Ciclo de vida del dato
© Ediciones Roble, S. L.

Indice

Ciclo de vida del dato

I. Introducción y objetivos

Es ya un tópico que el volumen de los datos disponibles en el mundo crece de manera exponencial. Sin embargo, ese sería un hecho irrelevante si la utilidad de ese gran volumen de datos no estuviera también en crecimiento.

Ciencia de datos

Los **mecanismos para hacer que esa utilidad se incremente** son el objeto de estudio de la ciencia de datos.

Proyectos de ciencia de datos

Y su **aplicación práctica** se articula mediante proyectos de ciencia de datos.

En esta unidad, se exponen los conceptos básicos de ciencia de datos y algunas consideraciones sobre su relación con otras disciplinas.

A continuación, se hace una descripción general del tipo de proyectos de ciencia de datos y se detallan los distintos tipos de actividades a desarrollar, de manera que se agrupan en una serie de "etapas" más o menos generales.

Los **objetivos** de esta unidad son los siguientes:

1

Entender el concepto de "ciencia de datos", sus características particulares y su relación con otras ciencias.

2

Reconocer el papel de la ciencia de datos en el contexto más amplio del estudio científico, de la gestión empresarial y del desarrollo de productos y servicios.

3

Adquirir una visión general del tipo de trabajo que realiza un científico de datos.

4

Comprender la necesidad e importancia de cada una de las tareas de distinta naturaleza en el desarrollo de un producto de ciencia de datos, como son:

- El papel de la ciencia de datos en la definición de problemas a abordar.
- El estudio de las fuentes de datos y la calidad de los datos disponibles.
- La limpieza de datos y el *feature engineering*.
- La creación y evaluación de modelos de la realidad a partir de los datos.
- La elaboración de conclusiones y la explotación de modelos analíticos para crear valor.

II. Definición de ciencia de datos

2.1. Por qué una ciencia sobre los datos

El término "ciencia de datos" se ha hecho muy popular en los últimos años, pero, a pesar de ello, o quizá precisamente por ello, resulta difícil de definir de manera específica.

Existen muchas opciones distintas, pero, siguiendo la práctica tradicional de **definir una ciencia por su objeto de estudio y su enfoque epistemológico**, se podría definir de la siguiente manera:

Ciencia de datos es el estudio del tratamiento de datos orientado a obtener conclusiones útiles sobre el dominio al que se refieren esos datos.

Vale la pena aclarar que el objeto de estudio de la ciencia de datos no son los datos en general, sino específicamente el tratamiento de esos datos, ya que, si no, se podría afirmar, como hacen algunos autores, que "cualquier ciencia es ciencia de datos"; más bien al contrario, la ciencia de datos es una **ciencia sin un objeto específico particularmente interesante**, y su utilidad siempre viene dada como complemento a otros campos de estudio, mediante la colaboración con ellos.

Esto es importante tanto desde el punto de vista teórico, para clarificar qué es y qué no es ciencia de datos, como desde el punto de vista práctico, para entender la naturaleza colaborativa e interdisciplinar del trabajo de los científicos de datos.

Si la ciencia de datos es simplemente un complemento a otras disciplinas, ¿por qué es necesaria una ciencia separada, por qué existe como disciplina específica?

La razón está en que **existen muchas técnicas, herramientas, habilidades, conceptos**, etc. que son genéricos al tratamiento de datos de dominios muy distintos.

Ahora bien, si siempre ha hecho falta tratar con datos para hacer ciencia, ¿por qué precisamente ahora se ha popularizado tanto la ciencia de datos? ¿No debería haber sido popular al menos desde la Ilustración?

La razón hay que buscarla en el **aumento exponencial tanto de los datos disponibles como de las posibilidades de tratamiento**, que hacen que el conocimiento necesario para explotar al máximo las oportunidades que nos brindan esos datos no esté, muchas veces, al alcance de los especialistas del dominio correspondiente.

Ni los médicos ni los analistas financieros ni los ingenieros son necesariamente especialistas en el tratamiento de datos, y sin embargo sus disciplinas se han visto completamente transformadas por ese tratamiento.

Dado que es una disciplina cuya popularidad es reciente, la ciencia de datos se nutre de otras con más tradición. No solo se basa en sus métodos o sus resultados, sino que mucho conocimiento que tradicionalmente podía considerarse parte de esas disciplinas ahora puede verse, de manera **más integrada y sistemática**, como parte de la ciencia de datos.

Figura 1. La ciencia de datos en relación con otras disciplinas

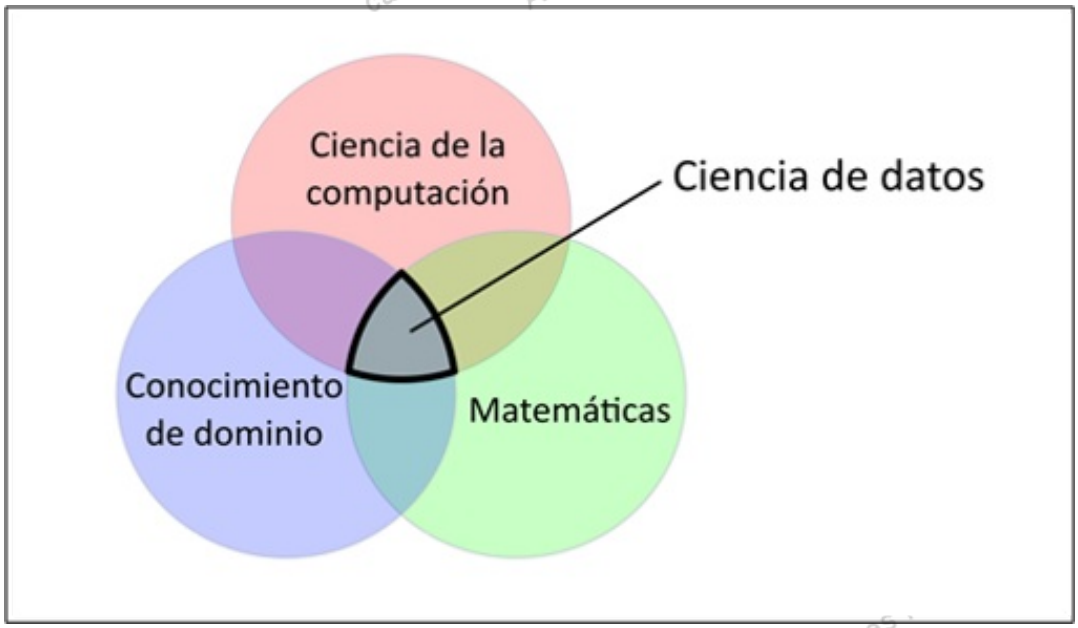

Figura 1. La ciencia de datos en relación con otras disciplinas.
Fuente: elaboración propia.

Las tres disciplinas principales que contribuyen a la ciencia de datos son la **ciencia de la computación, las matemáticas** y, de manera más general, **el conocimiento específico de ciertos dominios**, tal como muestra la figura 1.

En los siguientes apartados, se desarrollan un poco esas relaciones.

2.2. La ciencia de datos y la ciencia de la computación

El tratamiento de datos al que hace referencia la definición previa de "ciencia de datos" es, hoy en día, **automatizado**, y, por lo tanto, **la ciencia de la computación es un ingrediente fundamental** de la ciencia de datos.

Es imprescindible para los científicos de datos sentirse cómodos con el manejo de grandes cantidades de datos por ordenador y ser capaces de colaborar activamente con profesionales de otras áreas como la ingeniería de datos o el desarrollo de software.

Algunas de las **áreas de la ciencia de la computación** que son particularmente relevantes para la ciencia de datos son las siguientes:

Los lenguajes de programación

Buena parte del trabajo de ciencia de datos se realiza mediante **desarrollo de código en diversos lenguajes de programación**, ya sean de uso general (Python, SQL, Scala, Julia, etc.) o específicos para ciencia de datos (R, SAS, etc.).

Generalmente, el tipo de código a desarrollar no es particularmente complejo, de manera que es posible hacer ciencia de datos sin un conocimiento muy profundo de esos lenguajes, y no es infrecuente en los proyectos de ciencia de datos la colaboración con expertos específicos de programación (por ejemplo, para resolver problemas de rendimiento o de escalabilidad, para mejorar la reusabilidad del código o facilitar su integración en sistemas de producción, etc.).

El ciclo de vida del software

Siempre que se esté trabajando con código es necesario tener en cuenta **el ciclo de vida del software**, ya sean las prácticas de control de versiones y de configuración, de documentación, de pruebas y verificación, de integración y despliegue, de colaboración en proyectos de software más amplios, etc.

Típicamente el tipo de desarrollo de software que llevan a cabo los científicos de datos es **menos extenso y más exploratorio** que en otros casos, incluso muchas veces se trata de análisis a realizar una única vez, lo que requiere una adaptación de esas prácticas a las necesidades específicas de los proyectos la ciencia de datos.

Los protocolos de comunicación, formatos de datos, sistemas de almacenamiento, etc

El mundo en el que se manejan los datos es un **entorno técnico que el científico de datos debe conocer**.

