Glosario del MOOC Big Data sin misterios

Índice

1.1.4 El Big Data como uno de los detonantes de la 4ª revolución industrial

- **Cuarta Revolución Industrial/Industria 4.0.** Es la tendencia actual de automatización y el intercambio de datos dentro de las tecnologías de manufactura.
- ***Big Data*.** Es un concepto que hace referencia a un conjunto de procesos, tecnologías y modelos basados en el almacenamiento masivo de datos, procesamiento y transformación de los mismos en conocimiento, para anticipar lo que sucederá en un mundo complejo y con muchas interacciones.
- **Inteligencia artificial (en inglés: *Artificial Intelligence* o "*AI*", por su sigla en inglés).** Es la inteligencia exhibida por máquinas. Una máquina inteligente ideal es un agente racional y flexible, que percibe su entorno y lleva a cabo acciones que maximicen sus posibilidades de éxito en algún objetivo o tarea. Coloquialmente, el término "inteligencia artificial" se aplica cuando una máquina imita las funciones "cognitivas" que los humanos asocian con otras mentes humanas, como, por ejemplo: "aprender" y "resolver problemas".
- I**nternet de las cosas (en inglés: "*Internet of Things*" o "*IoT*", por su sigla en inglés)**. Es un concepto que se refiere a la interconexión digital de objetos cotidianos con Internet, convirtiéndose así en objetos inteligentes.
- **Fabricación aditiva (FA).** Es un proceso por el cual un archivo 3D es convertido en un objeto físico, mediante la adición capa por capa de material. Este material puede ser plástico, resina, metal, papel y muchos más. El término más comúnmente usado como sinónimo de "fabricación aditiva" es "impresión 3D".
- **Realidad aumentada (RA).** Es el término que se usa para definir la visión de un entorno físico del mundo real, a través de un dispositivo tecnológico. Este dispositivo o conjunto de dispositivos añaden información virtual a la información física ya existente. De esta manera; los elementos físicos tangibles se combinan con elementos virtuales, creando así una realidad aumentada en tiempo real. La realidad aumentada es diferente de la realidad virtual: en la realidad aumentada el usuario percibe una mezcla de las dos

realidades (real y virtual). En cambio, en la realidad virtual, el usuario se aísla de la realidad material del mundo físico para sumergirse en un escenario o entorno totalmente virtual.

- **Redes inteligentes.** Es aquella que puede integrar de forma eficiente el comportamiento y las acciones de todos los usuarios conectados a ella, de tal forma que se asegure un sistema energético sostenible y eficiente, con bajas pérdidas y altos niveles de calidad y seguridad de suministro.
- ***Community manager*.** Es la persona encargada de gestionar, construir y moderar comunidades en torno a una marca en Internet. Esta profesión se perfila dentro de las empresas que descubren que las conversaciones sociales en línea son cada vez más relevantes y que necesitan un profesional que conozca sobre comunicación en línea, haciendo uso de los nuevos canales de comunicación a través de herramientas sociales.
- **Científico de Datos (en inglés: *Data Scientist*).** Dentro de una organización, es el encargado de extraer conocimiento y valor de los datos, siendo capaz de descubrir relaciones ocultas y patrones entre los datos de diferentes líneas de acción o áreas funcionales; y tecnología, mediante el uso de herramientas avanzadas para el análisis de datos. Son profesionales altamente cualificados que se encargan de interpretar y responder a las nuevas estrategias desde el punto de vista tecnológico y analítico. Deben tener fuertes conocimientos de estadística, matemática e informática, Al mismo tiempo, deberán adaptarse de manera sencilla a los entornos necesarios para el desarrollo de los modelos.

