UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Diseño de un modelo para la creación de conocimiento y toma de decisiones en el área sanitaria

TRABAJO FIN DE GRADO

Grado en Ingeniería Informática

Autor: Carlos Almela Seller

Tutor: Eva María Cutanda García

Curso 2018-2019

Resum

La salut ha experimentat una millora en els processos assistencials i a nivell d'investigació, i també una creixent implantació d'equips informàtics promoguda per la informatització dels sistemes d'administració a nivell global en tot els sectors. Això ens porta a què aspectes quotidians com les dades d'una consulta, expedients sanitaris electronics, altes mediques o gestió de recursos humans en hospitals quede reflectida en documents digitalitzats amb multitud de dades que poden explotar-se per a obtindre informació i generar coneixement.

El nostre objectiu és mostrar i explicar quals poden ser els beneficis d'explotar les dades dels sistemes d'informació en un hospital, creant a partir d'ells informació valuosa tant per a qualsevol àmbit de l'hospital, ja siga mèdic o enfocat a la gestió del mateix, amb la meta de millorar l'atenció als pacients. Per a això s'abordaran diverses tècniques i termes que engloben aquest àmbit, així com també s'elaborarà un model que ens permeta la creació de coneixement i la presa de decisions.

Paraules clau: Intel·ligencia de negoci, macrodades, salut, presa de decisions

Resumen

La salud ha experimentado una mejora en los procesos asistenciales y a nivel de investigación, y también una creciente implantación de equipos informáticos promovida por la informatización de los sistemas de administración a nivel global en todo los sectores. Ello nos lleva a que aspectos cotidianos como los datos de una consulta médica, expedientes sanitarios electrónicos, altas médicas o gestión de recursos humanos en hospitales quede reflejada en documentos digitalizados con multitud de datos que pueden explotarse para obtener información y generar conocimiento.

Nuestro objetivo es mostrar y explicar cuales pueden ser los beneficios de explotar los datos de los sistemas de información en un hospital, creando a partir de ellos información valiosa para cualquier ámbito del hospital, ya sea médico o enfocado a la gestión del mismo, con la meta de mejorar la atención a los pacientes. Para ello se abordarán diversas técnicas y términos que engloban dicho ámbito, así como también de elaborará un modelo que nos permita la creación de conocimiento y la toma de decisiones.

Palabras clave: Inteligencia de negocio, macrodatos, salud, toma de decisiones

Abstract

Health has experienced an improvement in the assistance processes and in a research level, and also a recently growing implantation of informatic equipment promoted by the computerization of the management systems in a global level in all sectors. Thus take us to the quotidian aspects like the data of a medical consult, electronic health records, medical releases or Human Resources Management in hospitals that are recorded in digitalized documents with a a lot of data that can be mined to obtain information and generate knowledge.

Our objective is to show and explain which could be the benefits of mining the date of the information systems in a hospital, creating from them valuable information for any scope of the hospital, weather in a medical one or focused to the management of it, with the goal of improving the attention given to the patients. For it different techniques and terms related with the scope will be approached, as well as it will be elaborated a model that allows the creation of knowledge and the decision-making.

Key words: Business Intelligence, Big Data, health, decision making

Índice general

Índice de figuras

CAPÍTULO 1
Introducción

1.1 Motivación

La motivación que me lleva a la realización de dicho proyecto es el conocimiento de la existencia de diversas técnicas de procesamiento y análisis de datos en el entorno empresarial, que pueden ser potencialmente aplicadas a cualquier ámbito u organización para la obtención de distintos enfoques y descubrimiento de tendencias, lo que nos facilita la implantación de mejoras en la estructura de la organización y en la toma de decisiones. Tras haber descubierto, a grandes rasgos las mejoras que se podían obtener gracias a una asignatura enfocada en *business intelligence*, descubrí por medio a la asignatura de 'Gestión de la innovación y tecnologías en la salud' la influencia de la tecnología y técnicas vanguardistas sobre la salud, y los grandes beneficios que se pueden obtener en el sistema sanitario gracias a la utilización de las mismas.

La realización de varios trabajos y tareas realizadas en ambas asignaturas me proporcionó el punto de vista desde el que se plantea la realización de este proyecto, en el que convergen las esencias de ambas asignaturas para el desarrollo de un modelo que nos ayude a ofrecer un mejor servicio sanitario a los usuarios.

1.2 Objetivos

Nuestro objetivo es mostrar y explicar cuales pueden ser los beneficios de explotar los datos de los sistemas de información en un entorno sanitario a partir del análisis de los mismos. La transformación de los datos provenientes de las distintas fuentes de información en conocimiento es el estado clave al que queremos llegar, implementando dicho conocimiento en la toma de decisiones y estructura de las bases de la sanidad. El objetivo final es el de mejorar la atención proporcionada a los pacientes a través del apoyo de la tecnología, proponiendo una modernización de la sanidad y por ello de sus servicios asistenciales.

Se diseñará un modelo de creación de conocimiento a partir de los conceptos estudiados en el estado del arte, siendo capaces de darles uso y ponerlos en práctica. El modelo diseñado tiene como propósito formar parte del núcleo del sistema de información sanitario, aportando las mejoras que se describen a lo largo del proyecto.

Se materializará una pequeña herramienta en torno a datos anonimizados de Atención Primaria de 2016, proporcionados por el Ministerio de Sanidad, Consumo y Bienestar Social. Con ello se pretende recrear, a pequeña escala, cuales serían los beneficios de tener un sistema *big data* en un nivel sanitario basado en la asistencia al paciente, como podría ser una consulta de especialización de un hospital general.

1.3 Estructura de la memoria

El estado del arte comenzará en el segundo capítulo, en el que se introducirá el mundo del *big data*, tratando desde sus aspectos más relevantes como sus características, usos y definición del mismo.

Posteriormente se presentará el ámbito sanitario, el cual es objeto de un breve análisis en este proyecto. Se describirá su sistema de información actual y se explicará detalladamente las posibilidades del uso del *big data* en el ámbito sanitario, así como la repercusión de la tecnología en la salud.

Con ello, habiéndose introducido ambos temas centrales del proyecto, y habiéndoles dado la importancia necesaria, se desarrollará un modelo para la creación de conocimiento y toma de decisiones. A lo largo del capítulo 4 se describirá el modelo analizando cada una de las partes del mismo y exponiendo la solución que se plantea. Se tratará de temas relacionados con el diseño del mismo así como también varios temas relacionados con su implementación en sanidad.

Posteriormente, en el capítulo 5, se propone la realización una herramienta para la creación de conocimiento y toma de decisiones a través de la explotación y análisis en tiempo real de los datos obtenidos por el Ministerio de España, y podremos observar un pequeño prototipo de la herramienta desarrollada. A través de la herramienta observaremos las distintas características del modelo diseñado en este proyecto, plasmando el potencial de la propuesta del proyecto.

Para terminar en el capítulo 6 se habla de los posibles futuros trabajos que puedan desarrollarse en futuras etapas académicas.

CAPÍTULO 2

Datos masivos

Desde los inicios de la informática, cuando hablamos de análisis de datos nos referimos a dicho análisis como un proceso en el que se ven implicadas las capacidades teóricas de un experto sobre las técnicas de análisis y la parte analítica que suele ser llevada a cabo por un computador o procesador de datos, debido a la eficiencia temporal que nos otorga el uso de los mismos. A lo largo de los años, la capacidad de procesamiento de instrucciones y operaciones de lo que hoy en día conocemos como ordenadores o computadores ha sufrido un aumento exponencial debido a la ley de Moore. Éste hecho ha tenido una repercusión favorable en distintos aspectos en la sociedad, aportando muchas ventajas a las personas a la hora de desarrollar diversas tareas.

Además también ha llevado consigo el desarrollo de nuevas tareas estrechamente enlazadas con la obtención de dichos análisis, con el uso o con el perfeccionamiento del mismo análisis para conseguir un objetivo muy concreto: observar tendencias que nos ayuden a mejorar cualquier tipo de servicio o producto.

En el ámbito más organizacional, se trata de obtener nuevos puntos de vista que nos permitan comprender cuándo, cómo, por qué y más concretamente qué ofrecer a nuestro cliente. Éstas técnicas, las cuales serán descritas a lo largo del proyecto no sólo están enfocadas a la mejora de los ingresos de las organizaciones si no que son una fuerte herramienta para mejorar servicios o aumentar la satisfacción del cliente.

Actualmente, se utiliza con mucha frecuencia el termino *big data*, el cual vamos a tratar en profundidad para entender mejor las bases de este proyecto, ya que está relacionado estrechamente con éste.

2.1 Big data

Algunas organizaciones y científicos indican que el primer uso del término *big data* lo acuñó la NASA para señalar la creciente cantidad de datos que suponía un problema para los sistemas informáticos de aquel entonces. Otras fuentes afirman que es Jhon Mashey, ingeniero informático, el que hace uso del término por primera vez en una reunión de *Sillicon Valley Inc*. Ambas definiciones aparecen a mediados de 1990, año en el que ya se observaba una gran presencia de las comunicaciones bidireccionales y un aumento de la implementación de sistemas de administración automatizados dentro de las empresas, además también se había creado la WWW y las bases de datos ya estaban expandidas.

El término *big data* es un concepto usado hoy en día en multitud de sectores de la sociedad, habiendo sufrido un crecimiento exponencial en su uso desde que aparece el término. A lo largo de los años dicha definición permuta hasta conocer la que conocemos hoy en día, debido en parte al entorno cambiante de la información y las tecnologías que envuelven al mundo de la informática.

Big data se utiliza actualmente para referirse a la explotación masiva de datos, a partir de los cuáles podemos extraer información importante para nuestro negocio o simplemente tendencias y predicciones de los datos. Más allá de la definición global que nosotros le podemos dar a partir de nuestro entendimiento, podemos encontrar dichas definiciones [1]:

> *Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.*
>
> — Gartner IT Glossary, n.d.

> *Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.*
>
> — TechAmerica Foundation's Federal Big Data Commission, 2012

Si a partir de las mismas creamos una definición conjunta, y en castellano, lo que obtendríamos seria una definición más completa y unificada, mostrada de la siguiente manera:

> *Big data es el término que describe grandes volúmenes de datos a alta velocidad y con una gran variedad y complejidad, la cual requiere de innovadoras y avanzadas técnicas y tecnologías para su captura, almacenamiento, distribución, gestión y análisis.*

2.1.1. Características

Como ya hemos adelantado en la definición del mismo, el análisis de datos masivos se compone principalmente de tres características conocidas como "3V" y presentadas por Gartner en 2001 que son las que, por decirlo de alguna manera, le proporcionan el nombre que tiene debido a que no se puede tratar con los sistemas tradicionales de bases de datos.

Estas características son volumen, velocidad y variedad:

- **Volumen** es una magnitud que indica el tamaño de los datos a analizar, normalmente especificados en terabytes o petabytes, y que también puede expresarse en registros o tablas. A

su vez, volumen es un término confuso debido a que es bastante vago a la hora de cuantificar el tamaño que clasifica los datos como *big data* o no.

En base al estudio hecho por TWDI [2] las empresas actualmente suelen tratar un volumen medio de 10-100 TB, con una visión de futuro en la que incluyen un aumento del volumen a tratar hasta casi una cantidad similar a los datos que dice producir la compañía Boeing por vuelo. Estas ingentes cantidades de datos se deben a la capacidad de almacenar todos los datos que podemos guardar.

- **Velocidad** se refiere al ritmo con el que los datos son generados, un aspecto a tener cada vez mas en cuenta debido a la aparición y masificación de los dispositivos móviles, *wearables* o dispositivos comprendido en el término anglosajón *Internet of Things*. Así también cuando hablamos de velocidad se hace referencia a la rapidez con la que se actúa sobre los datos generados.

  Cada vez la velocidad con la que se generan los datos es mucho mayor, y también la necesidad de realizar análisis en tiempo real. A veces identificar una tendencia, problema o oportunidad antes que tu competidor es cuestión de minutos o segundos, así como analizar los datos antes de que se vuelvan obsoletos por la aparición de datos mas recientes. Uno de los datos relevantes en cuanto a la velocidad como magnitud, es la periocidad mediante la cual las empresas renuevan los datos que utilizan para los análisis, algo a tener muy en cuenta en *business intelligence*.[3]

- **Variedad** es el término que nos indica las diferentes fuentes y tipos de datos que son generados, en los que podemos incluir fuentes como sensores o redes sociales, y tipos de datos como texto, vídeo o *data logs*. Dichos datos no siempre se refieren a los tradicionales datos estructurados o relacionales, de hecho solo el 20 % de los datos que se almacenan suelen ser relacionales. Por ejemplo XML es un lenguaje textual para intercambio de datos que se trata de un formato semi-estructurado, incluido en un estudio del artículo 'Big Data Analytics' de TDWI [2], en el cual se concluye que el 92 % de las empresas trabajan con datos estructurados y el 54 % con datos semi-estructurados.

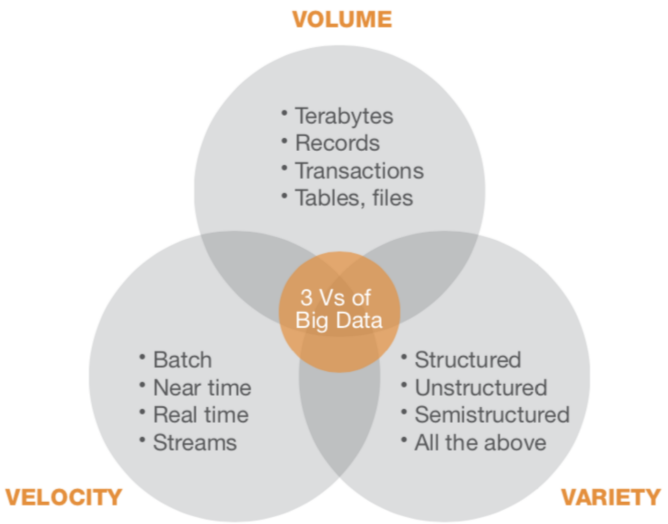


Figura 2.1: Las 3V del big data. [2]

Como característica añadida al *big data*, algunos especialistas y organizaciones añaden las siguientes como la cuarta V:

- **Veracidad**. IBM presenta la veracidad como la cuarta V, la cual se centra en la fiabilidad de los datos. Existe la necesidad de lidiar con los datos imprecisos e inciertos, y por ello es una cara más de los macrodatos.[4]

- **Variabilidad**. SAS introduce dicho concepto acompañado de la complejidad, constituyéndose como dos dimensiones más del *big data*. La variabilidad se refiere a la variación de los flujos de datos, y la complejidad a la diversidad en la generación de los datos, los cuales deben ser transformados para poder tratarlos correctamente en algunos casos.[5]

- **Valor**. Oracle lo introduce como un atributo esencial en el *big data*, a través del cual podemos indicar si un conjunto de datos tiene un valor de alta densidad, el cual puede transformar su valor tras ser tratado y analizado. Consiste básicamente de extraer información del conjunto de datos recogido.[6]

El potencial del valor, como característica del *big data*, es muy grande a la hora de proporcionar información para la toma de decisiones. Permite a las organizaciones crear procesos eficientes que pueden transformar grandes volúmenes de datos, que se generan a gran velocidad y proceden de diversas fuentes, en información valiosa que nos pueden aportar enfoques relevantes. Para ello deben llevarse a cabo diferentes procesos en los datos, dichos procesos pueden ser básicamente tecnológicos en los que contemplaríamos la extracción, transformación y almacenamiento de los datos, o analíticos lo cual implica una gran cantidad de opciones en cuanto a técnicas o plataformas.

2.2 Fuentes de datos

Las fuentes de información a día de hoy siguen ampliándose, llegando a ofrecer un gran abanico de alternativas para las empresas que desean explotar datos masivamente. Las diferentes clasificaciones según Kapow Software (figura 2.2) que podemos otorgar a los datos basándonos en su naturaleza son las siguientes:

- **Archivos**. Este tipo de archivos pueden ser registros médicos, documentos escaneados, archivos en papel o correspondencia de usuarios. Suelen tener una baja velocidad, variedad y también volumen, y no son fácilmente integrables en el sistema de almacenamiento de la empresa ya que los métodos de consulta son muy variados y no todos los archivos están digitalizados.

- **Documentos**. Quedan incluidos todo tipo de documentos como XML, JSON, CSV, etc, así como también de tipo ofimático como Word, Powerpoint y PDF. También encontramos correos electrónicos y archivos de texto plano. La diversidad de tipos de documentos nos indica que podemos observar una notoria variedad.

- **Multimedia**. La generación de este tipo de fuente suele ser muy alta, implicando una gran volumen y velocidad de creación de los datos. Imágenes, audios, vídeo, como los generados a partir de una radiografía, una ecografía, o los registros de audios de las llamadas de los *call centers*, son los que se incluyen en este tipo de fuente de información.

- **Almacenamiento de datos**. Aquí quedarían recogidos los datos más comunes de las organizaciones, como son las bases de datos, ficheros de sistema, repositorios o HDFS. Éstos suelen ocupar un alto volumen del total de la información almacenada por las empresas y están totalmente implementados para otorgar rapidez al momento de proporcionar consultas e informe.

Figura 2.2: Fuentes de datos. [7]

- **Aplicaciones de negocio**. Las empresas suelen tener a menudo diversas aplicaciones para la gestión de proyectos, recursos humanos, CRM, ERP o programas financieros. Todas estas aplicaciones generan una cantidad media de datos, que puede tener un valor significativo para la empresa. Suelen ser fáciles de explotar debido a su estructuración.

- **Web**. Hace referencia a la web pública, a todos esos datos que se pueden extraer de páginas de bases de datos como Wikipedia o IMDb, así como también datos de índole publica como los gubernamentales, trafico, tiempo, economía, etc. Éstos suelen ser totalmente externos a las organizaciones, y se contempla una gran variedad y volumen de información.

