Trabajo de Fin de Grado

“La analítica de Big Data en el sector sanitario”

Germán Linares Vallejo

Grado de Ingeniería Informática

Xavier Martínez Fontes

11 de enero de 2016

FICHA DEL TRABAJO FINAL

Título del trabajo:	"La analítica de Big Data en el sector sanitario"
Nombre del autor:	Germán Linares Vallejo
Nombre del consultor:	Xavier Martínez Fontes
Fecha de entrega (mm/aaaa):	01/2016
Área del Trabajo Final:	Business Intelligence
Titulación:	Grado de Ingeniería Informática

Resumen del Trabajo (máximo 250 palabras):

El presente trabajo tiene como objetivos realizar un análisis de las posibilidades que ofrece Big Data al sector sanitario a nivel global y del estado actual de la adopción de esta tecnología en nuestro sistema sanitario. Para lograr ambos objetivos se ha seguido un enfoque metodológico basado preferentemente en la revisión y estudio de recursos bibliográficos. La exposición de contenidos se ha estructurado en los siguientes apartados:

— Descripción del concepto de Big Data desde una doble vertiente, que se refiere tanto a las características de los datos masivos como a las herramientas tecnológicas que posibilitan su captura, almacenamiento y análisis.

— Definición de analítica predictiva. Técnicas, modelos y métodos de analítica predictiva. Fases de un proyecto de construcción y mejora de un modelo predictivo. Herramientas de software para el análisis predictivo.

— Adopción empresarial y estrategias de integración de Big Data en un sistema de información para trabajar de forma complementaria con una solución Business Intelligence.

— Posibilidades que ofrece Big Data al sector sanitario, reparando en la práctica clínica y en la investigación biomédica. Situación actual de la adopción de esta tecnología en nuestro sistema sanitario, considerando los retos y barreras que se deben afrontar y superar, particularmente en la práctica clínica. Presentación de algunos casos de éxito.

— En el último apartado se exponen las conclusiones del trabajo realizado. Entre ellas, cabe destacar que la adopción de Big Data está consiguiendo ya logros interesantes en el campo de la investigación biomédica.

Palabras clave:

Big Data, analítica predictiva, sistema sanitario, Business Intelligence, práctica clínica, investigación biomédica

Abstract (in English, 250 words or less):

This study analyzes the possibilities Big Data offers the healthcare sector on a global level as well as the present state of its implementation in the Spanish healthcare system. Our methodology is based on the review and study of existing literature. The contents are structured as follows:

- Description of the concept of Big Data from two aspects, which refers both to the characteristics of the big *data* as to *technological tools* that enable their capture, storage and analysis.
- Definition of predictive analytics. Techniques, models and predictive analytics methods. The stages of a project to construct and enhance a predictive model. Software tools for predictive analytics.
- Implementation of the model and strategies for the integration of Big Data into an information system that complements a Business Intelligence solution.
- The possibilities Big Data offers the healthcare system from the double standpoint of clinical practice and biomedical investigation. The present status of the implementation of this technology in the Spanish healthcare system, taking into account the challenges to be faced and the barriers to be overcome, especially in the field of clinical practice. Presentation of some successful cases.
- The last section contains our conclusions. Among others, it is worth noting that the use of Big Data technology is already obtaining interesting results in the biomedical research.

Key words:

Big Data, predictive analytics, healthcare system, Business Intelligence, clinical practice, biomedical research

Índice

Lista de figuras

1. Introducción

1.1. Contexto y justificación del Trabajo

La expansión reciente de las últimas tendencias en Tecnologías de la Información (Redes Sociales, Movilidad, Cloud Computing, Big Data e Internet de las Cosas) supone importantes avances, pero también grandes desafíos, que exigen un considerable esfuerzo de adaptación por parte de las organizaciones.

El área de la salud es un sector donde la implantación de estas tecnologías ofrece mayores oportunidades. No obstante, para que éste pueda beneficiarse de las posibles ventajas de su adopción es necesario previamente realizar una serie de cambios en la mayoría de los sistemas de información de los servicios sanitarios que existen en la actualidad. Estos cambios deben estar orientados, entre otros objetivos, a conseguir gestionar y analizar grandes volúmenes de datos procedentes de fuentes muy diversas y registrados en formatos muy heterogéneos.

Los médicos y el personal sanitario pueden recopilar datos de los pacientes a partir de historias clínicas electrónicas, aparatos de telemedicina, análisis clínicos o dispositivos portátiles que monitorizan constantes vitales. Por otra parte, la investigación biomédica también aporta numerosos datos relacionados con la salud de los pacientes, como es el caso de la genómica, que hace posible obtener el perfil genético de una persona y, consecuentemente, predecir el riesgo de padecer determinadas enfermedades. Otros datos relevantes pueden provenir de encuestas de salud pública, estudios epidemiológicos o resultados procedentes de ensayos clínicos que evalúan la eficacia y seguridad de un determinado fármaco.

La capacidad de Big Data de capturar, gestionar y procesar datos masivos en un tiempo razonable, unida al uso de técnicas estadísticas y de minería de datos orientadas al análisis predictivo, puede ofrecer información valiosa que, a su vez, sea utilizada en numerosas aplicaciones dentro del campo de la medicina (prevención de enfermedades, descubrimiento de nuevos fármacos, diagnósticos, tratamientos, mejora de la atención al paciente, reducción de costes sanitarios...).

No obstante, para que los usuarios que forman parte de los diferentes entornos sanitarios puedan tomar las mejores decisiones a partir de esta información, es necesario también disponer de una infraestructura de Business Intelligence (BI), que integre las herramientas adecuadas para facilitar la extracción y visualización de información de Big Data. En este sentido, los principales proveedores de software BI (Oracle, IBM, Microsoft, Pentaho...), conscientes de las limitaciones que presentaban las soluciones tradicionales, han comenzado a incorporar este tipo de herramientas en sus últimas versiones.

1.2. Objetivos del Trabajo

En este trabajo se plantean los siguientes objetivos:

- Definir los fundamentos de Big Data y las principales técnicas de analítica avanzada que puede incorporar esta tecnología para operar sobre datos masivos. Dentro de estas técnicas se tratará de profundizar en el concepto de analítica predictiva, debido a la importancia que adquiere su aplicación dentro del sector de la salud.
- Explicar las diferencias y la posible complementariedad entre Business Intelligence y Big Data, así como las posibilidades de integración de esta última tecnología dentro de un sistema de información.
- Analizar las posibilidades que ofrece la adopción de Big Data en diferentes áreas del sector sanitario, especialmente como apoyo a la operativa clínica y a la investigación biomédica, describiendo algunos casos de éxito.
- Exponer las principales conclusiones que se hayan podido deducir del estudio realizado, tratando de presentar una visión realista sobre el estado actual de la adopción de Big Data en el sector de la salud y las limitaciones que existen en nuestro sistema sanitario para implantar soluciones basadas en esta tecnología.

1.3. Enfoque y método seguido

De acuerdo con los objetivos planteados en el punto anterior y las características de este trabajo, se ha procurado enfocar el estudio partiendo de una amplia revisión bibliográfica sobre los temas que se han propuesto para su desarrollo, de modo que se

pueda obtener una visión actualizada de los mismos antes de proceder a la redacción de cada apartado.

El método seguido comienza por un proceso de búsqueda, especialmente en la web, de los recursos bibliográficos específicos (libros, artículos en revistas o en blogs especializados, informes y otros materiales) que han de servir para documentar cada uno de los aspectos que se han de abordar en el proyecto. De las fuentes de información encontradas se seleccionan aquellas que tengan mayor valor e interés. Una vez analizados los contenidos, se procede a clasificar los documentos para acceder a ellos y poder ser utilizados y citados debidamente cuando convenga en la redacción del trabajo. A partir de ese momento, figurarán en la bibliografía general.

Por otro lado, conviene dejar constancia de la dificultad que ha supuesto recabar información valiosa sobre el tema principal del trabajo, Big Data en el sector sanitario, debido a su enorme dispersión. Por esta razón, la fase de búsqueda y selección bibliográfica en torno a esta materia ha sido un proceso continuo y laborioso a lo largo del proyecto.

1.4. Planificación del Trabajo

El trabajo se ha distribuido en cuatro bloques, los tres primeros coincidiendo con los límites temporales y los hitos parciales para la realización y entrega de las PECs, más un cuarto bloque para la realización de la Memoria y Presentación virtual y el hito de entrega final. La primera actividad en cada uno de los tres primeros bloques será la de búsqueda, selección y estudio de la bibliografía. No obstante, aunque en el cronograma tenga un tiempo reservado para estas tareas, teniendo en cuenta lo expuesto en el apartado anterior, el proceso se realizará de forma continuada.

El bloque primero corresponde temporalmente a la realización de la PEC1, con una duración de dos semanas. En él se define la propuesta del trabajo que se va a realizar. A partir de la información preliminar, se describe el estado de la cuestión, los objetivos, el alcance y se establece la planificación detallada, en tareas e hitos, para este y los restantes bloques.

En el bloque segundo, PEC2, con una duración de 5 semanas, el tiempo se reparte inicialmente por igual entre cuatro tareas, la primera de estudio de documentación y

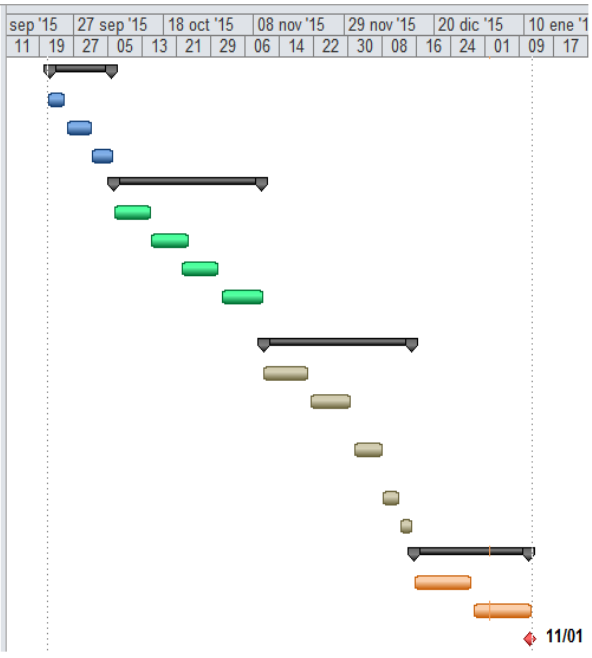
las otras tres de exposición de contenidos: “Introducción al concepto de Big Data”, “Analítica de Big Data” y “Adopción e integración de Big Data y BI en las organizaciones”.

El bloque tercero, PEC3, también de 5 semanas de duración, comprende cinco tareas. En esta parte se prevé un predominio de documentación en inglés para abordar la tarea de redacción de las “Posibilidades que ofrece Big Data en el sector sanitario”. Es por ello que a estos dos apartardos se les haya dado un mayor peso (semana y media para cada uno). Las otras tres tareas están dedicadas a exponer la “Adopción de Big Data en nuestro sistema sanitario”, la “Presentación de casos” y la redacción del “Primer borrador de conclusiones”.

Para el bloque final, de cuatro semanas, se distribuyen por igual las dos principales actividades: la elaboración de la “Memoria final” y la preparación de la “Presentación Virtual”.

Tal como se han descrito, las tareas quedan reflejadas en el diagrama de Gantt que se presenta a continuación. No obstante, cabe añadir que el tiempo de dedicación al proyecto será relativamente variable en el día a día, dependiendo de los compromisos puntuales con otras asignaturas, aunque es esperable que en el cómputo semanal se alcance un mínimo de 25 horas (un promedio de 3 a 4 horas/día).

	Nombre de tarea	Dura	Comienzo	Fin
1	**PEC 1 (Plan de Trabajo)**	**15 días**	**lun 21/09/15**	**lun 05/10/15**
2	Búsqueda de información preliminar	4 días	lun 21/09/15	jue 24/09/15
3	Definición del trabajo	6 días	vie 25/09/15	mié 30/09/15
4	Planificación	5 días	jue 01/10/15	lun 05/10/15
5	**PEC 2**	**36 días**	**mar 06/10/15**	**lun 09/11/15**
6	Búsqueda información	9 días	mar 06/10/15	mié 14/10/15
7	Introducción al concepto de Big Data	9 días	jue 15/10/15	vie 23/10/15
8	Análisis predictivo aplicado a Big data	9 días	jue 22/10/15	vie 30/10/15
9	Adopción e integración de Big Data y BI en las organizaciones	10 días	sáb 31/10/15	lun 09/11/15
10	**PEC 3**	**36 días**	**mar 10/11/15**	**lun 14/12/15**
11	Búsqueda de información	11 días	mar 10/11/15	vie 20/11/15
12	Posibilidades de Big Data en el sector sanitario	10 días	sáb 21/11/15	lun 30/11/15
13	Adopción de Big Data en nuestro sistema sanitario	7 días	mar 01/12/15	lun 07/12/15
14	Presentación de casos	4 días	mar 08/12/15	vie 11/12/15
15	Redacción de conclusiones	3 días	sáb 12/12/15	lun 14/12/15
16	**Memoria Final y Presentación Virtual**	**28 días**	**mar 15/12/15**	**lun 11/01/16**
17	Elaboración Memoria Final	14 días	mar 15/12/15	lun 28/12/15
18	Elaboración Presentación Virtual	14 días	mar 29/12/15	lun 11/01/16
19	**Entrega final**	**0 días**	**lun 11/01/16**	**lun 11/01/16**



1.5. Breve sumario de productos obtenidos

La Memoria del Trabajo y la Presentación virtual del Trabajo serán los productos resultantes.

1.6. Breve descripción de los otros capítulos de la memoria

En los capítulos restantes de la memoria se desarrollan los siguientes contenidos, en relación con el trabajo global.

En el primero de ellos, capítulo 2, se plantea una introducción al concepto de Big Data, bajo la doble vertiente de la descripción de los principales rasgos que distinguen los "datos masivos" y de la utilización de las nuevas herramientas tecnológicas que permiten su captura desde diferentes fuentes y sistemas, así como su transformación, almacenamiento, análisis y visualización, de modo que sirvan para adquirir conocimiento y toma de decisiones.

El capítulo 3 se ocupa de la analítica predictiva, o analítica avanzada de Big Data, aplicación muy importante dentro del sector de la salud. Se define conceptualmente como una rama de la minería de datos que permite operar sobre los "datos masivos" con el propósito de construir modelos predictivos de calidad. Se presentan sus técnicas, modelos y métodos; las fases de un proyecto de construcción y mejora continua de modelos; así como las soluciones disponibles en el mercado y las principales aplicaciones.

En el capítulo 4 se plantean algunas cuestiones generales relacionadas con la adopción e integración de Big Data y Business Intelligence en las organizaciones, desglosadas en varios puntos, entre los que destacan las fases de adopción empresarial de Big Data y las estrategias de integración de Big Data dentro de una solución BI.

El capítulo siguiente, el más extenso, está dedicado íntegramente a Big Data en el ámbito sanitario, y se encuentra dividido en tres apartados. A partir de dos estudios fundamentales, los informes *Big Data: The Next Frontier for Innovation, Comptetition and Productivity* y *Big Data Healthcare Hype and Hope,* en el primer apartado se analizan las posibilidades de la analítica de Big Data en el sector de la salud. En el

siguiente, se expone el estado en que se encuentra actualmente su adopción en nuestro entorno sanitario. A continuación se presentan algunos casos de éxito.

Por último, en el capítulo sexto se presentan las conclusiones resultantes del trabajo realizado.

2. Introducción al concepto de Big Data

En este capítulo se define el concepto de Big Data bajo un doble significado: el de la descripción de los principales rasgos que distinguen los denominados “datos masivos” y el de las herramientas tecnológicas que permiten manejar esos datos de modo que sirvan para adquirir conocimiento y tomar decisiones.

2.1. ¿Qué se entiende por Big Data?

Durante los últimos años se ha intentado explicar la complejidad que encierra el término Big Data o “datos masivos”, aludiendo tanto a las características de los datos como a las herramientas de las tecnologías de la información para poder manejarlos en provecho de las organizaciones. En palabras de Mayer y Cukier, “el volumen de información había aumentado tanto que la que se examinaba ya no cabía en la memoria que los ordenadores emplean para procesarla, por lo que los ingenieros necesitaban modernizar las herramientas para poder analizarla” (2013, pág. 17). Esta sería la razón para impulsar el desarrollo de nuevas tecnologías que lo hicieran posible.

