Módulo 3

3.1 ¿Qué, cómo y por qué?

*Por **Alberto Fernández Hilario***
Catedrático de la Universidad de Granada. Instituto Andaluz Interuniversitario en Ciencia de Datos e Inteligencia Computacional (DasCI)

1. ¿POR QUÉ SON IMPORTANTES LA CIENCIA DE DATOS Y EL MACHINE LEARNING?

La importancia asociada a la Ciencia de Datos y el Machine Learning (también denominado "Aprendizaje Automático") es su habilidad para resolver numerosos problemas actuales en áreas como la ingeniería, medicina, telecomunicaciones, finanzas y muchos otros. La función principal de la Ciencia de Datos consiste en descubrir conocimiento oculto y de gran interés incluido en los datos del problema. Este conocimiento, permite la creación de sistemas automáticos que ayudan a los expertos a tomar decisiones que resultan directamente en el beneficio humano.

Mediante el uso de distintos algoritmos y programas informáticos, la Ciencia de Datos detecta patrones no triviales a simple vista, utilizando para ello la información que reside en los datos. Su principal ventaja es que su uso es independiente de la tarea a realizar. A continuación, se destacan tres ejemplos distintos de la potencia de las técnicas de Machine Learning:

1. Diagnóstico por imagen para determinar, a partir de radiografías de pecho, si un paciente está infectado por coronavirus, o tiene otro tipo de patología respiratoria. En concreto, las herramientas actuales asisten a los expertos médicos a identificar, en ocasiones incluso con mayor precisión, características comunes a otros casos similares de una enfermedad respiratoria.

2. Descubrimiento de firmas o paneles genéticos para el diagnóstico de un cierto tipo de cáncer. Actualmente, resulta imposible analizar manualmente el nivel de expresión sobre las decenas de miles de genes que se obtienen a través del proceso de secuenciación. Gracias a la potencia computacional de los equipos informáticos, y mediante el uso de tests estadísticos y técnicas de Machine Learning es posible determinar de manera eficiente cuáles son los genes que juegan un papel más relevante sobre la enfermedad.

3. Desarrollo de fármacos, identificando grupos de componentes con principios activos similares, u optimizando la composición del fármaco para mejorar su efectividad o eficiencia. Los algoritmos

informáticos son muy potentes a la hora de encontrar propiedades comunes (semejanzas) entre un gran volumen de datos y crear grupos similares.

Uno de los objetivos principales de la conexión entre Machine Learning y los sistemas de salud es el desarrollo de la medicina personalizada de precisión, que se introdujo en la cápsula 2 del Módulo 1, que tiene el objetivo de que cada paciente reciba un tratamiento dirigido, de acuerdo a sus características individuales. La clave reside en la acumulación exponencial de datos biomédicos, principalmente gracias al uso masivo de aplicaciones como registros médicos electrónicos, secuenciación genómica o sensores móviles.

2. ¿QUÉ ES LA CIENCIA DE DATOS?

Si has llegado a este curso desde un área de estudio distinta a la ingeniería informática, matemáticas o estadística, posiblemente te estés preguntando en qué consiste exactamente la Ciencia de Datos y el Machine Learning. No existe una respuesta universal, si bien resulta válido definir Ciencia de Datos como un campo de estudio para extraer conocimiento de interés que está oculto en los datos, y el Machine Learning como la herramienta necesaria para alcanzar dicho objetivo.

Ser un auténtico científico de datos requiere una serie de conocimientos y habilidades que, lamentablemente, no es posible abordar en este curso tan compacto. No obstante, entre los objetivos si se encuentra el ser capaz de manejar los procedimientos básicos para introducirse en la temática, y así poder resolver los problemas dentro del campo de estudio de interés, sea la bioinformática, o cualquier otra área de biosalud o biociencia. De este modo, a lo largo del presente curso el objetivo es conocer cuál es el tipo de técnicas de Machine Learning disponibles para resolver todo caso de estudio que se pueda plantear en estas áreas.