- **Redes Sociales**. Todas las redes sociales son recogidas en este bloque, las más conocidas como Facebook, LinkedIn o YouTube, y se destaca la velocidad con la que se generan los datos debido a su gran numero de usuarios a nivel mundial. El análisis en este tipo de datos suele ser más complicado debido a la implicación del factor del lenguaje natural y de sentimientos, para los cuales deben llevarse a cabo procesos específicos que nos clasifiquen dichos datos correctamente para su uso posterior.

- **Registros de máquinas**. Se incluyen todas las comunicaciones entre servidores, aplicaciones y máquinas, así como los registros que generan las mismas máquinas como, por ejemplo, registros de eventos, localización móvil o *clickstream*. Este tipo de datos tienen una muy alta variedad, volumen y velocidad de creación de los mismos, suelen ser fáciles de implementar en nuestro sistema de almacenamiento para posteriormente realizar consultas o informes sobre los mismos.

- **Sensores**. Aquí encontramos todo tipo de sensores como cámaras de tráfico, dispositivos médicos, etc. Podríamos incluir en este grupo todos los dispositivos relacionados con *Internet of Things* y *wearables*, los cuales están teniendo una creciente expansión en todo tipo

de organizaciones. Debido a ésto la velocidad, el volumen y la variedad de los datos suele ser muy alta, pero a pesar de ello tienen una muy buena estructuración que nos otorga un acceso más fácil a los datos.

A parte de dichas clasificaciones también encontramos otras que pueden englobar diversos grupos dentro de ellas, como son las de datos estructurados, semi-estructurados o datos no estructurados [8]:

- **Datos estructurados**. Dependen de la existencia de un modelo de datos, en el que cada campo es accesible y está relacionado con los demás. Un ejemplo claro serian los archivos de Excel o las bases de datos, en los que están basadas las aplicaciones de negocio, datos almacenados, registros de máquinas y datos de sensores. Una representación gráfica sería de la siguiente forma:

```
<height>185</height>
<weight>76</weight>
<color>blue</color>
<zip>90458</zip>
```

- **Datos no estructurados**. No tienen etiquetas ni campos que estén definidos por un modelo de datos, normalmente se trata de texto plano con fechas y números. Su uso está extendiéndose cada vez más debido al análisis de imágenes, vídeo o PDF. Estos tipos de archivos suelen prevenir de fuentes como las redes sociales, multimedia o archivos y documentos. Este tipo de datos son los más difíciles de procesar.

```
Lorem ipsum dolor sit amet consectetur adipiscing elit purus
accumsan posuere metus litora fringilla fermentum, condimentum
quis suspendisse aptent dui nostra nibh auctor odio nullam in
augue velit.
```

- **Datos semi-estructurados**. Es un tipo de dato estructurado que no está asociado a ningún modelo de datos pero que, sin embargo, tiene etiquetas para separar elementos semánticos y hacer cumplir las jerarquías de campos y registros. Los más comunes son XML y JSON.

```
<note>
    <to>Tove</to>
    <from>Jani</from>
    <heading>Reminder</heading>
    <body>Don't forget me this weekend!</body>
</note>
```

Además, también seria conveniente hacer una distinción entre los datos que proceden del sistema de almacenamiento de información de nuestra propia empresa como claramente pueden ser los referentes al almacenamiento de datos, archivos y en la mayoría de casos aplicaciones de negocio y ficheros *log*.

2.3 Almacenamiento de la información

Tras haber capturado toda la información que es relevante para nuestra empresa y que servirá para proporcionar diferentes enfoques en distintos campos de la organización, estos datos deben ser almacenados y transformados para su posterior análisis. Aquí básicamente encontramos dos tipos de alternativas:

- **Data warehouse**. Se trata del modelo tradicional de almacenamiento de información, enfocado al almacenamiento de datos estructurados y bases de datos relacionales. Dicho almacén suele utilizarse para la realización de informes, *business intelligence* y visualización de datos.

  El desarrollo de un *data warehouse* (DW) para el almacenamiento de una gran cantidad de datos supone una inversión muy elevada, pero tiene la ventaja de que es un sistema muy maduro y con una alta seguridad debido a su longevidad en el mercado. El esquema de funcionamiento es *schema on-write*, lo que nos indica que sigue el proceso tradicional *extract, transform and load* (ETL) y que los datos ya han sido transformados una vez se desea acceder a los mismos, es decir para que puedan ser cargados en el almacén de datos debemos transformar los datos previamente.[9]

  Esta transformación de los datos estructurados se paraleliza, e incluye procesos de limpieza a través de los cuales se llevan tareas como comprobación de valores válidos, aseguración de la consistencia de los datos, eliminación de datos duplicados o aplicación de reglas o procesos de negocio, así como también se llevan a cabo procesos de conformación en los que se realiza fusiones de datos de distintas fuentes, como por ejemplo cruzar los datos de un banco en cuanto a transacciones en cajero y vía internet mediante el código de identificación del cliente.

  Aquí también encontramos los *data mart*, un tipo de *data warehouse* enfocado a departamentos específicos de la organización, como marketing, ventas, etc.

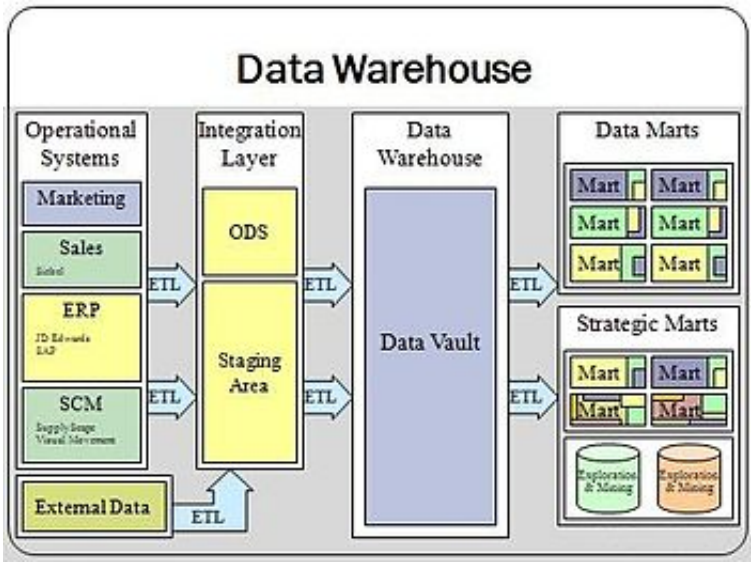


Figura 2.3: Estructura de un DW [10]

- **Data lake**. Éste actúa como la zona de llegada de los datos potencialmente valiosos y sin procesar. El objetivo del *data lake* es lidiar con los datos con baja densidad de valor para incrementar dicha densidad. Una de las grandes ventajas de esta alternativa es que se admiten

todo tipo de datos, ya sean estructurados, semi-estructurados o no estructurados, y que los datos siempre son conservados en su estado inicial. [11]

La forma de trabajo del *data lake* es del tipo *schema on-read*, lo que significa una vez extraídos los datos éstos serán cargados y transformados cuando se requiera de su uso. Entre los métodos mas comunes de análisis podemos encontrar el conocido ETL, el cual incluye operaciones en los datos sin procesar para estructurarlos, sesionalizarlos, eliminar etiquetas XML y extraer palabras clave. Para ello se utilizan plataformas como Hadoop, debido a que son ágiles, flexibles y tienen un alto grado de paralelismo a través de soluciones de bajo coste. Esta tecnología está formada por muchos componentes, entre los cuales encontramos dos que son muy importantes, el sistema de archivos Hadoop Distributed File System (HDFS) el cual está diseñado para trabajar con grandes *data sets* así como para ofrecer a un alto rendimiento en el momento de acceder a los datos de aplicaciones y almacena su información en distintos servidores con una alta fiabilidad y una alta tolerancia a fallos.[12]

La otra parte principal de la que se compone Hadoop es MapReduce, un modelo de programación creado por Google que se centra en el procesamiento de datos a través de la computación paralela, y el cual nos permite agrupar los datos de una forma rápida debido a que su función es dividir los grandes conjuntos de datos entre todos los servidores disponibles [13] . MapReduce es utilizado en multitud de campos como la genómica en la bio-informática [14] o predicción del tiempo, y algunas empresas como Facebook y Google lo utilizan para minería de datos, PageRank, detección de *spam* o búsqueda de rutas en mapas.

Otra de las figuras importantes en el *framework* de Hadoop es la presencia de componentes relacionados con las consultas SQL como puede ser Hive, así como también podemos encontrar varios motores de optimización de tareas y procesos. A parte de estas herramientas que hemos comentado podemos encontrar una amplia multitud de alternativas dentro y fuera de la plataforma de Apache Hadoop, que pueden ayudarnos en todos los procesos que se llevan a cabo en los datos en el *data lake*.

Una de las opciones que están implementando recientemente las empresas para mejorar el tiempo de respuesta de sus sistemas de información es la implementación de los mismos en la memoria de los servidores, obteniendo así una gran optimización del tiempo de respuesta, pues los datos se encuentran en memoria principal y no en memoria secundaria dónde la velocidad de lectura y escritura es significativamente más alta.

Hoy en día existe una creciente tendencia de implementación de los *data lakes* en multitud de empresas, según TDWI [15] aproximadamente un cuarto de las empresas tienen actualmente *data lakes* implementados en su sistema de almacenamiento de datos en su empresa, y cerca del 50 % piensa implementarlo en menos de 3 años. Su uso principal se da en empresas cuyo objetivo es utilizar los datos para *machine learning*, análisis predictivos, ciencia de datos o herramientas BI.

Lo mejor de ambas plataformas es que no son excluyentes de si mismas, de hecho se complementan cubriendo las necesidades de las empresas que no podrían ser cubiertas con la implementación de sólo uno de una de ellas. La mayoría de organizaciones implementan sistemas híbridos (figura 2.4) en los que el *data lake* se utiliza bajo Hadoop para ingerir toda la información de forma rápida, y cuando sea necesaria el uso de la misma se estructura y se manda al *data warehouse*. En éste también se observa la presencia de los datos tradicionales provenientes principalmente de las aplicaciones empresariales para la gestión de la organización y la nuevas fuentes resultantes del mundo del *big data* como las redes sociales o datos creados por dispositivos móviles.

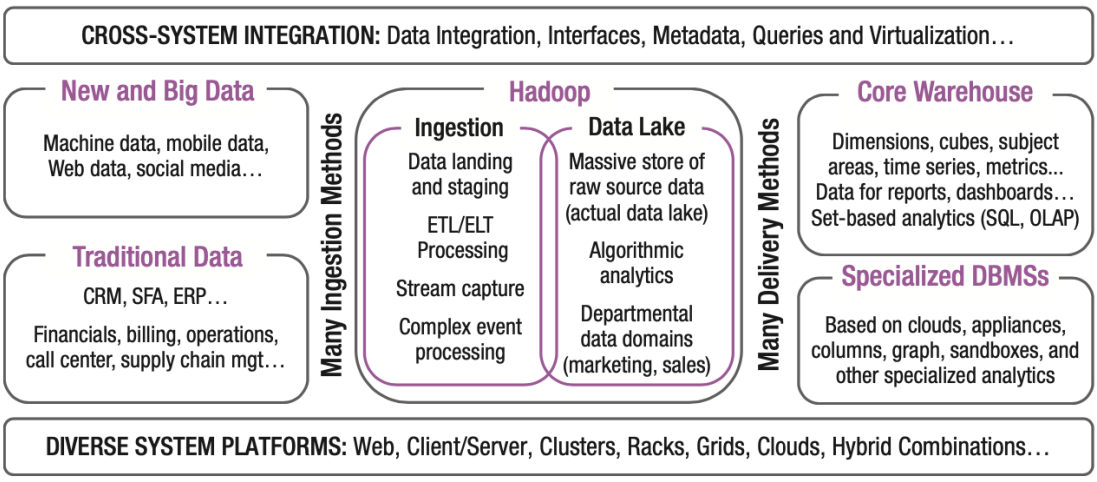


Figura 2.4: Convergencia de múltiples técnicas y sistemas de almacenamiento. [15]

2.4 Análisis

Habiendo sido ya transformados todos los datos que tienen un valor potencialmente significativo a través de diversas plataformas como Hadoop o procedimientos ETL, en la que los datos han sido limpiados y estructurados con el objetivo de poder almacenarlos nos preparamos para realizar el estudio de los datos que más se ajuste a nuestros requisitos.

Hoy en día existen múltiples tipos de análisis en *big data* debido a diversidad de objetivos de todas las empresas que hacen uso de la explotación masiva de datos, los diferentes tipos de datos que almacena cada organización y también debido a la aparición de nuevas tecnologías y métodos que facilitan el análisis de datos, dejando atrás de esta manera, técnicas que han sido muy usadas durante años para ser remplazadas por mejores y más potentes métodos de análisis.

A continuación exponemos, de una forma simplificada, y dándole importancia a las técnicas con más potencial, una agrupación en la que podríamos clasificar la mayoría de las técnicas de análisis masivos de datos.

- **Análisis avanzado**. En esta opción se engloban los análisis predictivos, *data mining*, *machine learning*, análisis estadísticos, algoritmos heurísticos, inteligencia artificial, procesamiento de lenguaje natural, análisis del sentimiento y métodos de bases de datos avanzados. Estos métodos van mas allá de la generación de informes y la creación de archivos preprocesados, planteándose como una de las opciones prometedoras del *data science*.

  Este tipo de análisis es el que más fuerza tiene en los últimos, siendo muy relevante en aspectos científicos y de vanguardia, en el que se necesitan técnicas muy complejas para analizar datos principalmente no estructurados.

- **Visualización avanzada de datos**. Es la técnica de análisis de la cual se espera mas crecimiento en los siguientes años, en parte debido a su potencia de representación de datos de todo tipo. A día de hoy, la mayoría de aplicaciones y herramientas de visualización avanzada de datos son compatibles con la mayoría de fuentes de datos permitiendo así a los analistas de negocios que puedan explorar una gran variedad de conjuntos de datos, generalmente en tiempo real. Además la mayoría de las herramientas han tenido una gran evolución debido a la creciente demanda de las mismas.

  El *operational business intelligence* es una práctica que se encarga de medir y monitorizar en tiempo real el desempeño de las operaciones de negocio. En los últimos años la utilización

de esta técnica ha aumentado considerablemente, y con el paso del tiempo los paneles de información son cada vez más analíticos. En este tipo de análisis la velocidad con la que se crea la información es muy relevante.

- **Técnicas de análisis tradicional**. Las consultas SQL son un análisis muy común en el mercado debido al amplio conocimiento del lenguaje por los analistas de negocio, llegando a poder crear complejos programas de SQL más conocidos como SQL Extremo. Este tipo de análisis suele ser aplicado a fuentes de datos con un alto nivel de detalle, que apenas han tenido que ser transformados.

  Por otro lado tenemos también el conocido formato OLAP, el cual hace referencia al procesamiento de análisis en línea. Este tiene una amplia aceptación en el mercado, y nos permite relacionar datos de diferentes partes de nuestra organización creando estructuras de datos diversas y multidimensionales, poniendo a nuestra disposición la posibilidad de realizar consultas de una forma más rápida.[16]

2.5 Herramientas big data

Las configuraciones de las sistemas de información que hacen uso de técnicas relacionadas con *big data* no tienen una configuración estándar, y la aparición de herramientas desarrolladas de forma pública o privada plantean una gran cantidad de posibilidades para la formación de un sistema de información. A continuación vemos algunas de las herramientas y plataformas más representativas del entorno de los macrodatos [17] :

- **Hadoop** es una de las herramientas más conocidas y que más se asocian al término *big data*, y que ha sido explicada con anterioridad. Se trata de una herramienta *open source* que actúa como *framework* para nuestro sistema de almacenamiento orientado al gran volumen de datos. Éste mismo ofrece un gran abanico de posibles compatibilidades con multitud de plataformas que nos permite construir lo que se conoce como 'ecosistema Hadoop', ya que es éste el que se encarga de orquestar el funcionamiento de las herramientas y su interacción con nuestro sistema de almacenamiento de datos.

  Como ya se ha comentado, dispone de un sistema de archivos distribuido (HDFS) en cada nodo del clúster y se ayuda de MapReduce para repartir la carga proveniente de la computación paralela de tareas de minería de datos.

- **Hive**. Creado por Facebook, es un software bajo la estructura de Hadoop que facilita la gestión y agrupación de grandes grupos de datos. Todo ello a través de la transformación de consultas SQL a procesos MapReduce o tareas Spark, y también mediante la indexación de los datos para la aceleración de ejecución de tareas, almacenamiento de metadatos y ejecución de múltiples funciones definidas por usuarios.

  Actualmente forma parte de la organización sin animo de lucro Apache Software Fundation, conocida también por el desarrollo de Hadoop.

- **Spark** está desarrollado también por Apache, se trata de un motor de procesamiento de datos con una destacada velocidad de análisis de los mismos. Spark hace uso de HDFS para el almacenamiento distribuido, pero se plantea como alternativa a Hadoop Mapreduce siendo hasta 100 veces más rápido que Hadoop en términos de memoria. Permite la programación de aplicaciones a través del uso de diversos lenguajes como Java, R, Python o Scala.

  Spark puede ser implementado con Hadoop mediante YARN, un manejador de recursos y programador de trabajos que funciona como intermediario entre ambos sistemas.

- **MongoDB**. Propio de las bases de datos no relacionales, también conocidas como NoSQL. Se trata de una herramienta orientada a sistemas distribuidos con una alta escalabilidad horizontal. Sus características principales son la indexación, consultas por campos y rangos, replicación y *sharding* para la escalabilidad horizontal, y agrupaciones similares a las del lenguaje SQL.