A pesar de la importancia y popularidad que ha alcanzado el fenómeno Big Data y la abundante bibliografía que se ha originado en torno al tema, según ha señalado Joyanes Aguilar, “no existe unanimidad en la definición de Big Data, aunque sí un cierto consenso en la fuerza disruptiva que suponen los grandes volúmenes de datos y la necesidad de su captura, almacenamiento y análisis”. (2014, pág. 2). En su estudio selecciona algunas de las definiciones más significativas propuestas por distintas instituciones relevantes en este ámbito. Entre ellas figura la de la consultora tecnológica IDC:

> “Big Data es una nueva generación de tecnologías, arquitecturas y estrategias diseñadas para capturar y analizar grandes volúmenes de datos provenientes de múltiples fuentes heterogéneas a una alta velocidad con el objeto de extraer valor económico de ellos”.

Otra de las seleccionadas por Joyanes es la que propone la consultora Gartner:

> “Big Data son los grandes conjuntos de datos que tienen tres características principales: volumen (cantidad), velocidad (velocidad de creación y utilización) y variedad (tipos de fuentes de datos no estructurados, tales como la interacción

> social, vídeo, audio, cualquier cosa que se pueda clasificar en una base de datos". (2014, pág. 3). [1]

La pregunta ¿qué es Big Data? sigue suscitando nuevas respuestas con la intención de matizar mejor el concepto. Así, para Barranco Fragoso (2012), el término Big Data se refiere a "la tendencia en el avance de la tecnología que ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semiestructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a una base de datos relacional para su análisis."

El interés, cada vez más generalizado, por desentrañar lo que significa Big Data y la importancia que puede tener su adopción para las organizaciones no cesa. Así lo evidencian los numerosos artículos, informes y monografías publicados sobre el tema, especialmente a partir de 2012.

2.2. Características de Big Data

Como ya se ha anticipado, las características principales que definen los datos masivos son las llamadas originalmente las "tres V", las dimensiones volumen, velocidad y variedad. Sin embargo, con posterioridad se han añadido algunas más, como veracidad y valor.

Volumen. Es la característica que mejor se identifica con las cantidades masivas de datos que deben ser procesados y analizados mediante la tecnología Big Data. La cuestión es a partir de qué volumen se considera necesario emplear esta tecnología. En el informe de IBM *Analytics: el uso de big data en el mundo real*, que recoge los resultados de la encuesta "Big Data @ Work Survey" realizada en 2012, algo más de la mitad de los 1.144 profesionales encuestados opinaron que conjuntos de datos entre un terabyte [10^{12} bytes] y un petabyte [10^{15} bytes] ya deben considerarse Big Data. No obstante, todos los participantes estaban de acuerdo en que debería ser un "volumen alto", que en un futuro lo sería aún más. Se espera que entre 2015 y 2020 se pase del límite del exabyte [10^{18} bytes] al zettabyte [10^{21} bytes].

[1] Estas características de Big Data, conocidas como "las tres Vs" fueron definidas por primera vez por Doug Laney en su artículo publicado en febrero de 2001 titulado "3-D Data Management: Controlling Data Volume, Velocity and Variety". El propio Laney lo explica, en otro artículo más reciente, de febrero de 2012. Véase la referencia en la bibliografía final.

Variedad. Se refiere a los diversos tipos y fuentes de datos que Big Data es capaz de admitir y gestionar. Los tipos de datos, además de incluir los estructurados (que son los ya incorporados a los *data warehouse* corporativos tradicionales, bien definidos para su gestión en bases de datos relacionales), pueden ser también semiestructurados y no estructurados. Cuando se han de analizar juntos exigen el empleo de las nuevas técnicas.

En cuanto a las fuentes de los datos también son diversas, y en una proporción mayoritaria los datos suelen ser no estructurados. Dichas fuentes pueden ser internas o externas a la organización, tradicionales o no: de operaciones transaccionales, generados por personas (llamadas telefónicas, correos electrónicos, documentos digitalizados...), información obtenida de redes sociales, información biométrica, de máquina a máquina (conexión con otros dispositivos, como sensores o medidores), etc. Más adelante se tratará este punto con un poco más de detalle.

Velocidad. Esta cualidad se asocia a los datos en movimiento, es decir, a la rapidez del flujo de los datos, desde que se crean hasta que se presentan como información útil para la toma de decisiones, pasando por las distintas fases intermedias (almacenamiento en la base de datos, procesamiento y análisis). Para determinados propósitos, como la detección del fraude o el marketing multicanal instantáneo, puede resultar crítico que la rapidez con la que se realicen estos procesos se acerque al tiempo real.

Veracidad. Esta característica, añadida por IBM, se refiere a la incertidumbre que comporta la presencia de cierto tipo de datos que se encuentran en el conjunto extraído. Por mucho esfuerzo que se dedique a la limpieza de los mismos para obtener una mejor calidad, en algunos de ellos no se puede suprimir ni la imprevisibilidad (del tiempo, de la economía…) ni la incertidumbre (de los sentimientos, de la sinceridad de las personas o de sus reacciones ante determinados hechos). No obstante, esta dimensión de los datos masivos debe ser tenida en cuenta. Los directivos necesitan abordar esta incertidumbre y determinar cómo planificarla para el beneficio de su organización.

Valor. Big Data podría entenderse conceptualmente como una combinación de las dimensiones anteriormente mencionadas. Cada empresa u organización podrá adoptar esta tecnología bajo diferentes enfoques, pero con un objetivo común: mejorar el rendimiento y la toma de decisiones. En definitiva, se trataría de obtener valor mediante el

uso estratégico de la información que pueden proporcionar el proceso y análisis de los grandes conjuntos de datos que se han ido almacenando en su entorno.

Mayer y Cukier, en su documentado estudio (2013), dedican todo un capítulo a poner en evidencia el valor de los datos masivos. Consideran básicamente que dicho valor no ha de verse sólo a la luz del que tienen en su uso primario, y que justifica su almacenamiento, sino también teniendo en cuenta las posibilidades que ha abierto la tecnología para su reutilización, pudiéndose procesar una y otra vez con propósitos múltiples. Dicho de otro modo, "los datos pasan de unos usos primarios a otros secundarios, lo que los vuelve mucho más valiosos a lo largo del tiempo" (pág. 131). Proponen tres vías para desencadenar el llamado "valor de opción" de los datos: la reutilización básica, la fusión de conjuntos de datos y el hallazgo de combinaciones "dos por uno" (pág. 132).

2.3. Tipos de datos

En función del manejo de los datos masivos, se suelen distinguir tres tipos: los datos estructurados, los semiestructurados y los no estructurados.

Los **estructurados** son los que tienen un esquema definido, en formato y longitud, para poder ser incluidos en un campo fijo (fechas, números, cadenas de caracteres, etc.) y almacenados en tablas, por ejemplo las de una hoja de cálculo o una base de datos relacional. Los datos **semiestructurados** carecen de formato fijo o de campo determinado, pero están dotados de marcadores que permiten diferenciar los distintos elementos dato. Un ejemplo lo constituyen las etiquetas de lenguajes HTML y XML. En el ámbito de la salud, ejemplo de estas modalidades serían los datos de contabilidad, facturación electrónica, algunos datos de actuario o datos clínicos.

Los datos **no estructurados** no tienen un formato específico ni se pueden asignar a un campo fijo, por lo que no es posible su almacenamiento en una tabla. Se tratan como documentos u objetos. Ejemplos de este tipo de datos son documentos de audio, vídeo, fotografías, e-mails o archivos PDF. En el sector sanitario, cabe citar imágenes de radiografías, resonancias magnéticas, recetas en papel, etc.

2.4. Fuentes de los datos masivos

Atendiendo a los datos que las empresas deben analizar, según diferentes propósitos y con arreglo a las fuentes donde se originan, Sunil Soares (2012) estableció una clasificación, que se ha convertido en una referencia, en la que distingue cinco categorías básicas de fuentes de datos, comprendiendo cada una de ellas distintos tipos de información. La que se presenta a continuación se basa en la versión publicada por Barranco Fragoso (2012). A los ejemplos que cita se incluirán algunos referidos al ámbito sanitario:

- ***Web and Social Media*.** Incluye datos procedentes de los flujos de clicks, entradas de Twitter, Facebook, LinkedIn, blogs y diversos contenidos Web. Por ejemplo, Big Data puede recoger información cada vez más abundante para el sector de la salud desde este tipo de fuentes, como las redes sociales temáticas para profesionales médicos o para comunidades virtuales de pacientes.[2]

- ***Machine-to-Machine* (M2M)**. Se refiere a las tecnologías que permiten conectarse a otros dispositivos, como sensores o medidores que capturan un evento en particular (humedad, velocidad, temperatura, presión, etc.). Entre los datos procedentes de estos dispositivos se encuentran: lecturas de medidores inteligentes, de RFID, de sensores de plataformas petroleras y señales GPS. En servicios de salud, los sistemas que recogen los datos procedentes de sensores en dispositivos *weareables* o de smartphones en pacientes monitorizados en teleasistencia.

- ***Big Transaction Data*.** Incluye registros de facturación, de telecomunicaciones y registros detallados de llamadas (CDR). Los datos transaccionales pueden ser semiestructurados y no estructurados.

- ***Biometrics*.** Hace referencia a datos de información biométrica, como huellas digitales, de reconocimiento facial, de escaneo de retina, de genética (ADN), etc. Son importantes en el área de seguridad e inteligencia para las agencias de investigación.

[2] Sobre este punto, véase el artículo de José María Cepeda "Redes sociales y salud: las comunidades virtuales de pacientes" (2013).

- ***Human Generated*.** En este apartado se incluyen datos generados por personas, como los que se guardan en un *call center* al establecer una llamada telefónica, las notas de voz, correos electrónicos, documentos, estudios y registros médicos electrónicos o recetas médicas.

2.5. Plataformas de Big Data

Como ya se ha anticipado, el uso de los datos masivos requiere la utilización de nuevas herramientas tecnológicas para su captura desde las diferentes fuentes y sistemas, así como su transformación, almacenamiento, análisis, visualización, etc. La plataforma de código abierto Apache Hadoop es la que ha liderado desde un principio los distintos proyectos de software especializado en Big Data.[3] Ha sido adoptada tanto por la comunidad de desarrolladores de aplicaciones de software libre como por los principales proveedores de software propietario de bases de datos (Oracle, IBM y Microsoft).

El proyecto Hadoop consta de tres componentes fundamentales: *Hadoop Distributed File System* (HDFS), *Hadoop MapReduce* y *Hadoop Common.*

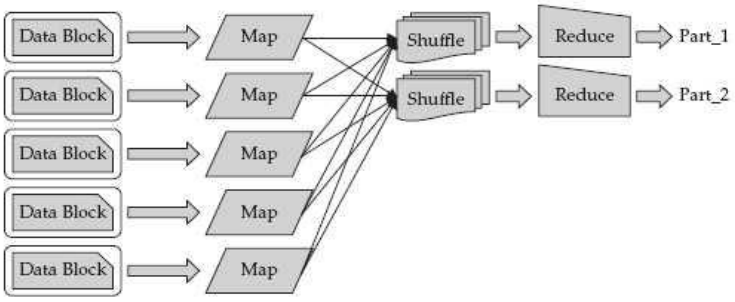
- ***Hadoop Distributed File System* (HDFS).** Los datos en el clúster de Hadoop se dividen en piezas pequeñas llamadas *bloques* que son distribuidas a través del clúster; de este modo, las funciones `map` y `reduce` pueden ser ejecutadas en pequeños subconjuntos, y esto proporciona la escalabilidad necesaria para el procesamiento de grandes volúmenes.

- ***Hadoop MapReduce.*** Es considerado el componente nuclear de Hadoop. El término MapReduce se refiere a dos procesos separados que ejecuta Hadoop. El primero de ellos, ***`map`***, toma un conjunto de datos y lo convierte en otro conjunto, donde los elementos individuales son separados en `tuplas` (pares de clave/valor). El segundo proceso, ***`reduce`***, obtiene la salida de `map` como datos de entrada y combina las tuplas en un conjunto más pequeño de las mismas. Existe una fase intermedia, denominada `shuffle`, que obtiene las `tuplas` del proceso `map` y determina qué nodo deberá procesar estos datos, dirigiéndolos hacia una salida

[3] Para elaborar este apartado se han seguido las siguientes fuentes bibliográficas: "El ecosistema Hadoop", en Joyanes Aguilar (2014: págs. 211-33); "¿Qué es Big Data", IBM, (18/6/2012) y el sitio Web de Apache Hadoop: http://hadoop.apache.org/.

específica para una tarea `reduce`. En la siguiente figura se presenta un ejemplo del flujo de datos en un proceso MapReduce.

- ***Hadoop Common*.** Constituye un conjunto de librerías que soportan varios subproyectos de Hadoop.



Ejemplo de MapReduce

Fuente:IBM https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

Además de estos tres componentes principales, existen otros proyectos relacionados que se describen a continuación:

- ***Avro***. Es un proyecto de Apache que provee servicios de serialización. Cuando se guardan datos en un archivo, el esquema que define ese archivo es guardado dentro del mismo; de este modo es más sencillo para cualquier aplicación leerlo posteriormente puesto que el esquema está definido dentro del archivo.

- ***Cassandra.*** Es una base de datos no relacional (NoSQL) distribuida y basada en un modelo de almacenamiento de <clave-valor>, desarrollada en Java. Permite grandes volúmenes de datos en forma distribuida. Su arquitectura está diseñada para ser altamente escalable. Twitter es una de las empresas que utiliza Cassandra dentro de su plataforma.

- ***Chukwa.*** Es un sistema de colecciones de datos para la gestión de grandes sistemas distribuidos.

- ***Flume.*** Su tarea principal es dirigir los datos de una fuente hacia algún otro lugar, en este caso hacia el entorno de Hadoop. Existen tres entidades principales: `sources`, `decorators` y `sinks`. Un **`source`** es básicamente cualquier

fuente de datos, **`sink`** es el destino de una operación en particular y un **`decorator`** es una operación dentro del flujo de datos que transforma esa información de alguna manera, como por ejemplo comprimir o descomprimir los datos o alguna otra operación en particular sobre los mismos.

- ***HBase.*** Es una base de datos NoSQL columnar (*column-oriented database*) que se ejecuta en HDFS. Cada tabla contiene filas y columnas como una base de datos relacional. HBase permite que muchos atributos sean agrupados en las denominadas *familias de columnas,* de modo que los elementos de una misma familia de columnas sean almacenados en un solo conjunto. Eso es lo que las distingue de las bases de datos relacionales orientadas a filas, donde todas las columnas de una fila dada son almacenadas en conjunto. Facebook utiliza HBase en su plataforma desde noviembre del 2010.

- ***Hive.*** Es una infraestructura de *data warehouse* que facilita administrar grandes conjuntos de datos que se encuentran almacenados en un ambiente distribuido. Hive tiene definido un lenguaje similar al SQL llamado Hive Query Language (HQL). Estas sentencias HQL son separadas por un servicio de Hive y son enviadas a procesos MapReduce ejecutados en el clúster de Hadoop.

- ***Jaql.*** Donado por IBM a la comunidad de software libre. Es un lenguaje de consultas, funcional y declarativo, que permite la explotación de datos en formato JSON (Javascript Object Notation), diseñado para procesar grandes volúmenes de información. Jaql posee una infraestructura flexible para administrar y analizar datos semiestructurados como XML, archivos CSV, archivos planos, datos relacionales, etc.

- ***Lucene.*** Es un proyecto de Apache bastante extendido para realizar búsquedas sobre textos. Proporciona librerías para indexación y búsqueda de texto. Los documentos (*document*) se dividen en campos de texto (*fields*) y posteriormente se genera un índice sobre estos últimos. La indexación es el componente fundamental de Lucene, ya que le permite realizar búsquedas rápidamente con independencia del formato del archivo, ya sean PDFs, documentos HTML, etc.

- ***MongoDB.*** Es una base de datos de documentos, NoSQL, diseñada para reemplazar a las SQL tradicionales. De código abierto y escrita en C++, proporciona más funcionalidad que las bases de datos NoSQL de tipo clave-valor, permitiendo sistemas de consultas *ad-hoc* e indexación de ciertos valores dentro del documento. Mantiene un rendimiento elevado y facilita la distribución de la carga en varios servidores.

- ***Oozie.*** Es un proyecto de código abierto que simplifica los flujos de trabajo y la coordinación entre cada uno de los procesos que deben ser ejecutados en distintos momentos. Permite que el usuario pueda definir acciones y las dependencias entre dichas acciones.

- ***Pig.*** Es un lenguaje de programación que simplifica las tareas comunes de Hadoop, como son la carga de datos, la expresión de las transformaciones sobre los datos y el almacenamiento de los resultados finales.

