en *grid* o en distribuciones uniformes sobre los parámetros del espacio de búsqueda, aunque existen algunas sofisticaciones basadas en métodos de regularización o enfoques bayesianos que evitan el *overfitting* incorporando condiciones de parada.

Competición de AutoML Management Solutions

Objetivo y definición

Con un espíritu similar, en Management Solutions se ha diseñado y llevado a cabo una competición, dirigida a los profesionales de la Firma, con el objetivo de generar un algoritmo de AutoML que sea capaz de realizar predicciones en diferentes *datasets* sin realizar modificaciones en el código, con una limitación temporal para incentivar la eficiencia computacional. El ejercicio planteado se basó en la resolución, a través de la aplicación de enfoques supervisados, de problemas de respuesta binaria bajos las siguientes condiciones:

- 3 *datasets* con distintos tamaños (<100 kb, <1 Mb y <5 Mb), todos ellos con una muestra balanceada.
- Sin valores *missi*ngs, con variables tanto categóricas como continuas, e incluyendo variables irrelevantes.
- Con una limitación de recursos computacionales: equipo con Windows 10, procesador Intel Core i5-6300 CPU @ 2.40GHz 2.50GHz y 8 Gb de memoria RAM, y con un tiempo de ejecución máximo de 20 minutos para cada *dataset*.

Evaluación

Se ha evaluado la función enviada con tres *datasets* diferentes, similares a los enviados como muestras de entrenamiento. Para ello, se han tenido en cuenta los siguientes aspectos:

- Métrica área bajo la curva (AUC) (50%).
- Calidad y limpieza del código y utilización de estándar PEP8 (20%).
- Utilización de Programación Orientada a Objetos (10%).
- Originalidad (20%).

Resultados

El número de participantes ha sido superior a cien, integrados en más de setenta equipos, con perfiles muy diversos; tanto físicos, matemáticos, ingenieros y economistas. Muchos de los participantes, poseen o están cursando algún postgrado en *data science*. El origen geográfico de los participantes es, asimismo, muy diverso: Perú, Chile, Colombia, Brasil, Alemania, Estados Unidos y España.

A lo largo de la competición los participantes se han enfrentado a diversas elecciones referidas al tratamiento de datos, a la obtención de modelos y la optimización de los hiperparámetros. La mayoría de los equipos ha realizado un tratamiento de los datos basado en la eliminación de posibles *outliers* y variables correlacionadas, la normalización de variables, la reducción de la dimensionalidad y la imputación de *missings*. Algunos equipos han utilizado técnicas WOE o *one-hot-encoding* para las variables categóricas y un tratamiento

específico para el caso de que existieran conjuntos de datos desbalanceados, así como la consideración de interacciones entre las variables para aumentar la capacidad predictiva; o la eliminación de variables irrelevantes, como pueden ser variables constantes, con muy poca varianza o variables categóricas con un número de categorías muy grande con respecto al número total de entradas.

El objetivo detrás de estos tratamientos es claro: por un lado, prepara los datos para que sean correctamente leídos por los modelos utilizados, y, por otro, reduce la dimensionalidad del espacio de búsqueda, de forma que se requiera menos tiempo para encontrar una configuración óptima. Otros participantes han adoptado un enfoque distinto para tratar este problema, limitando el número de ejecuciones del algoritmo a un número específico y constante o limitando el número de modelos que se evalúan por el sistema.

La optimización de hiperparámetros se ha enfocado, en la mayoría de casos, a través del uso de búsquedas *grid*. Algunos equipos han utilizado *random search*, algoritmos genéticos o búsqueda bayesiana. Cabe destacar que un participante ha implementado el uso de un *random search* para posteriormente realizar una búsqueda en un entorno de la configuración óptima encontrada, para tratar de mejorar la métrica con dichas configuraciones en el caso de que el resultado dado por el *random search* no superase una puntuación determinada.