  Uno de los principales inconvenientes de este sistema es que no asegura el cumplimiento de las propiedades ACID en toda la base de datos, si no que simplemente se limita al mismo documento.

- **Elasticsearch** es una potente motor de búsqueda que proporciona una búsqueda distribuida y a lo largo de todos los campos de los datos almacenados. Dicha herramienta permite la realización de consultas a gran velocidad debido a la indexación de los datos, las búsquedas de texto complejas y visualización de la evolución en tiempo real de datos.

- **Python**. Un lenguaje de programación con reciente popularidad en el sector pero que uno de sus puntos a favor es que es relativamente fácil para multitud de usuarios que no están familiarizados con la informática. Sin embargo, la ejecución de las instrucciones hacen que ésta se muestre como una herramienta lenta para tareas complejas frente a sus competidores.

- **Lenguaje R** se trata de un entorno y lenguaje de programación para cálculo estadístico y representación de datos mediante gráficos. Éste es usado por multitud de profesionales para la minería de datos, investigación en bioinformática y también relevante para matemáticas.

  Al igual que Python, ambos se tratan de lenguajes muy valorados por la comunidad con multitud de librerías a disposición de los mismos.

- **Tableau** es una potente herramienta de visualización de datos, principalmente enfocada al *business intelligence*. Dentro de ésta podemos encontrar multitud de formas para la representación de datos estructurados provenientes de nuestro sistema de almacenamiento, como bases de datos relacionales, cubos OLAP y hojas de cálculo.

  Se trata de una herramienta propietaria. El uso de la misma está planteado para que no sean necesarios altos niveles informáticos, a pesar de que se pueden añadir algunas dimensiones y medidas mediante pseudocódigo a las que ya vienen predefinidas por los datos.

2.5.1. Business analytics y toma de decisiones

Es desde 1970 cuando se empiezan a utilizar sistemas para la toma de decisiones y su uso se ha prolongado hasta la actualidad. De esta forma las empresas han transformado y perfilado las técnicas de análisis de datos debido a la necesidad de extraer conocimiento de datos no estructurados, mucho más complejos y con un mayor volumen que en épocas anteriores. Todos estos términos tienen un objetivo común, el de obtener nuevos puntos de vista sobre nuestra organización con el propósito de ayudarnos en la toma de decisiones e implementar cambios en la organización. [16]

A pesar de que dichos términos han surgido por que las necesidades de entorno habían cambiado, y se necesitaban nuevos métodos de análisis, éstos a día de hoy hacen un trabajo excelente a la hora de combinarlos, cubriendo los requisitos de las organizaciones.

El término de *business intelligence* es acuñado a mediados de 1990 por Gartner Group, y desde entonces se ha sometido a un cambio constante debido a la inclusión de nuevas técnicas y capacidades, como por ejemplo la de la inteligencia artificial o potentes métodos de análisis. La definición que podemos encontrar a día de hoy es la siguiente:

> *La combinación de tecnología, herramientas y procesos que me permiten transformar datos almacenados en información, esta información en conocimiento y este conocimiento dirigido a un plan o una estrategia comercial. La inteligencia de negocios debe ser parte de la estrategia empresarial, esta le permite optimizar la utilización de recursos, monitorear el cumplimiento de los objetivos de la empresa y la capacidad de tomar buenas decisiones para así obtener mejores resultados.*
>
> — Data Warehouse Institute

Como se comenta en la definición, el objetivo del BI es proporcionar acceso, en algunas ocasiones en tiempo real, a los datos y ofrecer herramientas para la manipulación de los mismos dando así a los analistas de negocios la habilidad de realizar análisis apropiados a través del procesamiento de datos actuales, históricos, situaciones e incluso *Key Performance Indicators* (KPI). De esta forma, se transforma la información en decisiones, y las decisiones directamente en acciones.

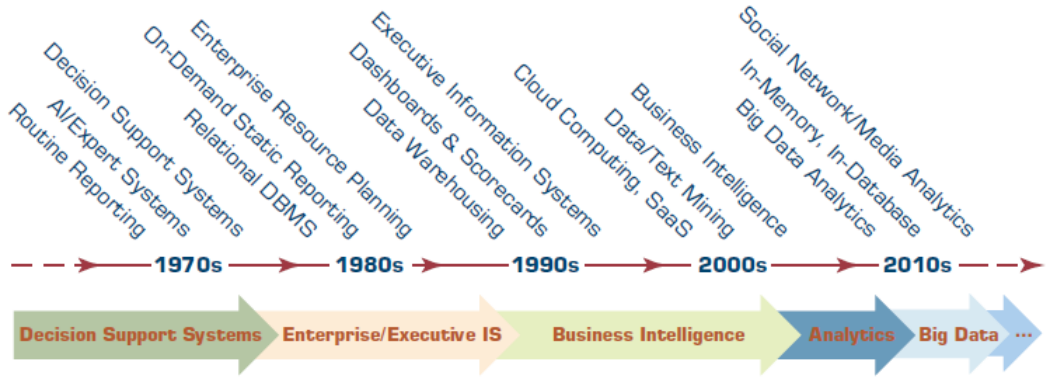


Figura 2.5: Evolución de los soportes de toma de decisiones.[18]

Pero a día de hoy, y como hemos podido ver en la figura 2.5, los términos son cambiantes. Actualmente un gran numero de expertos, consultores y autores de referencia están empezando a sustituir el termino BI por *business analytics* o simplemente *analytics*. Dentro de este nuevo término se definen 3 tipos distintos de análisis [19] :

- **Análisis descriptivo**. Éste hace referencia a lo qué está pasando en la organización, destacando las tendencias de los datos y las causas de las mismas. Ello implica el afianzamiento de las fuentes de datos y la disponibilidad de los mismos en un sistema de información que permita la elaboración de informes y análisis pertinentes. Dicho sistema de información hace referencia al DW.

  A partir de dicha infraestructura podemos elaborar diferentes tipos de informes, consultas, alertas y *dashboards*, una tecnología muy relevante a día de hoy que nos permite la visualización de distintos datos, como los KPI, con el objetivo observar el rendimiento de la organización.

- **Análisis predictivo**. Se centra en determinar qué es probable que suceda en el futuro y también el porqué. Los métodos que se usan en este tipo de análisis son técnicas de minería de datos, minería de textos y también minería de datos web y multimedia, combinadas con técnicas de clasificación como la regresión logística, árboles de decisiones, redes neuronales o clusterización.

  Un ejemplo de este análisis es el estudio de la posibilidad de que un cliente se cambie a una empresa competidora, estudiando los motivos de dicha acción y presentando todos los casos asociados, descubriendo así como se puede evitar que suceda.

- **Análisis de decisión**. El objetivo es el de, ante una situación determinada, reconocer lo que está sucediendo, explorar pronósticos probables y tomar las mejores decisiones posibles. Las técnicas empleadas en dicho análisis son métodos enfocados a la optimización, simulación de casos *what-if*, modelos de decisiones o sistemas expertos.

  Generalmente dicho análisis se centra en la optimización del rendimiento de sistemas u organizaciones, proporcionando recomendaciones especificas que deben llevarse a cabo ante tal situación.

2.6 Aplicaciones del big data

El análisis de grandes volúmenes de datos se ha convertido en tendencia en los últimos años y cada vez más empresas lo incorporan en su estrategia organizacional para mejorar sus servicios, productos o atención al usuario. Esta técnica polivalente debido a su gran espectro de técnicas que lo componen, el *Big Data*, se ha extendido a lo largo de multitud de sectores en la sociedad, desde su uso más comercial y enfocado en el marketing y las ventas, hasta su uso para la ciencia, la mejora de los servicios sanitarios o incluso a la seguridad.

De hecho gracias al análisis masivo de datos se están creando novedosos sistemas y conceptos que antes no existían hasta la fecha, como el de *Smart city* o *Smart health*. La predisposición es la de convertir cualquier sector de la sociedad en inteligente, y convertirlo en un recurso optimizado. Esto es debido a que los beneficios de su implementación son muy importantes:

- **Gestión del cambio**. Desde la creación de nuevos productos debido a la detección de nuevas necesidades de los usuarios, búsqueda de nuevas oportunidades de negocio, segmentación de usuarios y soporte de tomas de decisiones automáticas, hasta mejoras implementadas en la estrategia organizacional.

- **Anticipación de problemas**. Análisis predictivos, simulación de escenarios y datos cruzados nos permiten anticiparnos a situaciones problemáticas que podrían producirse en el futuro. De esta forma estamos preparados para que en esta situación la repercusión de los problemas sea la menor posible.

- **Mejoras de procesos**. Incluidos algunos como la reducción de costes, optimización de tiempos, reducción de riesgos, detección de fraudes, simplificación de procesos empresariales.

A continuación veremos algunos de los ejemplos en la sociedad en los que el *big data* se ha aplicado con mayor éxito [20]:

- **E-commerce y social media**. Éste es el sector principal en el que el *big data* empieza a utilizarse con el objetivo de comprender al usuario web. Desde sus comienzos se han producido importantes avances en el comercio electrónico desde la aparición de la Web 2.0 en distribuidores como Amazon, Google o Facebook. Éstos son una referencia en la implementación de plataformas digitales y con ello sistemas de recomendación de productos, análisis web, *cloud computing*, plataformas sociales.

  La aplicación del *big data* en este campo se lleva a cabo a través de técnicas como análisis del sentimiento, análisis de redes, minería de datos y de grafos, segmentación y *clustering* de usuarios. Estos cuatro últimos son técnicas clave para los sistemas de recomendación de productos. La mayoría de estos análisis pueden llevarse a cabo a través de Gephi, una herramienta *open source* de análisis y visualización de redes. Dichos análisis ejecutados a gran escala proporcionan a las empresas correspondientes una gran herramienta para comprender a sus clientes a través de una gran cantidad de datos no estructurados como opiniones y

reviews, y algunos estructurados como *clickstream* y contenidos generados durante la navegación, dando lugar a aplicaciones de recomendaciones de productos y monitorización de redes sociales, entre otras muchas.

Los impactos directos de esta aplicación son visibles en la clasificación de los productos pertenecientes a la regla de larga cola en marketing, recomendaciones personalizadas y dirigidas a clientes específicos y también en el incremento de las ventas y satisfacción de los clientes.

- **Política y Gobierno**. Éste es un campo medianamente reciente debido a la digitalización de los partidos políticos a causa de la incorporación de las redes sociales en los mismos. Actualmente los partidos políticos se vuelven más transparentes y conectados ello significa un nuevo y amplio acceso a datos a los que anteriormente no se tenía acceso. Específicamente, ésta aplicación del *big data* en política es muy popular en los Estados Unidos de América.

  Se pueden utilizar técnicas como análisis del sentimiento, análisis de redes, monitorización de redes sociales, análisis semánticos de datos gubernamentales, etc. Mediante estas técnicas, el análisis de datos como regulaciones y leyes, información gubernamental y de servicios, *feedback*, comentarios, e interacciones sociales de los ciudadanos nos pueden proporcionar aplicaciones de compromiso y participación ciudadana, campañas políticas y votaciones electrónicas, sistemas de acceso igualitario y servicios públicos. Los tipos de datos comentados suelen ser no estructurados, en texto plano, fragmentos legales o conversacionales.

  En esta aplicación se puede ver reflejado algunos rasgos de las *Smart cities*, en las que se utiliza la información para mejorar los servicios e infraestructuras en la ciudad, dando voz a los ciudadanos, mejorando la transparencia de los gobiernos, e impulsando la participación ciudadana y la igualdad.

- **Ciencia y Tecnología**. Diversas áreas se ven implicadas en la aplicación de *big data* para procesamiento de datos procedentes de sensores o instrumentaciones especificas del campo, como la astrofísica, oceanografía, o algunos más investigacionales como la genómica o ciencias ambientales. El objetivo de dichos medios científicos y tecnológicos es la administración, análisis, visualización y extracción de información útil de grandes conjuntos, diversos y distribuidos con el fin de acelerar el progreso de los descubrimientos científicos y la innovación.

  El impacto directo de dichos usos se basa en la comprensión de procesos e interacciones humanas y sociales, promoción de la salud y calidad de vida, descubrimientos científicos y tecnológicos, a través del uso de modelos analíticos y matemáticos propios de cada dominio.

- **Smart health**. Es más común cada vez el uso de técnicas de *big data* para el procesamiento de información proveniente principalmente de dos tipos de fuentes, la primera es gnómica en la que se incluyen datos provenientes de genotipado, expresión y secuenciación genética, y la otra fuente es la referente al entorno médico, como los registros de salud electrónicos, recetas electrónicas, reseñas de pacientes, etc. Dicha información es de tipo sensible, y la aplicación del *big data* para analizarla supone la implementación de un sistema seguro y ético, a partir del cual no se puedan producir fugas de ningún tipo de información. Ésta es una de las causas del atraso de la implantación del análisis masivos de datos en la salud respecto a la del *e-commerce*.

  Las técnicas usadas para la extracción de conocimiento son análisis secuencial y genómico, minería de registros de salud electrónicos, monitorización y análisis de la salud en las redes, análisis de efectos adversos a medicamentos, minería para conservar la privacidad,

etc. Éstas son usadas para la creación de plataformas de ayuda de toma de decisiones en salud, análisis de comunidades de pacientes así como otras enfocadas al estudio del genoma humano y de plantas.

Las repercusiones más visibles son la obtención de un sistema de salud mejorado y optimizado, optimización de recursos e incremento de importancia respecto al paciente.

- **Seguridad**. Debido a los múltiples sucesos ocurridos en la ultima década en algunos de los países mas importantes del mundo, el *big data* se ha implementado exponencialmente en sistemas de seguridad. Además, dicho uso del análisis masivos de datos en seguridad no tiene como único objetivo la prevención de atentados terroristas, si no también el análisis de criminológico y la ciberseguridad.

  El análisis mediante técnicas de asociación de reglas, minería y *clustering*, análisis de redes criminales, procesamiento de lenguaje natural o análisis de ciberataques provenientes de datos extraídos directamente de bases de datos de criminales, mapas de criminalidad de ciudades, noticias y contenidos web, bases de datos de terroristas y datos de ciberataques, virus y la web nos permite incrementar y mejorar la seguridad pública de las ciudades.

  Un ejemplo destacable es el del sistema de vigilancia que tiene implementado China, a través del cual se puede detectar en cuestión de segundos la ubicación de un sospechoso simplemente introduciendo su foto en el sistema.

CAPÍTULO 3

Ámbito sanitario

El sector sanitario es uno de los ámbitos más importantes a nivel nacional, que crea multitud de empleos en el sector privado y público. Es reconocido a nivel internacional y se sitúa como el epicentro del estado de bienestar, el cual tiene una gran repercusión sobre la sociedad. El Sistema Nacional de Salud se crea a principios del siglo XX, y es a finales de 1989 en el que se completa la cobertura sanitaria universal y gratuita en todo el territorio nacional.

Es tarea de todos los implicados en comunidad sanitaria la de contribuir a la mejora constante del sistema sanitario para ofertar una mejor asistencia al ciudadano. Para ello la superación de algunos retos como el del incremento de la medicina preventiva y la inclusión de la tecnología en el sistema son requeridos para una mejora del sector que se plantea como necesaria.

A lo largo del siguiente capítulo veremos cómo una de las aplicaciones más prometedoras del mundo del análisis de los datos es la salud. En el apartado posterior profundizaremos en este tema, observando la posibilidad de los beneficios que se puede proporcionar a los pacientes mediante el uso del *big data* debido al gran crecimiento tecnológico y la generación de datos del ámbito de la sanidad.

3.1 La salud

La salud ha cambiado con el paso del tiempo como consecuencia de la ampliación del conocimiento del ser humano respecto a la medicina. Este concepto se definía anteriormente como un estado de ausencia de enfermedades biológicas. Actualmente el concepto de salud más certero y que más se ajusta a la sociedad actual es el definido por la OMS:

> *La salud es un estado de completo bienestar físico, mental y social, y no solamente la ausencia de afecciones o enfermedades*
>
> — Organización Mundial de la Salud, 1946

Los factores de los que depende la misma se basan en los estados biológicos y socioculturales, estados fisiológicos de equilibrio, la alimentación o relaciones familiares y hábitos cotidianos. La OMS califica el buen estado de la salud como precursor de una mejor calidad de vida, concluyendo que un sistema eficaz de salud es necesario para la prosperidad económica y la obtención de una sociedad más satisfecha en general. [21]

En cambio el término de salud digital es relativamente moderno, creado como consecuencia de los avances tecnológicos que dan lugar a que dicho concepto sea posible. Ésta se plantea básicamente como la aplicación de las tecnologías y los sistemas de información al sector de la salud con la intención de mejorar el acceso, la eficacia, la optimización y la calidad de los procesos clínicos y empresariales relativos a la sanidad para alcanzar el objetivo de mejorar el estado de salud de los pacientes y la satisfacción de los mismos.

Pero esta definición no solo supone una mejora técnica en cuanto al sistema sanitario, si no que plantea una nueva manera de pensar, un compromiso con las nuevas tecnologías y las redes sociales, procesos educativos, intercambios de información y rompe los límites de atención de la salud tradicional, transformándose en una salud más ética y equitativa. La salud digital supone el empoderamiento de los pacientes debido a la cantidad de información y decisiones que están a su disposición, creando una gran concienciación sobre el bienestar, el ejercicio físico y la salud en la ciudadanía.