- ***ZooKeeper*.** Proporciona una infraestructura centralizada de servicios que pueden ser utilizados por aplicaciones para asegurarse de que los procesos a través de un clúster sean serializados o sincronizados.

La lista que se ha presentado no pretende ser exhaustiva, ya que el interés por la tecnología Big Data sigue en aumento y no dejan de surgir continuamente proyectos nuevos o evoluciones de los ya conocidos.

2.6. ¿Qué factores han influido en el auge actual de Big Data?

En los últimos años han sido numerosos los autores que han tratado de explicar la enorme popularidad que ha experimentado recientemente la utilización de Big Data por parte de las organizaciones en comparación con el uso de las tecnologías precedentes.

José Ramón Rodríguez se preguntaba en 2012, en su artículo *Octubre, el mes de los grandes datos*: "¿Qué hay de nuevo con los Big Data comparado con lo que siempre hemos llamado inteligencia de negocio, y más recientemente inteligencia analítica?".

La respuesta que ofrece es que la diferencia principal de esta tecnología con respecto a las anteriores radica en "la multiplicación de la velocidad, el volumen y la variedad y tipología de datos y la división del coste de su producción, tratamiento y almacenamiento".

Por otro lado, IBM, en su Informe ejecutivo ya citado de 2012, señala dos tendencias que explican en buena medida las diferencias entre las actividades actuales de Big Data y las anteriores:

- "La digitalización de prácticamente *todo*", que ha generado en diferentes sectores nuevos tipos de grandes datos en tiempo real, muchos de ellos no normalizados (datos en *streaming*, geoespaciales o generados por sensores) que no pueden ser correctamente procesados por "los *warehouses* relacionales, tradicionales y estructurados".
- "Las tecnologías y técnicas de análisis avanzado de hoy en día", que hacen posible que las organizaciones extraigan conocimientos de los datos "con un nivel de sofisticación, velocidad y precisión nunca antes visto" (pág. 3).

Más recientemente, Luis Joyanes, en su libro *Big Data. Análisis de grandes volúmenes de datos en organizaciones* (2013, pág. 11), en el apartado "¿Cómo se ha llegado a la explosión de Big Data?" explica este fenómeno a partir de la diferencia entre los datos estructurados que se utilizaban antes y el enorme volumen de datos de carácter no estructurado manejado en la actualidad, proveniente de todo tipo de fuentes (redes sociales, dispositivos móviles, Internet de las cosas...).

3. Análisis predictivo aplicado a Big Data

En este apartado se define conceptualmente la analítica predictiva. Se presentan sus técnicas, modelos y métodos; las fases de un proyecto de construcción y mejora continua de modelos; así como las soluciones disponibles en el mercado y las principales aplicaciones.

3.1. Definición de analítica predictiva

Dentro de las posibilidades que ofrece la analítica de Big Data (análisis en tiempo real, grandes volúmenes de datos, datos no estructurados...), en el presente trabajo se repara especialmente en el estudio del análisis predictivo, debido a la importancia que tiene su aplicación en el sector de la salud, como se pondrá de manifiesto en posteriores capítulos.

Joyanes define la analítica predictiva como “una rama de la minería de datos centrada en la predicción de las probabilidades y tendencias futuras”, que “trata de analizar hechos actuales o históricos con el propósito de hacer predicciones sobre sucesos futuros” (2014, pág. 374).

Conviene precisar, como señala Alex Guazzelli, que la analítica predictiva existía ya desde unas décadas antes de que su relevancia en la industria se incrementara por efecto de las posibilidades que ofrecía la tecnología Big Data: “cantidad de datos que se capturaban de las personas (por ejemplo, de transacciones online y redes sociales) y sensores (por ejemplo de dispositivos GPS móviles) así como la disponibilidad de poder de procesamiento costeable, ya sea basado en la Nube o en Hadoop.” (2012). Se podría decir, por tanto, que existe un antes y un después en la analítica predictiva con respecto a Big Data.

Según explica Eric Siegel, el hecho de aprender a predecir a partir de datos en el ámbito científico ha recibido en ocasiones la denominación de "aprendizaje automático", una disciplina basada en la informática y la estadística a la que se dedican conferencias y programas académicos, mientras que la aplicación de esa rama de la ciencia al mundo donde tiene lugar la acción real, en el ámbito comercial, industrial, político..., ha cambiado el nombre por el de "analítica predictiva" para referirse a “una tecnología que aprende de la experiencia (los datos) para predecir el futuro comportamiento de los individuos con el propósito de tomar mejores decisiones”. Para este autor, la AP es “la pionera de la

tendencia que existe actualmente para tomar decisiones *basadas en datos*, confiando menos en el instinto personal y más en una evidencia empírica y palpable" (2014, págs. 35-36).

Por su parte, Mayer-Schönberger y Cukier entienden que lo esencial de los datos masivos consiste en que permiten hacer predicciones. No obstante, a pesar de que se los engloba en la ciencia de la computación llamada inteligencia artificial y, más específicamente, en el área llamada aprendizaje automático o de máquinas, esta caracterización puede inducir a error. Para dichos autores, la utilización de los datos masivos "no consiste en intentar *enseñar* a un ordenador a *pensar* como un ser humano. Más bien consiste en aplicar las matemáticas a enormes cantidades de datos para poder inferir probabilidades". Asimismo, estos autores ponen de relieve que el buen funcionamiento de estos sistemas radica precisamente en que "están alimentados con montones de datos sobre los que basar sus predicciones. Es más, los sistemas están diseñados para perfeccionarse solos a lo largo del tiempo, al estar pendientes de detectar las mejores señales y pautas cuando se les suministran más datos" (2013, págs. 23-24).

Guazzelli concluía de este modo el primero de sus cuatro artículos dedicados a la analítica predictiva: "En un mar de datos en expansión constante, recolectados a partir de personas y sensores, la analítica predictiva proporciona herramientas de navegación esenciales para que las compañías e individuos alcancen exitosamente su destino. Lo hace al pronosticar lo que está por suceder, de forma que uno pueda responder del modo más conveniente para mantenerse en el curso más preciso, seguro, repetible, rentable y eficiente" (2012).

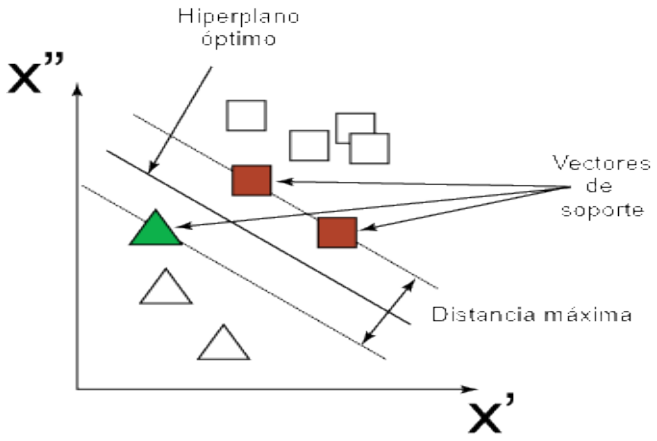
3.2. Técnicas, modelos y métodos de análisis predictivo

La analítica predictiva se materializa mediante la creación de modelos de conocimiento predictivos. Siegel define el modelo predictivo como un "mecanismo que predice un comportamiento de un individuo, como un clic, una compra, una muerte, o una mentira. Toma como datos de entrada las características del individuo y genera como salida una puntuación predictiva. Cuanto mayor sea la puntuación, más probable será que el individuo exhiba el comportamiento predictivo" (2014, pág. 51).

Los modelos predictivos son funciones matemáticas o algoritmos, capaces de determinar y aprender la correlación entre un conjunto de variables de datos de entrada, por lo general

empaquetadas en un registro, y una variable de respuesta o de destino. Estos algoritmos forman parte de las técnicas y métodos de minería de datos. El creciente uso de la analítica predictiva para aplicarla sobre un componente importante de datos no estructurados ha impulsado a los desarrolladores de software a incluir en sus aplicaciones específicas un número cada vez mayor de algoritmos para cubrir un amplio espectro de posibles soluciones de modelado predictivo, de forma que el modelo óptimo se encuentre en la combinación de métodos. No obstante, según señala Guazzelli, hay un reducido grupo de algoritmos genéricos que suelen incluir tanto los fabricantes de software de código abierto como comercial, como los señalados a continuación:[4]

Máquinas de vectores de soporte (**SVM**). Se trata de un conjunto de algoritmos de aprendizaje supervisado que dan solución a problemas de clasificación y regresión. Una SVM construye uno o varios hiperplanos en un espacio de dimensión mayor que el conjunto hallado calculando aquel que proporcione la mayor separación entre dos subconjuntos diferenciados, que será el "hiperplano óptimo". Eso los proveerá de una etiqueta de clase y de una función de regresión que le otorgue valor predictivo. La predicción será que los puntos de un nuevo conjunto analizado por el modelo construido serán clasificados correctamente.



Hiperplano óptimo que separa datos y vectores de soporte.

Imagen extraída de: http://www.ibm.com/developerworks/ssa/industry/library/ba-predictive-analytics2/

[4] Véase, Alex Guazzelli (2012, 17 de diciembre) "Técnicas de modelado predictivo". También se han tenido en cuenta las siguientes entradas de Wikipedia: "Minería de datos", "k-means" y "K-vecinos más cercanos" y "reglas de asociación".

Redes neuronales (**NN**). Las redes neuronales representan una estructura de aprendizaje automatizado inspirada en el funcionamiento del sistema nervioso de los animales. Está compuesto de una capa de entrada, con tantos nodos como número de campos y características que se están considerando; de una capa de salida, con un solo nodo que representa el campo predicho; y de una o más capas intermedias de nodos ocultos. Deberá establecerse una función de correlación entre los campos de entrada y de destino. Cuando hay más de una capa interpuesta de nodos ocultos puede aprender mejor la función un modelo de red neuronal de retropropagación, que busca el ajuste de los valores intermedios desde el valor de salida.

Árboles de decisión. Dado un conjunto de datos, se construyen diagramas de construcción lógica, similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que se presentan de forma sucesiva para la resolución de un problema. Al contrario que los modelos anteriores, resulta más fácil de usar y entender.

De agregación o *clustering*. Es un procedimiento que trata de agrupar de modo cercano a grupos de individuos con características semejantes. Entre los métodos más utilizados para establecer el agrupamiento figura el de los centroides, o *k-means* y el de los vecinos más cercanos, o *K-nn* (*K nearest neighbors*). El algoritmo *k-means* trata de obtener la partición de un conjunto de *n* observaciones en *k* grupos, en el que cada observación pertenece al grupo más cercano a la media. Los elementos más próximos a la media son los centroides. En el conjunto inicial se pueden elegir aleatoriamente los centroides o bien las particiones. Se calcula la media de cada grupo y se repite el proceso hasta que la asignación de sus centroides no varía. El algoritmo *k*-nn es un método de clasificación supervisada, que se basa en establecer previamente ejemplos de entrenamiento ya clasificados. Un nuevo elemento que se desea clasificar se asignará a la clase a la que pertenezca el mayor número de vecinos más cercanos de un grupo de *k* elementos.

Reglas de asociación. Se utilizan cuando una variable de destino o una medida similar no es importante, pero sí lo son las asociaciones entre los elementos de entrada. Por ejemplo, qué pueden tener en común las personas que además de

comprar pañales y leche compran también cerveza. Esto sería un análisis de la cesta de la compra, que puede utilizarse para decisiones de marketing. El modelo se usa en otras muchas áreas, entre ellas la investigación en biología molecular.

3.3. Fases de un proyecto de construcción y mejora continua de un modelo de análisis predictivo

Toda vez que se haya determinado un objetivo para el análisis predictivo, un proyecto de construcción y mejora continua del modelo de análisis predictivo, de acuerdo con la metodología señalada por la consultora Forrester, debería progresar del siguiente modo:[5]

- **Identificación de los datos necesarios y de su variedad de fuentes.**

  Los datos potencialmente valorables a veces están localizados en lugares de difícil acceso, tanto internos (silos de datos en aplicaciones de la propia empresa) como externos (medios sociales, datos del gobierno y otras fuentes de datos públicos). Las herramientas de visualización pueden ayudar a explorar los datos desde varias fuentes para determinar lo que puede ser relevante para el proyecto.

- **Preparar los datos para el análisis**

  Calcular los campos agregados, limpiar los caracteres extraños, rellenar los datos que faltan, fusión de múltiples fuentes de datos, etc.

- **Construir el modelo predictivo**

  Es la fase central del proceso. Mediante la herramienta elegida, se escogen, entre los mejores, uno o más algoritmos, dependiendo del tipo y la integridad de los datos y el tipo de predicción deseado. Los analistas ejecutan el análisis en un subconjunto de los datos llamados "datos de entrenamiento" y dejan aparte otro conjunto, el de "datos de prueba" que servirá para evaluar el modelo.

- **Evaluar la eficacia y exactitud del modelo**

  El análisis predictivo no pretende llegar tanto a la exactitud como a un nivel alto de probabilidad. Para evaluar la capacidad de predicción del modelo, los analistas de datos utilizan el modelo para predecir el conjunto de "datos de prueba". Si el modelo predictivo puede predecir con el conjunto de datos de prueba, es un candidato para su implementación.

[5] Véase Gualtieri y Curran (2015, 1 de abril)

- **Entregar copias del modelo para ponerlo en uso en el entorno operativo de sus compañeros de negocio**

  Hay poco valor en una predicción si no permite aprovechar una oportunidad de predicción o evitar un evento negativo. Los compañeros del negocio deben aprender a confiar en las predicciones de los modelos y los que crean los modelos tienen que aprender de sus socios en el negocio de lo que pueden ser las ideas más viables.

- **Monitorizar y mejorar la eficacia del modelo**

  Los modelos predictivos son tan exactos como lo son los datos introducidos en ellos, y con el tiempo se pueden degradar o aumentar su eficacia. Para supervisar la eficacia de los modelos y el valor en curso, los datos recién acumulados se vuelven a ejecutar a través de los algoritmos. Si el modelo se vuelve menos preciso, los profesionales analistas tendrán que ajustar el modelo (por ejemplo, mediante el ajuste de los parámetros en los algoritmos) y / o solicitar datos adicionales.

3.4. Herramientas de análisis predictivo

Como se expondrá más adelante, a propósito de la acogida de Big Data por parte de las empresas, las encuestas realizadas entre los años 2012 y 2014 revelaban cómo una de las barreras para la adopción de esta tecnología era la escasez de personas con conocimientos avanzados para el aprovechamiento eficaz de su uso, especialmente en el campo de la analítica predictiva. Los fabricantes de software tomaron buena nota de ello y han procurado seguir ofreciendo continuas mejoras en sus productos, respondiendo así a un importante aumento de la demanda en todos los sectores. Los usuarios del negocio cuentan ahora con herramientas más fáciles de manejar. De este modo recoge Wikipedia uno de los titulares del informe "The Forrester Wave™: Big Data Predictive Analytics Solutions, Q2 2015"[6]:

> "Organizaciones de todas las industrias se están animando a dar valor al análisis predictivo. Con el crecimiento de la demanda, los proveedores de análisis predictivo

[6] Gualtieri y Curran, Rowan (2015, 1 de abril)

están proveyendo herramientas que reducen la barrera y aumentan el atractivo a aquellos con menos habilidades estadísticas."[7]

En ese mismo informe, la consultora Forrester publica la nueva evaluación de trece fabricantes de software comercial de analítica predictiva:

- **Líderes**

  En esta categoría coloca a IBM, SAS y SAP.

  IBM reúne un impresionante conjunto de prestaciones, poniendo la predicción en el centro. No importa cómo una organización desee empezar con el análisis predictivo, ya que cuenta con una opción para ellos. Su software ofrece prestaciones para construir modelos, realizar análisis y desplegar aplicaciones predictivas, tanto en las instalaciones como en la nube. El análisis predictivo de IBM puede tomar datos realmente grandes y proporcionar información decisiva.

  Por su parte, **SAS** continúa siendo una potencia en analítica. Con un enfoque estratégico desde 1973, SAS dispone de soluciones de análisis predictivo que ofrecen casi todas las características que un científico de datos o usuario de negocios podría desear. También se mantiene al día respecto a las necesidades cambiantes de los usuarios. SAS Visual Analytics proporciona a los científicos de datos una solución "todo-en-uno" de herramientas de visualización y análisis predictivo. Las soluciones de SAS también se integran con las de código abierto R, Python y Hadoop.