Aun cuando los análisis se realicen en herramientas específicas y en entornos controlados, los datos a analizar pueden provenir de multitud de fuentes o tener usos finales muy distintos, de manera que es necesario ser capaz de integrarse con ese entorno más amplio. No solo eso: en muchas ocasiones los mayores problemas en un proyecto real tienen que ver con la obtención de datos, la integración de distintas fuentes o la limpieza de esos datos, y esos problemas no pueden entenderse sin conocer el entorno de origen de los datos.

Un área tecnológica que tiene mucho contacto con la ciencia de datos es la **ingeniería de datos**, pero es importante separar las responsabilidades de cada una.

La ingeniería de datos

La **ingeniería de datos da respuesta a los problemas derivados de la necesidad de tratamiento masivo de datos**: cómo almacenarlos y acceder a ellos, cómo poder procesarlos de manera masiva y paralela, cómo manejar flujos constantes de datos asegurando su integridad y el rendimiento y disponibilidad de las aplicaciones, etc.

Pero la ingeniería de datos no se preocupa por el contenido de estos; lo único que importa son cosas como el volumen, el tipo de datos o los requerimientos operativos sobre esos datos.

La ciencia de datos

La **ciencia de datos**, por el contrario, **se preocupa estrictamente por el contenido de los datos**, ya que su objetivo es obtener a partir de ellos conclusiones útiles para el mundo real.

Aun así, es frecuente, sobre todo en los proyectos de ciencia de datos más pequeños o en organizaciones menos sofisticadas respecto al tratamiento de datos, que ambas áreas se entremezclen y los científicos de datos desarrollen algunas tareas que estrictamente se podrían considerar ingeniería de datos, y viceversa.

2.3. La ciencia de datos y las matemáticas

Las **matemáticas** son el soporte fundamental de muchas ciencias, y la ciencia de datos sin duda es una de ellas. Sin las matemáticas la ciencia de datos sería poco más que un conjunto de trucos y habilidades prácticas, sería “artesanía de datos” más que una ciencia.

Esto no quiere decir que todos los científicos de datos deban tener un conocimiento matemático muy profundo o centrarse en esos aspectos, pero **deben al menos ser capaces de manejar el lenguaje matemático y entender los conceptos fundamentales**. El tipo de proyecto específico determinará muchas veces el nivel de conocimientos matemáticos necesarios.

Son muchas las **áreas de las matemáticas** que toca la ciencia de datos, pero las dos principales son la estadística y la teoría de la probabilidad.

La estadística

La estadística ha ocupado tradicionalmente una parte importante del espacio que hoy ocupa la ciencia de datos, pero no puede decirse ni que toda la estadística sea ciencia de datos, ni mucho menos que toda la ciencia de datos sea estadística.

La estadística surgió de una necesidad práctica, la recogida y análisis de datos sociales, para luego extenderse a otros tipos de datos, y solamente en el siglo XX comenzó a formalizarse matemáticamente. Sin embargo, mucha de su terminología y de sus herramientas han pasado al acervo de la ciencia de datos, y es necesario conocerlas.

Se suele dividir la estadística en dos ramas:

1

La estadística descriptiva, que se ocupa de la descripción de los datos, ya sea resumiendo la información, generando indicadores, utilizando visualizaciones, etc.

2

La estadística inferencial, que se ocupa de la extracción de conclusiones a partir de los datos, teniendo en cuenta los posibles errores en los datos, el comportamiento aleatorio, etc.

La teoría de la probabilidad

La teoría de la probabilidad se ocupa de cuestiones como el modelado de procesos aleatorios, el manejo de la incertidumbre y la información incompleta, o la estimación de parámetros de un sistema a partir de datos generados por ese sistema, todas ellas extremadamente relevantes para la ciencia de datos.

La teoría de la probabilidad es especialmente relevante en áreas como el **aprendizaje profundo o la inteligencia artificial**.

2.4. La ciencia de datos y el conocimiento de dominio

Los datos sobre los que trabaja la ciencia de datos son siempre **datos *sobre algo***, ya sean datos médicos, datos financieros o cualquier otra cosa (como dice la famosa cita de Seymour Papert, uno de los pioneros de la inteligencia artificial: "*you cannot think about thinking, without thinking about thinking about something*").

Evidentemente la gran cantidad de conocimientos específicos sobre el tema concreto sobre el que tratan los datos no pueden considerarse parte de la ciencia de datos, pero es importante disponer de algunos de esos conocimientos para poder desarrollar correctamente cualquier proyecto.

Algunos ejemplos de conocimientos de dominio que suele ser necesario conocer para desarrollar proyectos de ciencia de datos en ese dominio, y que pueden considerarse parte de la misma, son los siguientes:

La terminología específica

Es habitual que **cada área del conocimiento tenga su propia terminología** y en muchos casos usa términos distintos para los mismos conceptos.

Lo que un geógrafo llama altitud, un piloto lo llama elevación; lo que un dentista llama pieza, su cliente lo llama diente; y así sucesivamente.

Es muy habitual que un científico de datos se encuentre en la posición de tener que intermediar entre distintas áreas de conocimiento y hacer de "traductor" de conceptos entre una y otra.

Las métricas e indicadores habituales en ese dominio

No solo facilitan la comunicación, sino que muchas veces son **reflejo de las prioridades o las ideas preconcebidas** en ese dominio.

Las técnicas de procesamiento para tipos concretos de datos

Muchos tipos de datos requieren **técnicas específicas de tratamiento**, como pueden ser las imágenes médicas, los datos de audio o vídeo, la información espacial, etc.

Aunque muchas técnicas de procesamiento de datos son simplemente aplicaciones concretas de técnicas más generales, no es extraño encontrase con que en un área de conocimiento concreta tienen sus propias técnicas desarrolladas específicamente.

El estado del arte de la ciencia de datos en ese dominio

Hay áreas del conocimiento donde la ciencia de datos se lleva aplicando mucho tiempo y de manera muy sofisticada (por ejemplo, la economía o la genética), y otras en las que apenas está entrando (como el derecho o la psicología).

El conocer el estado del arte de la ciencia de datos en el área en que se está trabajando puede ayudar en gran medida a **orientar correctamente los esfuerzos y gestionar las expectativas**.

III. El ciclo de vida de los datos

3.1. Tipos de proyectos de ciencia de datos

Con un objetivo tan vago como “obtener conclusiones útiles” y un objeto de estudio tan amplio como “los datos”, es evidente que existen muchos tipos de proyectos en los que participa la ciencia de datos, y muy diferentes entre ellos.

Sin embargo, desde un punto de vista práctico es posible describir algunos de los **tipos de proyectos más comunes** en los que la ciencia de datos tiene un papel fundamental:

Estudios puntuales

En muchas ocasiones, se trata simplemente de **analizar los datos para llegar a una conclusión final**.

Algunos ejemplos de proyectos de ciencia de datos de este tipo podrían ser:

1

Los estudios clínicos, en los que a partir de los datos de la evolución de pacientes tratados de distinta manera trata de establecerse la efectividad de un tratamiento, su seguridad, las dosis a utilizar, etc.

2

El análisis financiero de un mercado o una compañía concreta, para tomar decisiones de inversión.

3

El uso de la ciencia de datos como ayuda a la investigación forense.

Creación de “productos de datos”

Entendemos aquí por “producto de datos” el hacer disponible a unos usuarios finales una capacidad de análisis útil para ellos, para que la utilicen de manera recurrente.

El componente distintivo es que **existan modelos reutilizables y capacidades de ejecutar el mismo análisis con distintos datos de entrada**.

Algunos ejemplos podrían ser:

1
La creación de un sistema de cálculo de riesgos de una aseguradora.
2
Un sistema de optimización operativa a partir de predicciones de demanda.
3
Un cuadro de mando integral para la gestión de un negocio.

Creación de nuevas capacidades

Yendo un paso más allá, muchos proyectos de ciencia de datos tratan de crear nuevas capacidades que antes no existían.

En este caso ya no se trata de un “producto de datos” como tal, sino que se vuelve invisible, **se convierte en parte del mecanismo interno que hace funcionar otros productos o servicios** que sin él simplemente no eran posibles.

Ejemplos de nuevas capacidades podrían ser:

1
Un sistema de reconocimiento del habla, que a partir de unos datos de sonidos y su correspondiente transcripción es capaz generar un sistema capaz de transcribir a texto otras locuciones, y se convierte en parte integrante de sistemas como Alexa.
2
Un sistema de control de la navegación de un vehículo que pasa a formar parte de un coche de conducción autónoma.

Todos estos tipos de proyectos, aun siendo muy diferentes entre sí, tienen en común una serie de **pasos**, unas tareas **características** y unas **técnicas a aplicar**.

3.2. Etapas de un proyecto de ciencia de datos

La figura 2 muestra las **diferentes etapas de un proyecto de ciencia de datos**, que serán desarrolladas con detalle en los siguientes apartados.