La aplicación del *big data* al sector sanitario, tanto en el ámbito puramente más médico como el referido a la investigación, se ha convertido en un movimiento inevitable e imparable. Los beneficios de esta técnica de análisis masivos de datos son claros y aportarán una mejora del sistema, pero las limitaciones y riesgos deben ser estudiadas concienzudamente ya que a diferencia de otros sectores estamos tratando con datos de alta sensibilidad que son claves para la toma de decisiones muy importantes y que tienen consecuencias graves respecto a la vida de las personas.

3.1.1. Actualidad de la salud digital

Historia Clínica Electrónica

El referente mas claro de la digitalización de la sanidad en el ámbito nacional es el de los historiales médicos electrónicos o historia clínica electrónica del paciente, más conocida como EHR o HCE en español. Este concepto se define como el conjunto de información personal relativa de un paciente sobre su historia médica, síntomas y enfermedades, resultados de pruebas médicas, alergias, datos médicos básicos, consultas médicas pasadas, historial de medicaciones y recetas. [22]

Este tipo de registro sirve de gran ayuda a los profesionales médicos debido a que su formato digital permite la consulta de la información anteriormente comentada, bajo el acceso controla-

do y limitado, otorgando un mejor servicio asistencial al paciente y proporcionando información clave para la atención primaria, especialidades, ingresos hospitalarios y en situaciones dónde la agilidad y rapidez médica es vital como en situaciones de urgencias dónde la consulta de dicha información en papel podría suponer graves consecuencias en el estado del paciente por problemas de legibilidad, de ordenación o de alteración de datos.

La historia clínica electrónica es el registro unificado y personal, multimedia, en el que se archiva en soporte electrónico toda la información referente al paciente y a su atención. Es accesible, con las limitaciones apropiadas, en todos los casos en los que se precisa asistencia clínica [22] (urgencias, atención primaria, especialidades, ingresos hospitalarios y demás).

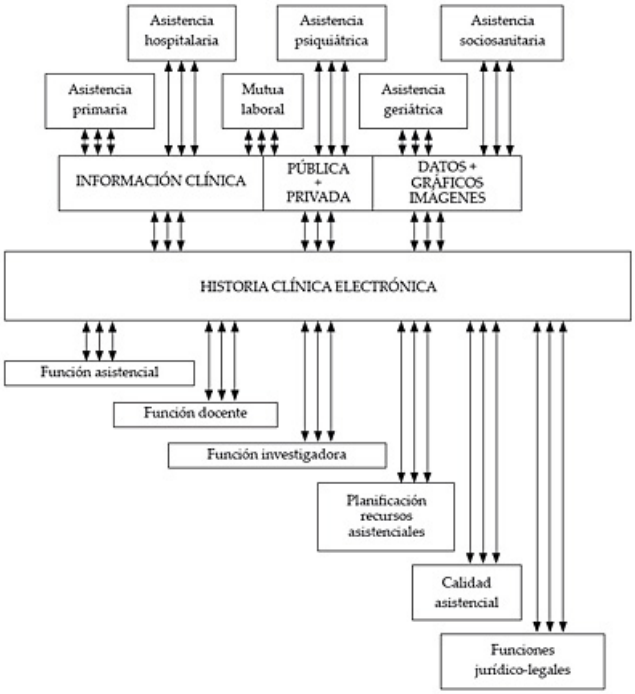


Figura 3.1: Historia Clínica Electrónica [22]

En total, hay cerca de 180 millones de documentos sanitarios electrónicos accesibles provenientes de pacientes de todo el territorio español [23]. La implantación de la historia clínica electrónica hace necesario que los sistemas de información que se usan en las instituciones sanitarias implementen estándares informativos internacionalmente reconocidos, con el objetivo de poder garantizar la interoperabilidad, la integridad y disponibilidad de la información. El estándar informático más conocido y más desarrollado se trata de HL7 cuyas características principales son la visualización de documentos médicos, modelo funcional de historia clínica electrónica, anexión de informes para tramitación así como funciones centradas en la mensajería.[24]

Otras de las mejoras que tuvieron un despliegue paralelo a la de la digitalización de los historiales clínicos fue la incorporación de la receta electrónica y la petición de citas médicas online.

mSalud y salud 2.0

El concepto de mSalud se basa en el uso de dispositivos móviles como apoyo para realizar el ejercicio de la medicina. Tienen una alta presencia el uso de *wearables* y sistemas de monitorización para personas de avanzada edad o con enfermedades especificas y permite que el uso de los mismos suponga la extracción de información relevante para el diagnóstico médico. Estos dispositivos hoy en día tienen multitud de funciones, como detección de constantes vitales, realización de electrocardiogramas, registros de tomas de insulina, sensores de monitorización para

parkinson o complejos prototipos de robots que se encargan de la monitorización y cuidado de personas mayores.

Gracias también a la utilización de los smartphones como centro médico de muchos seguros de vida y seguros de salud privados, algunas compañías de este tipo han implementado la detección de pasos en sus aplicaciones para ofrecer descuentos a sus usuarios en función de los pasos que caminan cada día, lo que repercute notoriamente en incremento de la salud de las personas. A través de este tipo de aplicaciones como la de Adeslas o DKV, también se oferta la asistencia médica remota [25], una característica en auge desde la última década que basa en la atención del paciente de forma telemática y a través de cualquier dispositivo conectado a internet como los smartphones o tablets. [26]

Y es que los dispositivos móviles han tenido una gran influencia también en el desarrollo de aplicaciones y plataformas con el objetivo de ayudar los usuarios, donde la Salud 2.0 se plantea como una forma de asistencia médica centrada en el ciudadano en la que se emplea de manera activa las redes sociales, aplicaciones y herramientas web con el objetivo de mejorar la calidad de vida de las comunidades. Gracias a ella se ha producido un incremento de la concienciación social e impulsando a multitud de personas a adoptar una forma de vida más saludable. Podemos encontrar aplicaciones simples para el control de las calorías que consumimos al día, creación de hábitos más saludables, motivación y seguimiento del ejercicio físico, recordatorios para diabéticos o para la hidratación, toma de medicaciones, realización controles a nuestra piel o control de alérgenos, hasta aplicaciones más especificas para ayudar a ejercitar la mente de personas con Alzheimer a través de recuerdos personales.

La presencia de portales web también es notoria, predominando páginas como PatientLikeMe con multitud de comunidades especificas para varios tipos de enfermedades que se centran en proporcionar consejos y poner en contacto a personas con la misma enfermedad para hacer esta más llevadera en la medida de lo posible, y otras más especificas y centradas en una enfermedad en concreto como tudiabetes.org o stupidcancer. [27]

La presencia redes sociales para personal médico también se encuentra en auge, sobre todo en territorio estadounidense, dónde aproximadamente la mitad de profesionales utilizan algún tipo de aplicación, como Doximity o Sermo, en la que se comparten casos de diagnósticos anonimizados para dar a conocer nuevas perspectivas y diagnósticos, así como para también facilitar la obtención de opiniones de otros expertos sobre un mismo diagnóstico.

3.2 Big data en salud

Como hemos podido ver la salud digital es muy prometedora, y el uso del *big data* explicado en la sección anterior, combinado con el sector de la salud abre un abanico de posibilidades muy interesantes que benefician tanto al paciente y personal médico e investigadores, como al sector en sí debido a la mejora de procesos y optimización de recursos.

3.2.1. Las 4V

Al igual que en el apartado del capítulo 2 (2.1.1) encontramos las 3V características del *big data* y también otras tres que son complementarias. Esta clasificación puede observarse en la figura **??**, en la cual observamos todos los datos que se generan a lo largo de nuestra vida a causa del uso de los servicios sanitarios. A continuación se explican las V más características de los datos masivos en el ámbito sanitario [28]:

- **Volumen**. A lo largo de la historia la creación de datos en salud ha sido enorme, pero hasta ahora la tendencia era almacenar dichos datos en papel. Actualmente la mayoría de todos estos datos tienen su versión digitalizada: Imágenes 3D, radiológicas, resonancias magnéticas y ecografías, lecturas de sensores biométricos, historiales médicos, datos sobre medicamentos farmacológicos o secuenciaciones genéticas son, tan solo, algunos de la gran lista de datos que podemos almacenar para posteriormente realizar un procesamiento masivo de dichos datos.

- **Velocidad**. La sanidad es un servicio usado por miles de millones de ciudadanos diariamente a nivel global, lo que significa un flujo constante de creación de datos. Esto supone una velocidad abismal en la creación de datos, y por ello también es necesaria una gran agilidad en el tiempo de análisis de los datos, donde los sistemas de procesamiento de datos en tiempo real toman mayor relevancia y pueden significar la diferencia entre la vida y la muerte. Un ejemplo en el que la velocidad de los análisis es primordial es en el control de epidemias o detección de infecciones en neonatos.

- **Variedad**. La medicina se trata de un campo de la ciencia muy amplio, en el que podemos encontrar muchos tipos de especialidades y con ello, sus respectivas pruebas y métodos de análisis de enfermedades: Análisis de sangre, electrocardiogramas, resonancias magnéticas, wearables, etc. La mayoría de éstos se tratan de datos semi-estructurados pero a ello hay que añadirle el genoma humano y todos los datos del paciente que están almacenados en su historial clínico electrónico, en el que no sólo encontramos datos estructurados como pueden ser nombre del paciente, datos de nacimiento, dirección, tratamientos, si no que a ello debemos añadir los campos en los que encontramos anotaciones médicas realizadas por profesionales médicos y que se encuentran en lenguaje natural. Al igual que los datos que podemos recoger de las redes sociales centradas en la salud 2.0, éstos se tratan de datos no estructurados.

- **Veracidad**. Es una de las características más importantes de los datos en la medicina, ya que la calidad de los datos es vital en la toma de decisiones. Un estado incorrecto de los datos puede repercutir negativamente sobre los pacientes y esto se produce con más probabilidad sobre datos no estructurados.

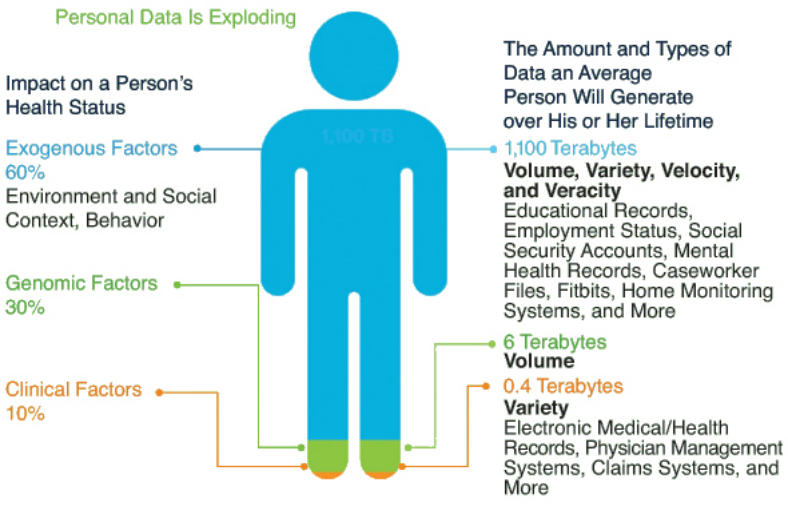


Figura 3.2: Cantidad de datos generados por las personas [29]

3.2.2. Fuentes de información

Las fuentes de información como hemos podido observar brevemente son muy amplias [30]. Las podemos clasificar de la siguiente manera:

- **Salud 2.0**. Datos provenientes de las redes sociales, páginas web y aplicaciones dedicadas a la mejora de la salud de los pacientes.
- **Registros de máquinas**. Como ya vimos en el apartado de *big data* éste se trata de uno de las fuentes de información cuyo volumen de datos es mayor. Se incluyen datos como las lecturas de sensores de monitorización.
- **Registros diversos**. Quedan recogidos, por ejemplo, las reclamaciones de pacientes, registros de facturación, etc. Este tipo de datos suelen ser semi-estructurados o no estructurados.
- **Datos biomédicos resultantes de pruebas**. Multitud de datos son recogidos a lo largo de nuestras estancias en los hospitales y pruebas médicas como datos provenientes de análisis genéticos, resultados de imágenes médicas, constantes vitales y presión arterial, entre otros.
- **Datos del paciente**. Datos semi-estructurados como el historial clínico electrónico (HCE), y otros no estructurados como notas en papel de los médicos o correos electrónicos.

Todas estas fuentes de información sobrepasan la capacidad de entendimiento que el ser humano puede poseer, por ello es necesario la intervención del análisis masivo de datos y las tecnologías en el ámbito sanitario, ayudando así a extraer conocimiento de tantas y variadas fuentes de información. [29]

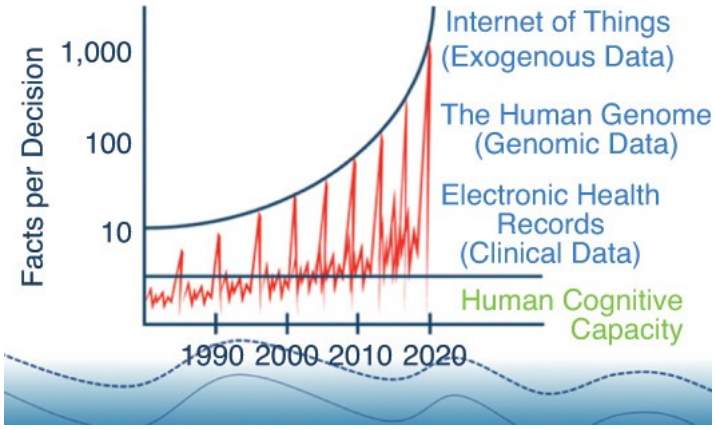


Figura 3.3: Gráfico de la creciente tendencia de fuentes y volumen de información [29]

3.2.3. Privacidad y conciencia de datos

Los aspectos éticos y legales de cualquier proyecto de investigación en el que se utilicen datos biológicos personales deben ser evaluados por comités de ética de la investigación, los cuales se encargan de cercionarse que se cumplen mecanismos de protección de los derechos de las personas [32]. Es necesario concretar quién tendrá acceso a los datos recogidos, sabiendo con total certeza las actividades de tratamiento que se vayan a realizar sobre los datos, y si dichas operaciones se externalizarán a terceros, quedando los datos expuestos. Para ello existen contratos de confidencialidad entre empresas, al igual que los que firman los investigadores que desarrollan proyectos bajo las herramientas de datos masivos.

En este proceso hay que tener en cuenta los intereses de empresas y particulares, que disfrazados de investigación pueden estar promovidos por la monetización de los datos personales de salud, los cuales son de carácter muy sensible y requieren una especial protección. Así mismo tratar de evitar el mal uso de los datos, el uso desproporcionado y parar la acumulación de datos sin un objetivo claro, ya que ello puede suponer un problema de seguridad. [33]

Como hemos comentado los datos de salud se generan cada vez con más facilidad y a lo largo de más dispositivos, pero los datos genómicos tienen un valor extraordinariamente sensible, y la mala practica de la seguridad de la información puede tener consecuencias irreversibles para el paciente del cual se ha filtrado dicha información.

Para asegurar todos estos aspectos éticos y legales deberán de cumplirse los siguientes puntos:

- **Consentimiento del paciente**. El paciente debe de consentir de forma expresa o bien siendo informado del uso que se hará de sus datos. Así también debe de tener la opción de ejercer el derecho a la rectificación y la cancelación de sus datos con fines investigacionales o para el análisis de los mismos. El paciente debe de tener en todo momento la opción de elegir la compartición de sus datos, sin formar éste parte de un acuerdo obligatorio para la obtención de ningún tratamiento. En todo momento, los datos deben de ser propiedad de los pacientes y no de la entidad sanitaria.

- **Confidencialidad**. La anonimización de los datos es un proceso obligado para el análisis de los mismos, con el objetivo de ocultar la identidad de los pacientes, relativos a unos datos de carácter muy sensible que pueden ser usados con intenciones negativas. La confidencialidad debe de mantenerse si los datos son enviados a otras empresas u organismos colaboradores del análisis.

- **No Discriminación**. Aun así las técnicas de anonimización pueden no ser suficientes, ya que a través del análisis de los datos se pueden acabar consolidando estereotipos que desemboquen en problemas sociales como la exclusión, o penalizaciones por parte de los seguros privados por la tendencia a contraer de forma más probable una enfermedad.

- **Transparencia**. Se debe de proporcionar una total transparencia del objetivo del análisis de los datos, así como aclarar los métodos de gestión de los mismos y si fuera posible los algoritmos que se utilizan para el procesamiento de los datos.

Recientemente la Unión Europea estableció un nuevo reglamento de de protección de datos llamado General Data Protection Regulation o en español Reglamento General de Protección de Datos. En este reglamento se recogen leyes que no permiten el almacenamiento de datos personales sin un consentimiento explícito en el que se explique con claridad: con qué fines y qué tipos de datos van a ser almacenados para su análisis, las empresas solo podrán recopilar y editar datos con unos objetivos específicos, se deberá almacenar la menor cantidad de datos posibles del usuario, las empresas deberán de ser responsables de establecer medidas de seguridad informáticas para la protección de los datos. Éste es un pequeño listado de las obligaciones a las que están sometidas las empresas, y es necesario tener en cuenta que el incumplimiento de cualquier punto será objeto de sanción.

3.3 Aplicaciones del big data en salud

- **Operativa clínica**. El análisis de datos para la realización de decisiones estratégicas, de planificación y gestión de recursos es un campo importante que puede proporcionarnos muchos beneficios. Gracias al *big data* tenemos a nuestro alcance la transformación de la operativa clínica a una versión más efectiva y eficaz de la misma, proporcionando un mejor

servicio a los pacientes y siendo más eficiente en cuanto a gastos. La reducción de gastos es un aspecto muy presente en la sanidad, y a partir de la operativa clínica se pueden estudiar los flujos de ingresos, número de enfermeros disponibles, tiempos de espera o triaje y multitud de procesos que se llevan a cabo en la salud.