En efecto, existen muchas aplicaciones diferentes en Ciencia de Datos, desde sistemas de recomendaciones, toma automática de decisiones, analítica predictiva, o descubrimiento de patrones, entre muchos otros. Para llevar a cabo con éxito cualquiera de las tareas anteriores, es necesario aplicar lo que se conoce como el "**ciclo de vida**" de la Ciencia de Datos. Como se observa en la Figura 1, este ciclo de vida consiste en 7 etapas bien diferenciadas, donde las fases 1 a 4 se estudiaron en el Módulo anterior, y las fases 5 a 7 son el objetivo del presente Módulo.

Figura 1. Ciclo de vida de Ciencia de Datos

1) **Comprensión del problema e identificación de objetivos**: disponer de cierto conocimiento experto para entender el caso de estudio o problema sobre el que se va a trabajar y el conocimiento que se desea obtener. En otras palabras, se define el contexto del problema, así como los objetivos que se desean alcanzar. Éstos pueden ser identificar un objeto entre varios tipos, o dar un valor numérico a una predicción futura.

2) **Recopilación de datos:** capturar toda la información posible del problema, en base a recopilación de muestras (por ejemplo, mediante técnicas NGS), anotaciones (información clínica), o cualquier otro dato que sea de interés para resolver el caso de estudio. En esta fase se crea un conjunto de datos formado por instancias, cada una representa un objeto o ejemplo correspondiente al problema.

3) **Exploración y análisis de los datos**: realizar un estudio los datos, por ejemplo, mediante el cálculo de estadísticos descriptivos o la visualización de los datos para observar sus propiedades. Esta exploración de datos consiste en calcular estadísticos como medias, desviaciones típicas, valores mínimos y máximos, entre otros. De esta forma, se comprueba si existen valores perdidos o anómalos, si los datos son representativos del problema que se está estudiando, o la correlación entre variables, entre otros. Esta fase ayuda a decidir qué método y modelo son más adecuados para el proceso de extracción de conocimiento mediante las herramientas de Machine Learning.

4) **Preparación de los datos**: pre-procesar el conjunto de instancias recopiladas. Uno de los motivos es que los datos pueden proceder de diversas fuentes y estar en formatos distintos; por tanto, es importante “limpiar los datos”, a partir de la información obtenida en la etapa anterior. El procedimiento más común es normalizar los datos, eliminar variables o ejemplos no relevantes, imputar valores perdidos, entre otros.

5) **Aprendizaje de modelos**: decidir qué algoritmos o técnicas de Machine Learning son adecuados para el problema que queremos resolver. La salida o resultado de este paso es lo que se conoce como “modelo”, definido en términos generales como una representación simplificada de los datos.

6) **Obtención de resultados y validación de modelos**: estimar la calidad del modelo aprendido, utilizando para ello un conjunto de los datos distinto al del paso 5. Un modelo permite realizar hacer predicciones sobre nueva información, o tomar decisiones en base a la información y conocimiento condensada en el dicho modelo. Si no se obtienen buenos resultados, habrá que volver a los pasos anteriores y ampliar o cambiar los datos (etapa 2), refinar el pre-procesamiento (etapa 4) o generar modelos distintos (etapa 5).

7) **Interpretación e implementación de modelos**: realizar predicciones sobre nuevos datos, crear informes que permitan realizar distintas consultas, actualizar los modelos cuando entren más datos, entre muchos otros. El objetivo final es el de interpretar los resultados y extraer información útil para resolver el problema. En otras palabras, analizar el conocimiento extraído por los modelos para determinar si aportan información útil, y sobre todo con un sentido asociado al problema que se está resolviendo.

3. EL "DATO" COMO EJE CENTRAL DEL CONOCIMIENTO

Los datos son la materia prima con la que se trabaja para extraer información valiosa y tomar decisiones informadas. La cuestión que surge es ¿qué es un conjunto de datos? La respuesta más directa sería definirlo como una colección de información organizada. Puede contener una amplia variedad de información, desde números, texto, imágenes, sonidos hasta cualquier otro tipo de registro. Estos conjuntos de datos son como tesoros de información, y el trabajo de la Ciencia de Datos es extraer conocimiento de ellos.