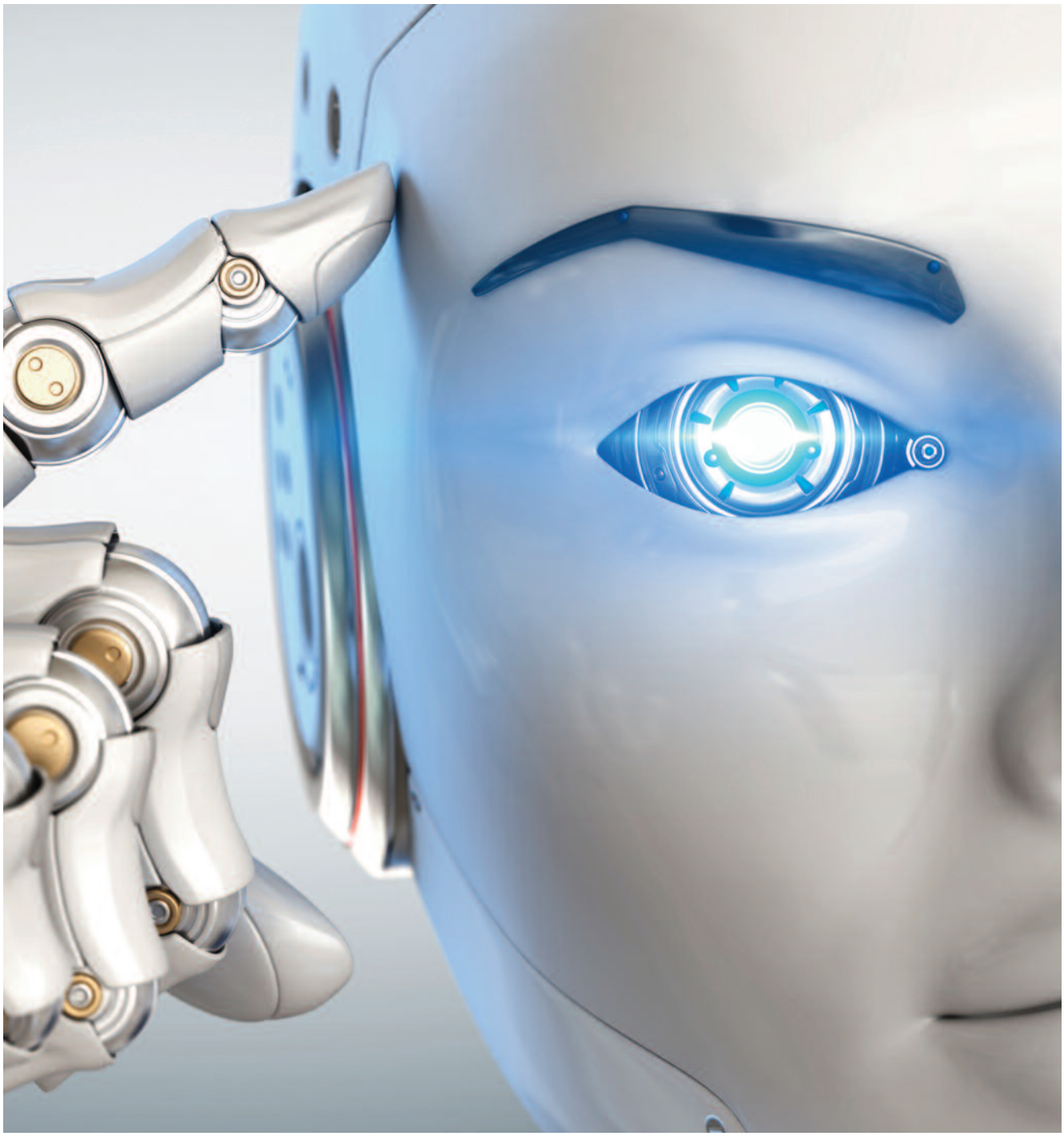
Para evaluar el desempeño de la configuración propuesta, los equipos han utilizado cross validation, y los modelos implementados fueron, en su mayoría, obtenidos de la librería *scikit-learn,* salvo algunas excepciones como el uso de *keras, lightgbm* o *xgboost*. Para optimizar tiempo computacional, algunos participantes realizaron un estudio previo de las variables más predictivas para trabajar solamente con ellas, mientras que otros evaluaron una lista de modelos y dejaron de evaluar cuando se alcanzó el tiempo máximo definido, pudiendo existir modelos estimados pero no evaluados en la muestra.

En general se ha llevado a cabo una optimización de todo el pipeline, la utilización de modelos *stacking* o la paralelización de las tareas en diversos núcleos, así como la inclusión de módulos de interpretabilidad por parte de algunos de los equipos participantes, ya sea para interpretar los *datasets* o para interpretar el proceso de AutoML, como puede ser la elección de un modelo frente a otros.

La limitación en el tiempo de ejecución de cada *dataset* no ha tenido un gran impacto en general, puesto que los archivos de evaluación eran pequeños y los AutoML de los participantes ejecutaban sin problema en ese tiempo. Solo algunos participantes limitaron el número de modelos a evaluar para evitar superar el tiempo estipulado.