Una característica a destacar de los proyectos de ciencia de datos es que generalmente tienen un **carácter iterativo, no lineal**: no se van completando etapas sucesivamente hasta llegar al resultado esperado, sino que hay frecuentes ciclos entre actividades de distinta naturaleza.

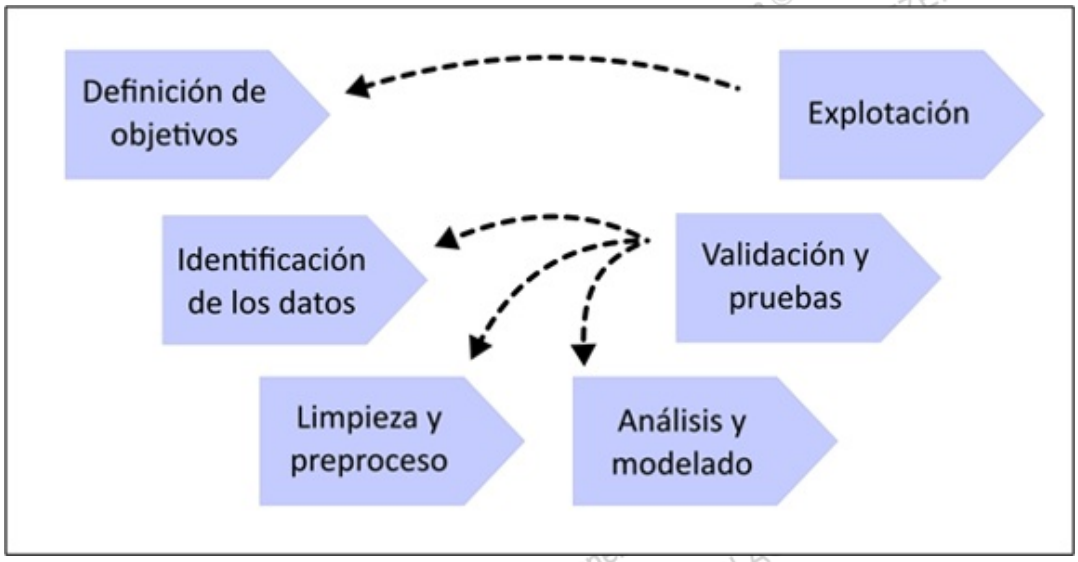

Figura 2. Etapas típicas de un proyecto de ciencia de datos.
Fuente: elaboración propia.

3.3. El ciclo de vida del proyecto frente al ciclo de vida de los datos

Es necesario señalar en este punto que **el ciclo de vida de los propios datos en la mayoría de los casos no coincide con el ciclo de vida del proyecto**.

Hay datos que se generan específicamente en el contexto de un proyecto, por ejemplo, haciendo algún tipo de experimento o recogida de datos, pero **lo común es que los datos ya existan previamente y el proyecto lo que haga es extraer nuevas conclusiones a partir de ellos.**

El **abaratamiento del almacenamiento de datos** ha sido aún más rápido que el aumento en la velocidad de generación de datos. Eso hace que, de manera cada vez más habitual, sobre todo en entornos corporativos, sea posible guardar todos los datos generados, independientemente de que en el momento de su generación se sepa si se van a utilizar para algo o no.

Esto tiene implicaciones en el **formato y soporte** en que se guardan esos datos:

1
Cuando los **datos se guardan para una finalidad concreta**, lo natural es guardarlos ya transformados, de la manera que sean más útiles y fáciles de usar para esa finalidad.
2
Por el contrario, si los datos se guardan pensando no solo en los usos presentes sino en **posibles usos futuros**, lo normal es guardarlos de la manera menos procesada posible, en el formato original en que se generaron.

Eso se traduce en **diferencias tanto tecnológicas como metodológicas**.

Por otro lado, cada vez es más común la existencia de **datos como mercancía**.

1
En ocasiones estos datos pueden estar **disponibles de manera gratuita**, que es lo que se conoce como datos "abiertos", como pueden ser los generados por gobiernos u otras organizaciones sociales.

2
En otras esos datos tienen un **coste económico**, como las bases de datos usadas en marketing directo o los datos de analistas financieros o sectoriales. Estos datos pueden aportar muchísimo valor al proyecto de ciencia de datos, o incluso ser los únicos que se utilicen, de manera que siempre es necesario tenerlos en cuenta.

Como puede verse, es posible y muchas veces necesario desacoplar el ciclo de vida de los datos del ciclo de vida de los proyectos.

IV. Definición de objetivos en un proyecto de datos

4.1. Papel de la ciencia de datos en la definición de objetivos

Un proyecto se define simplemente por un **objetivo y unos recursos destinados a la consecución de ese objetivo**.

Sin embargo, a menos que se trate de un proyecto de investigación sobre la propia ciencia de datos, al comienzo del proyecto ese objetivo no estará definido en términos de ciencia de datos, sino en términos de necesidades de negocio, de objetivos de una investigación, etc.

El primer paso, por lo tanto, para tener un "proyecto de ciencia de datos" es **generar ese objetivo**.

En general, pueden darse dos casos distintos:

Que el proyecto sea, fundamentalmente, un proyecto de datos

En este caso, la labor consistirá principalmente en **traducir los objetivos del proyecto a unos objetivos propios de la ciencia de datos y hacerlos más concretos**.

Si bien no todas las actividades a desarrollar serán de ciencia de datos, son las etapas y los ritmos de la ciencia de datos los que dan forma al proyecto y articulan el resto de actividades.

Por ejemplo, el proyecto Genoma Humano tuvo como objetivo "mapear y comprender toda la información genética de los humanos". Este objetivo se tradujo en los objetivos concretos de secuenciar todo el ADN humano y hacer mapas de localización de los genes en los cromosomas (en el momento en que se inició este proyecto, existían métodos para secuenciar pequeños fragmentos de ADN al azar y de manera muy rudimentaria, mientras que el genoma humano completo contiene unos tres mil millones de bases y unos treinta mil genes, de manera que el principal problema práctico era de correlación y manejo de unas cantidades de datos gigantescas para la época).

Que sea necesario definir un "subproyecto" de datos dentro de un proyecto más amplio

En este caso, es **el análisis del objetivo general del proyecto el que determinará las necesidades de ciencia de datos**.

El **objetivo del subproyecto de ciencia de datos estará supeditado al objetivo principal** y no es extraño que vaya cambiando según se desarrolle el proyecto, algo que es muy necesario tener en cuenta.

Por ejemplo, en el proyecto de un coche de conducción autónoma pueden hacer falta unas ciertas capacidades de visión artificial, que serían un proyecto de ciencia de datos, y estas pueden ir cambiando según se usen unas u otras tecnologías para conseguir el objetivo general del proyecto.

Tanto en un caso como en otro, es muy importante la **colaboración de expertos del dominio del problema con expertos en ciencia de datos**.

Los primeros podrán aclarar el problema y aportar información sobre el estado del arte, los datos existentes, etc.; mientras que los segundos podrán ayudar a establecer objetivos concretos y realistas.

En muchas ocasiones, dada la velocidad a la que se mueve la ciencia de datos, los expertos del dominio simplemente desconocen las posibilidades existentes y, o bien no se atreven a establecer objetivos suficientemente ambiciosos, o, por el contrario, ven la ciencia de datos como una alquimia milagrosa capaz de obtener resultados imposibles en la realidad.

También muy frecuentemente, los resultados potenciales del trabajo de ciencia de datos son inciertos, ya que, *a priori*, no se conoce la calidad de los datos que van a estar disponibles, lo predecibles o no que son los fenómenos a modelar, etc.

4.2. Consideraciones prácticas

Existen muchas formas de expresar los objetivos de un proyecto de ciencia de datos, y hay que **buscar el equilibrio entre conseguir un objetivo que oriente y organice todas las actividades en la dirección correcta**, pero, al mismo tiempo, sea lo bastante amplio como para **permitir el grado de exploración e innovación que sea apropiado**.

"Definición de éxito"

Una fórmula muy popular es la de hacer una "definición de éxito": describir, con todo el detalle que sea posible, pero sin nada accesorio, que se ha de haber conseguido para que el proyecto se considere un éxito.

Esta fórmula tiene la ventaja, entre otras, de **centrarse en el producto final del proyecto**, no en las actividades a realizar (el objetivo no es "hacer", es "conseguir", y esto es especialmente crítico cuando lo que hay que "hacer" puede no estar nada claro al principio del proyecto).

En los proyectos en los que el resultado final sea una conclusión sobre una pregunta que antes estaba abierta, es muy importante **no confundir el objetivo del proyecto con el resultado más deseable.**

Por ejemplo, si se está haciendo un estudio de mercado de un producto en diversos segmentos, el objetivo no debe ser encontrar los segmentos que encuentran atractivo el producto, sino encontrar cómo de atractivo encuentra el producto cada segmento.

Existe el llamado sesgo de confirmación, que nos lleva a dar más importancia y credibilidad a las informaciones o los datos que confirman nuestra hipótesis frente a los que la niegan, y esto está en el origen de muchos errores en proyectos de datos.