Los datos provienen de diversas fuentes, lo que hace que su diversidad sea prácticamente ilimitada. Algunas fuentes comunes serían repositorios públicos y privados, es decir, lugares con acceso desde la Web donde se almacena información que abarca una amplia gama de temas, desde genomas (como el NCBI GenBank o Ensembl) hasta datos climáticos (como NASA Earthdata o el Open Climate Data). También podemos recopilar datos en tiempo real a través de sensores de temperatura, cámaras, GPS y muchos otros. Otro ejemplo canónico se encuentra en campos como la medicina, donde los historiales de pacientes contienen datos valiosos que se usarán para tomar decisiones bien informadas. Finalmente, un lugar desde el que actualmente se obtienen datos de alto interés para la creación de perfiles de usuario es, sin lugar a dudas, las redes sociales.

Como se indicaba anteriormente, es la diversidad de los datos la que ofrece un potencial absoluto para su aprovechamiento. Por defecto, en la mayoría de aplicaciones de Machine Learning se considera el uso de los denominados como **datos tabulares**, es decir, conjuntos de datos organizados en filas y columnas, similares a una hoja de cálculo. Así, cada fila identificará a una instancia (a veces denominado ejemplo o muestra) y cada columna será una variable o propiedad (característica) que describe a la muestra, que puede tomar valores categóricos o numéricos, tal como se indicó en la Cápsula 1 del

Módulo 2 (El problema - ¿Cómo obtener y preparar los datos?). Además de lo anterior, cada vez es más frecuente hacer uso de datos multimedia, como las imágenes médicas, las fotografías de la naturaleza o los sonidos del océano; si bien también es muy usual en los últimos años trabajar con datos en formato texto, desde documentos escritos hasta tweets en redes sociales, mediante el uso del denominado como procesamiento del lenguaje natural. En estos dos últimos casos, es necesario transformar la información a un formato tabular, mediante un procedimiento conocido como "extracción de características" que, por su relativa complejidad, queda fuera del ámbito de este curso.

Finalmente, queda la discusión al respecto de la "temporalidad" de los datos. Por definición, los conjuntos de datos tabulares se reconocen como estáticos, en el sentido de que no cambian con el tiempo y no tienen un componente temporal inherente. En contraste, los datos dinámicos se refieren a datos que cambian con el tiempo y están vinculados a lo que se denomina como serie temporal. Cada punto de datos en una serie temporal se registra en función de un instante específico en el tiempo, y los valores pueden variar con el mismo. Estos datos se usan a menudo para representar fenómenos como la temperatura diaria, el precio de las acciones o la producción de energía a lo largo de los días o años. La elección entre cada tipología de datos dependerá de la aplicación y el tipo de análisis que se va a realizar.

4. MACHINE LEARNING Y MODELOS DE DATOS

La era digital y la revolución de los datos han potenciado la investigación y el desarrollo de las técnicas de Machine Learning. Éstos aprenden lo que se denomina **modelos de datos**, que permiten representar el conocimiento y dar valor y utilidad a los propios datos. Dos cuestiones surgen al respecto. La primera ¿qué es exactamente un modelo de datos? La segunda ¿qué se entiende por Machine Learning?

Trabajar con los datos "en crudo" no es útil ni viable para resolver una tarea, tal como se expuso en las Cápsulas 1 y 2 del Módulo 2 (*Análisis bioinformático sobre un problema en Ómicas*). Por este motivo, el término "modelo" implica una visión simplificada o condensada de los datos. Éste se construirá a partir de las variables que describen el problema, es decir, sus propiedades. Para crear dicho modelo, se pueden utilizar ecuaciones matemáticas, reglas tipo "si-entonces", similitud entre muestras, y muchas otras. El modelo se convierte por tanto en un "intermediario" para extraer el conocimiento que reside en los datos.

Machine Learning es el área de estudio que se encarga de diseñar algoritmos (una especie de programa informático) que sean capaces de aprender o construir modelos de manera totalmente **automática** a partir de los datos. Este concepto puede generar nuevas dudas, en concreto cómo un programa informático es capaz de realizar tareas de aprendizaje, siendo ésta una tarea que requiere de cierta capacidad racional. En realidad, el procedimiento es muy similar al de la cognición en los seres vivos, es

decir, en base a la práctica o la experiencia. ¿Y cómo dotamos de esta experiencia a nuestro sistema computacional? El principal recurso para realizar el aprendizaje en Machine Learning es, como ya habrás imaginado, a través de los datos.