Reflexiones finales

¿Y bien? ¿Qué opinas de mi nuevo poema?
Una vez leí que, dado un tiempo infinito, un millar de monos con máquinas de escribir podrían, con el tiempo, escribir las obras completas de Shakespeare
Pero, ¿qué hay de mi poema?
Tres monos, diez minutos
– Scott Adams[44]



Situación actual y retos del AutoML

Las configuraciones utilizadas para la obtención de modelos de *machine learning* dependen significativamente de los apriorismos de los analistas y de ajustes manuales, lo que genera que el desarrollo de modelos mediante técnicas de *machine learning* requiera programar explícitamente una gran cantidad de código. De esta manera, la elección y por tanto el desempeño de muchos de los métodos de *machine learning* utilizados depende de una gran cantidad de decisiones sobre su diseño que se toman de manera manual o basadas en hipótesis previas y, por tanto, se puede incurrir en subóptimos, cometiendo *overfitting* en modelos desarrollados con *datasets* pequeños y *underfitting* para *datasets* de mayor tamaño[45], lo que indica que aún se requieren mejoras para poder garantizar un uso adecuado de estos sistemas para su aplicación industrial.

Si bien los enfoques de AutoML han llegado a un grado de desarrollo que puede competir y batir a los expertos humanos en *machine learning*, aún existen muchas cuestiones que deben resolverse para que puedan aplicarse correctamente. El principal reto al que se enfrentan los actuales sistemas de AutoML es que las decisiones de diseño se tomen con un enfoque *data-driven,* de manera objetiva y automática.

En todo caso, lo anterior no es incompatible con que el usuario tenga la posibilidad de interactuar con el sistema y poder modificar y sobreescribir las decisiones que este tome. En este sentido, el desarrollo de modelos de *machine learning* se realiza a través de una industria artesanal donde los expertos abordan los problemas mediante un diseño de soluciones manuales, que en muchos casos se toman *ad-hoc* para ese proyecto, así como las preferencias y apriorismos de los expertos, pero muchas veces no incorpora la sensibilidad de los expertos en negocio. Una interfaz entendible para los *business analysts* que permita ejecutar un sistema de AutoML evita la decisión manual en las configuraciones, y a su vez posibilita la incorporación de decisiones de negocio respecto al signo o la importancia de las variables o la selección de modelos con base en la interpretación de las proyecciones, la sensibilidad frente a escenarios o la complejidad e interpretabilidad de los modelos obtenidos.

Otra cuestión abierta es cómo tratar los elementos que son *black boxes,* ya que limitan su interpretabilidad y pueden conllevar problemas tanto legales, como éticos y técnicos en su incorporación en las decisiones. En esta línea, algunas de las principales líneas de investigación van dirigidas al *Explainable AI,* la interpretabilidad y la mejora en la trazabilidad y transparencia de los modelos.

Por último, algunos aspectos, como la eficiencia de los procesos de búsqueda, están siendo constantemente mejorados, como puede verse en las distintas competiciones[46].

Grado de desarrollo

Los avances en AutoML son desiguales: la mayor parte de las innovaciones están dirigidas a algunas técnicas de *feature engineering* y a la selección de modelos frente al tratamiento y preparación de datos[47], donde aún queda mucho recorrido de mejora. Esto tiene efectos tanto sobre el tipo de tareas que se deben desarrollar en las organizaciones, como en el volumen de empleo.

Por un lado, sustituye las tareas con mayor complejidad y menos relacionadas con el negocio relativas al diseño de *pipelines* para cada problema específico, lo que permite el diseño de un *pipeline* completo por parte de perfiles con menos conocimiento de *machine learning* y por tanto dedicar a estas tareas a expertos en el negocio con una formación menos profunda en ámbitos *data science*.

Por otro, requiere una infraestructura que permita llevar a cabo estos procesos, así como mantenerlos actualizados, bien sea a través de la externalización o la contratación de servicios de AutoML bien a través de la generación de un AutoML propio que requiera de equipos especializados para que el *workflow* funcione correctamente.

Además, existen cuestiones que apenas han sido tratadas por parte de los sistemas de AutoML, de forma que tareas como la integración o la limpieza de datos, la generación de variables o el tratamiento de estas, así como algunos enfoques de *machine learning* como, por ejemplo, el aprendizaje no supervisado o el *reinforcement learning,* no están habitualmente integrados en estos sistemas.