Por lo tanto, es siempre preferible establecer el objetivo del proyecto en términos neutros. A la hora de planear las actividades y los análisis, deberíamos **dar prioridad a aquellos orientados a probar las alternativas menos deseables**, no al revés, ya que eso será lo que nos permitirá alcanzar conclusiones más sólidas y evitar sesgos inconscientes.

V. Identificación de los datos necesarios

5.1. La tarea de identificación de los datos necesarios

La "materia prima" con la que se trabaja en ciencia de datos son **los datos**, sin ellos no hay proyecto posible, ellos son los que determinan el curso del proyecto.

Esto, que parece una obviedad, no siempre lo es: por ejemplo, no es infrecuente ver proyectos de ciencia de datos que toman decisiones sobre el tipo de modelos a utilizar o los resultados esperados sin ni siquiera haber visto los datos disponibles.

La primera tarea a realizar una vez el objetivo del proyecto está claro (y aún antes en ocasiones, para poder definir objetivos realistas) es **analizar las necesidades de datos y determinar su disponibilidad**.

Distintos tipos de procesamiento de datos o de modelos pueden requerir tipos o cantidades de datos muy diferentes, y puede ser necesario un **proceso iterativo** para encontrar el punto óptimo.

Por ejemplo: muchos modelos de *deep learning* o aprendizaje profundo requieren una cantidad inmensa de datos para entrenarlos correctamente, mientras que, en ocasiones, se pueden conseguir resultados similares con modelos diferentes y muchos menos datos.

Solo el entendimiento profundo de las necesidades reales y las posibilidades técnicas hará posible decidir entre invertir más en conseguir más datos o invertir más en conseguir modelos que generen resultados válidos a partir de los datos existentes.

5.2. Los tipos de fuentes de datos

A la hora de identificar las fuentes de datos a utilizar en un proyecto de ciencia de datos hay que tener en cuenta muchas consideraciones prácticas.

En primer lugar, hay que plantear la cuestión de **si los datos se pueden utilizar para el objetivo del proyecto o no**, independientemente de otras consideraciones.

Razones típicas por las que pudiera no ser posible utilizar unos datos para un cierto proyecto son las limitaciones legales, éticas o de seguridad de la información, o la falta de autorización expresa cuando esta es necesaria.

Dependiendo del objetivo del proyecto, puede ser posible utilizar otros datos que sí estén dentro de las posibilidades de uso, como pueden ser datos anonimizados en lugar de datos personales, datos agregados en lugar de datos granulares, etc.

A continuación, está la cuestión de **si los datos existen o hay que crearlos como parte del proyecto**. En el caso de que no existan, el proceso de generación de esos datos se definirá en función de las necesidades del proyecto.

Por ejemplo:

- En un proyecto de simulación de una maquinaria industrial para mejorar el mantenimiento predictivo, pueden no existir los datos necesarios para realizar esa simulación de manera suficientemente realista, pero, conociendo el funcionamiento de la maquinaria y el tipo de simulación a realizar (una vez más, colaboración entre expertos de dominio y expertos en ciencia de datos), puede ser posible definir los nuevos requerimientos de sensorización de esa maquinaria.
- En un estudio de mercado puede no existir la información necesaria sobre demografía, preferencias, etc. y que sea necesario generarla mediante encuestas, estudios de mercado, experimentos, recopilación manual de datos existentes, etc.
- Otra posibilidad es que no existan los datos específicos que se necesitan, pero sea posible inferirlos a partir de otros datos que sí existan en la medida necesaria para cumplir el objetivo del proyecto. Este es un caso muy común en econometría, entre otras materias (por ejemplo: no conozco las horas en las que una población está en sus casas, pero puedo inferirla a partir de datos del consumo eléctrico).

Si los datos existen, habrá que plantearse **si esos datos son internos a la organización en la que se desarrolla el proyecto o externos a la misma**. Si son internos ya se determinó que es posible usarlos, solo quedará asegurase de tener acceso a los mismos. En el caso de los datos externos, es posible que sea necesario seleccionar la mejor fuente para los datos que se buscan, tanto desde el punto de vista de la calidad de datos como del posible coste que pudieran tener.

La figura 3 resume los distintos tipos de orígenes de datos y las estrategias a seguir en cada caso.

Figura 3. Tipos de fuentes de datos

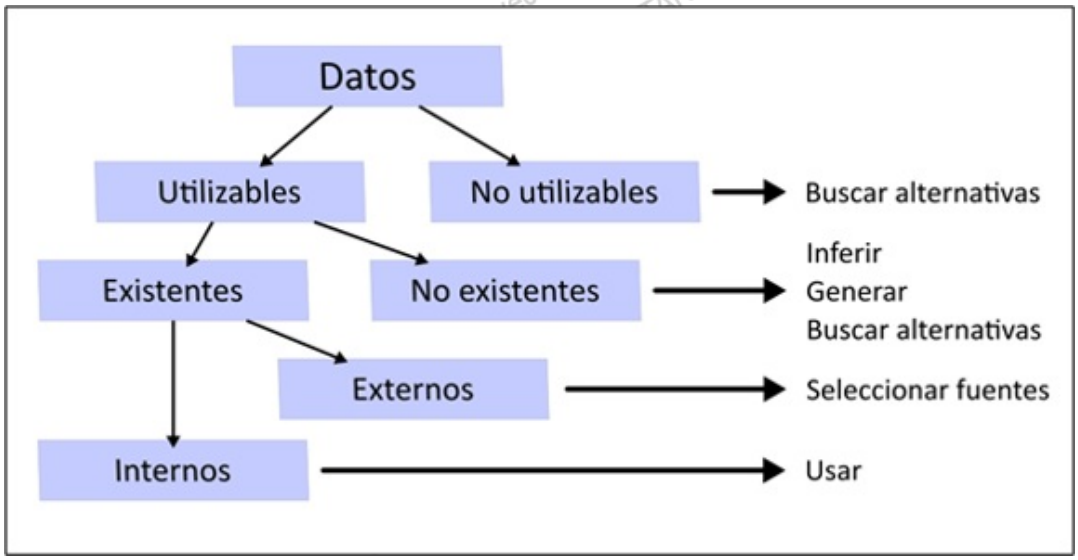

Figura 3. Tipos de fuentes de datos.
Fuente: elaboración propia.

VI. Preparación y preproceso

6.1. El acceso a los datos

Una vez identificados los datos a utilizar para el análisis, el siguiente paso es **acceder a ellos y hacerlos disponibles**. Esta es una tarea que puede requerir esfuerzos muy diferentes en diferentes situaciones.

Datos ya disponibles

El caso más sencillo es el de los **datos ya disponibles**, ya sea porque se han utilizado para análisis anteriores o porque existen mecanismos y procedimientos para acceder a ellos.

Por ejemplo, en una empresa, sería el caso de los datos disponibles en un *almacén de datos*.

Estos sistemas están diseñados específicamente para facilitar el acceso a esos datos, de manera que muchas veces el acceso se puede hacer directamente desde los mismos sistemas o lenguajes de programación utilizados para el análisis. También suele ser trivial el acceso a los datos generados *ex profeso* para el análisis en cuestión.

Datos existentes

El siguiente nivel sería el de los **datos existentes**, internos a la organización, pero no disponibles automáticamente para su análisis.

El ejemplo más común serían las bases de datos de los sistemas operacionales internos.

El acceso a esos datos, sobre todo si es masivo o necesita repetirse periódicamente, puede requerir de desarrollos tecnológicos específicos, como procesos de replicación o acceso automático en horas de baja actividad.

En estos casos es muy conveniente disponer de un almacén intermedio donde se guarden los datos en bruto, para separar operativamente los procesos de transformación y análisis de los datos de los procesos de extracción, a menos que se trate de análisis que deban realizarse en tiempo real.

Datos externos

Los **datos externos** también suelen requerir un trabajo de acceso. Una vez identificadas las mejores fuentes disponibles (que no siempre es una tarea sencilla), es necesario obtener esos datos y guardarlos localmente para su procesamiento.

Los tres mecanismos de acceso más comunes a datos públicos son los siguientes:

Ficheros de datos

En los **sitios dedicados específicamente a facilitar datos para su análisis**, es habitual que estos estén disponibles como ficheros a descargar, en formatos más o menos estándar.

Para los datos tabulares son comunes los ficheros de texto en formato CSV o similares, o los datos en formato JSON.

Para otros tipos de datos hay multitud de formatos, muchas veces específicos de un dominio concreto, que es necesario conocer y decodificar.

Interfaces de aplicación (API)

Muchos servicios y fuentes de información ofrecen acceso mediante **interfaces públicas**, ya sea para volcado de datos masivos o para acceso específico.

Es el caso, por ejemplo, de las redes sociales o servicios de mensajería (Twitter, Slack, WhatApp…) o de muchos proveedores de información financiera.

Web scrapping

El *web scrapping* es el **uso de sistemas automáticos para recopilar sistemáticamente información disponible en la web**, generalmente de manera masiva.