2.2.5 Aplicaciones de Big Data en el sector privado

- ***Logs* de navegación.** Los *logs* son aquellos archivos que contienen el registro de toda la información referente la navegación del usuario. Se envían al servidor y son procesados mediante software de analítica especializados.
- **Motor de recomendación.** Producto de la tecnología y de la necesidad de personalización de la oferta, nacen los "motores de recomendación". Éstos utilizan uno o más algoritmos para tomar decisiones en tiempo real basadas en datos conductuales. Deben procesar a gran velocidad información como datos de *cookies*, historial de transacciones, datos de navegación, parámetros de productos y otros datos externos.
- **Técnica de *collaborative filtering***: Esta es una de las técnicas que utilizan los sistemas de recomendación para resolver el problema de cómo saber qué elementos de un catálogo de productos se pueden recomendar a un usuario, basándonos en sus valoraciones de otros productos y en las valoraciones que todos los usuarios en general han ido emitiendo sobre el total del catálogo.
- **Microsegmentación.** La microsegmentación es una técnica que consiste en dividir al público objetivo de un producto o servicio en partes lo más pequeñas posibles, partes casi únicas; es decir, en segmentos realmente muy, muy pequeños.
- **Data Lake o Lago de Datos.** Un Data Lake es un repositorio de almacenamiento que contiene gran cantidad de datos en bruto (estructurados, semi-estructurados y desestructurados).
- ***Key Performance Indicator* (*KPI*, por su sigla en inglés).** También conocido como "indicador clave", "medidor de desempeño" o "indicador clave de rendimiento", es una medida del nivel del rendimiento de un proceso o producto concreto.
- **Analítica predictiva.** Es una forma de análisis avanzado, que utiliza datos nuevos e históricos para predecir la actividad futura, comportamientos y las tendencias.

- **Detección de anomalías.** Es una tarea de vital importancia en muchos sectores, como pueden ser la seguridad, finanzas o la salud; y consiste en identificar cuándo ocurre un suceso o evento que no es habitual, con el fin de realizar una acción asociada.

2.4.4 De la catástrofe a la acción: cómo *Twitter* puede salvar vidas

- **Variable numérica.** Es una magnitud cuantificable que puede tomar distintos valores numéricos dentro de un conjunto de números especificado (por ejemplo, números naturales, como: 1,2,3, entre otros.).
- **Algoritmo.** Conjunto de reglas que, aplicada sistemáticamente a unos datos de entrada apropiados, resuelven un problema en un numero finito de pasos elementales. Los algoritmos utilizados para el análisis de las redes sociales, permiten, por ejemplo, categorizar los mensajes publicados por los distintos usuarios, localizar grupos o analizar el sentimiento de las distintas opiniones.
- **Red neuronal convolucional.** Es un tipo de red neuronal artificial donde las neuronas corresponden a campos receptivos de una manera muy similar a las neuronas en la corteza visual primaria de un cerebro biológico.
- **Procesamiento del lenguaje natural.** Es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre los computadores y el lenguaje humano.

3.3.8 Gobierno del dato: la importancia de ordenar nuestro *Data Lake*

- **Gobierno del dato (en inglés: *Data Governance o Big Data Governance*).** Son estructuras y metodologìas para gestionar la información, que tienen por objeto asegurarse de que los datos son siempre fiables y válidos en cada contexto empresarial, que su calidad se mantiene a lo largo del tiempo y que existen mecanismos de control sobre quién puede hacer qué con los datos en cada momento. Para lograr estos objetivos es necesario establecer un conjunto de estándares, procesos y políticas que rijan el desarrollo y la utilización de los datos a nivel corporativo.

4.3.1 Data Wrangling: conceptos básicos y operaciones

- ***Data Wrangling*.** Es un proceso, que consiste en el correcto almacenamiento de los datos, su procesamiento y la preparación de los mismos para las consultas, hecho que permite disponer los datos con la mejor calidad y accesibilidad para su análisis.