- **Actores Fuertes (*Strong Performers*)**

  Pertenecen a este grupo Alpine Data Labs, Alteryx, Angoss, Dell, FICO, KNIME, Oracle y RapidMiner. Los evaluadores consideran que todos ellos aportan cierto atractivo en sus productos, lo cual los mantiene como excelentes opciones para las empresas. Con mejores puntuaciones en estrategia, Alteryx, Angoss, FICO, Oracle, y RapidMiner serían también líderes.

[7] Wikipedia, "Análisis predictivo"

- **Contendientes**

  Microsoft y Predixion Software entran en esta categoría. Ambos comienzan desde un nicho, pero tienen mucho espacio de funcionamiento para crecer y ser un valor único para las empresas. Microsoft se centra exclusivamente en los servicios en la nube y Predixion Software permite a Excel capacidades de análisis predictivo también en la nube.

En cuanto a las **soluciones de código abierto**, el informe emite el siguiente comentario:

> "En la opinión mayoritaria de los programadores, la comunidad de software de código abierto (OSS) es un potente motor de análisis predictivo. El lenguaje de programación de código abierto **R**, para estadísticas y análisis predictivo, está omnipresente en los entornos universitarios y todos los proveedores evaluados en este Forrester Wave lo apoyan. Los desarrolladores de aplicaciones también tienen una gran cantidad de bibliotecas API disponible para preparar los datos y construir modelos predictivos utilizando Java, Python y Scala. Apache Mahout y WEKA tienen APIs de Java. Apache Spark MLlib incluye API para Java, Python y Scala. Los desarrolladores de Python pueden utilizar NumPy y SciPy para preparar los datos y construir modelos predictivos."

3.5. Principales aplicaciones de la analítica predictiva

El análisis predictivo puede aplicarse a muchas circunstancias y entornos, en los que se trata de resolver problemas de riesgos o de oportunidades que se puedan presentar en el futuro. Los más frecuentemente citados inciden en los siguientes campos:

- En la gestión de relación con los clientes (CRM), interesa conocer:
  - A qué clientes se puede dirigir una campaña publicitaria con mayor opción de respuesta.
  - Quiénes pueden estar a punto de abandonar e interesa retenerlos.
  - Quiénes son buenos compradores, para premiar su fidelidad.
  - Cuáles son las preferencias de compra para proponer ofertas personalizadas, ventas cruzadas, etc.
  - En qué circunstancias pueden consumir una mayor cantidad de productos, para poder mantener el abastecimiento.

- Relacionados con la asistencia sanitaria[8]
    - Conocer cuáles son los factores de riesgo de los pacientes para desarrollar problemas crónicos: asma, diabetes, enfermedades cardiovasculares, etc., con el fin de establecer un mejor control en la asistencia ambulatoria, evitando así la hospitalización.
    - Identificar patrones de riesgo de infección en pacientes monitorizados.
    - Detectar riesgos relacionados con el patrón genético, que permitan desarrollar medidas de prevención y tratamiento personalizados.
    - Conocer el riesgo de reingreso en el hospital antes de dar un alta anticipada.
    - Detectar fraudes y abusos en el cuidado de la salud.
- Análisis de cobros
    - Conocer qué clientes pueden resultar morosos, con el fin de evitar el pago a crédito o de gestionar los cobros mediante agencias de recaudación externas.
- Detección del fraude
    - Reconocer qué créditos son inapropiados, qué transacciones, cobros de prestaciones, reembolsos, reclamaciones son fraudulentas por suplantación o robo de identidad u otros motivos; se podrán tomar medidas a tiempo, evitando la continuación del proceso de forma automática y filtrándolo hacia una auditoría humana.

[8] Este punto se tratar con más detalle y extensión en el capítulo 5.

4. Adopción e integración de Big Data en las organizaciones

En este apartado se plantean cuestiones relacionadas con la adopción e integración de Big Data y Business Intelligence en las organizaciones a nivel general, tales como las fases de adopción empresarial de Big Data o las estrategias de integración de Big Data dentro de una solución BI.

4.1. Adopción empresarial de Big Data

Diversos estudios aparecidos en los últimos años ponen de manifiesto la progresiva adopción de Big Data por parte de las organizaciones y empresas. La información publicada en los mismos nos permite conocer el proceso de implantación y afianzamiento de Big Data, desde las fases iniciales hasta el comienzo de una etapa de consolidación. Por otro lado, esta información también hace referencia a las barreras que han obstaculizado su adopción empresarial. A continuación se enumeran algunas de las principales conclusiones recogidas en diferentes informes desde el 2012 hasta la actualidad.

En el informe de IBM, ***Analytics: el uso de big data en el mundo real,*** citado anteriormente (Schroeck y otros, 2012), los autores del estudio sugieren cuatro fases principales en el proceso de adopción y evolución de Big Data, que denominan *educar*, *explorar*, *interactuar* y *ejecutar*, y que se resumen a continuación:

- En la fase **educar** se trata de crear una base de conocimiento y observaciones del mercado. Las empresas que se encuentran en esta primera etapa estudian las posibles ventajas de las tecnologías y la analítica de Big Data, e intentan entender cómo estas pueden ayudarles a abordar oportunidades de negocio en sus correspondientes sectores y mercados.

- En la fase **explorar** se debe desarrollar una estrategia y una hoja de ruta basándose en las necesidades de negocio y los retos empresariales. Las organizaciones incluyen entre sus principales objetivos desarrollar un caso de negocio y crear un proyecto de Big Data, donde tendrán en cuenta los datos, la tecnología y las habilidades existentes, para luego establecer cuándo y dónde comenzar y cómo desarrollar el plan conforme a la estrategia de negocio de la empresa.

- En la fase **interactuar** las empresas comienzan a comprobar el valor de negocio que proporciona Big Data, así como sus tecnologías y habilidades. En esta etapa las organizaciones desarrollan pruebas de concepto (POCs) o proyectos piloto que les permiten interactuar con la nueva tecnología, dentro de un ámbito definido y limitado, y de esta forma evaluar si cumple con los requisitos y los resultados esperados.

- La última fase, **ejecutar**, tiene como objetivo implementar Big Data a escala. El nivel de operatividad e implementación de las funciones analíticas es mayor dentro de la empresa, habiendo desplegado dos o más iniciativas de Big Data. Son empresas que ya están aprovechando estas tecnologías para transformar sus negocios y obtener el máximo valor de sus activos de información.

Los resultados de la encuesta en la que se basó este informe concluyeron que la mayoría de las empresas de los 1.144 encuestados se encontraban en aquel momento en las primeras fases del desarrollo de Big Data (24% en la fase *educar* y 47% en la de *explorar*), mientras que el resto, pertenecientes a empresas de vanguardia, estaban ya en la fase *interactuar* (22%) y en la de *ejecutar* (6%).

Por otro lado, con relación a los desafíos para la adopción de Big Data, los participantes de la encuesta señalaron como principal obstáculo “articular un caso de negocio atractivo”. Los siguientes retos destacados fueron superar las pruebas de concepto y encontrar personas con las "habilidades necesarias" a nivel técnico, analítico o de gobierno de datos.

* * *

En otra encuesta realizada al año siguiente, en 2013, por EMC España a 510 profesionales de TI a nivel nacional, una amplia mayoría de los encuestados (77%) consideraba que Big Data podría mejorar la toma de decisiones en sus organizaciones, mientras que un 55% opinaba que las empresas e instituciones que hicieran uso de esta tecnología podrían alcanzar mayor éxito[9].

Por otra parte, la mayoría de los participantes citaba como principal barrera para la adopción de Big Data la falta de habilidades y competencias. Además, un 74% consideraba

[9] Véase Paniagua (2013, 16 de octubre)

un reto adaptar las capacidades de los equipos de TI al ritmo de la innovación tecnológica que se esperaba alcanzar hasta 2016.

* * *

Al año siguiente, en septiembre de 2014, Accenture publica un informe basado en una encuesta realizada a CIOs y otros directores de negocio y altos ejecutivos de empresas de 19 países.[10] Más del 60% de los encuestados aseguraban que sus empresas ya habían adoptado con éxito una solución de Big Data, mientras que el 36% declaraba que no había iniciado ningún proyecto ni, por el momento, tenía previsto hacerlo; el 4% restante había comenzado su primer proyecto pero sin completarlo todavía.

Según el estudio, los ejecutivos realizaban un uso moderado o amplio de Big Data con el propósito de identificar nuevas fuentes de ingresos (94% de media), de aumentar los índices de adquisición y retención de clientes (90%) y de desarrollar nuevos productos y servicios (89%). A tenor de estos objetivos, más de la mitad de los encuestados reconocían ya resultados tangibles, especialmente en la búsqueda de nuevas fuentes de ingresos, en el desarrollo de nuevos productos y servicios y en la mejora de la experiencia del cliente.

Como ya se venía haciendo en años anteriores, la encuesta recogió también las principales barreras que habían tenido que superar las organizaciones para la adopción de Big Data. En primer lugar figuraban la seguridad (51%) y las limitaciones presupuestarias (47%), seguidas de la falta de expertos en Big Data (41%), en el uso continuado de Big Data y analítica (37%) y en la integración en los sistemas existentes (35%).

Teniendo en cuenta que la falta de profesionales con experiencia constituye uno de los principales obstáculos para la adopción de Big Data, entre las recomendaciones que se indican a las empresas en dicho informe se encuentra la mejora de los conocimientos de sus empleados mediante programas de formación y desarrollo.

* * *

Por último, el reciente estudio de septiembre de 2015, publicado por la consultora Gartner[11], revela que la inversión en Big Data por las empresas a nivel global, sin mencionar

[10] Véase Sánchez. (2014, septiembre)
[11] Véanse Heudecker, Nick; Kart, Lisa (2015, 3 de septiembre) e Iglesias (2015, 28 de septiembre).

explícitamente el ámbito sanitario, ha superado ya la fase inicial de rápido crecimiento de años anteriores, para entrar en una etapa de consolidación con un crecimiento bastante más lento. Se espera que en los próximos dos años el 75% de las empresas invierta en Big Data, lo que supone solo un 3% de incremento respecto a 2014.

Asimismo, este informe señala un cierto cambio hacia la especialización en el análisis de datos. Un 70% de las compañías que han invertido en Big Data están analizando o planificando analizar datos de localización, y un 64% lo están haciendo para analizar texto en formato libre.

También la finalidad de inversión en Big Data está cada vez más definida. El 64% de las empresas tiene como principal objetivo mejorar la experiencia del cliente, un 47% la racionalización de los procesos existentes y la reducción de los costes de la organización y un 23% la seguridad de la propia información.

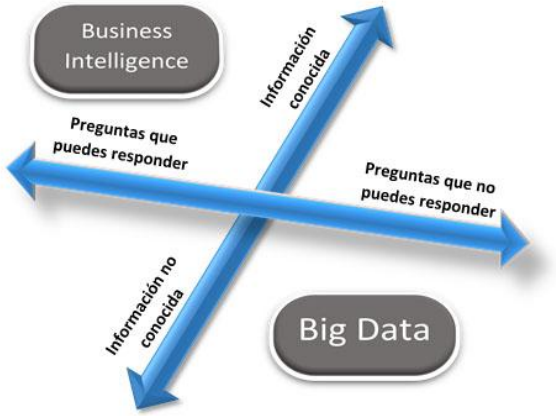
Otra cuestión importante que presenta la encuesta es la incertidumbre con respecto a la posible rentabilización futura de las inversiones en Big Data. El 43% de las empresas que planifican invertir y el 38% de las que ya lo han hecho no saben todavía si el ROI va a ser positivo o negativo.

4.2. Business Intelligence y Big Data, ¿tecnologías complementarias o sustitutivas?

La progresiva implantación de Big Data en los diferentes entornos empresariales que ha tenido lugar durante los últimos años ha suscitado un debate que continúa abierto acerca de la posible coexistencia de esta tecnología junto con Business Intelligence (BI) dentro de una misma organización. Una de las razones que explica esta controversia se encuentra en que ambas soluciones tratan de ofrecer a los usuarios de la organización, y especialmente a los profesionales que ocupan puestos directivos, herramientas que ayuden a optimizar el proceso de toma de decisiones.

Por otro lado, mientras que BI se ha centrado desde sus orígenes en tratar de estructurar información útil y relevante para la toma de decisiones a partir de datos consolidados (información conocida), Big Data es capaz de facilitar el análisis de una cantidad mucho

mayor de datos de todos los tipos, y especialmente no estructurados (información desconocida), los cuales han ido adquiriendo un valor estratégico cada vez mayor para las organizaciones.



Diferencias Big Data – Business Intelligence

Imagen extraída de: http://necsia.es/big-data-vs-business-intelligence-complemento-o-sustituto/

La convergencia en cuanto a los objetivos y los posibles avances que presenta Big Data con respecto a BI han dado lugar a que algunos autores afirmen que esta última tecnología se verá con el tiempo desplazada por la primera, considerando a Big Data como una evolución de BI. Sin embargo, a pesar de los argumentos que respaldan la visión anterior, son muchos los autores que defienden la posibilidad de que ambas tecnologías coexistan dentro de un mismo sistema de información, y señalan los beneficios que se pueden obtener a partir de la complementariedad de ambas soluciones: mientras que Big Data puede aportar a Business Intelligence información procedente de datos no estructurados, BI es capaz de realizar un análisis avanzado de esta información y de ofrecer al usuario una solución visualmente atractiva.

No obstante, hay que señalar que cada organización posee características y necesidades particulares, las cuales se deben identificar con precisión antes de tomar la decisión de implantar cualquiera de estas dos tecnologías (o una combinación de ambas). No hay que descartar, por lo tanto, que existan situaciones en las que resulta desaconsejable la adopción conjunta de ambas soluciones, si se considera que los beneficios que pueda reportar su implantación son inferiores a los costes que esta supone.

En aquellos casos en los que sí puede ser adecuada la implantación de ambas tecnologías, es necesario realizar una correcta integración de la solución planteada dentro del sistema de información de la organización. A continuación se describen, de forma general, los aspectos fundamentales que se deberían considerar para llevar a cabo con éxito esta integración.

4.3. Grados de madurez de las empresas con respecto a Business Intelligence

El primer paso para la integración de una solución que incorpore BI y Big Data dentro del sistema de información de una organización consiste en definir una correcta estrategia de inteligencia de negocio. Como punto de partida para la definición de esta estrategia, Conesa Caralt y Curto Díaz, en su libro "Introducción al Business Intelligence" (2010, págs. 17-19), aconsejan realizar un análisis del grado de madurez de la organización con relación a la inteligencia de negocio, y recomiendan la utilización del BIMM (*Business Intelligence Maturity Model*) como herramienta para realizar este estudio.

Según los autores mencionados, los resultados del análisis obtenido a partir de la utilización del BIMM servirán para clasificar el nivel de madurez de la organización en una de las siguientes fases:

- **Fase 1**: No existe BI. Los datos se encuentran en los sistemas de procesamiento de transacciones en línea (OLTP, *On-Line Transaction Processing*), repartidos en otros soportes o en algunos casos únicamente en el *know-how* de la organización. Las decisiones no están basadas en datos consistentes, sino en la intuición o la experiencia. No se percibe la importancia del uso de datos corporativos a través de las herramientas adecuadas en la toma de decisiones.

- **Fase 2**: No existe BI, aunque los datos se encuentran accesibles. Para la toma de decisiones no se realiza un procesado formal de los datos, aunque algunos usuarios pueden acceder a información de calidad y justificar decisiones con dicha información. A menudo este proceso se lleva a cabo mediante Excel o algún tipo de *reporting*. Aunque se tiene conciencia de las limitaciones de este proceso se desconocen las posibilidades de BI.

- **Fase 3**: Se llevan a cabo procesos formales de toma de decisiones basada en datos. Existe un equipo que controla los datos, a partir de los cuales es posible elaborar informes y tomar decisiones fundamentadas. Los datos se extraen directamente de los sistemas transaccionales sin realizar limpieza de datos ni modelización. No existe un almacén de datos o *data warehouse*.

- **Fase 4**: Se dispone de *data warehouse*, aunque el *reporting* se realiza de forma personal.

- **Fase 5**: Se amplían las funcionalidades del *data warehouse* y se formaliza el *reporting* a nivel corporativo. Se plantea la adopción de OLAP.

- **Fase 6:** Ni el *reporting* ni el acceso al *data warehouse* permiten responder de forma eficiente a preguntas complejas, por lo que se decide la implantación de OLAP. Las decisiones tienen un peso cada vez mayor en los procesos de negocio de toda la organización.