Para que un algoritmo o técnica de Machine Learning sea capaz de procesar adecuadamente los datos, éstos deberán estar representados en forma tabular, es decir, como una matriz u "hoja de cálculo". En muchos problemas que se desean resolver, existe una variable de salida, que será el objetivo que se desea predecir, y que también puede ser categórico o numérico. En estos casos, se determina que el aprendizaje es de tipo "**supervisado**", mientras que cuando no existe tal variable de salida, el aprendizaje que se realiza es de tipo "**no supervisado**". Estudiaremos con más detalle ambos tipos de aprendizaje en las siguientes cápsulas del presente Módulo.

5. ALGORITMOS DE MACHINE LEARNING

Como se introdujo anteriormente, el Machine Learning se basa en algoritmos para construir modelos de datos. Un algoritmo no es más que una secuencia de operaciones para transformar una serie de entradas, en una salida. Imagina una receta de cocina, donde las entradas son los ingredientes, y la salida es el plato. Cada receta te permite obtener un plato diferente, incluso con los mismos ingredientes. En este caso particular, cuando se desea aplicar un algoritmo de Machine Learning sobre un problema concreto, se utiliza un software o programa informático. Este software contiene una serie de pasos escritos en un lenguaje de programación (R, Python, entre otros) que permiten aprender a partir de los datos de entrada un modelo para cualquier problema, y por tanto no es necesario diseñar o implementar uno distinto para cada caso de estudio.

Como se aprecia en la Figura 2, un algoritmo de Machine Learning recibe como entrada el conjunto de datos, y una serie de parámetros que nos permiten configurar el funcionamiento del algoritmo a las características de nuestro problema. La salida de este programa será justamente el modelo o función matemática que establece la relación entre las variables de entrada y la de salida. Así, podrá usarse posteriormente para analizar con detalle la información extraída de los datos, o hacer predicciones sobre nuevas muestras del mismo tipo que el conjunto de entrada disponible.

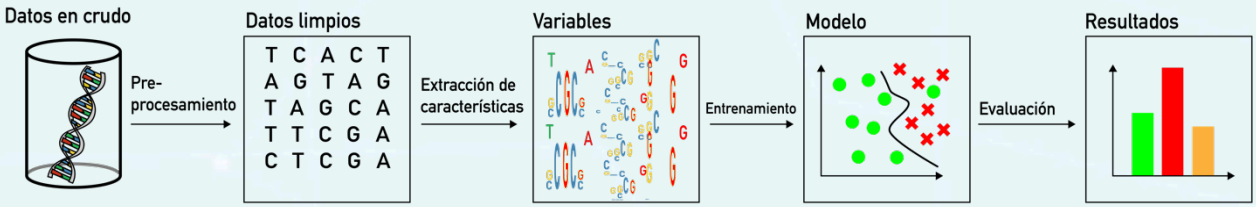


Figura 2. Flujo de un proceso en Machine Learning

No existe un programa único para aprender, sino que hay multitud de algoritmos diseñados a tal efecto, dependiendo tanto de la tarea que se quiera llevar a cabo (aprendizaje supervisado o no supervisado), como el tipo de modelo que se desea extraer (funciones matemáticas, reglas, o incluso redes neuronales). Estos programas se encuentran disponibles en diferentes plataformas para su uso directo. Por un lado, aparecen como bibliotecas o paquetes de distintos lenguajes de programación, como Scikit-Learn para Python, o Caret para R. Por otro lado, existe software de propósito específico para trabajar en modo ventana, como Knime, Weka, u Orange Data Mining (ver Figura 3).