Por ejemplo, un sitio web de anuncios inmobiliarios podría ofrecer una página con el listado de ofertas y una página con el detalle de cada oferta: un sistema de *web scrapping* navegaría automáticamente a cada una de esas páginas siguiendo los enlaces en el listado, y guardaría los datos relevantes como el precio, la localización, las fotos, etc.

Muchos sitios web no permiten el *scrapping* masivo de sus datos, tanto por mecanismos técnicos como legales. Aun así, es una técnica que puede resultar muy valiosa en ciertas ocasiones, pero, en general, las alternativas anteriores son preferibles si existen.

Los detalles técnicos de cómo utilizar cada uno de estos tipos de fuentes de datos se desarrollan en unidades posteriores.

La tabla 1 resume algunas de las **características del acceso a diversos tipos de fuentes de datos**.

Tabla 1. Características técnicas del acceso a diferentes fuentes de datos

		Necesario almacenamiento intermedio	Necesario desarrollo	Complejidad técnica de acceso
Datos internos	Generados *ex profeso*	**SÍ**	**NO**	**MUY BAJA**
	Disponibles para análisis	**NO**	**NO**	**BAJA**
	No disponibles para análisis	**RECOMENDABLE**	**SÍ**	**MEDIA a ALTA**
Datos externos	Ficheros	**SÍ**	**NO**	**BAJA a MEDIA**
	API	**MUY RECOMENDABLE**	**SI**	**MEDIA**
	Web scrapping	**SÍ**	**SÍ**	**MEDIA a ALTA**

Tabla 1. Características técnicas del acceso a diferentes fuentes de datos.
Fuente: elaboración propia.

6.2. Calidad vs. utilidad de los datos

Un concepto fundamental en la ciencia de datos es el de **calidad de los datos**, que se desarrolla en unidades posteriores.

Respecto al **ciclo de vida de un proyecto de ciencia de datos**, basta decir aquí que el conocimiento del nivel de calidad de los datos se va adquiriendo progresivamente, a lo largo de las distintas etapas, y puede tener un impacto muy grande en el resultado.

Identificación de los datos

Es posible tener una idea de la calidad de los datos durante la **fase de identificación de los datos**, pero es en esta fase, una vez se ha accedido a los datos y se pueden analizar, cuando empieza a estar clara esa calidad.

Verificación y pruebas

La otra fase del proyecto de datos en la que se adquiere una visión todavía más completa de la calidad de los datos con los que se está trabajando es la de **verificación y pruebas** (una vez hecho un análisis o un modelo, el grado de precisión de este puede darnos pistas sobre problemas de calidad de los datos).

La calidad de los datos disponibles no puede mejorarse a menos que se acceda a otros datos, pero sí hay muchos mecanismos para lidiar con los posibles problemas de calidad de los datos y aumentar su utilidad.

Aunque la división no siempre es precisa, los **mecanismos para aumentar la utilidad de los datos** se pueden agrupar en dos categorías: limpieza de los datos y *feature engineering*.

Limpieza de datos

La limpieza de datos hace referencia a las actividades orientadas a mejorar la calidad de los datos en general. Incluye transformación de formatos, estandarización, imputación de valores faltantes, homogeneización de datos, correlación, mapeo, detección y corrección de errores, etc. Estas tareas, en muchas ocasiones, constituyen una parte muy significativa del esfuerzo de un proyecto de ciencia de datos, o pueden incluso ser el objetivo único del proyecto si esos datos se quieren usar para otros fines.

Feature engineering

Se refiere a las transformaciones de los datos que se realizan con el objetivo de facilitar alguna tarea específica de análisis o modelado. Son muy dependientes del tipo de tarea a realizar y del tipo de modelo, siendo innecesarias o triviales en algunos casos, y muy sofisticadas y completamente críticas en otros. El objetivo final es disponer del conjunto de indicadores que mejor caractericen los datos y permitan a los modelos funcionar de manera más eficiente.

A lo largo de este módulo se irán desarrollando ampliamente los conceptos y técnicas que tienen que ver con la limpieza de datos y su tratamiento.

VII. Análisis y modelado

7.1. El concepto de modelo en la ciencia de datos

Si bien las actividades que más tiempo llevan en un proyecto de ciencia de datos suelen ser las relacionadas con la recopilación y tratamiento previo de los mismo, las tareas más especializadas y muchas veces las más diferenciales son las de **análisis y modelado**.

Más adelante, a lo largo del módulo, veremos con detalle las técnicas a utilizar, tipos de modelos, etc.

“Análisis de datos”

Cuando se habla de **“análisis de datos”**, aunque no siempre se mencione, siempre se está considerando un modelo de la realidad, ya sea implícito o explícito. Sin un modelo de la realidad a la que se refieren los datos, lo único que se puede hacer es “describir” los datos, no analizarlos.

Por ejemplo: cuando se hace un análisis de unos datos de ventas por geografía y periodo temporal, implícitamente se está asumiendo que hay una cierta regularidad en el comportamiento temporal de las ventas en cada región.

Si adicionalmente se comparan las ventas en cada región con algún indicador como la población o el PIB de esa región y no otra, implícitamente se está asumiendo que existe una relación entre las ventas y el tamaño de la región, y se quiere intentar entender esa relación (si una región “vende menos” de lo que sugiere su tamaño, ¿es por alguna característica especial de esa región o porque el esfuerzo de venta no está siendo el adecuado?).

El reconocer estos “modelos implícitos” es de gran utilidad a la hora de orientar el análisis exploratorio y de presentar los resultados de manera eficiente.

Los distintos modelos que se usan en ciencia de datos, ya sean explícitos o implícitos, pueden **clasificarse en función del objetivo del análisis** que se esté realizando.

La figura 4 muestra esta clasificación, que se explica en los siguientes apartados.

Figura 4. Tipos de análisis en ciencia de datos

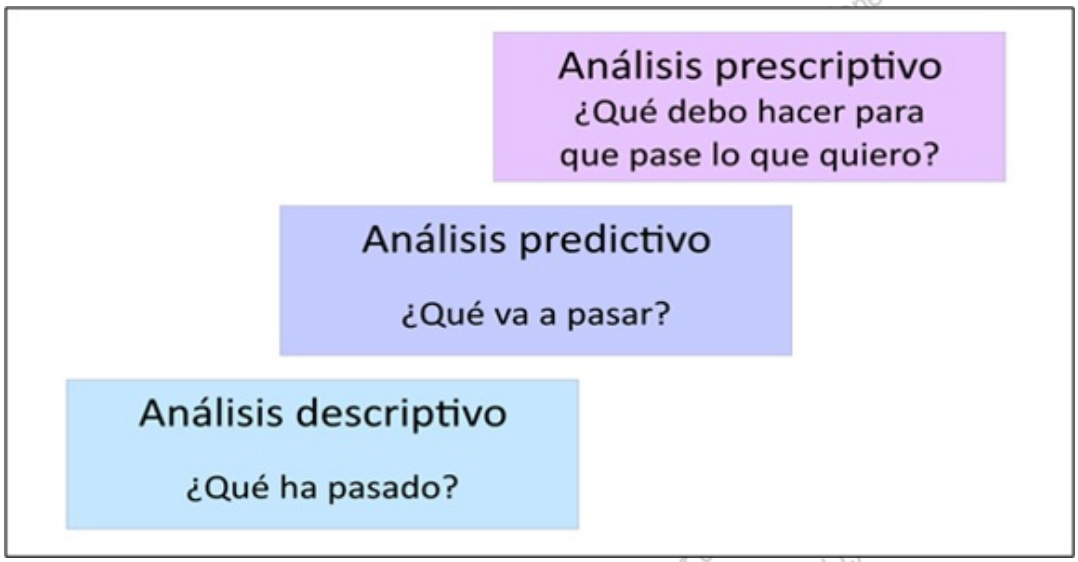

Figura 4. Tipos de análisis en ciencia de datos.
Fuente: elaboración propia.

7.2. Análisis descriptivo

El **análisis descriptivo** es aquel que trata de responder a preguntas acerca de “qué ha pasado”. No solamente trata de describir unos ciertos hechos representados por unos datos, sino de comprender las relaciones entre ellos, los mecanismos causales, la existencia de tendencias generales o de casos excepcionales, etc.

Podría parecer que este tipo de análisis es muy limitado y de menos valor que los que se describen a continuación, pero es justo lo contrario. La razón principal para hacer análisis de datos, y generalmente **lo más complicado de conseguir, es comprender la realidad**.

Esa comprensión de la realidad se utilizará luego para hacer previsiones o para decidir cursos de acción, pero esos pasos muchas veces pueden hacerse sin necesidad de la ciencia de datos si existen una buena compresión de la realidad.

De manera recíproca, da igual lo sofisticados que sean los modelos predictivos o prescriptivos: **basándose en una comprensión incorrecta de la realidad es imposible hacer predicciones precisas** o tomar decisiones correctas, excepto por azar.