- **Fase 7**: Se formaliza el uso de BI. Se hace necesaria la implantación de otros procesos de inteligencia de negocio como *Data Mining* o *Balanced ScoreCard*. Los procesos de calidad de datos tienen un impacto significativo en los procesos llevados a cabo por los módulos de CRM (*Customer Relationship Management*) o de SCM (*Supply Chain Management*). A nivel corporativo se diferencia claramente entre sistemas OLTP y DSS (*Decision Support System*).

4.4. Estrategias de integración de Big Data dentro de una solución de Business Intelligence empresarial

Una vez identificada la fase en la que se encuentra la organización a partir de la aplicación del BIMM, el siguiente paso consistirá en plantear el diseño de una plataforma que integre BI y la analítica de Big Data. En el artículo de Peter J. Jamack *Analítica de inteligencia de negocios de big data* (2013) se abordan desde una perspectiva general los aspectos más importantes que intervienen en este diseño.

Según este autor, una de las cuestiones principales que deben considerarse en el diseño de una plataforma integrada consiste en la extracción, la transferencia y la carga (ETL). Este aspecto tiene una importancia decisiva en proyectos de integración, ya que, de no realizarse correctamente, se corre el riesgo de utilizar datos incorrectos y poco fiables, que

afectarán a la calidad y al uso del sistema. Otros factores que también deben ser tenidos en cuenta en este diseño, según Jamack, son el almacenamiento de datos y las capas de GUI (*Graphical User Interface*) y de *front-end* de usuario.

Para cada uno de estos aspectos el autor propone diferentes soluciones:

- En el caso de los procesos ETL, existen diferentes herramientas y metodologías, como Sqoop, que permite procesar datos de sistemas de gestión de base de datos relacionales, u otras aplicaciones *open source,* como Flume o Scribe, que pueden ser útiles en el procesamiento de los registros. También se dispone de herramientas más visuales y que integran Big Data, como Talend o IBM InfoSphere DataStage. La selección de una u otra solución deberá estar en función de los sistemas, las fuentes, los datos, el tamaño y el personal de la organización.

- Para el almacenamiento de datos existen diferentes alternativas, como HBase, dentro del sistema de Hadoop, o bien Cassandra, Neo4j, Netezza o el sistema de almacenamiento de archivos HDFS. Jamack recomienda para la integración de Big Data el uso de una plataforma integrada que consista en Hadoop y Cassandra para datos no estructurados o semiestructurados e IBM Netezza Customer Intelligence Appliance, que utiliza en la capa del usuario el software de BI IBM Cognos, para datos estructurados.

- Para las capas de la GUI y de *front-end* de usuario también se dispone de distintas opciones. Por ejemplo, los usuarios avanzados pueden utilizar herramientas como IBM SPSS Statistics o R, que realizan tareas de minería de datos, modelado predictivo, aprendizaje automático y desarrollo de algoritmos y modelos complejos. Otras, como Cognos, permiten a distintos tipos de usuarios explorar los datos o visualizar informes simples. Jamack también sugiere Apache Mahout, para tareas de aprendizaje automático, o Apache Hive, que permite realizar consultas de tipo SQL (*Structured Query Language*) sobre Hadoop. En cualquier caso, el autor recomienda no perder de vista el interés del usuario final, que consiste básicamente en disponer de los datos correctos en el momento adecuado.

Además de los ya citados, Jamack advierte de la importancia de otros aspectos que se deben considerar en el diseño de una plataforma integrada, como el traslado de los datos. En función del tamaño y la distribución geográfica de la compañía, las operaciones de

traslado de datos de un lugar a otro pueden tener una gran complejidad, como en el caso de organizaciones que cuentan con sedes ubicadas en varios países, en donde los datos, que pueden ser no estructurados o semiestructurados, pueden estar repartidos en distintas bases de datos en múltiples *datacenters* a nivel global, y protegidos con diferentes mecanismos de seguridad.

Las tecnologías de Big Data como Apache Hadoop tienen en cuenta esta complejidad, además del coste temporal que supone el traslado de los datos y los posibles riesgos, como la pérdida de datos, paquetes o archivos, y por ello fomentan el traslado del sistema hasta donde se encuentran los datos en lugar de llevar los datos hacia el sistema.

No obstante, el traslado de la aplicación de Big Data hacia los datos también puede ser una tarea muy compleja, sobre todo cuando se cuenta con diferentes sistemas de gestión de datos maestros (*Master Data Management* o MDM) para cada una de las áreas de la organización (productos, ventas, clientes...), y existen dificultades para integrar los distintos MDM.

Por otro lado, una aplicación de estas características requiere una gran cantidad de espacio, memoria y velocidad, que puede provocar fallos o no ser operativa si el sistema sobre el que se instala es antiguo o no está suficientemente preparado. Para abordar esta problemática, Jamack recomienda en primer lugar analizar los sistemas existentes, los casos de uso y el grado de experiencia y competencia del personal de la organización.

Una vez realizado este análisis, existen diferentes opciones, como el desarrollo de un sistema completo de código abierto partiendo de tecnologías como Vanilla Hadoop (Sistema de Archivos Distribuidos de Hadoop [HDFS] y MapReduce), Zookeeper, Solr, Sqoop, Hive, HBase, Nagios y Cacti.

Otra posible opción sería el desarrollo de un sistema que incorporara tecnologías que proporcionan un soporte mayor, como IBM InfoSphere BigInsights e IBM Netezza. Una tercera alternativa consistiría en la separación de datos estructurados y no estructurados y el desarrollo de una capa de interfaz gráfica de usuario (GUI) para usuarios, usuarios avanzados y aplicaciones.

Por último, el autor aconseja tener en cuenta la experiencia del personal de la organización en el trabajo con Big Data para llevar a cabo la integración de BI y la analítica de Big Data. En caso de que la organización cuente con personal experto, esta puede optar por un sistema de código abierto, ya que resulta mucho más rápido y menos costoso de implementar. En caso contrario, Jamack aconseja una aplicación de proveedor de Big Data, a pesar del mayor coste que esta supone.

5. Big Data en el sector de la salud

Este capítulo está dedicado íntegramente a Big Data en el ámbito sanitario, y está dividido en tres apartados. Partiendo de dos estudios fundamentales, en el primero se analizan las posibilidades de la analítica de Big Data en el sector de la salud. En el siguiente, se expone el estado en que se encuentra actualmente su adopción en nuestro entorno sanitario. A continuación se presentan algunos casos de éxito.

5.1. Posibilidades de la analítica de Big Data en el ámbito sanitario. Estudios fundamentales

En este apartado se incluyen algunos de los trabajos que más contribuyeron en su momento a dar a conocer las posibilidades que la tecnología Big Data ya había comenzado a ofrecer en el sector de la salud. Así lo pusieron de relieve de modo especial dos estudios de referencia, como son el extenso informe que el McKinsey Global Institute (MGI) publicó en junio de 2011 bajo el título *Big Data: The Next Frontier for Innovation, Competition and Productivity* y el informe *Big Data Healthcare Hype and Hope*, que la doctora y consultora Bonnie Feldman, en colaboración con Ellen Martin y Toby Skotness, publicó en octubre de 2012. En los subapartados siguientes se revisan las principales aportaciones de cada uno de ellos.

5.1.1. El informe *Big Data: The Next Frontier for Innovation, Competition and Productivity*

En este documento se destaca el potencial transformador de Big Data en cinco grandes sectores, que los expertos de la consultora MGI han estudiado en profundidad: la atención sanitaria y el comercio minorista en EE.UU., la administración del sector público en la Unión Europea, y, a nivel global, la fabricación y los datos de geolocalización de personas y entidades.

Los citados ámbitos representaban en 2010 cerca del 40% del PIB mundial. Para cada uno de ellos se identifican "palancas" o recursos mediante los cuales Big Data facilitará la creación de valor.

En el caso de la atención sanitaria en EE.UU., se estima que en el plazo de diez años el valor creado mediante el uso de Big Data podría alcanzar los 300.000 millones de dólares al año,

lo que supondría un ahorro anual de más de 200.000 millones en el gasto nacional en el sector.[12]

En el informe se distinguen quince recursos Big Data relacionados con el sector sanitario, que se distribuyen en diferentes categorías según ejerzan su impacto sobre la operativa clínica, la investigación y desarrollo, el sistema de pagos y fijación de precios, el impulso de nuevos modelos de negocio o la salud pública. La mayor parte de ellas pertenece a las dos primeras categorías, que son las que se comentan a continuación.[13]

1) **En la operativa clínica***,* Big Data podrá impulsar la creación de valor de cinco modos diferentes:

   - Mediante programas CER (*Comparative Effectiveness Research*) de **investigación comparativa de la eficacia** de los tratamientos aplicados a los pacientes. El análisis detallado de grandes conjuntos de datos de características de los pacientes, de los resultados de los tratamientos y de los costes que han supuesto, puede identificar cuáles son los tratamientos más eficaces para cada caso desde el punto de vista clínico y también bajo criterios de coste-eficacia. Implementando esta investigación, el sistema de salud podrá reducir la incidencia de tratamientos excesivos (intervenciones que perjudican más que benefician), o bien de tratamientos insuficientes (que se deberían haber prescrito pero no lo fueron). Tanto en un caso como en el otro derivarían en peores resultados para los pacientes y mayores costes de atención sanitaria a largo plazo. Estos programas ya han venido aplicándose con éxito en el Reino Unido, Alemania, Canadá y Australia.

   - Con **sistemas de soporte a las decisiones clínicas**, para mejorar la calidad de las prescripciones. Estos sistemas registran las órdenes de tratamiento de los médicos y realizan un análisis comparativo frente a las pautas recomendadas, alertando de potenciales errores o efectos adversos del medicamento. De este modo, se pueden reducir las reacciones adversas y la tasa de errores médicos, así como de reclamaciones de responsabilidad civil derivadas de los mismos. En un estudio realizado en una unidad de cuidados intensivos pediátrica en una gran área

[12] Véase el citado informe, págs. 49-51.
[13] Ibíd., págs. 43-49.

metropolitana de Estados Unidos, una herramienta de sistema de apoyo a las decisiones clínicas redujo las reacciones y eventos adversos por medicamentos en un 40% en tan sólo dos meses.

- **Creando transparencia sobre los datos médicos**. Los conjuntos de datos operacionales y de rendimiento del proveedor pueden ser analizados para crear mapas de procesos y cuadros de mando que permitan la transparencia informativa. El objetivo es identificar y analizar las fuentes de variabilidad y de pérdidas en los procesos clínicos con el fin de optimizar estos procesos. Simplemente, la publicación de datos de costes, calidad y rendimiento, incluso sin una recompensa financiera tangible, a menudo crea la competencia que impulsa mejoras en el rendimiento.

- La **monitorización de pacientes a distancia**. La monitorización remota recoge datos de pacientes con enfermedades crónicas (diabetes, insuficiencia cardiaca congestiva, hipertensión…). El análisis en tiempo real de los datos servirá para controlar la evolución, vigilar el cumplimiento (si los pacientes están realizando el tratamiento prescrito) y mejorar futuras opciones de tratamiento y de medicación. Los sistemas incluyen dispositivos que monitorizan el estado del corazón, envían información sobre los niveles de glucosa en sangre, transmiten las instrucciones de los cuidadores, o incluso pueden incluir la tecnología "chip-en-una-píldora", que deja constancia de que el paciente está tomando la medicación. Por lo general, el uso de datos de los sistemas de monitorización remota permite reducir la estancia del paciente en el hospital, disminuir las visitas al servicio de urgencias y mejorar la focalización de la atención de enfermería a domicilio y de las citas con el médico a nivel ambulatorio, reduciendo a largo plazo las complicaciones de la enfermedad que impliquen hospitalización.

- La **analítica avanzada aplicada a reconocer perfiles de pacientes** permitirá identificar personas que se beneficiarían de un cuidado o de un cambio proactivo en el estilo de vida. Por ejemplo, al identificar pacientes que están en alto riesgo de desarrollar una diabetes, estos podrán beneficiarse de un programa de atención preventiva.

2) **En la investigación y desarrollo en la industria farmacéutica**. Big Data podrá también mejorar la productividad en los departamentos de I+D relacionados con la industria farmacéutica. Mediante los recursos que se mencionan a continuación podría crear en EE.UU. más de 100.000 millones de dólares en valor, de los cuales alrededor de la cuarta parte resultarían de la reducción del gasto nacional de salud**.**

   - El **modelado predictivo**. En el desarrollo de nuevos medicamentos, la agregación de datos masivos a la investigación en las fases preclínicas y en fases tempranas de ensayos clínicos permitirá la construcción de un modelo predictivo que anticipe los resultados finales en términos de eficacia, seguridad y efectos secundarios potenciales del producto. Ello reducirá considerablemente el tiempo de investigación que habitualmente venía suponiendo largos y costosos ensayos clínicos.

   - **Herramientas y algoritmos estadísticos para mejorar el diseño de los ensayos clínicos**. Estos recursos utilizan técnicas de minería de datos para optimizar la selección de pacientes candidatos a estudiar en el ensayo clínico, los sitios donde hay mayor concentración de casos, el diseño de protocolos más eficaces, la gama de indicaciones que se aplicarán al producto, el modelado comercial, las posibilidades de aprobación regulatoria o la selección de investigadores que han de dirigir el proyecto.

   - **Análisis de los datos de ensayos clínicos**. Este procedimiento posibilitará la identificación de nuevas indicaciones del fármaco así como el descubrimiento de efectos adversos. Una vez que el producto ya esté en el mercado, la recogida en tiempo real de los informes de reacciones adversas hará posible una farmacovigilancia más ágil y eficiente al poder detectar los casos raros que han escapado a la insuficiente potencia estadística de un ensayo clínico.

   - **La medicina personalizada**. Este recurso Big Data tiene como objetivo examinar las relaciones entre una determinada variación genética en las personas, la predisposición a padecer determinadas enfermedades y las respuestas específicas que esas personas puedan tener con ciertos medicamentos. Seguidamente se tendrá en cuenta esa variabilidad genética de los individuos en el proceso de

desarrollo de fármacos. La medicina personalizada promete mejorar la atención de la salud de tres maneras:

a) Mediante la detección temprana y el diagnóstico antes de que un paciente desarrolle síntomas de la enfermedad.

b) Buscando terapias más eficaces teniendo en cuenta que pacientes con el mismo diagnóstico se pueden segmentar con arreglo a un perfil molecular diferente, es decir, no responden de la misma manera a la misma terapia, en parte debido a la variación genética.

c) Ajustando las dosificaciones del fármaco de acuerdo con el perfil molecular del paciente para reducir al mínimo los efectos secundarios y maximizar la respuesta beneficiosa deseada.

- **Análisis de patrones de enfermedades**. La identificación de patrones y tendencias de enfermedades permite modelar la demanda y los costes futuros, así como tomar decisiones estratégicas de inversión en I+D. Este análisis puede ayudar a las empresas farmacéuticas a optimizar el foco de sus actividades de I+D y a prever la asignación de recursos, incluyendo equipamiento y personal.

5.1.2. El informe *Big Data Healthcare Hype and Hope*

Este estudio constituye una valiosa aportación. Se basa en el análisis de proyectos y herramientas que ya están en funcionamiento, siguiendo diferentes estrategias. En palabras de Paniagua (2012), el trabajo de Feldman "explora con bastante precisión cómo Big Data se está convirtiendo en una creciente fuerza de cambio en el panorama sanitario".[14]

Dentro de su detallada exposición, se pueden considerar de especial interés los apartados que dedica a las cuatro principales dimensiones que caracterizan a Big Data aplicadas al sector sanitario y a las formas que propone para mejorar el cuidado de la salud.

[14] En su artículo "Big Data en sanidad para predecir, prevenir y personalizar", Soraya Paniagua deja constancia del valioso trabajo de Feldman y sus colaboradores, publicado el mes anterior.

5.1.2.1. Dimensiones Big Data aplicadas a los datos del sector sanitario

Volumen. Los datos de asistencia sanitaria están experimentando un crecimiento exponencial. En 2012 el volumen de los mismos era de 500 petabytes (10^{15} bytes) en todo el mundo. Se estima que en 2020 llegará a 25.000 petabytes, es decir, el resultado de multiplicar por 50 la cifra de 2012. Este crecimiento provendrá por una parte de la digitalización de los datos ya existentes en las organizaciones, pero sobre todo, de los nuevos datos generados ya en modo digital: registros médicos personales, imágenes radiológicas, ensayos clínicos, secuenciación genómica, lecturas de sensores biométricos o imágenes en 3D.

Variedad. La diversidad de fuentes comporta también una variedad de tipos de datos (estructurados, semi-estructurados y no estructurados), lo que confiere a esta dimensión un peso mayor que las demás en complejidad para el análisis de los mismos.