De esta manera, se prevé que los sistemas de AutoML se configuren como una herramienta fundamental, que puede modificar el tipo de trabajos desarrollados, de forma que los *data scientists* se distribuyan hacia tareas más relacionadas con el análisis tanto previo como posterior al desarrollo de los propios modelos, a la generación de los sistemas de AutoML, y a la resolución de problemas donde las herramientas genéricas de AutoML no permitan una adecuada configuración.

[44] Scott Adams en una viñeta de Dilbert de 1989. Dibujante, autor de la tira diaria homónima
[45] Por ejemplo, en el caso de los métodos HPO. Ver Hutter, Kotthoff, & Vanschoren, 2019.
[46] Hutter, Kotthoff, & Vanschoren, 2019.
[47] Ibídem.

Bibliografía

Abbasi, A., Kitchens, B., & Ahmad, F. (2019). The Risks of AutoML and How to Avoid Them. Harvard Business Review.

Bank of England. (2019). Machine learning in UK financial services. Bank of England.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. Journal of machine learning research.

Cátedra iDanae. (3T-2019). Interpretabilidad de los modelos de Machine Learning. Cátedra iDanae.

Cátedra iDanae. (4T-2019). Ética e Inteligencia Artificial. Cátedra iDanae.

CrowdFlower. (2017). Data Scientist Report. CrowdFlower.

Elsken, T., Metzen, J. H., & Hutter, F. (2019). Neural Architecture Search: A Survey. Journal of Machine Learning Research.

European Banking Authority. (2020). EBA report on Big Data and Advanced Analytics. European Banking Authority.

European Comission. (2020). White paper on Artificial Intelligence - A European approach to excellence and trust. European Comission.

Fröhlich, H., & Zell, A. (2005). Efficient Parameter Selection for Support Vector Machines in Classification and Regression via Model-Based Global Optimization. IEEE Xplore.

Gartner. (2019). How Augmented Machine Learning Is Democratizing Data Science. Gartner.

He, X., Zhao, K., & Chu, X. (2019). AutoML: A Survey of the State-of-the-Art. arXiv preprint arXiv:1908.00709.

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). Automated Machine Learning: Methods, Systems, Challenges. Springer.

Management Solutions. (2014). Model Risk Management: Aspectos cuantitativos y cualitativos de la gestión del riesgo de modelo. Management Solutions.

Management Solutions. (2018). Machine Learning, una pieza clave en la transformación de los modelos de negocio. Management Solutions. Obtenido de Management Solutions.

Management Solutions. (2019). De proyectos Agile, a organizaciones Agile. Management Solutions.

Michie, D., Spiegelhalter, D., Taylor, C., & Campbell, J. (1994). Machine Learning, Neural and Statistical Classification. Ellis Horwood.

Mitchell, T. M. (1997). Machine learning. McGraw-Hill.

Narayanan, A. (2019). How to recognize AI snake oil.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development. IBM J. Res.

Sciuto, C. &. (2019). Evaluating the Search Phase of Neural Architecture Search. Sciuto, Christian & Yu, Kaicheng & Jaggi, Martin & Musat, Claudiu & Salzmann, Mathieu.

Segal, M. R. (2004). Machine Learning Benchmarks and Random Forest Regression.