Aparte de la simple descripción de casos o eventos a partir de los datos, algunas de las cuestiones que interesan en el análisis descriptivo tienen que ver con:

1
El grado de aleatoriedad o de predictibilidad de la realidad estudiada.
2
El proceso subyacente a la producción de los datos, para entender fuentes de errores, limitaciones de los datos, etc.
3
La existencia de patrones y otras regularidades en los fenómenos estudiados.
4
Las relaciones causales, las correlaciones espurias entre variables y la existencia de causas comunes entre eventos.

7.3. Análisis predictivo

El **análisis predictivo** es aquel que trata de responder a preguntas acerca de “qué va a pasar en el futuro”, o, de manera más general, de predecir algo sobre datos no estudiados previamente. Es un tipo de análisis muy popular en la ciencia de datos actual por su evidente valor práctico.

El análisis predictivo se basa en dos supuestos:

1
Que el sistema sobre el que se quiere predecir algo se comporta de manera regular.
2
Y que hay una cierta comprensión de esa regularidad.

Cuanto mayor sea esa regularidad y esa compresión, mayores serán las capacidades predictivas. Esta es una idea a mantener en mente: **si no se puede llevar la predicción más lejos por la falta de regularidad en los datos, tal vez se pueda llevar más lejos mediante una mayor compresión de la realidad**.

Por ejemplo:

1

Los modelos que tienen en cuenta relaciones causales (o sea, un mayor entendimiento de la realidad subyacente) en vez de únicamente correlaciones, tienen más capacidad de generalización.

Por ejemplo: un modelo que calculara el alcance de un proyectil a partir de la velocidad inicial y el ángulo mediante una regresión puede ser muy preciso, pero si se intenta aplicar en la Luna fallaría estrepitosamente; mientras que un modelo basado en las leyes de la física, que tenga en cuenta la gravedad, la resistencia del aire, etc. podría adaptarse para funcionar igual de bien cualquier sitio.

2

Los modelos generativos (que asumen un cierto proceso detrás del fenómeno a estudiar), si son correctos, tienen más capacidad de predicción que los modelos “caja negra” (que no asumen nada acerca del fenómeno a estudiar, solo tratan de aprender a recrear los datos con los que se entrenan).

El ejemplo clásico de esto sería comparar lo que en inversión se conoce como “análisis técnico”, que se basa únicamente en la forma de los datos de cotización de una empresa, frente al “análisis fundamental”, que analiza y trata de modelar los datos que describen la actividad de esa empresa.

7.4. Análisis prescriptivo

Finalmente, el último escalón en la jerarquía de modelos es el de los utilizados para el **análisis prescriptivo**.

Estos modeles tratan de dar respuesta a preguntas del tipo “¿qué debo hacer para conseguir mi objetivo?”.

Evidentemente, para responder a esas preguntas **primero es necesario tener modelos que me permitan entender la realidad a la que se refiere ese objetivo**, y tener modelos que me permitan estimar el resultado de opciones alternativas entre las que escoger la mejor.

Dos ejemplos de análisis prescriptivo podrían ser:

1

Un sistema de optimización de la cadena de suministro. Analizando datos existentes de ventas el sistema podría estimar la demanda de cada producto, y conociendo los costes y duración de los procesos logísticos podría optimizarlos para dar respuesta esa demanda de la manera más eficiente.

2

Un sistema de fijación de precios de una aerolínea puede balancear las probabilidades de vender un billete a mayor precio con las de acabar dejando vacío ese asiento, de manera dinámica y en función del perfil del posible comprador.

VIII. Validación y pruebas

8.1. La validación de los modelos

El nivel más básico de validación de los resultados de un proyecto de ciencia de datos es la **prueba de los modelos**.

Sin embargo, el comprobar que el modelo funciona correctamente muchas veces no es suficiente para asegurar que es el adecuado. **Es necesario asegurar también que cumple otros requisitos en función del objetivo** para el que se quiera utilizar.

Algunos de estos objetivos a comprobar son los siguientes:

Robustez del modelo

Muchos modelos requieren **reentrenarse con nuevos datos** cuando estos estén disponibles.

Es posible que un modelo funcione muy bien con unos datos concretos pero su comportamiento varíe mucho con otros:

- La presencia de casos extremos o de datos fuera del rango habitual pueden hacer que el modelo deje de funcionar correctamente.
- Un modelo puede funcionar muy bien cuando hay suficientes datos para entrenarlo, pero es posible que luego sea necesario usarlo en escenarios en los que hay menos cantidad de datos y no funcione correctamente.

Capacidad de generalización del modelo

Más allá de la capacidad de un modelo de generalizar a nuevos datos dentro de su régimen de entrenamiento, que se estudia mediante métricas y técnicas propias del *machine learning*, en ocasiones es necesario **estimar hasta dónde puede llegar esa generalización.**

Por ejemplo: un modelo que estime las ventas futuras de productos de gran consumo puede funcionar muy bien para productos de alimentación, pero, cuando se trata de aplicar a productos de moda, deja de funcionar (porque no tiene en cuenta la influencia de las tendencias en los productos de moda).

Características de ejecución del modelo

El objetivo del proyecto puede **imponer requerimientos al modelo** que son estrictamente de ciencia de datos.

Por ejemplo, un modelo que va a ejecutarse en una aplicación para móviles puede tener un límite en la cantidad de memoria que utilice, o un modelo que se use para tomar decisiones en tiempo real tiene limitaciones respecto al tiempo de cálculo que puede tomar. Si bien todos estos factores hay que tenerlos en cuenta durante todo el proyecto, **es en este punto donde se debe validar que se cumplen correctamente**.

8.2. La validación de la solución

Incluso cuando los modelos definidos se comporten correctamente y cumplan los requisitos definidos durante el proyecto, **es posible que no den una respuesta satisfactoria al objetivo del proyecto**, y por eso, el último paso antes de pasar a explotar el resultado de un proyecto de ciencia de datos debe ser la **validación de la solución**.

Por ejemplo:

- En un proyecto de diseño de un coche autónomo, pueden haberse conseguido todas las capacidades definidas en el proyecto pero que no sean suficientes para asegurar la seguridad necesaria.
- En un proyecto de optimización de una cadena de suministro, puede que las acciones de optimización generadas por el sistema sean tan sensibles a retrasos individuales que las convierta en inutilizables en la práctica.

La referencia para determinar el nivel de progreso de un proyecto de ciencia de datos debe ser siempre la consecución del objetivo explícito del proyecto. Si después de todas las actividades del proyecto hasta este punto no se ha conseguido ese objetivo, es necesario **replantear las actividades o replantear el objetivo**.

IX. Explotación

9.1. Qué se entiende por explotación

Una vez realizados los análisis y modelos necesarios, y probados y validados, es necesario asegurarse de que sirven para conseguir el objetivo del proyecto. El trabajo del científico de datos, igual que el trabajo de cualquier otro integrante de un proyecto, no termina hasta que se consigue el objetivo deseado, independientemente de cual sea su colaboración en cada tipo de actividades.

Entendemos aquí por "explotación" todas las actividades necesarias para **asegurar que los resultados de las etapas anteriores dan sus frutos**.

Naturalmente, dado que los objetivos de los proyectos de ciencia de datos pueden ser de muy distinta naturaleza, también lo son las actividades englobadas en esta fase.

9.2. El conocimiento como producto de la ciencia de datos

Se podría decir que el objetivo más básico en proyectos de ciencia de datos es el de **"generar conocimiento"**.

En estos casos, la explotación de los resultados generados consiste en **asegurarse de que ese conocimiento es trasladado del ámbito especializado del proyecto a un ámbito más general**, de manera que pueda ser aprovechado para la acción y la toma de decisiones.

La ciencia de datos es, innegablemente, un área del conocimiento muy especializada. El uso de tecnologías avanzadas de procesamiento de datos y modelos matemáticos complejos resulta poco menos que indescifrable para cualquiera que esté fuera de ella, la terminología que se usa es compleja y confusa para quien no la conoce, etc. Por eso es necesario un esfuerzo consciente para "traducir" los resultados a términos que se puedan entender, sin simplificar las conclusiones, pero eliminando la terminología específica o la información irrelevante.

Las dos herramientas fundamentales para conseguir ese objetivo son las **técnicas de comunicación eficaz y el uso del objetivo inicial del proyecto como marco de referencia**.

Sobre las técnicas de comunicación eficaz, una de las habilidades más importantes para un científico de datos es la **capacidad de "contar historias con datos"**. Los datos y los modelos pueden tener mucho interés y mucha relevancia para los especialistas, pero ninguna en absoluto para quienes no lo son.

En el caso de las historias con datos, una de las maneras más eficaces de transmitir la información relevante es mediante el uso de visualizaciones, ya sean estáticas o interactivas, como se verá más adelante. La clave está en usar esas visualizaciones para explicar una historia, no simplemente para mostrar unos datos o unas conclusiones.

Por supuesto, para que la comunicación sea eficaz es imprescindible que esa historia no solo esté bien contada, sino también que sea relevante. La forma de asegurar esa relevancia es precisamente **recurrir al objetivo definido al comienzo del proyecto**.