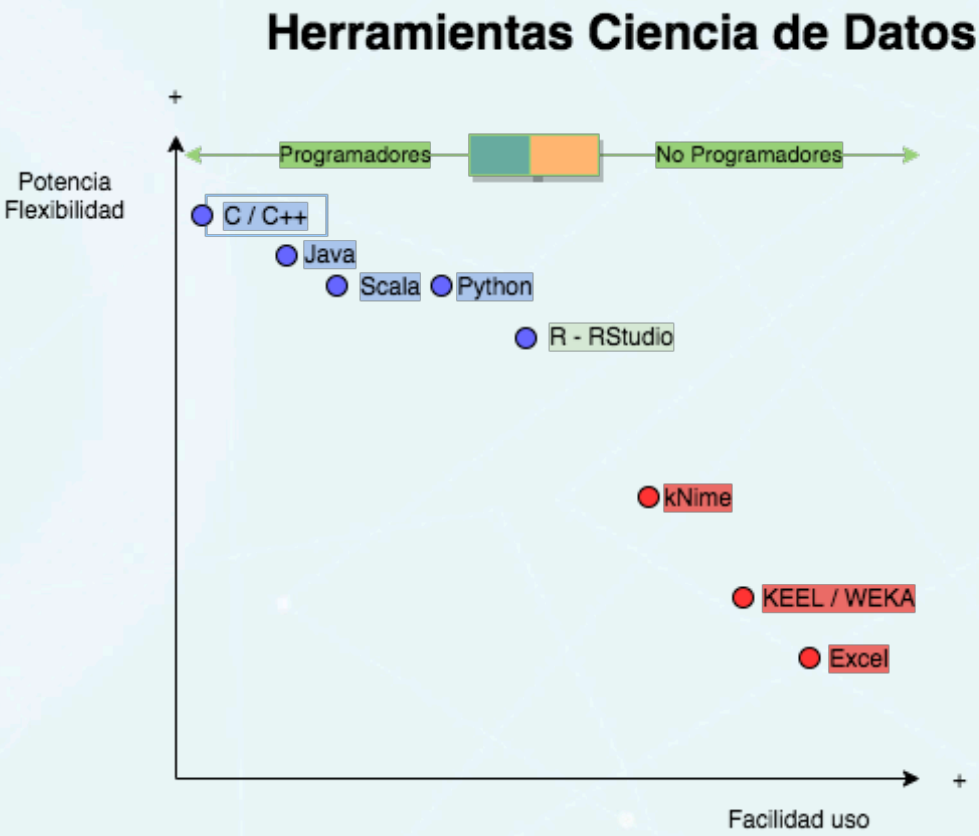


Figura 3. Herramientas para Ciencia de Datos

Utilizar las herramientas disponibles directamente sobre el lenguaje de programación, permite una mayor versatilidad y control sobre la solución que se pretende obtener. Como es natural, también implica cierta dificultad para aquellos usuarios que no posean un alto nivel sobre aplicaciones computacionales. Afortunadamente, estas herramientas software se han intentado hacer lo más amigables posibles al usuario, facilitando su utilización y ocultando al usuario detalles técnicos del algoritmo.

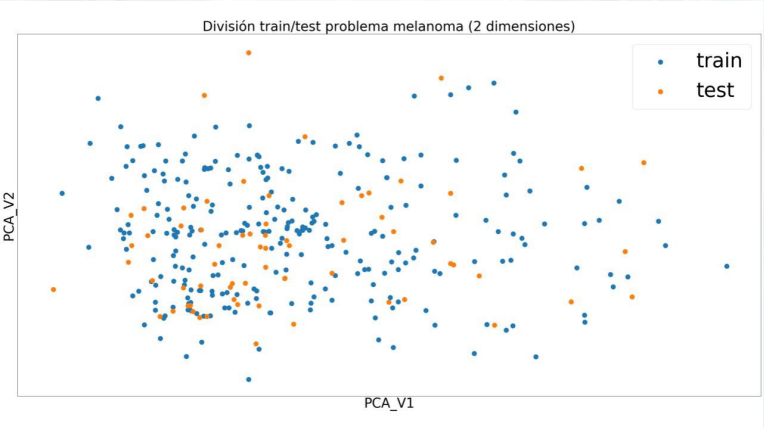
6. ENTRENAMIENTO Y VALIDACIÓN DE MODELOS DE MACHINE LEARNING

Es necesario recordar que en el ciclo de vida de la Ciencia de Datos se destacaron dos fases fundamentales en las que entran en juego los algoritmos de Machine Learning. Por un lado, el aprendizaje de modelos y, por otro lado, la validación de los mismos.