Los centros de atención médica han estado generando principalmente datos no estructurados: registros médicos, notas manuscritas de médicos y enfermeras, registros de admisión y de altas hospitalarias, recetas en papel, imágenes radiográficas, de resonancia magnética, TAC, etc. Por otro lado, los datos estructurados y semiestructurados proceden de la contabilidad, la facturación electrónica, algunos datos actuariales, clínicos o de lectura de instrumentos de laboratorio, así como de los generados por la conversión continua de registros de datos no estructurados en estructurados para su incorporación a la historia clínica electrónica.

Big Data podrá, además, capturar y almacenar nuevas formas de datos (estructurados y no estructurados) procedentes de otras fuentes, como los generados por sensores de dispositivos *fitness*, de las App de smartphones, de la investigación genómica o de los medios sociales.

Bonnie Feldman concluye esta parte afirmando que es precisamente en la capacidad de combinar datos tradicionales con las nuevas formas de datos, tanto a nivel individual como poblacional, donde radica el potencial de Big Data en el área sanitaria, y ya se está comprobando cómo los conjuntos de datos procedentes de multitud de fuentes apoyan de modo rápido y fiable la investigación y el descubrimiento. (2012, pág. 10).

Velocidad. Aunque en muchas ocasiones no se requiere una gran velocidad en los procesos de captación, análisis y entrega de información, hay algunas situaciones médicas que exigen un flujo de datos constante en tiempo real: monitorización de signos vitales en traumatismos, en la sala de operaciones durante la anestesia, en la UCI, etc.

Veracidad. Esta característica hace referencia a la incertidumbre de la calidad de algunos datos entrantes y del valor predictivo que se debe otorgar a los resultados del análisis en el que se hayan incluido esos datos. En la asistencia sanitaria es muy importante que el modelo predictivo resultante del análisis sea de calidad.

En algunos casos Big Data recurrirá a herramientas que filtren datos entrantes dudosos que podrían ser erróneos, para corregirlos o eliminarlos (textos manuscritos en recetas, anotaciones…). En otras ocasiones información que poseía escasa relevancia estadística o no se detectó al trabajar con conjuntos de datos insuficientes, sí la va a tener o será detectable al agregar conjuntos de datos mayores. Un ejemplo es someter a evaluación el modelo predictivo utilizado en los ensayos clínicos mediante la incorporación automatizada y en tiempo real de los datos de reacciones adversas a medida que son reportadas cuando el medicamento ya ha sido comercializado, donde se trabajará con una muestra prácticamente equivalente a la población total. Como ya se ha señalado, esto redundará en una mayor eficiencia del proceso de farmacovigilancia.

5.1.2.2. Formas de utilizar Big Data para mejorar el estado de salud

De la información que se obtuvo mediante más de 30 entrevistas con empresas y organizaciones del sector sanitario que ya han comenzado a utilizar Big Data para abordar diferentes retos de la atención médica, la Dra. Feldman distingue seis formas de utilización:

I. Apoyo a la investigación genómica y otras "-ómicas" [15]

   Diversas ramas de la biología molecular, y especialmente la genómica, han experimentado un gran avance mediante el uso de tecnología Big Data, con aplicaciones específicas (por ejemplo, para secuenciación de ADN) cada vez más rápidas y a menor coste. Diferentes soluciones (**Genoma Health Solutions**, **GNS Healthcare**,

[15] Hace referencia a otras ramas de la investigación en biología molecular, como "proteónica", "metabolónica", etc. Véase el concepto en https://es.wikipedia.org/wiki/Ómica

DNAnexus, **Appistry**, **NextBio**) buscan identificar el perfil genético o biomolecular de las personas con el fin de encontrar variaciones que lo relacionen con predisposición a padecer determinadas enfermedades y, consecuentemente, elegir las medidas más adecuadas de prevención y tratamiento. El traslado a la práctica clínica o al desarrollo de nuevos medicamentos "ad hoc" basándose en el conocimiento adquirido en la investigación es lo que hace posible la denominada medicina personalizada.

II. Transformar datos en información (y la información en datos)

Dada la creciente cantidad y variedad de datos que se generan en el ámbito de la salud, y la alta proporción de no estructurados (cercana al 80% según algunas estimaciones), un aspecto clave para la mejor gestión de los mismos y su conversión en información utilizable es hacer que los no estructurados pasen a ser estructurados y así facilitar su manipulación por la máquina. Herramientas como **Health Fidelity** utilizan un procesamiento de lenguaje natural (NLP) para convertir un registro médico narrativo en datos que sean comprendidos por la máquina. **Predixion Software** realiza análisis predictivo para identificar pacientes con riesgo de reingreso, para prevención de infecciones hospitalarias o de enfermedades crónicas.

III. Apoyo al auto-cuidado y colaboración ciudadana

Se asiste también al auge de nuevas herramientas que nos ayudan a cuidar de nosotros mismos. Mediante la recogida de datos procedentes de sensores instalados en dispositivos "*wearables*" y smartphones, Big Data puede devolver información a los usuarios sobre su estado de salud. Aplicaciones para móviles como **IBlueButton**, de Humetrix, permiten el intercambio rápido y seguro de información entre pacientes y registros médicos. La plataforma **Ginger.io**, basada en la nube, recopila datos en tiempo real desde los smartphones acerca del comportamiento, activo y pasivo, de los pacientes (movimiento, comunicación, uso del móvil…) para ayudar a los médicos, enfermeras, familiares y pacientes a gestionar su salud, comenzando por las enfermedades crónicas. Con el consentimiento del paciente, la recogida de datos y análisis quedarán a disposición de los proveedores de salud y los investigadores a través de un panel de control HIPAA.

IV. Apoyo a los proveedores de salud y a la mejora en la atención al paciente

Big Data se emplea también como herramienta para la inteligencia de negocio en la gestión sanitaria. Las plataformas actuales permiten explorar y medir en segundos una enorme cantidad de datos clínicos, de gestión financiera, redes sociales, etc., para medir el rendimiento y generar nuevos modelos de pago (pago por rendimiento). Servirán para mejorar la atención al paciente y reducir costes. Con distintos enfoques, soluciones integrales como las propuestas **OneHealth Solutions**, **Explorys** o **Humedica** abordan esta problemática.

V. Aumentar el conocimiento

Existen diversas herramientas que se utilizan para aumentar el conocimiento y resolver una gran variedad de problemas basándose en los datos. **Sproxil** utiliza Big Data para identificar medicamentos falsificados, protegiendo así la salud del paciente, y para que las empresas farmacéuticas puedan rastrear la distribución de medicamentos y prevenir el fraude. **Asthmapolis** recoge datos de los pacientes con asma y les proporciona retroalimentación que les ayuda a controlar mejor su enfermedad. Un dispositivo móvil mediante una aplicación para iOS / Android se conecta con inhaladores para el asma que disponen de un sensor que controla la hora y el lugar de los hechos en el momento de tomar la dosis y recoge datos de los posibles factores ambientales desencadenantes así como de los síntomas. **Sickweather LLC** escanea las redes sociales (Facebook, Twitter) para rastrear brotes de enfermedades y ofrece previsiones para los usuarios de forma similar a la predicción del tiempo.

VI. Agregación de datos para construir un ecosistema mejor

La agregación de muchos más datos dispares procedentes de fuentes externas al conjunto inicial disponible constituye una interesante aplicación de Big Data que permite realizar nuevos tipos de análisis y facilitar respuestas a las grandes preguntas que realizan los investigadores.

Qualcomm Life habilita una plataforma segura de interconexión sanitaria global basada en la nube, que mediante el nodo 2net™ recoge y almacena vía inalámbrica todos los datos biométricos que le son enviados desde smartphones y otros dispositivos

sanitarios (medidores de glucosa, tensiómetros, balanzas, etc.) de pacientes en teleasistencia. El ecosistema sirve de puente que conecta los pacientes con sus respectivos servicios de atención sanitaria de los que son clientes, pero en la plataforma quedará una versión anonimizada de todos los datos, con licencia para uso común. Los estudios basados en ese gran fondo de datos tendrán mayor relevancia estadística y valor predictivo.

Plataforma 2net

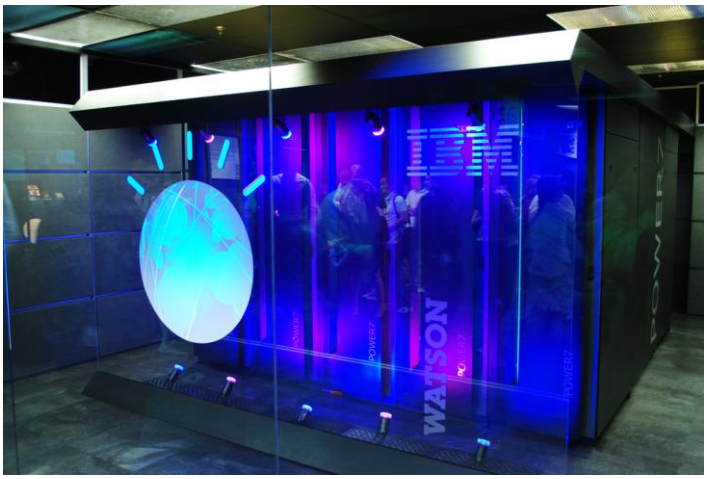
Imagen extraída de http://www.qualcommlife.com/wireless-health

Dentro de este apartado, el informe de la Dra. Feldman destaca el papel que ya está jugando el sistema **Watson** de **IBM** como el primero de una nueva clase de plataformas analíticas y de soporte a las decisiones que utiliza un análisis en profundidad de contenidos, un razonamiento basado en la evidencia y un procesamiento del lenguaje natural para apoyar el diagnóstico y la toma de decisiones clínicas de modo más rápido y preciso. Entre sus principales características señala las siguientes:

- Dotado con 21 subsistemas de computación, 16 terabytes de memoria RAM y del sistema de procesamiento del lenguaje natural más avanzado en el mundo, Watson puede almacenar una enorme cantidad de datos, desde los de las historias clínicas de los pacientes hasta todas las publicaciones de revistas médicas, incluyendo las que recogen los últimos avances en el diagnóstico y tratamiento, siendo capaz de leer toda esa información a razón de 200 millones de páginas en tres segundos y recordar cada palabra.

- Watson revisa los datos de la historia del paciente, antecedentes familiares, síntomas y resultado de pruebas, formula hipótesis que contrasta con la información almacenada y devuelve un listado de enfermedades clasificadas por nivel de confianza, que ayudan al médico en el diagnóstico y tratamiento.

- WellPoint ya está trabajando en un proyecto Watson conjuntamente con los oncólogos del Cedars-Sinai Medical Center de Los Ángeles, entrenándolo para que ayude en la práctica clínica a tomar decisiones en el diagnóstico y tratamiento de cánceres de mama, pulmón y colon.[16]



Superordenador Watson

Imagen extraída de https://es.wikipedia.org/wiki/Watson_%28inteligencia_artificial%29

[16] Véase también el artículo de Tom Groenfeld (2012, 21 de febrero) donde da más detalles de este trabajo en colaboración emprendido por IBM.

5.2. Big data en nuestro sistema sanitario

A partir de estos dos importantes y documentados estudios, como se ha pretendido poner de relieve en esta revisión, parece indudable que las posibilidades que ofrece la adopción de la tecnología Big Data al sector de la salud son muchas y ventajosas.

Por otra parte, se ha podido comprobar cómo en los dos últimos años ha aumentado sensiblemente la información publicada sobre el tema a nivel nacional. Son numerosos los artículos, entrevistas, reseñas, informes o noticias que ponen de manifiesto algunos de los logros alcanzados, así como los que se esperan conseguir en un futuro con la ayuda de Big Data en el sector de la salud, sin omitir, por otro lado, los principales obstáculos que será necesario ir superando. A modo de ejemplo, mencionaré algunos de ellos.

En un artículo publicado en abril de 2014, Julio Mayol, profesor de Cirugía y director de Innovación en el Hospital Clínico de Madrid, aborda con claridad los retos que debe afrontar nuestro sistema sanitario para aprovechar las ventajas que promete la tecnología Big Data, cuyos resultados, en su opinión, "no serán ni inmediatos ni siempre beneficiosos". Considera que la complejidad es grande y señala los siguientes seis retos significativos:

- Extraer conocimiento de **fuentes heterogéneas y complejas**, a veces no estructuradas.
- Comprender **notas clínicas no estructuras en su contexto** correcto.
- Gestionar adecuadamente gran cantidad de datos de imagen clínica y extraer **información útil** para **generar biomarcadores**.
- Analizar **los múltiples niveles de complejidad** que van desde los datos genómicos hasta los sociales.
- **Capturar los datos** de comportamiento de los pacientes, a través de **distintos sensores**, con sus implicaciones sociales y de comunicación.
- Evitar los **problemas de privacidad** y *profiling* que pueden generar riesgos para los individuos.

Por otro lado, Mayol afirma que para hacer frente a estos desafíos es necesario "gestionarlos desde los niveles político, tecnológico y sanitario." Propone que, entre tanto, "habrá que ir desarrollando nuevos marcos y sistemas de referencia que faciliten trabajar en las tres grandes fuentes de Big Data sanitaria: la Historia Clínica e Imagen Médica, las

redes sociales y sensores, y las 'ómicas' (genómica, transcriptómica, epigenómica, proteómica, metabolómica, microbiómica y exposómica)."

* * *

García Barbosa (2014) no duda en aventurar que la tecnología que utiliza los datos masivos será la gran aliada de la medicina del futuro, o "Medicina de las 4P: personalizada, predictiva, preventiva y participativa". Subraya el papel que esta nueva tecnología podrá desempeñar en cada uno de estos aspectos de la gestión clínica. Para sacar su máximo partido, "en la sanidad del futuro sería preciso capturar, almacenar y analizar todos los datos disponibles sobre ensayos clínicos, historiales médicos, secuenciación de ADN de pacientes, información procedente de redes sociales… Se debería disponer, por tanto, de una enorme base de datos compartida entre todos los hospitales y resto de agentes del sector de la salud".

Por otra parte, advierte que en la actualidad es necesario superar algunas trabas, entre las que figura la barrera tecnológica. En su opinión, la tecnología Big Data tiene que consolidarse todavía en el ámbito sanitario, por lo que cree necesario "que aumenten las inversiones públicas y privadas en este tipo de soluciones". No obstante, piensa que tal vez el factor más importante sea el humano, es decir, científicos de datos que sepan sacar provecho de la tecnología: "Es crucial contar con la presencia de analistas de datos expertos en el ámbito de la salud para que, a través del uso de tecnologías *big data*, puedan dar el soporte adecuado a los médicos en la toma de decisiones relativas a sus pacientes."

* * *

A finales de 2014, Mercedes Núñez publica un interesante artículo en el que recoge diez puntos clave del informe "Análisis de la eSalud en España" (2014) elaborado por Ametic, entre los cuales cabe destacar algunos de ellos:

- "La aplicación de las TIC a la sanidad contribuye a mejorar los resultados en salud, así como la eficiencia del sistema y reduce de forma significativa el consumo de recursos sanitarios y los costes."

- "Se han conseguido logros como que 20 millones de españoles tengan historia clínica digital, y que el 70 por ciento de las recetas sean electrónicas así como una amplia implantación de la radiología digital en nuestros hospitales."

- "Pero aún quedan muchos retos pendientes: no podemos hablar de historias clínicas integradas entre niveles, e interconectadas entre las diferentes comunidades autónomas y entre los sistemas públicos y privados de prestación sanitaria. Respecto a la receta electrónica hay también grandes diferencias y una evolución muy desigual en cuanto a prescripción y dispensación en las distintas CC.AA. Y en lo relativo a la cita previa sanitaria también quedan grandes retos por cubrir: multicanalidad, cita previa hospitalaria y en pruebas diagnósticas o servicios de valor añadido como la gestión de resultados y la orientación sanitaria."

- "Tanto para evitar el colapso del sistema como para ofrecer una sanidad moderna y adaptada a los nuevos tiempos es necesario que la inversión TIC en sanidad se intensifique. Las TIC son clave para garantizar la sostenibilidad del sistema sanitario: España es uno de los países más envejecidos del mundo, con una población mayor de 65 años superior al 25 por ciento y la cronicidad, debido en parte a este envejecimiento poblacional, representa cerca del 75 por ciento del gasto sanitario."