Puede haberse conseguido el objetivo o no, pero en cualquier caso esa es la historia. La información sobre las actividades del proyecto, los modelos utilizados, los resultados "técnicos", etc. únicamente deben servir como apoyo, nunca como hilo conductor de la historia. El uso de conceptos y terminología específica del dominio al que se refiere el objetivo del proyecto también ayuda mucho a asegurar esa relevancia.

9.3. La explotación continua de los modelos

Como ya se ha visto a lo largo de esta unidad, el coste de un proyecto de datos está fundamentalmente en las tareas a realizar para conseguir los datos, procesarlos, encontrar modelos que funcionen correctamente, etc., no en la ejecución del análisis de los datos.

Resulta muy común, por lo tanto, **reutilizar esos análisis de manera continua** una vez haya nuevos datos disponibles.

El hacer disponibles esos análisis para su uso recurrente es lo que se conoce muchas veces como **"poner en producción"** (tomando prestada terminología propia de la ingeniería de software, pero sin que quiera decir exactamente lo mismo).

La puesta en producción de resultados de ciencia de datos es, generalmente, un **proceso técnico, muy dependiente de las herramientas utilizadas y de la arquitectura de sistemas**.

Sin embargo, también hay consideraciones más específicas de la ciencia de datos a tener en cuenta, como puede ser el asegurarse de la robustez de los modelos (verificar un modelo con un conjunto de datos no necesariamente asegura que vaya a funcionar con otros datos distintos) o el utilizar técnicas compatibles con los requerimientos del sistema de explotación (por ejemplo: si un análisis se va a utilizar para la toma de decisiones en tiempo real, no tiene sentido utilizar un modelo que necesite horas para ofrecer un resultado).

9.4. La iteración en datos y la iteración en modelos

Una categoría muy importante de modelos son los conocidos como **modelos de aprendizaje automático o *machine learning***, y necesitan ser entrenados con datos reales antes de poder ser utilizados.

Existen, por lo tanto, dos actividades principales que se realizan con estos modelos:

Entrenamiento del modelo

Estas actividades están orientadas a **conseguir que el modelo funcione de la manera más eficaz posible** (por ejemplo, realizando las predicciones más precisas que sea posible).

Para ello, utiliza una gran cantidad de datos en los que la respuesta a la pregunta que trata de resolver el modelo ya se conoce: si es un modelo de diagnóstico a partir de imágenes, ya se conoce el diagnóstico por otros medios; si es un modelo de predicción temporal, se entrena con datos de hace dos meses para "predecir" los datos del mes pasado, que ya se tienen, etc.

Estas tareas pueden requerir una cantidad muy importante de procesamiento de datos y consumir un tiempo significativo.

Explotación del modelo

En este caso, el modelo ya entrenado se utiliza para hacer una **predicción a partir de datos para los que no se conoce la respuesta**.

En los casos antes mencionados, se trataría de analizar las imágenes de un paciente no diagnosticado todavía o de hacer previsiones para el mes posterior. Este uso de los modelos suele ser, comparativamente, mucho más sencillo desde el punto de vista del tiempo de cálculo.

Como se ve, una vez entrenado el modelo puede utilizarse para nuevos datos, en principio de manera indefinida. Sin embargo, el rendimiento del modelo puede disminuir en el tiempo a medida que los datos que se utilizaron para el entrenamiento se convierte en menos relevantes (por ejemplo, porque corresponden a eventos de un pasado cada vez más lejano, o porque los datos de entrada son cada vez más precisos, o en general porque la realidad subyacente al modelo ha cambiado desde el momento en que se generaron los datos de entrenamiento). Por esta razón puede ser necesario **reentrenar los modelos periódicamente con nuevos datos más relevantes.**

Existe todavía un **ciclo adicional de iteración en la explotación de los modelos**. En ocasiones, incluso mediante reentrenamiento del modelo es imposible mantener el rendimiento del mismo.

Por ejemplo, los algoritmos de *robo-trading* (algoritmos de toma de decisiones instantáneas de compra y venta de activos financieros) pierden efectividad cuando la estrategia que utilizan pasa de ser exclusiva a estar generalizada y que los demás también la usen. Por eso es necesario monitorizar el rendimiento de los modelos incluso aunque se reentrenen periódicamente, por si pierden rendimiento de manera irreversible y necesitan ser sustituidos.

En la terminología propia de la ciencia de datos, la causa del primer caso, de tener que entrenar los modelos periódicamente por los cambios en los datos, se denomina ***data drift***, mientras que la del segundo, de tener que reemplazar los modelos porque pierden efectividad con el tiempo, se denomina ***model drift***.

X. Resumen

La **ciencia de datos** es el estudio del tratamiento de los datos para obtener conclusiones útiles. Es una ciencia relativamente joven, cuya popularidad ha aumentado muy significativamente en los últimos años debido al aumento de la cantidad de datos disponibles y las capacidades para procesarlos.

La ciencia de datos se relaciona con otras disciplinas, particularmente con la **ciencia de la computación, las matemáticas y el conocimiento del dominio** al que se refieren los datos en cada caso. Es, por lo tanto, muy habitual que en los proyectos de ciencia de datos participen expertos en esas otras disciplinas.

Los **proyectos de ciencia de datos** son muy variados, pero, en general, pueden dividirse en tres tipos: los que tratan de obtener una conclusión concreta, los que tratan de generar análisis o modelos que se puedan utilizar de manera recurrente, y los que tratan de proporcionar una capacidad antes inexistente a un producto o servicio no relacionado con los datos.

Aunque las actividades a desarrollar en un proyecto de ciencia de datos son iterativas, pueden dividirse en seis fases:

1
Definición del objetivo del proyecto, en términos suficientemente concretos para orientar las actividades de ciencia de datos.
2
Identificación de los datos a utilizar, considerando las características de las distintas fuentes y los distintos tipos de datos. De esta fase puede surgir la necesidad de ejecutar actividades adicionales para generar los datos que no estén disponibles.
3
Preparación y preproceso de los datos, tanto para mejorar la calidad de los datos como para facilitar el trabajo de los algoritmos. Este tipo de tareas ocupa, en muchos casos, un porcentaje muy elevado del esfuerzo del proyecto.
4
Análisis y modelado, que permite aplicar las técnicas matemáticas y computacionales necesarias para extraer conclusiones útiles a partir de los datos, como pueden ser respuestas a preguntas concretas sobre lo que ha pasado, predicciones sobre lo que pasará en el futuro o indicaciones sobre la manera de actuar para conseguir un objetivo.
5
Verificación y pruebas, para conocer el grado de precisión de los modelos propuestos y asegurar su validez para conseguir los objetivos del proyecto.
6
Explotación, consistente en poner a disposición de los beneficiarios finales el resultado del proyecto, ya sea en forma de información, de sistemas accesibles para realizar análisis bajo demanda o de nuevos productos o servicios.

Ejercicios

Caso práctico 1

Datos

El **proyecto Genoma Humano** fue un esfuerzo científico internacional, de más de diez años de duración, para decodificar la información genética contenida en las células humanas. Puede considerarse uno de los mayores proyectos de ciencia de datos de la historia, y los avances que surgieron como consecuencia de ese proyecto no solo tuvieron un profundo impacto en la medicina y la biología, sino que también fueron, en buena medida, responsables de muchos avances técnicos en el campo del procesamiento de grandes cantidades de datos y de regulación del acceso a datos abiertos.

Algunos de los **hitos principales del proyecto** son los siguientes:

1988
Se funda HUGO ("Human Genome Organization") con el objetivo de "coordinar y organizar la investigación y las actividades técnicas para descifrar el genoma humano".
1990
Comienza oficialmente el proyecto Genoma Humano.
1994
Los científicos finalizan el mapeo de los genes en los cromosomas (una especie de mapa de todo lo que habría que descifrar).
1996
Se definen las reglas de acceso libre a los (futuros) datos resultantes del proyecto en el documento que se conoce como *Bermuda principles*.
1999
Comienza el secuenciamiento masivo del genoma humano gracias, entre otros, a los avances técnicos realizados por la empresa Celera.
1999
Se publica la secuencia (provisional) del cromosoma 22.
1999, 2000
Se hacen disponibles diversos sistemas de acceso público a los datos del genoma, lo que permite a muchas organizaciones empezar a utilizarlos.
2001
Se publica el primer borrador del genoma humano completo.

2003

Finaliza el proyecto Genoma Humano con la publicación del genoma humano corregido y se hace accesible a la comunidad científica.

Se pide

1

Clasificar de manera razonada al proyecto Genoma Humano en uno de los tres tipos de proyectos de ciencia de datos: estudio puntual, creación de un producto de datos o creación de nuevas capacidades.

2

Clasificar de manera razonada los análisis realizados como análisis descriptivos, predictivos o prescriptivos.

3

Para cada uno de los hitos mencionados, explicar brevemente a qué fase de un proyecto de ciencia de datos corresponde principalmente, ya sea poniéndola en marcha, consiguiendo un progreso significativo o finalizando la etapa (algunos hitos pueden corresponder a varias).

Solución

1

Se trata de un proyecto de creación de un producto de datos, ya que el resultado fue, fundamentalmente, información utilizable por terceros.