4.4.3 Una breve historia del *Machine Learning*

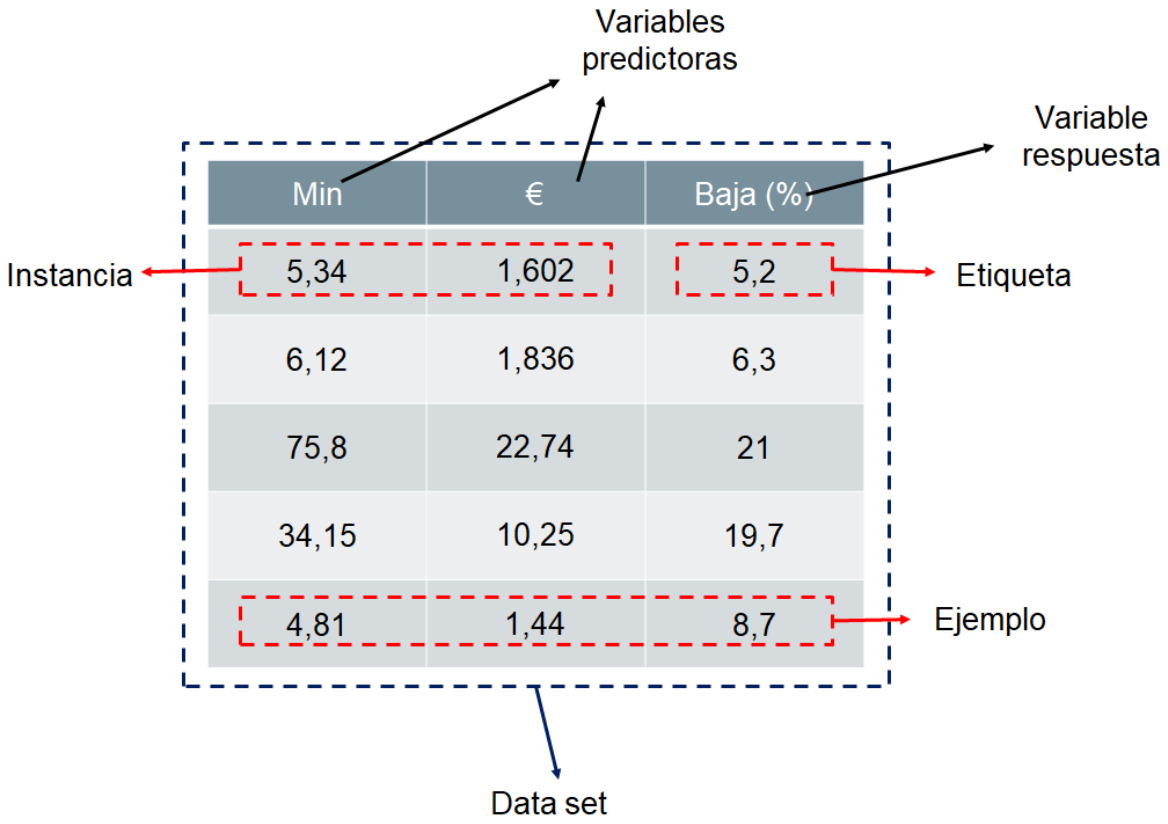
- **Aprendizaje automático (en inglés: *Machine Learning*).** El Aprendizaje automático (del inglés: *Machine Learning*) es una disciplina científica, procedente de la Inteligencia artificial, que estudia cómo los sistemas pueden ser programados para aprender y mejorar con la experiencia, sin intervención humana. Para abordar este problema, diariamente surgen nuevos paradigmas que permiten descubrir conocimiento a partir de datos específicos, basándose en sólidos principios estadísticos y computacionales.
- **Redes neuronales profundas (en inglés: *Deep Learning*).** Uno de los enfoques que está recibiendo más interés por parte de la comunidad científica hace referencia a las ***"Neural Networks"***. Estas redes se inspiran en el sistema nervioso animal, constituido por un sistema de células que colaboran entre sí para producir una respuesta a un estímulo externo. A medida que la topología de estos sistemas se complica, nos acercamos a lo que se conoce como ***"Deep Learning"***, un nuevo concepto marketiniano concebido para denominar a las "redes neuronales profundas". La idea bajo este nuevo paradigma es que, con un gran número de neuronas y muchos niveles de interconexión entre ellas, se pueden mejorar las predicciones en conjuntos de datos complejos.
- **Test de Turing.** Es una prueba de la habilidad de una máquina para exhibir un comportamiento inteligente similar al de un ser humano o indistinguible de este.
- ***Nearest Neighbor***. Es un algortimo, que ha sido ampliamente utilizado como una técnica no paramétrica en reconocimiento de patrones, debido a su sencillez y buen desempeño.