- **Investigación clínica**. El *big data* aplicado a la investigación clínica puede proporcionar a los profesionales sanitarios una mejor perspectiva con la que realizar diagnósticos más ajustados. Los laboratorios clínicos y farmacéuticos también se verían afectados de forma positiva en la calidad de la documentación científica y en la disminución del infradiagnóstico de patologías, repercutiendo directamente de forma positiva en los pacientes que podrían beneficiarse de diagnósticos más rápidos y precisos.

- **Análisis genómico**. El uso de técnicas de macrodatos a gran escala como la secuenciación nos permitiría el abaratamiento de análisis genómico, lo que haría posible que este tipo de análisis formase parte del núcleo principal de las decisiones en medicina. Esto sería una revolución de la medicina tal y como la conocemos hoy en día, desarrollándose una medicina más preventiva y personalizada.

  Por ejemplo, Somatic Mutations Finder se trata de una herramienta basada en *big data* que detecta de forma rápida y precisa los cambios genómicos causantes de la aparición de tumores [34]. La herramienta se encarga de analizar también la progresión del tumor a lo largo del tiempo.

- **Perfilación de pacientes**. Mediante métodos de análisis avanzados, como pueden ser la segmentación de usuarios, la modelización predictiva o la utilización de la inteligencia artificial, identificar a pacientes cuyo perfil encaja en nuevos fármacos, nuevos hábitos de vida o cambios que deben de realizar para poder llevar una vida más sana o prevenir el desarrollo de algunas enfermedades. Por ejemplo pacientes con un alto riesgo de desarrollar enfermedades con la diabetes podrían ser detectados para aplicar técnicas de medicina preventiva con el objetivo de minimizar las posibilidades de que esta enfermedad acabe desarrollándose.

- **Monitorización y seguimiento**. Captura y análisis de grandes volúmenes de datos producidos en tiempo real provenientes de dispositivos de hospitales y también dispositivos remotos como los *wearables* que acompañados de *machine learning* pueden predecir eventos o problemas que puedan suceder. La monitorización de dichos dispositivos supone una herramienta clave en la seguridad y la salud de enfermos crónicos.

  Help4Mood se trata de una herramienta para el apoyo en la depresión de los pacientes a través del seguimiento de la misma mediante una serie de sensores [35]. Estos sensores se encargan, a través del análisis de los datos recogidos, de identificar si el paciente está recayendo en la enfermedad para proporcionarle una atención inmediata.

- **Farmacología**. Los ensayos clínicos son el núcleo de la farmacología, sobre ellos la comunidad farmacológica y científica realiza estudios sobre los efectos de los fármacos en los pacientes. Este campo puede verse muy beneficiado de la recogida masiva de datos del entorno de los pacientes para el posterior análisis y comprensión de la efectividad de los medicamentos. La aplicación del *big data* supone también un abaratamiento de los fármacos y el desarrollo de otros nuevos a sectores de la población que están desatendidos.

- **Epidemiología**. Es un área de gran importancia en la salud, que estudia la propagación de enfermedades en la población. Mediante el análisis del geoposicionamiento de personas podemos saber cómo de rápido se está extendiendo un virus y en qué dirección, haciendo una mejor gestión de las técnicas de mitigación como la restricción de movimiento de

poblaciones. Queda contemplado también el abaratamiento de los estudios poblacionales, permitiendo un mayor desarrollo investigacional y mayor capacitación de los epidemiólogos.

3.3.1. Beneficios

- **Incremento de la calidad de la atención sanitaria**. La aplicación del *big data* es muy prometedora en cuanto a la mejora de la calidad de la atención médica, en lo que a investigación, diagnosis y tratamiento se refiere.

  En la medicina se observa una transición hacia un modelo basado en la evidencia, donde el análisis masivo de datos puede aportar un valor muy consistente a dicha transición a través de evidencias basadas en datos del mundo real. Los profesionales serían capaces de obtener otra fuente de conocimiento para construir conclusiones adecuadas que les lleven a una toma de decisiones más certera y con una visión global de la situación. [27]

- **Impacto económico y detección de fraude**. La optimización de recursos mediante el uso del *big data* puede proporcionarnos datos estadísticos y realizar una mejor planificación y aprovisionamiento. El uso de *dashboards* en tiempo real se sitúa como una de las opciones más relevantes para ello. [28]

  También se plantea la posibilidad de reducir los gastos a través de detecciones de abusos de la prestación sanitaria. El beneficio principal del análisis masivo de datos en este caso es la vigilancia constante y contrastación de múltiples factores en patrones de conducta para prevenir fraudes y posibles amenazas para la seguridad. [31]

- **Nuevos enfoques de la medicina**. Desarrollo de la medicina de las 4p [28] :

  - Medicina Predictiva: Gracias a la monitorización y análisis de datos se puede detectar la existencia de patologías antes de que aparezcan, así como la predisposición a las mismas.

  - Medicina Personalizada: La heterogeneidad de los pacientes presenta una oportunidad para comprender qué medicamentos pueden ser más efectivos en cada paciente. La revolución de la genómica junto a la farmacología forman una sinergia clave en la personalizaron de la medicina.

  - Medicina Preventiva: El estudio del genoma combinado con los factores del entorno nos permite entender mejor las enfermedades y por ello la prevención de las mismas toma una posición más fuerte. La recogida de datos a través de los nuevos dispositivos móviles, con multitud de sensores, permitirán el desarrollo de nuevas políticas de salud y promoción de estilos de vida más saludables.

  - Medicina Participativa: La inclusión de las nuevas tecnologías y el *big data* como fuente de conocimiento repercuten en un cambio de la relación entre médico y paciente, en el que la posición pro activa, informada y comprometida del paciente abre la puerta a la obtención de mejores resultados a la hora de combatir y prevenir enfermedades.

CAPÍTULO 4

Diseño de un modelo

En este capítulo nos centraremos en el desarrollo de un modelo para la ayuda de toma de decisiones y la creación de conocimiento en el ámbito sanitario, ante la observación de la falta de un sistema actualmente funcional que implemente los conocimientos propios de los macrodatos al ámbito sanitario. Aplicaremos conocimientos vistos y profundizaremos en conceptos que hemos visto en el estado del arte, formado por los capítulos 2 y 3. Previamente explicaremos cuál es la utilización actual de los datos recogidos en el Sistema Nacional Sanitario, y se expondrán algunos objetivos recogidos en documentos oficiales del SNS que van en consonancia con el desarrollo de éste proyecto.

Se detallará el funcionamiento del modelo explicando los requisitos de éste y la interacción de las distintas partes con el modelo en conjunto, plasmando de forma gráfica las distintas partes del modelo, especificando su funcionamiento y las características de las mismas. Se ahondará en el apartado tanto técnico, como los requisitos de la infraestructura tecnológica, como los aspectos más profesionales y relacionados con la protección de los datos de pacientes. Se establecerán las iniciativas de cara a futuro mediante la ilustración de un *RoadMap*, y integración de las partes implicadas en el modelo, así como también las posibles barreras que se tendrán que afrontar en lo referente a la protección de datos, interoperabilidad o la inclusión del modelo en la comunidad sanitaria.

Por ultimo se relacionarán los Objetivos de Desarrollo Sostenible con los objetivos y características de este mismo proyecto, poniendo en evidencia el cumplimiento de algunos de los objetivos de dicha iniciativa, y promoviendo así las mejoras que se desean alcanzar a nivel global a través de los ODS.

4.1 Uso actual de los datos

En 'El Código de Buenas Prácticas de las Estadísticas Europeas' [36] observamos el compromiso de calidad y los procesos estadísticos que la Unión Europea establece como un marco común. Mas allá de ello se nos hace difícil encontrar documentos facilitados de forma pública en los cuáles se establezca los usos actuales en concreto de los datos en sanidad, más allá del portal de Indicadores Clave del SNS que veremos a continuación.

Cuando accedemos a la pagina web del Ministerio de Sanidad, Consumo y Bienestar Social [37] podemos encontrar multitud de tipos de datos mediante varios portales para la consulta de información. Estos están configurados de forma muy efectiva ya que podemos aplicar varios filtros de información sobre los valores que deseamos obtener, y en la mayoría de los casos también se obtienen gráficos que nos ayudan a comprender dichos datos. Tenemos acceso principalmente a dos portales que son los que nos ofrecen este tipo de análisis que hemos comentado anteriormente.

El primero se trata del portal de Indicadores Clave del SNS, en el que como podemos observar en la figura 4.1, nos permite hacer una extensa exploración a través de los indicadores que se muestran a la izquierda de la herramienta. A través de estos podemos por ejemplo descubrir cuál es la comunidad autónoma en la que se frecuenta más los centros de atención primaria en el caso de las mujeres y en el año 2017, por ejemplo. Si indagamos profundamente en dicha aplicación podemos concluir que la cantidad de datos que se pueden extraer es muy grande, cuyo uso puede enfocarse en objetivos de concienciación ciudadana y mejora de la sanidad en ciertos aspectos de gestión pero no nos permite avanzar en el campo de la medicina personalizada y preventiva. El uso de los 247 indicadores según el ministerio es *ofrecer una imagen del estado de salud de la población de España, los factores determinantes de la misma, la respuesta del sistema sanitario a las necesidades de la población e información de contexto sociodemográfica para entender la imagen. Permiten monitorizar el funcionamiento del SNS, comparar diversas dimensiones y ver en qué medida el SNS está cumpliendo el objetivo para el que fue creado.*

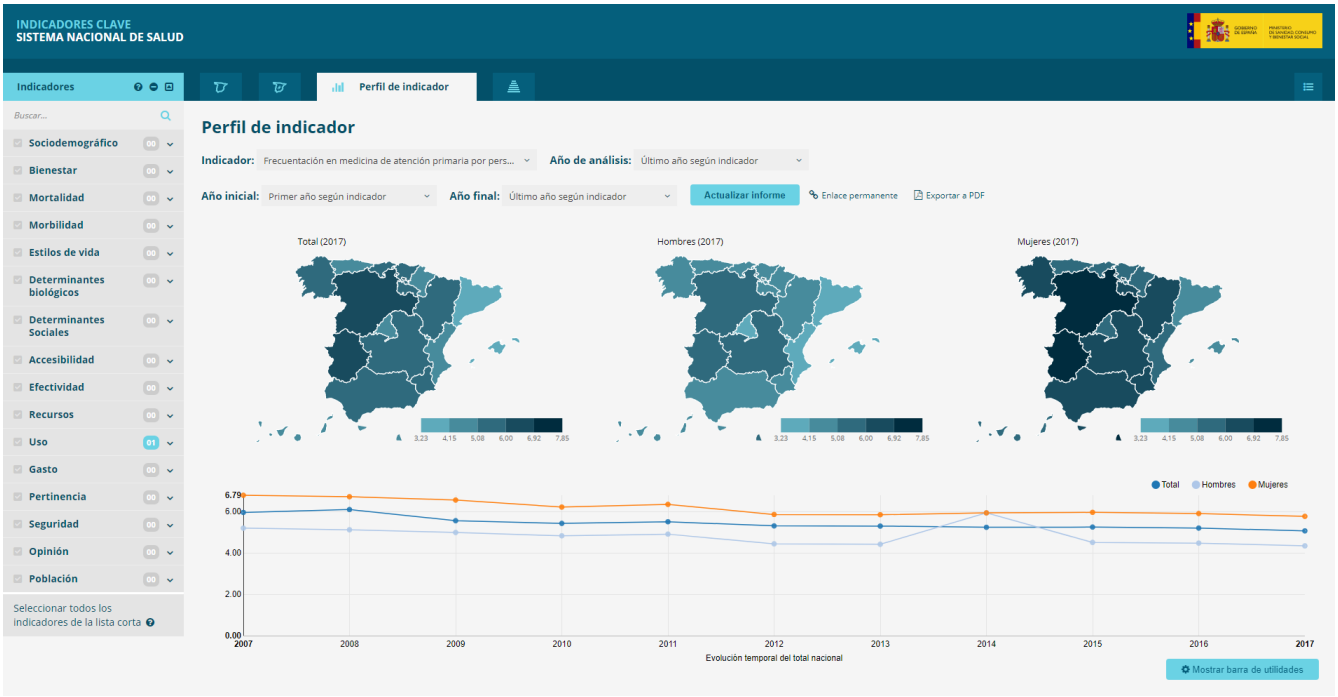


Figura 4.1: Portal de Indicadores Clave SNS

Por otro lado encontramos acceso a la aplicación de ICMBD (Indicadores y Ejes de Análisis del Conjunto Mínimo Básico de Datos), cuyos campos de estudio lista de los siguientes: estancia media, estancia media preoperatoria, frecuentación, hospitalizaciones potencialmente evitables, procedimientos realizados en hospitalización, tasa de ambulatorización quirúrgica, tasa de cesáreas, indicadores de seguridad de pacientes, tasa de infección nosocomial, tasa de mortalidad y tasa de reingresos. El objetivo principal de ésta aplicación es el análisis de datos procedentes las técnicas asistenciales en hospitales, y se realiza una explotación de datos con la intención de proporcionar conocimiento en el estudio de los diferentes casos que se producen y lo que engloba su desarrollo hospitalario. Se trata una aplicación muy potente, haciendo uso de diferentes tipos de análisis de los datos con su correspondiente visualización. Durante la visualización de la información se puede acceder en todo momento a documentos que explican la metodología y los tipos de datos que se analizan, facilitando con ello la comprensión de los datos.

Como vimos en el capítulo sobre *big data*, toda la información que podamos obtener es relevante y es interesante partir de una base solida a partir de la cual ya se realiza algún tipo de análisis sobre la información recogida. A pesar de ello la utilización de dichas aplicaciones no tienen el objetivo que nosotros queremos proporcionar al personal sanitario, siendo éste el de proporcionan de una forma clara y rápida información que sea clave para un diagnóstico mejorado y la toma de decisiones pertinentes, así como también ampliar la información que se le facilita al paciente en una consulta.

En el documento Marco Estratégico para la Atención Primaria y Comunitaria [38], fechado del 10 de Abril de 2019 podemos ver las mejoras que se desean implementar, las cuales están muy alineadas con este proyecto. Entre las acciones más importantes sobre la mejora del uso de los datos y la tecnología en el ámbito sanitario podemos encontrar los siguientes puntos que hemos decidido destacar:

> *Acción E.1.2 Garantizar la interoperabilidad de los sistemas y de los registros clínicos para facilitar el acceso seguro a la información por las y los profesionales de todos los ámbitos de atención. Medio plazo.*
>
> *Acción E.1.5 Impulsar el uso de tecnologías que permitan el registro de datos clínicos en el domicilio del paciente y su volcado a la Historia de Salud Digital.*
>
> *Acción E.2.1 Desarrollar sistemas de automatización que permitan integrar, en la Historia de Salud Digital, las recomendaciones de las Guías de Práctica Clínica del SNS, los protocolos de actuación y las Guías Farmacoterapéuticas de área, así como los protocolos para el desarrollo de las competencias enfermeras de indicación, uso y autorización de dispensación de medicamentos.*
>
> *Acción E.3.1. Impulsar las consultas telemáticas de ágil resolución, con identificación segura de la/del usuaria/o, que permitan la integración de la solicitud de consulta y respuesta en la Historia de Salud Digital.*

Dichos puntos se plantean como una mejora del marco estratégico actual, con la voluntad de llevar a cabo dichas acciones a medio y largo plazo. El objetivo común, como queda plasmado el en documento es el de la modernización de dicho servicio público.

4.2 Diseño del modelo

El objetivo principal del proyecto como ya hemos contado en numerosas ocasiones es la de proporcionar al personal sanitario un modelo de gestión de la información, el cual gire entorno

a la creación de conocimiento así como la ayuda en la toma de decisiones y gracias a ello llevándonos hacia la modernización del sistema sanitario. Este modelo que vamos a ver a continuación plantea una visión más sencilla y transparente del sistema sanitario, una medicina preventiva y personalizada a los pacientes, sintiéndose parte en todo momento de un modelo sanitario modernizado que les proporciona toda la información y ayuda posible. Todo ello a través de una tecnología y técnicas de análisis que existen desde hace años y que ha sido muy bien implementada en otros sectores, como en el comercio electrónico.

A continuación expondremos un esquema general de cuales son los componentes de nuestro modelo con los elementos claves del mismo, los cuales serán descritos detalladamente a lo largo de este apartado.

4.2.1. Análisis de los requisitos del nuevo modelo

Ampliación de las fuentes de datos

El primer cambio que debemos hacer es en el sistema de información, ya que éste repercute en las decisiones de los profesionales, y el objetivo principal es modernizar las fuentes de datos. Debemos mantener todas las fuentes de datos que tenemos actualmente y que son recogidas por el SNS. Estas son principalmente generadas por el sistema sanitario, como el HCE y los datos de pruebas médicas. Además de las fuentes tradicionales de información debemos implementar sistemas de recolección de datos provenientes de dispositivos móviles, wearables, aplicaciones, redes sociales. La aparición de estas supone una revolución del sistema sanitario tal y como lo conocemos actualmente, con una más que necesaria modernización en todos sus aspectos.