Aunque se crearon nuevas capacidades de secuenciación, de proceso de datos, etc., este no era el objetivo del proyecto ni son su resultado principal, sino que fueron pasos necesarios para alcanzar ese objetivo.

2

Los análisis realizados son análisis descriptivos.

No se genera explícitamente ninguna predicción ni se sugiere ningún curso de acción (aunque proyectos de ambos tipos se realizaron, más adelante, gracias a los resultados del proyecto Genoma, como pueden ser la predicción de la evolución de ciertas enfermedades genéticas o el establecimiento de recomendaciones de tratamiento en función de ciertos marcadores genéticos).

3

Las fases a las que corresponden los hitos descritos son las siguientes:

Fases

1988: se funda HUGO (Human Genome Organization) con el objetivo de coordinar y organizar la investigación y las actividades técnicas para descifrar el genoma humano.	**Definición de objetivos.**
1990: comienza oficialmente el proyecto Genoma Humano.	Comienza la fase de **identificación de los datos.**
1994: los científicos finalizan el mapeo de los genes en los cromosomas (una especie de "mapa" de todo lo que habría que descifrar).	Se avanza en la fase de **identificación de los datos** (aunque también se realizan ya tareas de acceso a los datos).
1996: se definen las reglas de acceso libre a los (futuros) datos resultantes del proyecto, en el documento que se conoce como *Bermuda Principles.*	Tareas de **explotación:** aunque todavía no están disponibles los resultados, ya se está trabajando en cómo se van a explotar.
1999: comienza el secuenciamiento masivo del genoma humano gracias, entre otros, a los avances técnicos realizados por la empresa Celera.	**Limpieza y preproceso, y análisis y modelado.**
1999: se publica la secuencia (provisional) del cromosoma 22.	**Limpieza y preproceso, y análisis y modelado.**
1999, 2000: se hacen disponibles diversos sistemas de acceso público a los datos del genoma, lo que permite a muchas organizaciones empezar a utilizarlos,	**Explotación de los datos.**
2001: se publica el primer borrador del genoma humano completo.	Principalmente, **validación y pruebas** (ya que la publicación del borrador tiene como objeto el poder corregirlo).
2003: finaliza el proyecto Genoma Humano con la publicación del genoma humano corregido y se hace accesible a la comunidad científica.	Finaliza la fase de **validación y pruebas** y comienza la **explotación** efectiva de los datos.

Caso práctico 2

Datos

Kaggle es un servicio donde diversas empresas e instituciones proponen problemas de ciencia de datos que les interesan y ofrecen premios a quienes den la mejor solución posible.

Un ejemplo real de competición en Kaggle fue la que propuso Netflix para generar recomendaciones de contenidos a sus clientes, a partir de los contenidos vistos con anterioridad y las valoraciones dadas.

Todas las competiciones funcionan de manera similar: se ofrecen unos datos a analizar y se pide a los concursantes que proporcionen la solución a una serie (suficientemente grande) de casos de prueba.

A partir de esas soluciones, el propio sistema de Kaggle calcula una métrica sobre cómo de correctas son esas soluciones, y genera una clasificación de participantes.

Los mejor clasificados se llevan un premio a cambio de, generalmente, compartir la solución utilizada con la empresa que propone la competición y, muchas veces, hacerla pública a los demás participantes.

Con el tiempo, **Kaggle se ha convertido en un entorno donde aprender sobre ciencia de datos** y se han empezado a publicar eventos del mismo estilo, pero sin intenciones competitivas y con fines puramente educativos.

Para la resolución del caso, se recomienda visitar el sitio web de kaggle para familiarizarse con el funcionamiento de este.

Se pide

1

¿Qué fases de un proyecto real de ciencia de datos lleva a cabo un participante en una competición de Kaggle y qué fases son realizadas por terceros?

2

¿Quién realiza más esfuerzo en una competición de Kaggle? ¿Los concursantes o la organización que propone el problema?

3

¿Qué tipo de análisis es el que se realiza más frecuentemente en las competiciones de Kaggle, a juzgar por las métricas de evaluación de los resultados?

4

Con el tiempo, las diferencias de puntuación entre los mejores clasificados de la mayoría de las competiciones en Kaggle ha pasado a ser muy muy pequeña, lo que ha hecho que la victoria se decida por pequeños detalles sobre el modelo o la forma de ejecutarlo.

Sin embargo, en las primeras competiciones, las diferencias entre los concursantes solían ser mucho mayores. ¿A qué puede deberse este fenómeno?

Solución

1

Las fases de definición de objetivos, identificación de los datos y limpieza y preproceso son llevadas a cabo, fundamentalmente, por la organización que publica la competición, con la ayuda en Kaggle en ocasiones.

Los participantes trabajan principalmente en tareas de **análisis y modelado**, aunque también realizan tareas de **preproceso de datos** (fundamentalmente, tareas de *feature engineering*) y de **pruebas**.

La **validación** de las soluciones la realiza Kaggle, y la **explotación** del modelo, en su caso, la realiza la organización que publica la competición.

2

El grueso del esfuerzo de un proyecto de ciencia de datos recae, generalmente, en las tareas de definición de objetivos, identificación de los datos y limpieza y preproceso; y, por lo tanto, elaborar una competición es significativamente más costoso que participar en ella.

Sin embargo, el gran número de participantes potenciales hacen que se extraiga el máximo valor posible de ese esfuerzo inicial, y esto es lo que hace que las competiciones valgan la pena para los organizadores.

3

Aunque típicamente los participantes comienzan realizando un análisis exploratorio de los datos, la inmensa mayoría de las competiciones corresponden a problemas de **análisis predictivo**, donde el competidor tiene que realizar predicciones (sobre qué contenido le va a gustar más a un cliente de Netflix, sobre si una radiografía corresponde a un tumor maligno o no, sobre el precio futuro de una propiedad inmobiliaria, etc.) y la métrica de evaluación compara esas predicciones con los valores reales, no publicados.

4

Fundamentalmente, por la popularización del conocimiento sobre ciencia de datos y las herramientas necesarias, que antes eran patrimonio exclusivo de entornos académicos o centros de investigación especializados, y ahora son accesibles a cualquiera que esté interesando.

Recursos

Bibliografía

- **“The future of data analysis”** :

  Tukey, J.*Annals of mathematical statistics. Vol. 33*; 1962.

- ***Data ccientist: the sexiest job of the 21st century*** :

  Davenport, T. H.; Patil, D. J. Havard Business Review; 2012.

- ***Principles of Strategic Data Science*** :

  Pevros, P. Ed. Packt Publishing; 2019.

- ***The data science handbook*** :

  Cady, F. Ed. Wiley; 2017.

- ***Thinking with data. How to turn information into insights*** :

  Shron, M. Ed. O’Reilly; 2014.

Glosario.

- **ALMACÉN DE DATOS (DATA WAREHOUSE)** : Es un sistema de información donde se almacenan datos ya preprocesados para su análisis, y que generalmente se usa para hacer informes y analizar distintos niveles de agregación de la información.

- **ANÁLISIS DESCRIPTIVO**: Actividades de ciencia de datos orientadas a describir y comprender la realidad descrita por unos datos.

- **ANÁLISIS PREDICTIVO**: Actividades de ciencia de datos orientadas a realizar predicciones basadas en los datos disponibles.

- **ANÁLISIS PRESCRIPTIVO**: Actividades de ciencia de datos orientadas a tratar de identificar el mejor curso de acción ante una situación.

- **CICLO DE VIDA DE LOS DATOS**: Secuencia de actividades que van desde la creación de los datos hasta sus distintos usos.

- **CICLO DE VIDA DEL PROYECTO DE DATOS**: Secuencia de actividades de un proyecto de ciencia de datos.

- **CIENCIA DE DATOS**: Estudio del tratamiento de datos, orientado a obtener conclusiones útiles sobre el dominio al que se refieren esos datos.

- **DOMINIO (EN EL CONTEXTO DE LA CIENCIA DE DATOS)**: Cualquier área del conocimiento o problema práctico que puede ser analizado a partir de datos y sobre el que existen conocimientos previos.

- **ESTADÍSTICA DESCRIPTIVA**: Técnica matemática que recolecta, organiza, analiza y presenta un conjunto de datos con el propósito de facilitar su uso.
- **ESTADÍSTICA INFERENCIAL**: Rama de la matemática que estudia los métodos para sacar conclusiones para toda una población a partir del estudio de una muestra.
- **PROYECTO DE CIENCIA DE DATOS**: Cualquier conjunto de actividades coordinadas orientadas a obtener información útil mediante el análisis de unos datos.
- **PRUEBAS (DE UN MODELO)**: Actividades orientadas a detectar errores y falta de adecuación a los fines de un modelo matemático.
- **ROBUSTEZ (DE UN MODELO)**: Capacidad de mantener el rendimiento ante la presencia de datos inexactos.
- **VALIDACIÓN (DE UN MODELO)**: Actividades orientadas a establecer la validez y corrección de un modelo matemático.