4.4.4 *Machine Learning:* técnicas básicas

- **Variable**. Una variable es una magnitude que puede tomar diferentes valores dentro de un conjunto posible de valores definidos para esa magnitud. Los datos, que constituyen información sobre un hecho y que permiten su conocimiento, se emplean como información de entrada en las variables correspondientes a un modelo.
- **Conjunto de datos (en inglés: *Data set*).** Es una colección de datos correspondiente a una o varias variables, habitualmente presentados en forma de tabla.
- **Atributo.** Es una variable del conjunto de datos o *data set*, incluyendo sus valores concretos.
- **Instancia.** Es un conjunto de todos los atributos o variables de un conjunto de datos o *data set*.
- **Variable predictora.** Es un atributo o atributos de partida en el estudio. Constituye la variables que se emplean como entrada para realizar la predicción.
- **Variable respuesta.** Es un atributo objetivo del estudio. Constituye la variable que se desea predecir.
- **Etiqueta o dato etiquetado.** Es el valor real concreto de la variable respuesta, obtenido a partir de un hecho sucedido en la realidad. Se emplea en el desarrollo de modelos de aprendizaje supervisado.
- **Dato no etiquetado.** Es un dato "no clasificado", es decir, que no tiene una etiqueta.
- **Ejemplo.** Es la unión de la instancia con su etiqueta correspondiente.

Para ilustrar los conceptos anteriores, se propone un ejemplo perteneciente a una compañía telefónica. En este ejemplo, se consideran datos de minutos ("min") y euros ("€"), a partir de los que se desea predecir la probabilidad de dada de baja de la compañía. Siguiendo la terminología vista, los "atributos" son las variables minutos ('mins') y euros ('€'); la "instancia" es el conjunto de todos los atributos, en este caso los minutos y euros; y la variable "respuesta"

es la variable “darse de baja” o “baja”. Las “etiquetas” son cada uno de los valores concretos de la variable baja (5,2; 6,3; 21; 19.7; 8.7), que es la variable respuesta a predecir. Por último, los “ejemplos” son cada una de las filas completas, lo que sería la unión de la instancia con su etiqueta correspondiente.



- **Modelo o “Hipótesis”.** Es la estructura obtenida tras el proceso de aprendizaje y resume un conjunto de datos para predicción o para clasificación.
- **Métrica.** Es la medida utilizada para evaluar el desempeño del modelo aprendido como, por ejemplo, la precisión o la cantidad de errores cometidos.
- **Objetivo.** Es la métrica que optimiza el algoritmo de aprendizaje.
- **Regresión lineal.** Es un modelo estadístico que permite predecir el valor de una variable cuantitativa (numérica), como una función lineal de las variables de entrada o predictores.
- **Regresión logística.** Es un modelo de regresión generalizado, utilizado como método de clasificación binaria, puesto que, en lugar de valores

numéricos, éste permite estimar la probabilidad de que ocurra (o no) un evento como función de otras variables.

- **Métodos basados en árboles.** Los métodos basados en árboles consisten en segmentar el espacio de predictores en varias regiones. Dentro de cada región, se utiliza la media o la moda de las observaciones de entrenamiento en esa región para hacer la predicción.

- ***Kmeans*.** Es un método que permite crear clusters de datos numéricos. La entrada del sistema son las mediciones numéricas de interés y la salida son los centroides de los *clusters* resultantes y la asignación de cada dato a un *cluster* determinado.

5.4.2 De la sensorización a la industria

- **Predicción de la demanda (en el sector industrial).** Es una actividad considerada de máxima importancia en el flujo de suministro entre proveedores y clientes. De hecho, es el primer paso para optimizar la cadena de producción. La empresa cliente necesita datos para estimar qué cantidades de producto debe adquirir con el objetivo de atender la demanda de sus consumidores, y el proveedor necesita datos para estimar el tiempo necesario para atender las necesidades requeridas por su cliente en función de sus procesos productivos. Esto se consigue mediante el análisis de los datos históricos recogidos en los sistemas de información.
- **Mantenimiento predictivo (en el sector industrial).** Su principal objetivo es minimizar el tiempo no productivo, a través de la instrumentación y el análisis de los datos, para identificar patrones que indican un fallo inminente. Para las cadenas de producción, la inactividad produce importantes pérdidas de ingresos. A través de la analítica avanzada, somos capaces de predecir un comportamiento anómalo, adelantándonos a posibles fallos y reduciendo, así, los tiempos de parada y de mantenimiento.