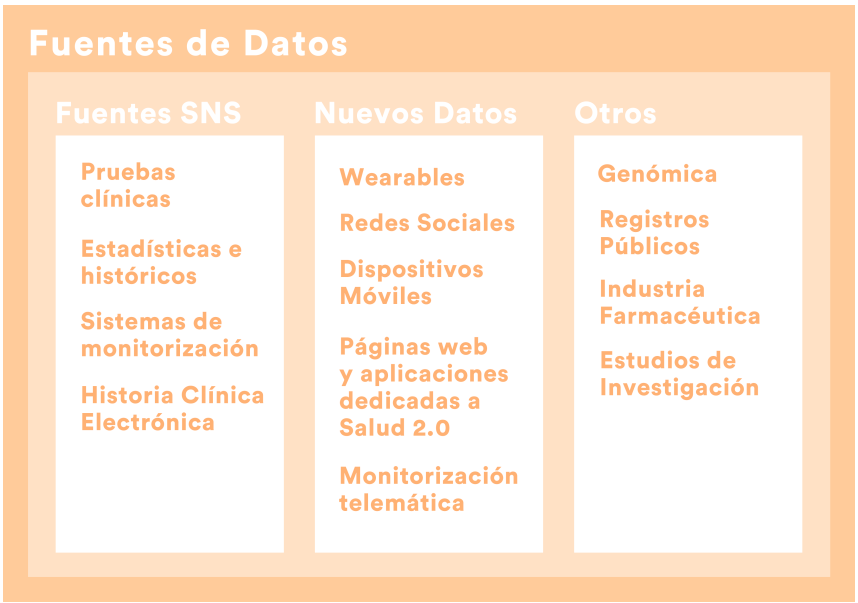


Figura 4.2: Fuentes de datos. **Fuente:** Elaboración propia

Gracias a la inclusión de nuevas fuentes de datos como vemos en el esquema de la figura 4.8 podemos analizar multitud de nuevas informaciones complementarias a las existentes. Como hemos visto en capítulos anteriores, el análisis de redes sociales supone un gran adelanto en el campo de la epidemiología o en el caso de la telemedicina el uso de dispositivos conectados a internet (IoT) es clave. En el modelo que proponemos todas las fuentes de información existentes son incorporadas al mismo, dando una gran importancia a aquellos dispositivos que generan multitud de datos debido a nuestra interacción con los mismos y que sirven para realizar un seguimiento en profundidad de nuestros hábitos de conducta.

Debe de existir un flujo de información que sea retro alimentado por si mismo, volviendo a tomar como información relevante los datos que se generan gracias al análisis de la información proporcionado por el mismo. Por ejemplo si gracias a la inclusión de los datos de wearables en personas mayores, como fuente de información de nuestro modelo, se crea conocimiento que sirve para crear técnicas de medicina preventiva, ésta será proporcionada como datos de entrada para otros pacientes con síntomas similares.

La velocidad de la creación de los datos en la actualidad es ingente, por ello se debe realizar una rápida captura de los datos para realizar el análisis de la forma más ágil posible, obteniendo una rápida herramienta para la gestión del cambio. Los datos que hemos analizado en el capítulo 5 son de 2016, tratándose éstos de los datos más recientes a los que podíamos acceder, eso quiere decir que debido al sistema de información actual la recogida y análisis de los datos no es tan ágil como se desearía, llevando un par de años de retraso en la publicación de los estudios estadísticos realizados sobre los datos recogidos por el SNS.

Sistema de información

Actualmente, de forma muy probable, el sistema de almacenamiento de la información del SNS esté basado en un *data warehouse* debido al tipo de datos que se puede acceder en el banco de datos de la pagina del Ministerio y también por los datos que hemos podido conseguir. Estos datos son procedentes de cuestionarios o formularios que son rellenados por profesionales médicos sobre datos de pacientes extraídos en consultas, urgencias, especialidades, etc. Junto al HCE conforman un gran grupo de datos estructurados, característicos de los DW, usados con fines estadísticos y de creación de informes.

Este tipo de almacenamiento es propio de las fuentes tradicionales que estamos acostumbrados a ver en el sector sanitario, pero debido a la necesidad de ampliar nuestras fronteras en lo que a recolección de información se refiere, necesitamos nuevas tecnologías de almacenamiento que sean capaces de mantener un sistema de información que acepte las nuevas fuentes de datos así como también las nuevas técnicas de *big data* para el análisis de datos.

La tecnología en la que se basa el modelo que planteamos es un *data lake* gobernado por el framework Hadoop (figura 4.3). En nuestro sistema de almacenamiento también se incluirá un DW donde se enviarán todos los datos que estén estructurados y que sólo necesiten de procesos ETL para clasificarlos o limpiarlos. Aquí también se enviará, tras ser procesada, parte de la información estructurada y almacenada en la parte de datos sin procesar del*data lake*. Con ello establecemos un flujo de información en el que el *data lake* se impone como la base de nuestro sistema de almacenamiento de datos, el cual acepta todos los datos provenientes de nuestra fuente de datos, incluyendo las fuentes tradicionales que serán almacenadas en la parte estructurada debido al tipo de dato relacional y estructurado. Estos datos son extraídos básicamente de las tablas almacenadas en las bases de datos del SNS.

Las nuevas fuentes de datos, como los provenientes de Salud 2.0 o mSalud necesitarán de herramientas específicas bajo el ecosistema de Hadoop para que dichos datos puedan ser capturados, almacenados, y procesados con el objetivo de generar información que sea consistente. Para ello vamos a describir cómo sería la configuración de dicho ecosistema con las plataformas Apache necesarias para la correcta configuración del sistema, a través de la figura 4.4 la cual formará parte de la imagen final de nuestro modelo.

Para empezar necesitaremos recolectar los datos de todas aquellas fuentes que no generan datos estructurados. Para ello utilizamos Flume, un potente servicio para la captura, agregación y movimiento de datos que se comporta muy bien con los datos no estructurados provenientes de la Salud 2.0 y mSalud, aunque también admite registros de aplicaciones, sensores y máquinas . Flume se encarga de entregar los datos sin procesar directamente al sistema de almacenamiento

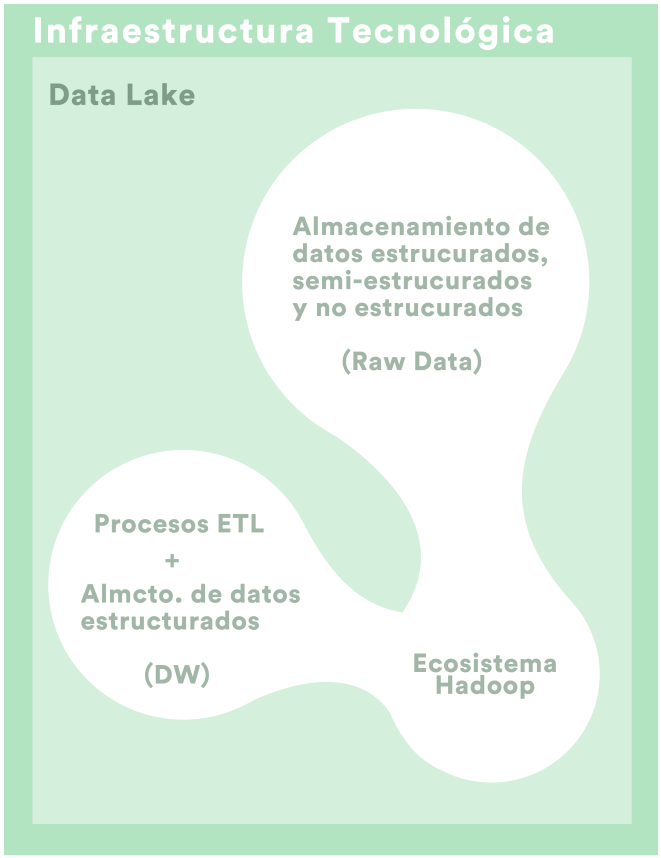


Figura 4.3: Infraestructura tecnológica. **Fuente:** Elaboración propia

HDFS, y permite una total integración con YARN el cual veremos más adelante. El otro servicio que planteamos es Sqoop, cuyo campo de uso se basa en el de *data warehouse* debido a que su funcionamiento está enfocado a la extracción de grandes cantidades de información proveniente de almacenes de datos y bases de datos relacionales, perfecto para la extracción de datos creados en los sistemas de gestión sobre la administración de los recursos sanitarios . Kafka se presenta como una rápida plataforma de análisis en tiempo real de transmisión de datos, usándose en muchos casos como un sistema para el control de flujos de datos mediante el sistema publicador/subscriptor. El uso de Kafka con Flume y Sqoop nos permite una gran flexibilidad en la captura de datos, así como también una gran velocidad en la ingestión de datos provenientes de múltiples tipos de fuentes.

La siguiente capa que podemos observar en nuestro esquema es la del almacenamiento y la gestión de recursos, la cual se basa en el almacenamiento distribuido de Hadoop y YARN como la arquitectura central del ecosistema. Gracias a HDFS podemos utilizar como almacenamiento clústers de bajas prestaciones con características de tolerancia a fallos, escalabilidad o una alta eficiencia de almacenamiento. Además añadimos la base de datos no relacional HBase, la cual nos permite el almacenamiento de grandes cantidades de datos multiestructurados y de gran volumen, ideales para datos genómicos, monitoriazación de pacientes o control de salud en poblaciones. YARN se encarga de la distribución de las cargas de trabajo y comunicaciones entre sus nodos y también entre las aplicaciones, al estilo de MapReduce pero utilizando Tez una versión más reciente y mejorada de éste, comportándose como procesador de datos para aplicaciones como Pig o Hive que veremos más adelante . Esta configuración se plantea como el núcleo sólido de la infraestructura tecnológica a partir de la cual podremos preparar todos los datos para ser analizados.

Para la configuración y mantenimiento del sistema, Oozie actúa como un planificador lógico para los trabajos, proporcionando un coordinador de trabajos y administrador del flujo de trabajo. Junto a YARN, se encarga de la coordinación de diversas plataformas incorporadas en el sistema para el procesamiento de los datos. ZooKeeper se incorpora al sistema para la configuración de los servicios en sistemas distribuidos y para coordinación de lectura/escritura en procesos distribuidos a través de la jerarquía entre los nodos. Por último, Ambari, una herramienta clave para la monitorización, mantenimiento y seguridad de los clusters de nuestro sistema. El conjunto de estas plataformas nos va a asegurar el correcto funcionamiento del sistema a través del mantenimiento del correcto funcionamiento de los clusters.

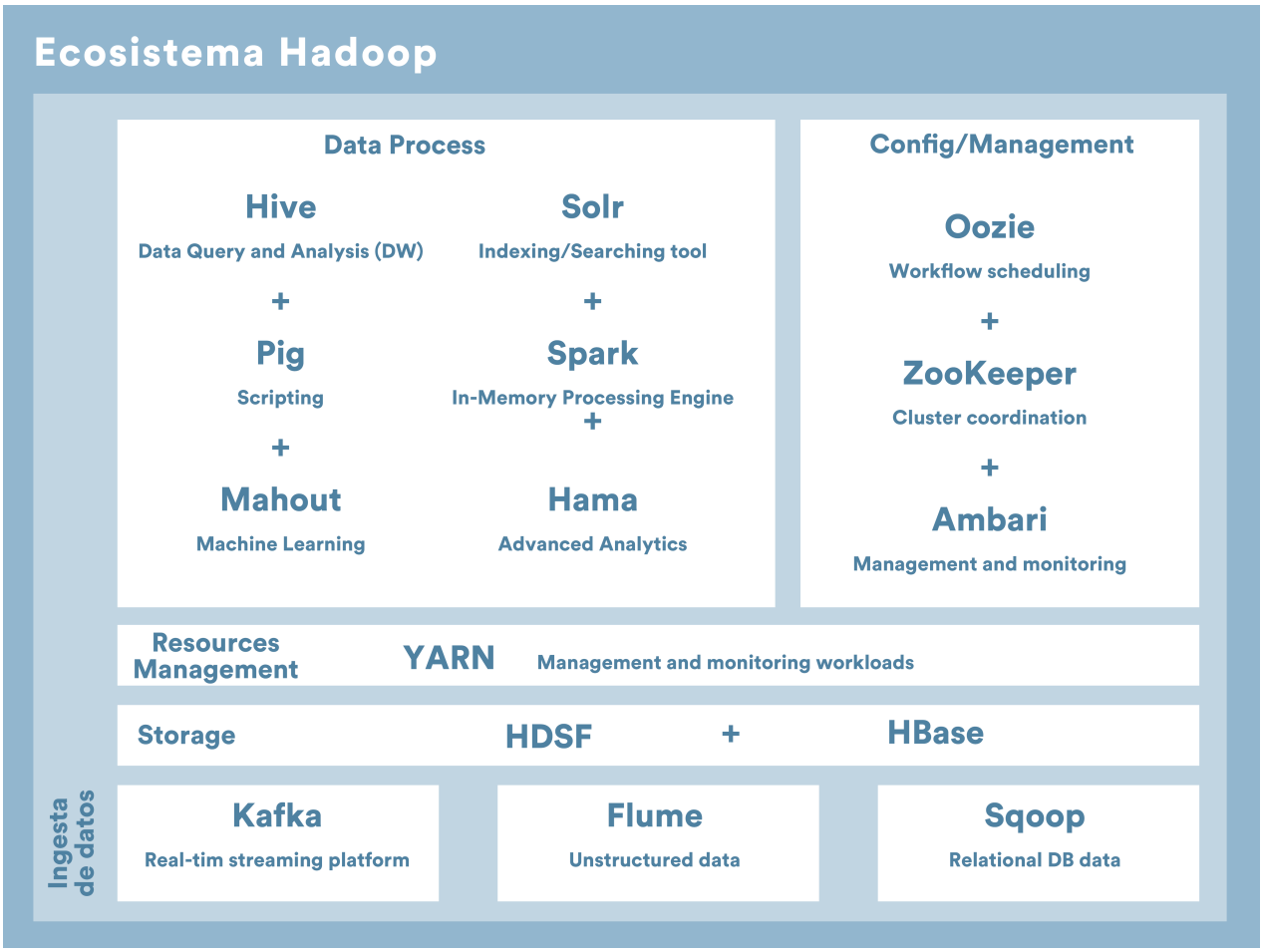


Figura 4.4: Ecosistema Hadoop. **Fuente:** Elaboración propia

Por último, nos encontramos con las herramientas para el procesamiento de datos. En este apartado encontramos plataformas que se encargan de hacer y gestionar cambios en los datos para aumentar la fiabilidad y calidad de los mismos, así como herramientas más avanzadas para el análisis de los datos. Spark se introduce como un motor de procesamiento en memoria enfocado a ejecución eficiente de transmisión de datos, *machine learning* o grandes cargas de trabajo SQL. Debido a la posibilidad del uso de multitud de librerias Python o Java podemos desarrollar aplicaciones enfocadas a ETL. Gracias a su procesamiento en memoria de los datos conseguimos velocidades muy superiores a las que obtendríamos sin éste. Solr se trata de una herramienta para la búsqueda e indexación de datos en HDFS en un tiempo cercano a lo real, con una optimización especializada para grandes volúmenes de datos provenientes de la web también es capaz de realizar lecturas de geolocalización o sensores, muy presentes en la telemedicina. Una de las plataformas que más tiempo lleva en Hadoop es Hive, debido a su estrecha relación con su uso destinado a *Data Warehouses*. Su ejecución junto a Tez nos permite una rápida y eficiente ejecu-

ción de tareas basadas en SQL, perfecto para la ejecución de procesos sobre datos provenientes de fuentes tradicionales como todas las que observamos en la figura 4.1. Pig se trata de una herramienta clave para la creación de funciones que cubran los requisitos específicos del tratamiento de datos, todo ello mediante la creación de *scripts* bajo el lenguaje PigLatin, aplicando procesos iterativos, ETL o búsquedas en datos sin procesar, todo ello a través de YARN y HDFS. Finalmente encontramos dos potentes herramientas para la ejecución de análisis matemáticos y avanzados. Mahout se trata de una plataforma de álgebra lineal distribuida y de expresiones matemáticas para el mundo del *data science*, permitiendo la propia implementación de algoritmos y con una alta presencia del *machine learning*. Su uso junto Spark nos proporcionará una consistente herramienta para la clasificación de grupos de pacientes, búsqueda de tendencias y patrones, recomendación de posibles tratamientos que puedan ajustarse a las necesidades y condiciones del paciente. Hama es una herramienta para el procesamiento de datos de forma paralela, a través de técnicas de grafos, algoritmos de redes y *deep learning*, muy útil para el campo de la epidemiología.

Generación de conocimiento

Hemos visto cuales son las herramientas que nos permiten la manipulación de los datos a través de la infraestructura tecnológica que hemos planteado y que está basada en el ecosistema de Hadoop, un proyecto de *open source* y de acceso gratuito que facilitaría en gran medida la implementación del mismo en el sistema sanitario. A continuación observamos una tabla con algunos de los análisis que engloban el mundo de los macrodatos y que nos serán muy útiles como parte de nuestro modelo para la obtención de algunos objetivos que se plantean a lo largo del proyecto, como la mejora de la satisfacción de los pacientes o el mejor diagnostico por parte del profesional médico. Este es el momento en el que a través de la información que hemos conseguido mediante el procesamiento de los datos en algunas de las herramientas de Hadoop, nos disponemos a crear verdadero conocimiento.