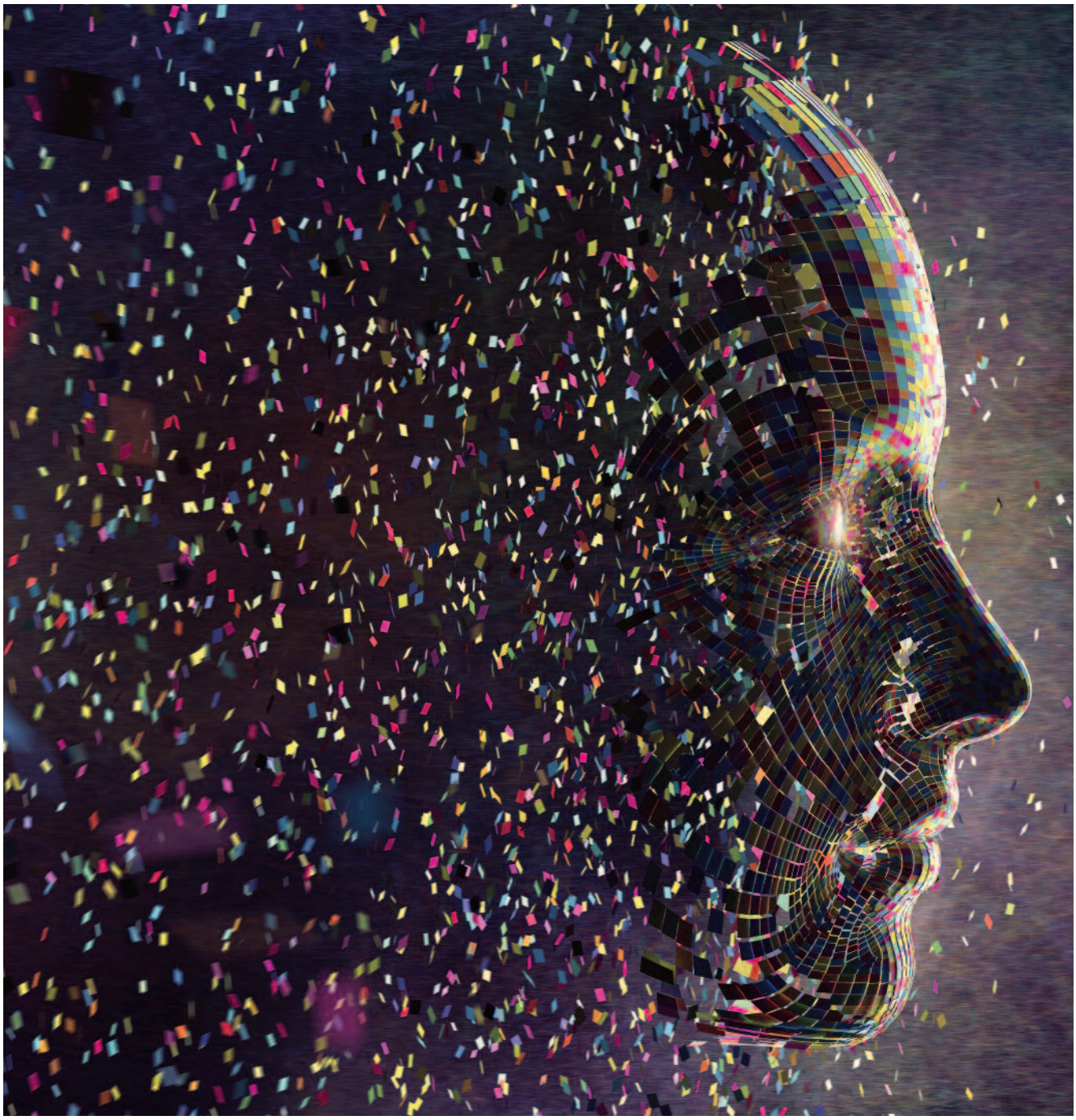
Stanford University. (2019). AI Index Report. Stanford University.

Statista (2019). Cost decreases from adopting artificial intelligence (AI) in organizations worldwide as of 2019, by function. Statista.

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems for Optimization. IEEE transactions on evolutionary computation.

Yao, Q., Wang, M., Chen, Y., Dai, W., Li, Y.-F., Tu, W.-W., . . . Yu, Y. (2018). Taking Human out of Learning Applications: A Survey on Automated Machine Learning. arXiv preprint arXiv:1810.13306.

Glosario



Cloud Computing: disponibilidad de recursos, como por ejemplo almacenamiento de datos y potencia computacional, sin una gestión activa por parte del usuario.

Configuración: las posibles combinaciones de valores que pueden tomar los hiperparámetros.

Cross Validation: proceso de validación de muestras cruzadas que consiste en dividir la muestra en k grupos, y de forma iterativa utilizar cada uno de los grupos para la validación y el resto para la construcción, cambiando el grupo de validación en cada iteración.

Early Stop: técnica que consiste en parar el proceso de búsqueda de forma anticipada a la planeada si se cumplen ciertos requisitos.

Espacio configuraciones: conjunto de todas las posibles configuraciones sobre el que se busca la configuración óptima para realizar la mejor predicción posible.

Feature Engineering: proceso de extracción de características de los datos mediante el uso de técnicas de data mining y conocimiento de un ámbito concreto.

Función de coste: función cuyos mínimos corresponden a las configuraciones óptimas. Buscar la configuración óptima equivale a encontrar los mínimos de la función de coste. Algunas funciones de coste pueden ser el error cuadrático medio o la entropía cruzada, entre otras opciones.

Grid / random / evolutivos / bayes / meta / transfer: diferentes métodos que se utilizan para buscar optimizaciones de hiperparámetros.