* * *

En la actualidad se puede decir que existe una opinión muy compartida sobre el valor que puede aportar la utilización de Big Data en el mundo de la salud. En palabras de Visitación López (2015), "los profesionales sanitarios cada vez entienden mejor que puede suponer un cambio de paradigma en la práctica de la Medicina, y las empresas farmacéuticas quieren utilizarlo para diseñar medicamentos cada vez más efectivos y con menor coste de investigación. A su vez, las administraciones quieren comprobar la eficacia de los nuevos medicamentos en lo que se denomina *Real World Data.*"

* * *

En el informe "Big Data, el poder de los datos", de la Fundación Innovación Bankinter, dado a conocer en mayo de 2015, los expertos consideran que Big Data "está contribuyendo a reinventar el sistema sanitario". Uno de los cambios observados hace referencia a la "transformación del tradicional seguimiento periódico del enfermo en consulta a un

proceso continuo en el que, a través de una plataforma digital o de una app, los profesionales pueden controlar minuto a minuto la evolución del paciente y realizar mediciones constantes" (Seyfert-Margoli, 2015, pág. 39).

* * *

También en la primavera de 2015 el mundo sanitario celebraba el anuncio por parte de IBM de la creación de su nueva unidad de negocio Watson Health, específicamente orientada a ese sector. Según se explica en la nota de prensa publicada el 14 de abril (Torralba, 2015), se trata de ofrecer una plataforma en la nube (IBM Watson Health Cloud) abierta y segura, que permita anonimizar, compartir y combinar los datos referentes a la salud. El enorme poder de Watson, con la tecnología de computación cognitiva, amplía con ello su oferta de soluciones para el ámbito de la salud, en colaboración con Apple, Johnson&Johnson y Medtronic, para la recopilación de información en tiempo real desde dispositivos médicos, de salud y bienestar personal. Apple integrará en la plataforma sus soluciones analíticas HealthKit y ResearchKit, y dará soporte para introducir datos desde aplicaciones iOS. Johnson&Johnson colaborará en el desarrollo de sistemas inteligentes para el cuidado pre y postoperatorio de pacientes. Medtronic usará la plataforma para crear soluciones personalizadas para pacientes con diabetes, con dispositivos de bombas de insulina y monitores de glucosa.

* * *

Muy recientemente, en diciembre de 2015, se ha dado a conocer un estudio de gran interés, el *Informe Big Data y Salud* elaborado por Planner Media y Prodigioso Volcán, en el que distintos expertos (médicos, farmacéuticos, gestores y otros profesionales) ofrecen su punto de vista acerca de las posibilidades de la recogida masiva de información en el sector de la salud.[17] Entre las ideas expuestas en este informe se puede destacar la opinión de Salvador Peiró, quien afirma que es en el campo de la "investigación clínica, farmacológica y epidemiológica" donde el análisis de los RWD (*Real World Data*) está suponiendo en la actualidad importantes beneficios para la población. Por otro lado, en lo que respecta a la

[17] Este informe, que recoge entrevistas y artículos de varios autores, figura en la bibliografía final con la referencia VV.AA. (2015, diciembre). *Informe Big Data y Salud.* Prodigioso Volcán y Planner Media (eds.). En línea. [Fecha de consulta: 13/12/2015: <http://www.plannermedia.com/downloads/informebigdataysalud.pdf >

práctica clínica, Bernardo Valdivieso opina que, aunque existe una percepción generalizada de las posibilidades de Big Data para cambiar el paradigma actual del modelo de cuidados, actualmente, “tanto aquí como en otros países, el Big Data supone más una oferta tecnológica que un proyecto real aplicado a la práctica clínica”.

En el informe se abordan diversos aspectos relacionados con la incidencia de Big Data en el sector sanitario, como las cuestiones relativas a la protección de la privacidad de los pacientes y los esfuerzos que se están realizando desde las distintas administraciones, en coordinación con instituciones europeas, para actualizar la legislación vigente en materia de protección de datos, con el objetivo de hacer frente a los nuevos desafíos que plantea el uso de esta tecnología en los diferentes entornos sanitarios. También hay que destacar la información relativa al proyecto europeo TrendMiner de monitorización de temas de salud en redes sociales, liderado por Paloma Martínez, así como el estudio Hakari patrocinado por Fujitsu y liderado por Julio Mayol, que actualmente se encuentra en fase inicial, y que tiene el objetivo de desarrollar una herramienta de apoyo para la toma de decisiones en salud mental.

5.3. Presentación de casos

Esta sección está dedicada a presentar algunos casos paradigmáticos de utilización de Big Data que pueden servir para ilustrar las posibilidades que esta tecnología ofrece al sector de la salud. Los dos primeros corresponden al año 2009, cuando todavía esta tecnología se encontraba en su primera etapa, mientras que los siguientes casos son de muy reciente publicación.

La aplicación *Flu Trends* de Google

Probablemente, el caso más referenciado que puso de manifiesto el poder predictivo que en minería de datos tiene el análisis con Big Data fue la aplicación Flu Trends, desarrollada por los ingenieros de Google, que permitió determinar casi en tiempo real por dónde se estaban propagando brotes de gripe y predecir su trayectoria posterior. Su eficacia quedó probada a raíz de la epidemia causada por el virus H1N1 entre 2009 y 2010. Mayer y Cukier comienzan su libro con un interesante comentario de este caso (2013, págs. 11-13). Para ellos, la originalidad que presentaba el método de Google fue que no estaba ligado a datos del diagnóstico de gripe procedentes del estudio médico, sino que se basaba "en los *big data*, los 'datos masivos': la capacidad de la sociedad de aprovechar la información de formas novedosas para obtener percepciones útiles o bienes y servicios de valor significativo". Su trabajo había consistido en hallar correlaciones entre la frecuencia de ciertas búsquedas de información y la propagación de la gripe a lo largo del tiempo y del espacio. Una enorme cantidad de modelos matemáticos compararon sus predicciones frente a los datos oficiales de años anteriores, lo que sirvió para identificar una combinación de cuarenta y cinco términos de búsqueda, como "gripe", "catarro", "fiebre" o "dolor de cabeza", que, en su conjunto, ofrecía una fuerte correlación con los datos oficiales. Teniendo en cuenta que los de los años anteriores, procedentes de encuestas o informes, habían tardado en conocerse hasta una o dos semanas después de su recogida, Google puso en 2009 su aplicación a disposición de los responsables sanitarios en más de dieciséis países para ayudarles a anticipar y combatir más eficazmente los brotes de gripe.[18]

[18] EFE (2009, 8 de octubre). "La nueva herramienta 'Google Flu Trends' ya está disponible en España" [Fecha de consulta: 28 de noviembre de 2015]
http://www.soitu.es/soitu/2009/10/08/info/1255003928_366743.html

El proyecto *Artemis* para la unidad de prematuros del Hospital For Sick Children de Toronto

Otro caso interesante al que también hacen referencia Mayer y Cukier (2013, págs. 80-81), a propósito del valor preventivo del análisis en tiempo real de datos masivos procedentes de dispositivos de monitorización, es el proyecto desarrollado por la doctora Carolyn McGregor al frente de un equipo de investigadores del Instituto de Tecnología de la Universidad de Ontario. Se implementó por primera vez en el Hospital para Niños Enfermos de Toronto en agosto de 2009 y ha estado funcionando continuamente desde entonces. Utilizando una plataforma de Big Data de IBM (*IBM InfoSphere Streams*) desarrolló una aplicación que recoge y analiza en tiempo real hasta dieciséis flujos continuos de datos de los dispositivos de monitorización de signos vitales de bebés prematuros (ritmo cardíaco, frecuencia respiratoria, temperatura, etc.) de modo que vigile la aparición de cambios apenas perceptibles (que la máquina ha aprendido a reconocer basándose en el análisis de datos históricos) que suelen acontecer 24 horas antes de que aparezcan síntomas de infección grave. La detección de esos cambios enciende la alerta que permite a los médicos comenzar el tratamiento de la infección o bien los advierte de que el tratamiento que se está haciendo no es efectivo.[19]

* * *

La utilización de tecnología Big Data ha proporcionado también éxitos en el campo de la investigación biomédica. Destacaremos algunos que se han realizado con la participación de equipos de investigadores españoles y que han tenido mayor repercusión mediática.

En Cataluña se encuentran importantes instalaciones como el Centro de Supercomputación de Barcelona (BSC) o el Centro Nacional de Análisis Genómico (CNAG) que juegan un papel fundamental en la biomedicina a nivel europeo.[20] Los investigadores del BSC son los creadores del método Somatic Mutation Finder (*SMuFin*)[21] que mediante la analítica de Big Data permite detectar variaciones (mutaciones) en el genoma de las células tumorales por comparación directa con el genoma de las células normales del mismo paciente.

[19] Véase también Horowitz, Brian T. (2013, 26 de septiembre).
[20] Ferrado, Mònica L. (2015, Junio).
[21] http://cg.bsc.es/smufin/,

Estudio del genoma de la leucemia linfática crónica

En su artículo "Mutaciones en la zona oscura del genoma causan leucemia", el Dr. Modesto Orozco, científico destacado del grupo de investigación Consorcio Español para el Estudio del Genoma de la Leucemia Linfática Crónica,[22] explica la importancia del análisis genómico que están realizando desde hace varios años sobre pacientes diagnosticados de leucemia linfática crónica (LLC), una vez que se ha superado el número de 500 casos estudiados, y con motivo de la publicación el día anterior de los resultados del trabajo en la revista *Nature*.[23] En palabas del Dr. Orozco, la LLC "es la leucemia más frecuente en los países occidentales, con más de mil nuevos pacientes diagnosticados cada año en nuestro país. Conocer qué cambios genéticos provocan el desarrollo tumoral y cómo influyen en el desarrollo de la enfermedad es un paso fundamental para mejorar el tratamiento del cáncer". El estudio les ha permitido identificar 60 genes diferentes en los que las mutaciones pueden provocar el desarrollo del tumor, así como descubrir que algunas de estas mutaciones ocurren en zonas por lo general poco exploradas del genoma. Consideran que esos resultados permitirán una mejor clasificación del paciente, así como un tratamiento basado en el tipo de mutaciones genéticas que sean detectadas en el tumor.

* * *

Predicción de respuesta al tratamiento en linfoma de Hodgkin

Otro enfoque de investigación oncológica mediante la analítica de Big Data es el realizado por un equipo de investigadores dirigido por el Prof. Juan Luis Fernández-Martínez de la Universidad de Oviedo, en colaboración con el Centro de Inteligencia Artificial de Gijón y los servicios de Hematología de la red pública del Principado de Asturias. En la entrevista que Covadonga Díaz le hace para *Diario Médico* (2015, 13 de julio), Fernández-Martínez hace una valoración del trabajo de su grupo, publicado pocos meses antes en la revista *Clinical and Translational Oncology* (Andrés-Galiana, 2015). Explica que el linfoma de Hodgkin (LH) es un tipo de cáncer cuyo porcentaje de curación es elevado, pero la respuesta al tratamiento es difícil de predecir y muy cambiante según los pacientes. La

[22] http://www.cllgenome.es
[23] La referencia que cita M. Orozco: Puente, XS, Bea S, Valdés-Mas R et al. (2015, 22 de julio) "Non-coding recurrent mutations in chronic lymphocytic leukaemia", *Nature* [En línea] http://www.nature.com/nature/journal/v526/n7574/full/nature14666.html#affil-auth

valoración de progresión de la enfermedad se realiza mediante imagen radiológica (TAC y PET), y para evaluar el pronóstico de respuesta al tratamiento se ha estado utilizando un conjunto de 35 variables biológicas. El trabajo ha consistido en someter a un nuevo estudio mediante técnicas de analítica predictiva el pronóstico de respuesta al tratamiento de 263 pacientes que habían sido diagnosticados de LH, utilizando ahora un gran conjunto de datos de valoración de progresión y de variables biológicas de pronóstico de respuesta, con el fin de descubrir cuáles de estas variables determinarán con mayor probabilidad qué paciente iba a responder y quién no. Mediante la aplicación de algoritmos de aprendizaje automático y de optimización global cooperativa[24] se logró un modelo predictivo que seleccionaba, entre las 35 variables evaluadas, únicamente a tres que eran altamente discriminatorias de progresión o de remisión de la enfermedad: la ferritina sérica, la alina transaminasa, y la fosfatasa alcalina. Con ello se simplificaba considerablemente el estudio para conocer anticipadamente si el tratamiento iba a ser efectivo, al tiempo que se mantenía el reto de mejorar el tratamiento para el grupo de no respondedores.

[24] En el "Abstract" del artículo se describen los métodos con bastante detalle.

6. Conclusiones

Se puede afirmar que Big Data surge como respuesta a la creciente necesidad de un número cada vez mayor de organizaciones de disponer de nuevas herramientas capaces de procesar de forma eficiente un enorme volumen de datos de diversa tipología procedentes de fuentes muy heterogéneas. A diferencia de otras soluciones tradicionales, Big Data permite gestionar grandes volúmenes de datos de todo tipo, especialmente no estructurados, a una velocidad muy superior y con un consumo de recursos mucho menor. Otra de las grandes ventajas que ofrece esta tecnología es la posibilidad de utilizar técnicas de analítica avanzada, como la analítica predictiva, con el objetivo de operar sobre datos masivos y construir modelos predictivos de calidad que sirvan de soporte a la toma de decisiones.

Recientemente se ha suscitado un debate acerca de la posibilidad de que a corto plazo Big Data sustituya completamente a las tecnologías precedentes. Sin embargo, en el caso de Business Intelligence, son numerosos los autores que defienden los beneficios de la coexistencia de ambas soluciones en un mismo Sistema de Información. En este sentido, aunque en algunos casos se puede plantear la conveniencia de utilizar Big Data para realizar las funciones tradicionalmente asignadas a Business Intelligence, no es descartable la utilización de las dos soluciones para desarrollar funciones complementarias. En estos supuestos, tras la correcta integración de ambas tecnologías, Big Data se centraría en el procesamiento de datos no estructurados, mientras que Business Intelligence aportaría el análisis avanzado de esta información a partir de herramientas consolidadas que ofrecen al usuario interfaces visualmente atractivas.

Los resultados de diversas encuestas acerca de Big Data realizadas durante los últimos tres años a numerosos ejecutivos de diferentes países y sectores empresariales ponen de manifiesto la progresiva adopción de esta tecnología por parte de la mayoría de organizaciones. También se puede observar cómo se ha pasado de una fase inicial de rápido crecimiento a una etapa actual de cierta consolidación, aunque con distintos grados de implantación en función del sector y del tamaño de la organización. Asimismo se puede afirmar que la opinión mayoritaria por parte de los encuestados sobre Big Data es bastante positiva. De hecho, existe prácticamente un consenso acerca de la importancia de esta tecnología en la

transformación digital de las organizaciones. También existe una percepción cada vez más generalizada de los beneficios que pueden obtenerse a partir de su implantación, especialmente en la búsqueda de nuevas fuentes de ingresos, en el desarrollo de nuevos productos y servicios y en la mejora de la experiencia del cliente.

Sin embargo, en estas mismas encuestas también se ponen de manifiesto los obstáculos que dificultan la adopción empresarial de Big Data, especialmente aquellos relacionados con la seguridad, las limitaciones presupuestarias y la falta de especialistas en el uso de este tipo de soluciones. Hay que destacar igualmente la incertidumbre que provoca en numerosas empresas la posibilidad de rentabilizar en el futuro las inversiones realizadas en esta tecnología.

Al igual que en el resto de sectores, el ámbito de la salud también debe hacer frente a las nuevas necesidades que plantea el incremento en volumen, variedad y velocidad de los datos, siendo necesaria la aplicación de soluciones como Big Data, cuyo potencial, como señala Bonnie Feldman, radica en la capacidad de combinar datos tradicionales con nuevas formas de datos, tanto a nivel individual como poblacional.

La inversión en nuevas tecnologías, entre las cuales Big Data juega un papel muy destacado, es clave además para garantizar la sostenibilidad a medio plazo de la sanidad pública en países como España, que deben enfrentarse al aumento de costes sanitarios derivados del envejecimiento poblacional. En este sentido, la inversión en soluciones TIC puede suponer un considerable ahorro en el presupuesto destinado al gasto sanitario, como ya se ha demostrado en el informe *Big Data: The Next Frontier for Innovation, Competition and Productivity*, aplicado a la atención sanitaria en EE.UU.

Además, la capacidad de Big Data de transformar datos no estructurados en datos estructurados fácilmente manipulables por sistemas informáticos resulta especialmente relevante en este sector, ya que permite aplicar técnicas de análisis predictivo en tareas como la identificación de pacientes con riesgo de reingreso o la prevención de infecciones hospitalarias o de enfermedades crónicas, obteniendo modelos predictivos de calidad.