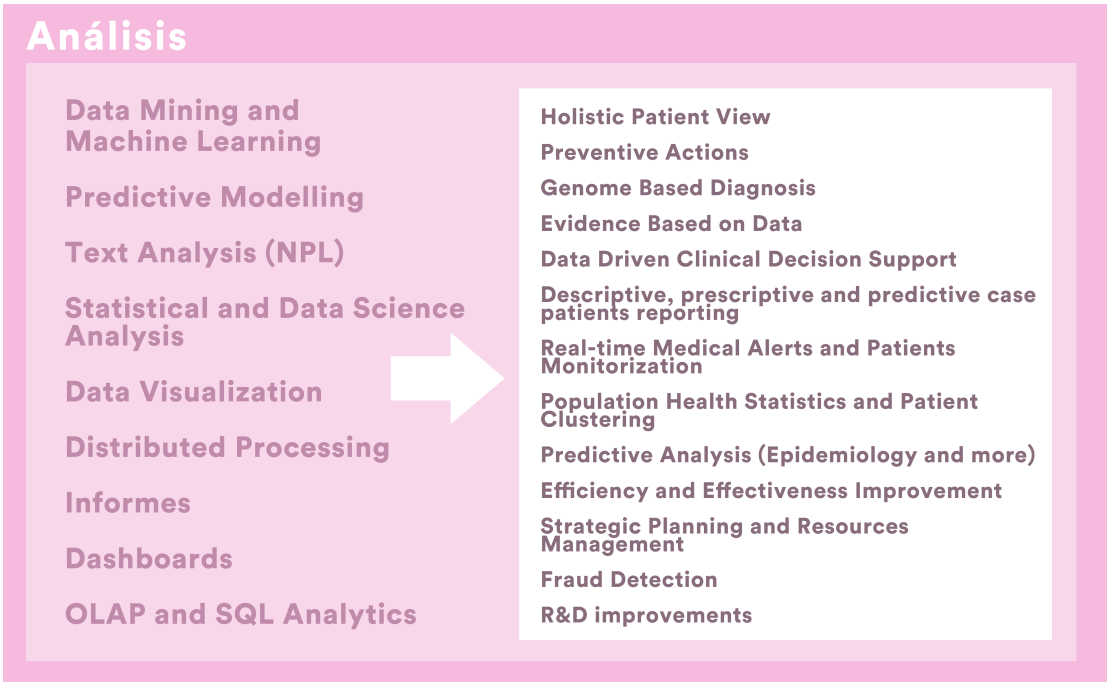


Figura 4.5: Análisis. **Fuente:** Elaboración propia

Mediante técnicas conocidas como *data mining* (DM), *machine learning* (ML), análisis estadísticos y matemáticos relacionados con *data science*, entre otras de las muchas técnicas, somos capaces de extraer información con un alto valor sobre la salud de la población, conocimiento basado en

las evidencias que los datos nos proporcionan y la detección de enfermedades. Esto nos permite ante el caso de una consulta saber si nuestro paciente encaja con síntomas de otro paciente cuyo diagnostico ha sido acertado, algo que un médico no puede saber sólo con su conocimiento debido al gran número de casos médicos que se producen anualmente. La obtención de los valores de salud de la población es un tipo de conocimiento que a día de hoy es accesible como hemos visto al principio del capítulo, pero la tecnología y las técnicas de análisis de datos masivos nos permiten un análisis de datos a una gran velocidad, lo que nos otorgaría una posición ventajosa ante la prevención de enfermedades detectadas en alguna parte de la población.

El DM y ML se plantean también como una técnica clave para la detección de fraude a través de la detección de patrones de conducta. Además de esto pueden servir de gran ayuda en el análisis de datos de localización, conteo de pasos, pulso cardíaco o actividades deportivas provenientes de dispositivos móviles para la extracción de hábitos de los pacientes, ofreciendo una medicina más preventiva desde la mejora de los hábitos de salud. El uso del análisis *natural processing language* (NPL) no sólo nos puede proporcionar una sólida herramienta para el tratamiento de datos de pruebas médicas, transcripción y análisis de anotaciones provenientes de Historiales Clínicos Electrónicos o análisis de textos, si no que también puede ser muy útil cruzar la información extraída de forma estadística sobre pacientes junto con los conocimientos publicados sobre medicina a través del procesamiento del lenguaje natural. En el análisis de textos incluimos también el análisis de sentimientos y de información proveniente de las redes sociales, lo cual nos puede proporcionar mucha información sobre la población y extraer hábitos de la misma.

Otros análisis a partir de técnicas predictivas o de visualización de información, así como procesamiento distribuido nos permite obtener una medicina preventiva y personalizada a través del análisis del genoma humano, y también realizar un gran avance en cuanto a epidemiología se refiere. Las técnicas predictivas y el *data mining* son de gran ayuda en el campo de la farmacología, pudiendo avanzar en estudios sobre los efectos adversos en los pacientes debido a la composición de los fármacos, una utilidad que junto a la secuenciación del genoma humano gracias al procesamiento distribuido y la aplicación de algoritmos avanzados nos permite profundizar en una selección de medicamentos más precisa.

La visualización de información y realización de informes nos proporciona visiones holísticas de los pacientes así como acciones que se pueden llevar a cabo para prevenir enfermedades o minimizar sus efectos y en gran medida la prevención de efectos adversos en el campo de la farmacología, planteándose así como uno de los nuevos enfoques de la medicina, la medicina predictiva y personalizada.

Por último encontramos técnicas relacionadas con *business intelligence*, como la realización de informes, *dashboards* o consultas SQL, las cuales nos proporcionan análisis de casos de los pacientes de una forma completa abarcando un análisis descriptivo de la situación, preventivo ante una posible situación que se formula como deseablemente evitable y de decisión que nos ayude a entender mejor la situación. Además el uso de cuadros de mando o *dashboards* está muy extendido en el ámbito empresarial en cuanto a gestión de objetivos o recursos, un uso que también podemos adaptar a la sanidad. Mediante el uso de *dashboards* podemos mostrar en tiempo real los recursos humanos de los que se está haciendo uso, sabiendo si estamos por encima o por la cuota necesaria para ofrecer un buen servicio. Por ejemplo en urgencias, si tuviésemos una gran carga de pacientes frente al numero de personal sanitario la visualización de dicha información a través de un cuadro de mando sería clara, y la inclusión de alertas proporcionaría una herramienta mucho más sólida.

Todo junto nos permite abordar una modernización del ámbito sanitario a través de los tres beneficios planteados en el capítulo 3, estos son: la reducción del impacto económico y detección de fraude, el incremento de la calidad de la atención sanitaria y por ultimo nuevos enfoques de la medicina gracias a las 4P.

Figura 4.6: Mejoras en la sanidad. **Fuente:** Elaboración propia

Gracias a ellos la modernización de la sanidad llevará a un sistema en el que el uso de los recursos es más eficiente y acertado, reduciendo tiempos de espera en urgencias, triaje u hospitalización, así como posibles accidentes que puedan producirse por la falta de personal en un área sanitaria especifica. La detección de casos fraudulentos que suponen abusos al sistema y un sobre coste en la sanidad serán atajados de manera mas eficiente, previniendo los mismos y estableciendo medidas que prevengan la aparición de los mismos.

La calidad de servicios asistenciales se verá impulsada por las nuevas herramientas como los sistemas de soporte de decisiones y conocimiento basados en datos que el modelo propone. El avance en investigación también será incrementado debido a la multitud de datos que sirven para la creación de nuevos estudios que ayuden a mejorar la vida de las personas a través de la sanidad.

Por último, el nuevo enfoque médico que nos proporciona la medicina basada en las 4P supone un nuevo entorno en el que el paciente toma una posición en la que sus decisiones a nivel médico toman más relevancia, y cuyo estado del bienestar se convierte en un actor principal para la prevención de posibles enfermedades. Por otra parte la evolución en la personalización de la medicina repercute positivamente en la eficiencia de los diagnósticos, así como también en la industria farmacológica la cual tendrá la posibilidad de desarrollar fármacos que se adecuen mejor a las necesidades de los pacientes.

Supongamos un caso ficticio, para visualizar de una forma más real la generación de conocimiento que se produciría: una visita de un paciente a urgencias.

Un paciente llega con dolores intermitentes en el pecho. Nuestro modelo, que ya lleva implantado dos años ha estado recogiendo y generando grandes cantidades de información que ha servido a profesionales médicos para realizar diagnósticos más precisos y eficaces. Además se han implementado con éxito la recogida de datos de numerosas fuentes que no se contemplaban anteriormente, procedentes de mSalud y Salud 2.0. Cuando el paciente llega a urgencias un profesional médico mediante el soporte informático que le facilita la información generada por los análisis comentados en nuestro modelo, le indica que dicho paciente tiene unos malos hábitos saludables. El profesional médico realiza un chequeo del paciente y introduce en el sistema los síntomas del paciente. Un proceso en el que DM y análisis del lenguaje natural se encargan de procesar los datos en tiempo real e introducirlos en la Historia Clinica Electrónica del paciente, así como en las bases de datos de nuestro sistema de información. A través de los datos recogidos por el sistema sobre el paciente observamos que él tiene una actividad física por debajo de lo normal, realiza desplazamientos esporádicos y poco habituales, y sus pulsaciones son altas. Además los datos extraídos a través del análisis de sentimiento del usuario el sistema detecta síntomas acor-

des con la depresión. El sistema nos manda una alerta ya que ha encontrado individuos cuyos síntomas eran los mismos en el historial de datos, gracias a técnicas de DM, ML y perfilación de pacientes. El sistema nos muestra dichos casos y nos elabora un informe descriptivo, predictivo y de decisión en el que ayuda a deliberar al profesional médico con un diagnóstico de episodios de ansiedad y trastornos depresivos.

La derivación del paciente a un especialista es inmediata, allí se le realizará un diagnóstico más avanzado y se le procederá a proporcionar una solución personalizada a través de un diagnóstico basado en el genoma. Mediante éste y técnicas predictivas podremos simular como será la aceptación del paciente de los fármacos que le proporcionarán para atajar su enfermedad. Además el sistema proporcionará una lista de actos de prevención y hábitos saludables que ayudarán a mejorar la vida del individuo. A través de la visualización de diferentes gráficas le mostramos al paciente que dicho diagnóstico es bastante común en individuos de su sexo y edad, y que la mayoría se recuperan en periodo X de tiempo.

Desde la atención del paciente hasta la deliberación de su diagnóstico hemos conseguido aplicar diferentes mejoras planteadas al final de nuestro modelo, como la medicina predictiva, preventiva, participativa y el incremento de la calidad del servicio ofrecido a través de un diagnóstico basado en los datos.

Protección de datos

Ante un modelo como el que planteamos, la protección de los datos del paciente es un apartado muy importante del diseño de nuestro modelo. Cabe recalcar que a pesar de que el objetivo de los datos analizados por las técnicas de *big data* la identificación de los usuarios no aporta ningún valor añadido al análisis, éste deber ser protegido a través de técnicas de anonimización que evitan cualquier proceso de reidentificación. El SNS como responsable de la puesta en marcha del nuestro modelo, es el encargado de facilitar y establecer una serie de bases para que la información de los pacientes nunca corra ningún peligro.

En la RGPD ya se recoge explícitamente el tratamiento de datos a gran escala, en éste se indican medidas organizativas y de seguridad que son necesarias llevar a cabo [39]. El análisis de los riesgos nos permite la toma de decisiones correctas para la reducción de riesgos en la protección de los datos. En dicho análisis se establecen medidas y controles de seguridad que garantizan las libertades y los derechos de los pacientes, entre las cuales podemos encontrar:

- **Flujos de información:** Debemos especificar como se van a recabar los datos, que personas tienen acceso a la información, objetivos, etc. En nuestro caso la información que recogeremos está especificada en la primera parte de nuestro modelo y sólo tendrán acceso a dicha información el personal médico que se encargue de su recogida en el SNS y los ingenieros que mantengan e implementen la infraestructura tecnológica en la que se realiza la captura de información.

- **Identificación de riesgos:** los riesgos más comunes son los incumplimientos de normas por parte del personal médico o informático que esté involucrado en el modelo. Estos recogen la comunicación de datos a terceros sin permiso, mantener los datos más tiempo del establecido, una incorrecta anonimización de datos, incumplimientos de leyes en protección de datos o brechas de seguridad.

- **Evaluación y mitigación de riesgos:** Establecer planes para la evaluación de los riesgos, para mitigar y evitar que los posibles riesgos acaben ocurriendo. Para ello es necesario mantener una buena seguridad en el sistema de información con el objetivo de evitar brechas de seguridad y amenazas externas, así como la formación del personal ante el tratamiento de

los datos y cumplimiento de leyes de protección de datos. Será obligatorio informar de las quiebras de seguridad a la Agencia Española de Protección de Datos (AEPD) en un máximo de 72 horas desde la detección de la misma.

Deberemos asignar un Delegado de Protección de Datos de procedencia interna o externa a la empresa para el cercioramiento del cumplimiento de las normas y leyes establecidas. Entre sus funciones también encontramos el asesoramiento y formación a los empleados que tratan directamente con los datos. Con ello podremos evitar posibles sanciones de la AEPD.

Figura 4.7: Protección de Datos. **Fuente:** Elaboración propia

Otro de los requisitos es el registro de las actividades sobre los datos. En éste debe de establecerse un encargado que lleve a cabo todas los puntos indicados por el Delegado de Protección de Datos, asegurando la correcta implantación de las normas. Los fines del tratamiento de los datos, así como los departamentos y el personal que tiene acceso a los mismos debe de quedar recogido. La posible transferencia de datos también debe de quedar recogida, la cual debe de efectuarse siempre bajo un contrato de confidencialidad.

A parte debemos cumplir los derechos ARCO se basan en 4 premisas:

- **Acceso:** Como paciente, éste tiene el derecho de saber en todo momento los fines del tratamiento de los datos, saber que datos se han obtenido de forma indirecta, cuando se transfieran datos a terceros, tiempo de almacenamiento de los datos.

- **Rectificación:** Se refiere al derecho que el paciente tiene a cambiar los datos personales que figuren de una forma inexacta.

- **Cancelación:** Los pacientes tienen derecho al olvido, es decir a que sus datos no sean recogidos y tratados nunca más, y que sean eliminados del sistema si así lo desean.

- **Oposición:** El paciente tiene derecho a negarse al tratamiento de la recogida y análisis de sus datos, como por ejemplo, los datos que se utilizan en nuestro modelo para numerosos análisis, como el de perfilación de pacientes.

Como ya comentamos en capítulos anteriores es el paciente el propietario de los datos extraídos por el sistema sanitario a lo largo del uso de sus servicios asistenciales y estancias hospitalarias, y el uso de dichos datos para su análisis debe realizarse ante un previo consentimiento explicito y recogido por escrito. Los pacientes deben de ser informados del objetivo del tratamiento de sus datos, qué datos serán tratados y por cuanto tiempo quedarán almacenados, ante la premisa de un plazo legal de 5 años como máximo. [40]

Para el cumplimiento de estos derechos, a partir de nuestro modelo planteamos la habilitación de una página dentro del servicio Çl@ve". Éste se trata de un servicio para el acceso electrónico de los ciudadanos a los servicios públicos. El acceso se realiza mediante un usuario y contraseña basados en el Documento Nacional de Identidad de cada ciudadano.

Interoperabilidad

A lo largo del estudio del proyecto se ha detectado un grave problema ante la falta de interoperabilidad de los sistemas de gestión de información en los distintos ámbitos. Existen fronteras actuales entre comunidades autonomas debido a la cesión de competencias sanitarias a las comunidades [28], entre departamentos que evitan la visualización de pruebas o informes médicos en otros departamentos y la comunidad sanitaria. El problema se complica aún más si comparamos sistemas de información de distintos hospitales o incluso el sector sanitario privado y el público. Esto es debido a la existencia de diversas versiones del mismo programa, o incluso el uso de distintos softwares Abucasis, Orion-Clinic, Sistema de Historia Salud Electrónica, Iris o Cordes en órganos públicos como urgencias, ambulatorios, consultas especializadas y hospitales.

Encontramos estudios [41] que plantean esquemas para la interoperabilidad en los intercambios de información y mensajes entre sistemas sanitarios. A pesar de ello nos consta, debido a la interacción con trabajadores del sector sanitario, que los sistemas de información actuales suponen un lastre para los profesionales sanitarios debido a la falta de integración de un modelo genérico que facilite la compartición de los datos entre órganos sanitarios para conseguir los objetivos anteriormente comentados.

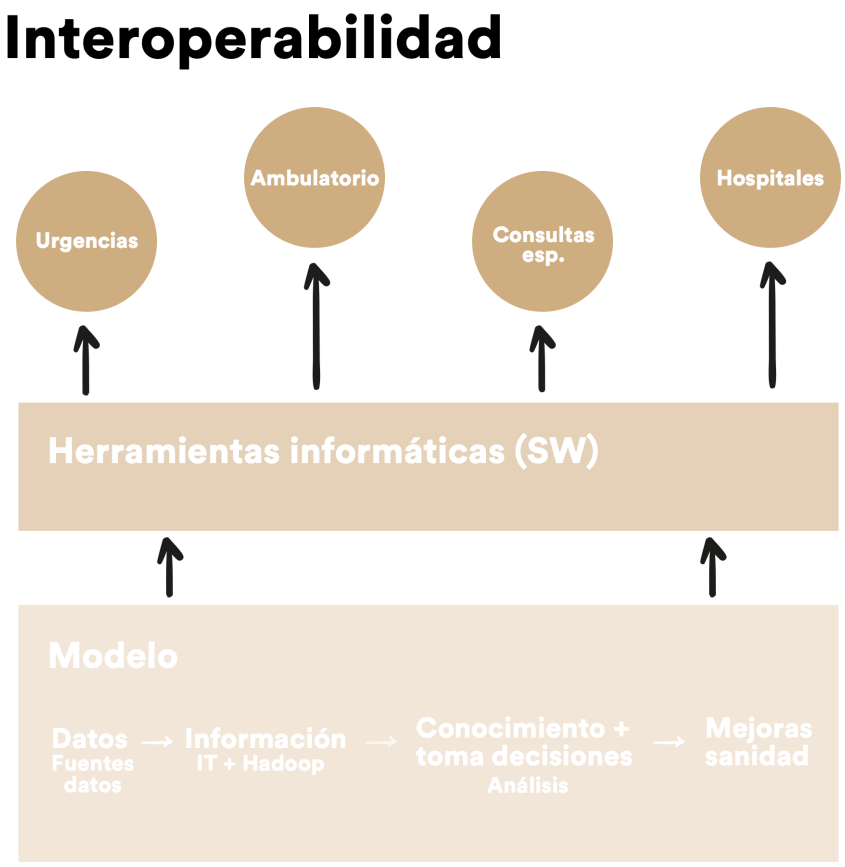


Figura 4.8: Interoperabilidad. **Fuente:** Elaboración propia

La creación de herramientas informáticas es un objetivo clave para la materialización del modelo. Nuestro modelo, como resultado de la aplicación de técnicas de análisis de datos masivos en salud, se plantea como el núcleo de un sistema informático para la gestión de la información médica. Para ello es inevitable la creación de una red en la que los distintos programas software

se comuniquen a través de un núcleo común gracias a la implementación de estándares que nos permitan una implantación real de un modelo de gestión de información interoperable, y que la compartición de datos no se limite a la información relacionada con las recetas electrónicas o los historiales clínicos electrónicos.