5.4.4 *Cloud computing*

- ***Cloud Computing*.** Es un modelo de servicios de *IT* (Informática), que pone a disposición de los usuarios un servicio, bajo demanda, de acceso a una red de ordenadores y servidores; dentro de un pool compartido de recursos (red, servidores, almacenamiento, aplicaciones y servicios).
- **Infraestructura como servicio (en inglés: *Infraestructure as a Service*. *"IaaS"*, por su sigla en inglés).** Consiste en la puesta a disposición, por parte de la empresa oferente, de sus servidores (hardware) para que otras empresas puedan utilizarlos como quieran; a través de un entorno virtualizado, que suele ser Internet.
- **Plataforma como servicio (en inglés: *Platform as a Service*. *"PaaS"*, por su sigla en inglés).** Supone la puesta a disposición de una plataforma y un entorno Big Data, a los que el cliente puede acceder por Internet y realizar sus propios desarrollos.
- **Software como servicio (en inglés: *Software as a Service*. *"SaaS"*, por su sigla en inglés).** Consiste en la puesta a disposición de aplicaciones software, a través de Internet (alojadas en la nube). Se trata de aplicaciones que pueden utilizarse para un amplio abanico de tareas, existiendo, incluso, soluciones empresariales para la gestión de recursos tales como la contabilidad, facturación, planificación, comunicaciones, etc.

6.3.2 Tableau: herramienta de creación de *dashboards*

- **Tableau.** Es una herramienta de creación de dashboards que ofrece una experiencia de exploración visual altamente interactiva e intuitiva para el negocio. Los usuarios puedan acceder fácilmente, preparar y analizar sus datos; sin necesidad de programar a través de sus productos.

6.3.4 Geovisualizaciones: CARTO

- **Carto.** Es una plataforma online (*Software as a service* o *SaaS*) que permite a los usuarios cargar datos y proporciona herramientas para visualizarlos en mapas interactivos. Los mapas realizados en CARTO pueden ser compartidos o incrustados en sitios web o aplicaciones.

6.3.7 Otras herramientas de visualización

- **Qlik.** *Qlik* ofrece *Data Discovery* y Análisis de datos gobernados ya sea como una aplicación independiente o (cada vez más) incrustado en otras aplicaciones. *QlikSense* es el producto líder de la compañía, mientras que *QlikView* continúa siendo mejorado y constituye una porción mayor de la base de clientes.
- **Power BI.** Microsoft ofrece una amplia gama de capacidades de BI y análisis con su suite Power BI, servido a través de su plataforma en la nube Azure. Power BI ofrece preparación de datos, descubrimiento de datos y paneles interactivos a través de una sola herramienta de diseño.
- **MicroStrategy.** *MicroStrategy* Versión 10 (lanzado en 2015) combina la preparación de datos de autoservicio, datos visuales, data-discovery y gran exploración de datos con BI empresarial.
- **Kibana.** *Kibana* es una herramienta open-source perteneciente a *Elastic*, que nos permite visualizar y explorar datos que se encuentran indexados en *ElasticSearch*, es decir, un plugin de *ElasticSearch*.
- **Ggplot2.** Es un paquete de código abierto programado sobre el lenguaje R, el cual permite generar gráficos 2D basados en capas, que se pueden adicionar en infinitas combinaciones, empoderando al científico para que pueda realizar gráficos imposibles para otros programas tradicionales.
- **Shiny.** Es un *framework* de *RStudio* para construir aplicaciones web programado en el lenguaje R.
- **Matplotlib.** Es una biblioteca para la generación de gráficos a partir de datos contenidos en listas o arrays en el lenguaje de programación *Python* y su extensión matemática *NumPy*.
- **D3.js.** Es una biblioteca JavaScript para manipular documentos basados en datos. D3 le ayuda a llevar datos a la vida usando HTML, SVG y CSS. El énfasis de D3 en los estándares web le brinda las capacidades completas de los navegadores modernos sin atarse a un marco propietario, combinando poderosos componentes de visualización y un enfoque basado en datos para la manipulación DOM.

- **DC.js.** Es una biblioteca de gráficos *javascript* con soporte de *crossfilter* nativo, que permite una exploración altamente eficiente en grandes conjuntos de datos multidimensionales.