La fase de aprendizaje se denomina comúnmente como "entrenamiento" (*en inglés train*) y claramente tiene una relación directa con el mismo proceso en las personas. Cuando se quiere aprender un deporte o música, se trabaja en base a prueba y error, y se practica mucho para alcanzar el nivel de calidad que se desea. En Machine Learning estos conceptos de nuevo están asociados a los datos. Como regla general, cuantos más datos se tengan, las técnicas de Machine Learning obtendrán un modelo más robusto y preciso. Por ejemplo, conforme más información se recopile sobre una enfermedad, más probabilidades habrá de generar un programa de diagnóstico con un mayor acierto.

Para realizar el aprendizaje se utiliza lo que se denomina como **conjunto de entrenamiento**. Lo ideal es que dicho conjunto sea suficientemente representativo del problema que se desea resolver. Es muy importante evitar **sesgos en los datos**, ya que el algoritmo procesa directamente toda aquella información que se le suministra. En otras palabras, no es capaz de interpretar por sí solo si los datos no son adecuados. De este modo, datos de baja calidad implicarán la generación de modelos que no sean de calidad. Por este motivo resulta tan importante la fase de exploración de los datos, para anticiparnos a las posibles carencias que puedan contener. Imagina que todas las muestras de un estudio sobre covid-19 pertenecen a varones de más de 65 años; el algoritmo de Machine Learning estará acotado al tipo de información que ha aprendido, y dará respuestas quizá poco útiles para mujeres o pacientes jóvenes.

Cuando el aprendizaje es de tipo supervisado, una cuestión muy importante es determinar si el proceso de entrenamiento ha sido adecuado. Resulta de vital importancia realizar un proceso de validación del modelo, para garantizar que el modelo funciona correctamente cuando recibe datos nuevos. Para ello, se utiliza el llamado **conjunto de test**, que contiene instancias del problema original que no fueron incorporadas en la fase de entrenamiento. Esto quiere decir que, en aprendizaje supervisado, las instancias del conjunto de datos inicial se deben dividir en dos conjuntos disjuntos, entrenamiento y test. Se lanzará el modelo sobre cada instancia de test, y si su respuesta coincide con el valor de la variable de salida, podremos determinar si el aprendizaje se ha realizado de manera satisfactoria.



Cuando el aprendizaje es de tipo no supervisado, entonces el conjunto de test carece de sentido. Ciertamente, si no se dispone de una variable de salida que permita determinar la bondad del modelo, es necesario guiarse por otro tipo de análisis sobre el propio conjunto completo de datos. Estos detalles se estudiarán con mayor profundidad en la Cápsula 3 de este módulo, donde se introduce con más detalle el aprendizaje no supervisado.

Existen distintas medidas para estimar si el modelo obtenido se ajusta adecuadamente a los datos. Estas medidas dependen del tipo de técnica: aprendizaje supervisado o no supervisado. A lo largo de este módulo, se analizará con más detalle qué tipo de medidas existen y cómo se calculan. En general, todas buscan determinar el error que comete el modelo al predecir la salida ante nuevos datos de entrada.

REFERENCIAS BIBLIOGRÁFICAS

- Alpaydin, E. (2014). Introduction to Machine Learning. The MIT Press. ISBN: 0262028182, 9780262028189
- Han, J., Kamber, M., Pei, J. (2011). Data Mining: Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers. ISBN: 0123814790, 9780123814791
- Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. (2017). Data mining: practical machine learning tools and techniques. Amsterdam; London: Morgan Kaufmann. ISBN: 9780128042915 0128042915

REFERENCIAS ADICIONALES

- Alpaydin, E. (2016). Machine Learning: The New AI. MIT Press. ISBN: 9780262529518
- Maimon, O. & Rokach, L. (eds.) (2005). The Data Mining and Knowledge Discovery Handbook. Springer. ISBN: 0-387-24435-2
- Shalev-Shwartz, S., Ben-David, S. (2014). Understanding machine learning: from theory to algorithms. ISBN: 9781107057135 1107057132