Hiperparámetro: parámetro que no puede ser obtenido durante el proceso, y debe ser fijado previamente. Los valores que deben tomar los hiperparámetros para resolver un problema específico son desconocidos.

Machine learning: campo de las ciencias de la computación que se centra en desarrollar técnicas que permitan que un programa pueda aprender a encontrar patrones en un conjunto de datos.

Métrica: medida para evaluar el desempeño de un modelo.

Missings: valores que faltan dentro de un *dataset.*

Modelo subrogado: modelo, generalmente más simple, que trata de emular un modelo más complejo en determinados entornos o situaciones.

Normalización: tratamiento de datos que consiste en hacer que la media de los valores de una variable esté centrada en cero y se mueva entre -1 y 1.

Outliers: valores que, por haber sido mal medidos, o por ser un comportamiento atípico, se encuentra numéricamente alejados del resto de datos.

Overfitting / underfitting: característica de un modelo que se da cuando este se ha ajustado demasiado / demasiado poco a la muestra de entrenamiento, de forma que no consigue resultados satisfactorios sobre muestras diferentes a esta (por ejemplo, sobre la muestra de validación).

Parámetro: propiedad interna al modelo que se aprende durante el proceso de aprendizaje, siendo necesario para la realización de predicciones.

Reducción de dimensionalidad: proceso por el cual se hace más pequeño el espacio de búsqueda, ya sea por combinación de variables, eliminación u otros métodos.

Regularización: técnica matemática que consiste en añadir un componente a la función de coste para detectar aquellas variables que no están aportando al modelo información significativamente diferente. Se utiliza para evitar problemas de overfitting (como, por ejemplo, el caso de las redes elásticas).

Variable continua / categórica: una variable continua es una variable numérica que puede tomar cualquier valor entre dos valores límite. Una categórica puede ser una variable numérica que sea discreta, o pueden ser palabras u otro tipo de variable.

Variables correlacionadas: variables que tienen un comportamiento similar.

WOE (Weight of Evidence): tratamiento de datos para variables categóricas.

Nuestro objetivo es superar las expectativas de nuestros clientes convirtiéndonos en socios de confianza

Management Solutions es una firma internacional de servicios de consultoría centrada en el asesoramiento de negocio, finanzas, riesgos, organización y procesos, tanto en sus componentes funcionales como en la implantación de sus tecnologías relacionadas.

Con un equipo multidisciplinar (funcionales, matemáticos, técnicos, etc.) de 2.500 profesionales, Management Solutions desarrolla su actividad a través de 31 oficinas (15 en Europa, 15 en América y 1 en Asia).

Para dar cobertura a las necesidades de sus clientes, Management Solutions tiene estructuradas sus prácticas por industrias (Entidades Financieras, Energía, Telecomunicaciones y Otras industrias) y por líneas de actividad (FCRC, RBC, NT) que agrupan una amplia gama de competencias: Estrategia, Gestión Comercial y Marketing, Gestión y Control de Riesgos, Información de Gestión y Financiera, Transformación: Organización y Procesos, y Nuevas Tecnologías.

El área de I+D da servicio a los profesionales de Management Solutions y a sus clientes en aspectos cuantitativos necesarios para acometer los proyectos con rigor y excelencia, a través de la aplicación de las mejores prácticas y de la prospección continua de las últimas tendencias en *data science, machine learning,* modelización y *big data*.

Javier Calvo Martín
Socio de Management Solutions
javier.calvo.martin@msgermany.com.de

Manuel Ángel Guzmán
Director de I+D de Management Solutions
manuel.guzman@managementsolutions.com

Daniel Ramos García
Supervisor de I+D de Management Solutions
daniel.ramos.garcia@managementsolutions.com

Segismundo Jiménez
Supervisor de I+D de Management Solutions
segismundo.jimenez@managementsolutions.com

Carlos Alonso Viñas
Consultor de Management Solutions
carlos.alonso.vinas@msspain.com

Management Solutions, servicios profesionales de consultoría

Management Solutions es una firma internacional de servicios de consultoría, centrada en el asesoramiento de negocio, riesgos, finanzas, organización y procesos.

Para más información visita **www.managementsolutions.com**

Síguenos en:

www.managementsolutions.com

Madrid Barcelona Bilbao Coruña London Frankfurt Paris Amsterdam Copenhaguen vOslo Warszawa Zürich Milano Roma Lisboa Beijing New York
Boston Atlanta Birmingham Houston San Juan de Puerto Rico San José Ciudad de México Medellín Bogotá Quito São Paulo Lima Santiago de Chile Buenos Aires