La operativa clínica es una de las áreas en donde Big Data podría tener mayor repercusión, proporcionando grandes beneficios en la investigación comparativa de la eficacia de los tratamientos, como sistema de soporte a las decisiones clínicas, creando transparencia sobre los datos médicos, facilitando la monitorización de pacientes a distancia y ayudando a reconocer perfiles de pacientes mediante el uso de técnicas de analítica avanzada.

Asimismo hay que subrayar el valor que Big Data puede aportar a la investigación y el desarrollo en la industria farmacéutica. Esta tecnología puede desempeñar un papel fundamental en la mejora de la productividad, aplicando recursos como el modelado predictivo en el desarrollo de nuevos medicamentos o herramientas y algoritmos estadísticos en el diseño de ensayos clínicos. También en esta misma área Big Data puede ser de gran utilidad en el análisis de patrones de enfermedades.

Entre las disciplinas médicas en donde Big Data puede ayudar a conseguir mayores avances sobresalen las diversas ramas de la biología molecular, y especialmente la genómica. Dentro de esta área, los conocimientos obtenidos a partir de los resultados de la investigación genómica aplicados a la práctica clínica y al desarrollo de nuevos medicamentos "ad hoc" pueden contribuir de forma decisiva al desarrollo del concepto de "medicina personalizada".

Big Data puede ayudar igualmente a fomentar el auto-cuidado y la colaboración ciudadana, recogiendo datos procedentes de sensores instalados en dispositivos móviles y devolviendo información a los usuarios sobre su estado de salud. También permite el intercambio rápido y seguro de información entre pacientes y registros médicos y hace posible que el personal sanitario pueda disponer de información en tiempo real sobre el estado de salud de los pacientes.

Todas las aplicaciones señaladas indican que la tecnología Big Data está destinada a jugar un papel fundamental en la transformación de los sistemas sanitarios, con el objetivo de acercarse cada vez más hacia la medicina del futuro, que algunos autores definen como la "Medicina de las 4P: predictiva, preventiva, personalizada y participativa".

No obstante, se puede afirmar que en la actualidad la implantación de Big Data en el sector sanitario a nivel global se encuentra todavía en una fase inicial. Este es el caso de España, en donde los expertos en tecnologías de la información de los principales

hospitales y centros asistenciales coinciden en afirmar que actualmente apenas existen proyectos de esta tecnología en la práctica clínica. Hay que destacar en este sentido los grandes retos que Big Data plantea a nivel tecnológico, administrativo y legal para su implantación en los diferentes entornos sanitarios, además de los obstáculos ya señalados para el conjunto de las organizaciones.

Desde el punto de vista tecnológico, es necesario realizar importantes inversiones con el objetivo de adaptar la arquitectura de las redes de comunicación y las infraestructuras tecnológicas actuales a los requisitos funcionales de Big Data. Por otro lado, la captura de datos de comportamiento de los pacientes, así como las cuestiones de privacidad que pueden generar riesgos para los individuos son otros retos importantes que se deben afrontar a nivel administrativo y legal. En este sentido, las diferentes administraciones españolas están llevando a cabo en la actualidad importantes esfuerzos por adaptar la legislación vigente a los nuevos desafíos que plantea el uso de esta tecnología, especialmente en materia de protección de datos, en colaboración con las instituciones europeas.

Además de los factores señalados, para aprovechar las ventajas que puede ofrecer Big Data a este sector también se requiere la formación de equipos multidisciplinares compuestos por personal sanitario y profesionales provenientes de otras disciplinas, como la informática o la estadística, que dispongan de los conocimientos y las habilidades necesarias para manejar las herramientas relacionadas con esta tecnología.

Teniendo en cuenta las dificultades expuestas, es previsible que durante los próximos años la implantación de esta tecnología en los diferentes entornos sanitarios se lleve a cabo de forma progresiva y gradual, tanto a nivel nacional como internacional, y que aquellos países e instituciones que en la actualidad destinen los recursos necesarios comiencen a beneficiarse a medio plazo de ventajas como una reducción sustancial de los costes sanitarios y una notable mejora de la calidad asistencial, derivadas del uso de Big Data.

7. Bibliografía

Andrés-Galiana, E.J.; Fernández-Martínez, J.L., et al. (2015, 21 de abril). “On the prediction of Hodgkin lymphoma treatment response”, *Clinical and Translational Oncology.* [Fecha de consulta: 28 de noviembre de 2015] <http://link.springer.com/article/10.1007%2Fs12094-015-1285-z>

Barranco Fragoso, Ricardo (2012, 18 de junio) “¿Qué es Big Data?”. *IBMdeveloperWorks*. [Fecha de consulta: 20 de octubre de 2015] <https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/ >

"Big Data y su fuerza para Business Intelligence" (2013, octubre). *Stefanini.* [Fecha de consulta: 12 de octubre de 2015] <https://stefanini.com/es/2013/10/big-data-y-su-fuerza-para-business-intelligence/>

"Business Intelligence y Big Data, ¿independencia o cooperación?" (2013, mayo), *Saima solutions*. [Fecha de consulta: 12 de octubre de 2015] <http://saimasolutions.com/business-intelligence-big-data/>

Camba, Pablo (2015, 14 de septiembre). “El ‘big data’ aplicado a la salud. [Fecha de consulta: 19 de noviembre de 2015] <http://www.diariomedico.com/2015/09/14/area-profesional/gestion/el-big-data-aplicado-a-la-salud>

Cepeda, José Mª (2013, 24 de Julio). “Redes sociales y salud: las comunidades virtuales de pacientes”. [Fecha de consulta: 12 de noviembre de 2015] < http://saludconectada.com/redes-sociales-comunidades-virtuales-de-pacientes/>

Conesa Caralt, Jordi, Curto Díaz, Josep (2010), *Introducción al Business Intelligence*, Barcelona: Editorial UOC

Díaz, Covadonga (2015, 13 de Julio) “*Big data* para predecir la respuesta al tratamiento” [Fecha de consulta: 28 de noviembre de 2015] <http://www.diariomedico.com/2015/07/13/area-profesional/gestion/big-data-predecir-respuesta-tratamiento>

Feldman, Bonnie; Martin, Ellen M.; Skotnes, Toby (2012, Octubre). *Big Data Healthcare Hype and Hope*. Disponible en pdf: < http://www.west-info.eu/files/big-data-in-healthcare.pdf>

Ferrado, Mònica L. (2015, Junio). “Líderes en ‘Big Data’ y Biomedicina”, *Barcelona Metrópolis*, nº 96. [Fecha de consulta: 25 de noviembre de 2015] <http://w2.bcn.cat/bcnmetropolis/es/dossier/liders-en-big-data-i-biomedicina/ >

García Barbosa, Julián (2014, 9 de octubre). “La medicina del futuro pasa por big data” [Fecha de consulta: 12 de noviembre de 2015] http://www.aunclicdelastic.com/la-medicina-del-futuro-pasa-por-big-data/

García González, David (2014, 25 de abril). "Big Data vs. Business Intelligence – ¿Complemento o Sustituto?" *Necsia IT Consulting.* [Fecha de consulta: 12 de octubre de 2015] <http://necsia.es/big-data-vs-business-intelligence-complemento-o-sustituto/>

Groenfeldt, Tom (2012, 1 de febrero). “IBM's Watson, Cedars-Sinai and WellPoint Take On Cancer” [Fecha de consulta: 28 de noviembre de 2015]. <http://www.forbes.com/sites/tomgroenfeldt/2012/02/01/ibms-watson-cedars-sinai-and-wellpoint-take-on-cancer/>

Gualtieri, Mike, Curran, Rowan (2015, 1 de abril). "The Forrester Wave™: Big Data Predictive Analytics Solutions, Q2 2015”. Forrester. [Disponible en pdf] <https://www.sas.com/content/dam/SAS/en_us/doc/analystreport/forrester-wave-predictive-analytics-106811.pdf>

Guazzelli, Alex (2011, 29 de noviembre) “Predictive analytics in healthcare. The importance of open standards”, *IBM developerWorks* [Fecha de consulta: 15 de octubre de 2015] <http://www 0.ibm.com/developerworks/library/ba-ind-PMML3/>

Guazzelli, Alex (2012, 26 de noviembre). “Predicciones sobre el Futuro. Parte 1: ¿Qué es la Analítica predictiva?”. *IBM developerWorks*. [Fecha de consulta: 5 de octubre de 2015] <http://www.ibm.com/developerworks/ssa/industry/library/ba-predictive-analytics1/>

Guazzelli, Alex (2012, 17 de diciembre). “Predicciones sobre el futuro, parte 2: Técnicas de modelado predictivo”. *IBM developerWorks*. [Fecha de consulta: 5 de octubre de 2015] <http://www.ibm.com/developerworks/ssa/industry/library/ba-predictive-analytics2/>

Guazzelli, Alex (2012, 24 de diciembre). “Predicciones sobre el futuro, parte 3: Cree una solución predictiva”. *IBM developerWorks*. [Fecha de consulta: 6 de octubre de 2015] <http://www.ibm.com/developerworks/ssa/industry/library/ba-predictive-analytics3/>

Guazzelli, Alex (2013, 14 de enero). “Predicciones sobre el futuro, parte 4: Ponga en marcha una solución predictiva”. *IBM developerWorks*. [Fecha de consulta: 6 de octubre de 2015] <http://www.ibm.com/developerworks/ssa/industry/library/ba-predictive-analytics3/>

Heudecker, Nick; Kart, Lisa (2015, 3 de septiembre) “Survey Analysis: Practical Challenges Mount as Big Data Moves to Mainstream”. *Gartner* [Fecha de consulta 20 de octubre]

http://www.gartner.com/document/3123817?ref=QuickSearch&sthkw=heudecker%20AND%20kart&refval=155652106&qid=2aa81339616939553482e7aae26523f4

Horowitz, Brian T. (2013, 26 de septiembre). “IBM InfoSphere, Big Data Help Toronto Hospital Monitor Premature Infants” *eWEEK* [Fecha de consulta: 30 de noviembre de 2015]. http://www.eweek.com/enterprise-apps/ibm-infosphere-big-data-help-toronto-hospital-monitor-premature-infants.html

Iglesias Fraga, Alberto (2015, 28 de septiembre). "El 75% de las empresas invertirá en Big Data durante los próximos dos años", *TICbeat.* [Fecha de consulta: 12 de octubre de 2015] <http://www.ticbeat.com/bigdata/el-75-de-las-empresas-invertira-en-big-data-durante-los-proximos-dos-anos/>

Jamack, Peter J. (2013, 4 de febrero). “Analítica de inteligencia de negocios de big data”. *IBM developerWorks.* [Fecha de consulta: 5 de noviembre de 2015] <http://www.ibm.com/developerworks/ssa/library/ba-big-data-bi/>

Joyanes Aguilar, Luis (2014). *Big Data. Análisis de grandes volúmenes de datos en las organizaciones*. Barcelona: Marcombo

Laney, Doug (2012, 14 de enero). “Deja VVVu: Others Claiming Gartner’s Construct for Big Data”. *Gartner Blog Network*. [Fecha de consulta: 10 de octubre de 2015] <http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/>

López López, Visitación (2015, 4 de mayo). “Big data sanitario: el acelerador del conocimiento y la decisión clínica”. [Fecha de consulta: 20 de noviembre de 2015] <http://www.aunclicdelastic.com/big-data-sanitario-el-acelerador-del-conocimiento-y-la-decision-clinica/>

Maqueda, Gema (2012, 16 de enero). “¿Qué puede hacer Big Data por el Sector Sanitario” [Fecha de consulta: 1 de noviembre de 2015] <http://www.aunclicdelastic.com/que-puede-hacer-big-data-por-el-sector-sanitario/>

Mayer-Schönberger, Viktor; Cukier, Kenneth (2013). *Big data. La revolución de los datos masivos*. Madrid: Turner.

Mayol, Julio (2014, 22 de abril). “Los 6 retos del Big Data en la sanidad” [Fecha de consulta: 15 de noviembre de 2015] < http://www.comparteinnovacion.philips.es/tendencias-en-salud/articulos/los-6-retos-del-big-data-en-la-sanidad >

McKinsey Global Institute (2011, Junio). *Big Data: The Next Frontier for Innovation, Competition and Productivity*. Disponible en pdf: <http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx>

Núñez, Mercedes (2014, 5 de febrero). "Las asombrosas cifras de la mHealth". [Fecha de consulta: 14 de noviembre de 2015] < http://www.aunclicdelastic.com/las-asombrosas-cifras-de-la-mhealth/>

Núñez, Mercedes (2014, 9 de diciembre). "Diez claves de la eSalud en España" [Fecha de consulta: 18 de noviembre de 2015] < http://www.aunclicdelastic.com/diez-claves-de-la-esalud-en-espana/>

Orozco, Modesto (2015, 23 de julio) "Mutaciones en la zona oscura del genoma causan leucemia" [Fecha de consulta: 20 de noviembre de 2015] https://www.irbbarcelona.org/es/news/mutaciones-en-la-zona-oscura-del-genoma-causan-leucemia

Paniagua, Soraya (2012, 12 de noviembre). "Big Data en sanidad para predecir, prevenir y personalizar" [Fecha de consulta: 1 de noviembre de 2015] <http://www.sorayapaniagua.com/2012/11/12/big-data-en-sanidad-para-predecir-prevenir-y-personalizar/>

Paniagua, Soraya (2013, 16 de octubre), "La falta de habilidades, principal barrera para la adopción de Big Data en España". [Fecha de consulta: 7 de octubre de 2015] <http://www.sorayapaniagua.com/2013/10/16/la-falta-de-habilidades-principal-barrera-para-la-adopcion-de-big-data-en-espana/>

Pérez García, Alejandro (2014, 28 de febrero), " Kettle no es una tetera, es la herramienta de ETL de Pentaho!". [Fecha de consulta: 13 de octubre de 2015] <http://www.adictosaltrabajo.com/tutoriales/kettle/>

Rodríguez, José Ramón (2012, 10 de octubre). "Octubre, el mes de los grandes datos". *Informática++*. [Fecha de consulta: 13 de octubre de 2015] http://informatica.blogs.uoc.edu/2012/10/10/octubre-el-mes-de-los-grandes-datos-2/>

Ros, Josep (2015, 14 de julio), "Introducción al Big Data con Pentaho", *El blog de encora.* [Fecha de consulta: 13 de octubre de 2015] <http://blog.ncora.com/2015/07/introduccion-al-big-data-con-pentaho.html>

Sánchez, José Luis (2014, septiembre). "Las empresas consideran Big Data fundamental para su transformación digital". *Informe ACCENTURE*. [Fecha de consulta: 13 de octubre de 2015] <https://www.accenture.com/es-es/company-big-data-fundamental-transformacion-digital.aspx>

Seyfert-Margolis, Vicki (2015). "Médicos y Pacientes", en *Big Data. El poder de los datos*. Fundación Innovación Bankinter. Disponible en línea: <https://www.fundacionbankinter.org/documents/11036/68724/Big+Data+ES+Completo+2015/4c567aad-92a4-4fc1-8b25-8cda397bb81f>

Siegel, Eric (2014). *Analítica predictiva. Predecir el futuro utilizando Big Data*. Madrid: Anaya

Schroeck, Michael; Shockley, Rebecca; y otros (2012). *Informe ejecutivo*. *Analytics: el uso de big data en el mundo real. Cómo las empresas más innovadoras extraen valor de datos inciertos.* IBM Global Business Services Business Analytics and Optimisation y Escuela de Negocios Saïd en la Universidad de Oxford. [Fecha de consulta: 8 de octubre 2015] <http://www05.ibm.com/services/es/gbs/consulting/pdf/El_uso_de_Big_Data_en_el_mundo_real.pdf>

Torralba, Patricia (2015, 14 de abril). "IBM crea una nueva unidad de negocio, Watson Health, orientada al sector sanitario". Nota de prensa. IBM. [Fecha de consulta: 15 de diciembre de 2015). <http://www-03.ibm.com/press/es/es/pressrelease/46621.wss>

VV.AA. (2015, diciembre). *Informe Big Data y Salud.* Ed.: Prodigioso Volcán y Planner Media. <http://www.plannermedia.com/downloads/informebigdataysalud.pdf >

Wikipedia, "Análisis predictivo". [Fecha de consulta: 16 de octubre de 2015] <https://es.wikipedia.org/wiki/Análisis_predictivo>

Wikipedia, "Minería de datos" [Fecha de consulta: 16 de octubre 2015] <https://es.wikipedia.org/wiki/Miner%C3%ADa_de_datos>