Como se ve en la figura 4.8, se genera un único sistema que evita los problemas de interoperabilidad planteados anteriormente a través de la alimentación de los datos a través del modelo planteado. De esta forma cualquier herramienta o aplicación leerá y tendrá acceso completo a los mismos datos que todas las demás, proporcionando un sistema consistente en el que toda la información es compartida para las mejoras asistenciales y también en investigación.

Alcanzar el objetivo de un modelo que figurase como núcleo de los sistemas de información proporcionándonos la interoperabilidad deseado nos traería consecuencias tan positivas como:

- **Historial único**. La obtención de un historial único supondría la eliminación de barreras entre servicios asistenciales que no utilizan los mismos sistemas de información y que por ello no comparten las pruebas o diagnósticos relativos a un paciente.

- **Acceso por profesionales sanitarios**. El acceso a la información por profesionales médicos y de enfermería a las distintas pruebas realizadas sobre el paciente pertenecientes a distintos departamentos sanitarios facilita el diagnóstico de dichos profesionales debido al acceso de todas las pruebas en el historial de un paciente.

- **Mejora sanitaria e investigacional**. La interoperabilidad abre las puertas a la gestión y análisis de mayor cantidad de información, lo que se traduce en mejores informes, diagnósticos mejorados, mayor eficiencia, reducción de costes y una mejor gestión de la sanidad. Todo ello nos conduce a la modernización sanitaria deseada.

4.2.2. Vista general del modelo

Una vez construido el modelo somos capaces de tener una vista más amplia del mismo, y en vez de presentarlo por apartados ahora podemos verlo como un todo. Como hemos visto anteriormente nuestro modelo empieza con la captura de información de diversas fuentes. Este es el primer punto de cualquier sistema que haga uso de técnicas de *big data*. Recordamos que esta tarea se realiza a partir de la capa de ingestión de datos del ecosistema de Hadoop, con el que somos capaces de almacenar todos los datos en nuestro *data lake*.

Como observamos en la figura 4.11 en nuestro *data lake* (DL). encontramos datos no procesados y otros datos que están previamente estructurados debido a que provienen de aplicaciones que usan bases de datos relacionales para el almacenamiento de los mismos. Estos datos serán enviados directamente al Data Warehouse que encontramos dentro del *data lake*., dónde enviaremos toda la información que ya hemos procesado y clasificado correctamente a través de las aplicaciones Hadoop. Por ello el DW tiene dos flujos de datos, el que le entra directo de las fuentes de datos y el proveniente del DL. Recordamos que el éste mantiene los datos sin analizar hasta que vayan a ser usados, pero nosotros procesaremos datos provenientes de fuentes cuyo tipo es no estructurado como, por ejemplo, notas médicas o anotaciones en los HCE para agilizar algunas de las informaciones de nuestro sistema de almacenamiento.

El ecosistema Hadoop nos blinda una gran plataforma que nos sirve para estructurar, limpiar y clasificar a gran escala todos los datos. Además nos permite ejecutar análisis en algunas de las plataformas que hemos presentado como Hama o Mahout. También tenemos la opción de procesar por nuestra cuenta la información que nos otorga Hadoop a partir de *scripts* en Pyhon, Scala o R, en los cuales podremos implementar algoritmos o nuestras técnicas de análisis y transformar dichos análisis de información en técnicas para la creación de conocimiento sobre los datos de

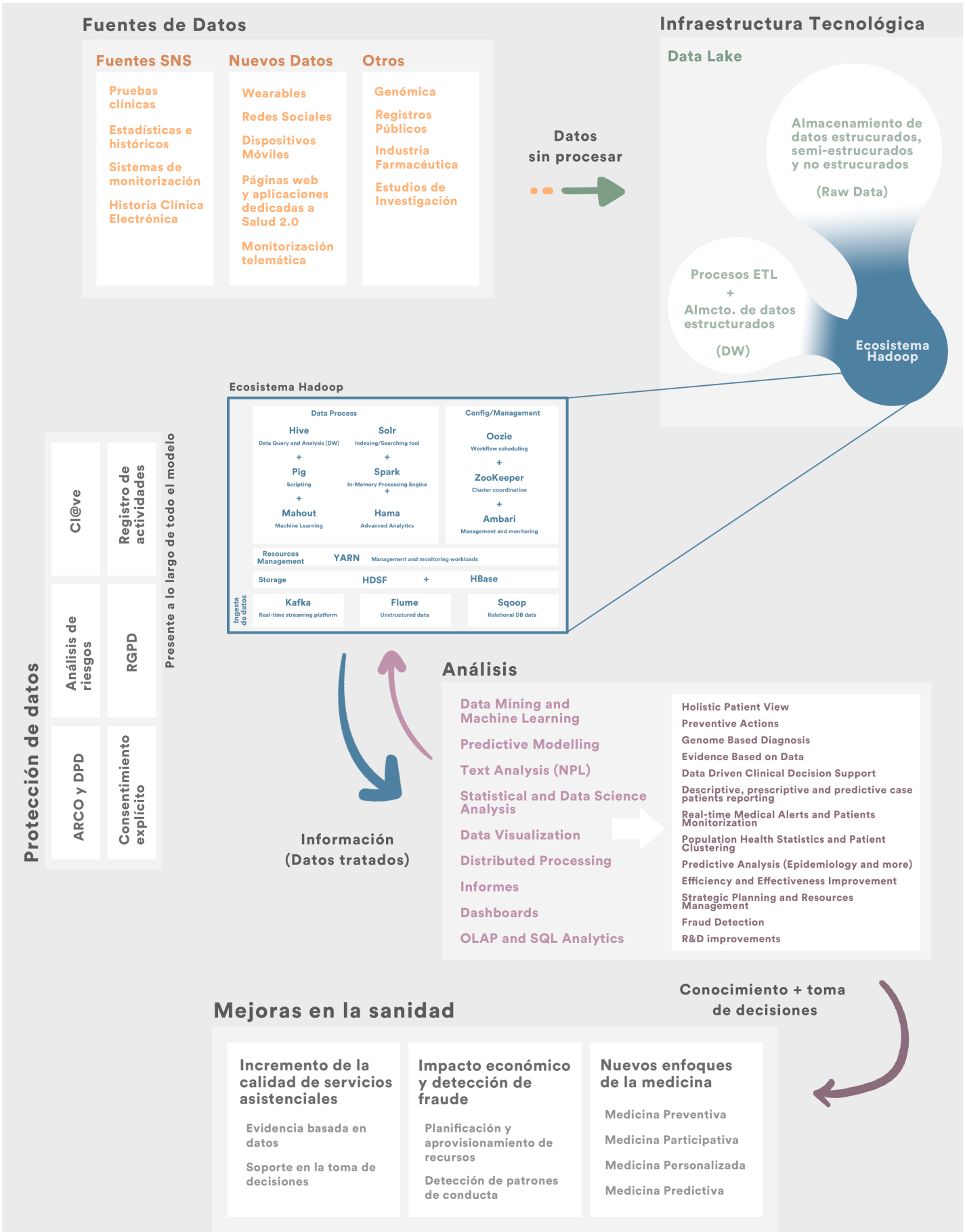


Figura 4.9: Modelo. **Fuente:** Elaboración propia

salud. Un ejemplo es a través del *data mining* y *machine learning*, analizando los datos de millones de pacientes y sus historiales médicos, crear acciones preventivas que pueden llevar a cabo los pacientes para mejorar su salud y evitar la aparición de ciertas enfermedades.

Una vez hemos generado el conocimiento su uso para la toma de decisiones es inevitable. Los nuevos puntos de vista que nos aporta la generación de un conocimiento ampliado y complementario al del profesional sanitario se plantea como una sinergia enriquecida entre el modelo y el personal médico. Ante este nuevo entorno en el que el conocimiento en tiempo real proporciona una nueva herramienta en diagnóstico del profesional médico encontramos diversos cambios en el ámbito sanitario que nos llevan al concepto de sanidad modernizada. Entre estos encontramos los nuevos enfoques de la medicina, el impacto económico y detección de fraude y también el incremento de la calidad de los servicios asistenciales. Aspectos claves para la transformación hacia una sanidad mas eficiente en cuanto a recursos y efectiva en cuanto a diagnósticos que derivan en un incremento del bienestar social.

Cabe señalar que a pesar de que el apartado de protección de datos esté indicado al final del diseño, en la siguiente figura 4.11 lo situamos a la izquierda del esquema con la intención de representar que la protección de datos y todos sus puntos deben de mantenerse a lo largo de todo el proceso, haciendo especial incapie en la recogida y análisis de los datos.

4.2.3. Integración y comunidad sanitaria

Hay que tener en cuenta que la implantación del modelo en el ámbito sanitario conlleva una evolución tecnológica y por ello hay que afrontar una serie de barreras que no solo afectan al personal médico, si no también a toda la población en calidad de pacientes. El modelo planteado supone una modernización del actual modelo sanitario y un cambio entre las relaciones paciente-médico, el cual ya se empieza a observar en la sociedad, en el que el paciente se sitúa en una posición ventajosa debido a toda la información que dispone a su alcance. A través de estos objetivos también queremos hacer llegar a los ciudadanos una visión más participativa en la salud ya que como se indica en el documento de satisfacción ciudadana del Ministerio de Salud y Bienestar, más del 50 % de la población percibe que la información recibida por el personal médico no es suficiente y el 30 por ciento no se siente participes a la hora de tomar decisiones sobre su salud.[42]

Para ello debemos crear guías ciudadanas de promoción del nuevo modelo que deseamos implementar, enseñando al paciente la posición ventajosa que obtendrá debido a la inclusión de las nuevas tecnologías en el ámbito sanitario. Además la formación del personal mediante manuales y cursos es una acción estratégica clave para el correcto funcionamiento del nuevo modelo, en el que a pesar de que toda la información sea generada y recogida de forma automática o casi automática, el uso del sistema informático que engloba nuestro modelo será utilizado por el personal médico bajo sus conocimientos sobre el mismo. Para obtener un mejor resultado y ofrecer un mejor servicio asistencial al paciente, así como trasladar el nuevo modelo de sanidad más efectivo y eficiente, debemos transmitir al personal médico que se les pone a disposición una herramienta con la que realizar una mejora inminente en la calidad de la sanidad.

Las acciones que deberemos de seguir para que se produzca una correcta integración del modelo en la sanidad por parte del personal médico serán las siguientes:

- **Creación de guías y manuales**. Aquí se incluye una presentación del nuevo modelo que se va a implantar inminentemente, transmitiendo al personal médico que ellos son el eslabón más importante del proceso. Debemos tener en cuenta las barreras culturales en cuanto a lo tecnológico se refiere, por ello se deben introducir los cambios que se realizarán de una forma breve y simple, sin llegar a utilizar tecnicismos tecnológicos que puedan confundir a

los profesionales médicos. Dichos cambios pueden ser visualizados a través de un esquema del funcionamiento del nuevo modelo que se desea implantar.

Una vez introducidos los cambios debemos especificar cual es la relación del personal con el sistema, indicando de qué forma interactuarán con el mismo y especificando los objetivos y beneficios de su uso. Sería importante incluir algunos esquemas representativos del mismo en el uso diario de los profesionales en la salud.

- **Cursos formativos**. La interacción del personal sanitario con el modelo se realizará a través de una herramienta informática, la cual no queda recogida en el proyecto y que se plantea como reto de futuro, cuyo funcionamiento deberá ser explicado a los profesionales médicos. La mejor solución es la disposición de cursos formativos del nuevo modelo, de la herramienta y de lo que supone la modernización del sistema actual.

  En éstos no solo se formará al personal en la utilización de la herramienta en los distintos ámbitos de la salud, si no que se introducirá también los nuevos enfoques se que quieren transmitir al paciente. Con ello nos aseguraremos de que la evolución del sistema no repercute solo en el incremento de la calidad de los servicios asistenciales en cuanto a diagnósticos, si no también en la forma de interacción y esa nueva relación paciente-médico que hemos comentado anteriormente.

- **Vídeos explicativos**. La elaboración de vídeos informativos para el personal puede suponer la mejor asimilación de los cambios que se introducirán con el nuevo modelo. En estos quedarán plasmado el funcionamiento de lo que puede ser un caso simulado de un paciente ficticio. A través de dicho sistema se espera una mejor aceptación y una completa utilización del nuevo sistema.

Los cambios introducidos en el ámbito sanitario también deberán ser trasladados a los ciudadanos a través de campañas de información que se encarguen de enseñar el nuevo modelo y los beneficios que supone para los pacientes. El nuevo enfoque de la medicina de las 4P proporciona un cambio sustancial, involucrando al paciente en las decisiones de su propio bienestar y haciéndolo más participe del sistema sanitario, otorgando una mayor relevancia a los hábitos saludables en calidad de prevención de enfermedades.

4.2.4. Iniciativas futuras

RoadMap

A continuación presentamos una hoja de ruta del desarrollo e implementación del modelo que se ha plantado en este proyecto. No se ha podido determinar el intervalo de tiempo con el que se deben de cumplir los objetivos que se señalan en el *RoadMap* debido a que no tenemos constancia de la implementación de ningún sistema parecido bajo la normativa Europea, y como bien es sabido son las leyes sobre el tratamiento de datos y el cumplimiento de normas sanitarias las que más pueden retrasar el proyecto debido a a procedimientos ajenos al simple desarrollo e implementación del sistema.

Por ello presentamos una breve y simplificada hoja de ruta con los objetivos que se consideran clave para la posible puesta en marcha de dicho modelo, un plan que se deja para iniciativas futuras.

Dichos objetivos quedarán comprendidos dentro de las siguientes fases del proyecto, como también se ha indicado en la figura 4.11.

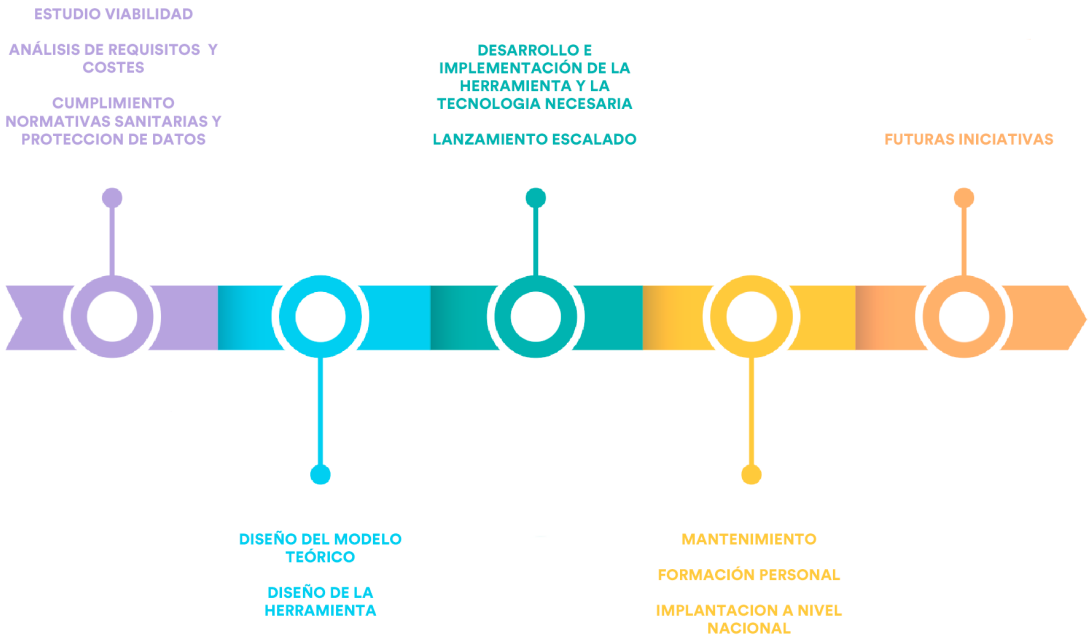


Figura 4.10: RoadMap. **Fuente:** Elaboración propia

- **Planificación**. En esta fase nos encargaremos de los primeros pasos para poner en marcha el proyecto que podría llevarse a cabo a raíz de nuestro modelo. Necesitaremos una selección de personal y creación de grupos de trabajo entre los que encontramos informáticos encargados de la implementación del sistema o el Delegado de Protección de Datos. El cercioramiento de que se cumplen las leyes y normativas en calidad de protección de datos es primordial para la puesta en marcha del proyecto. Es importante establecer los requisitos del sistema que se quiere desarrollar, tanto en cuanto a el hardware que se necesitará para implementar la herramienta *big data* como también profundizar en los requisitos del modelo que se ha planteado en este proyecto.

- **Diseño**. Partimos de la ventaja que ya tenemos el diseño del modelo en el que se basa la herramienta, a pesar de ello es solo un diseño teórico en el que hay que profundizar a la hora de la implementación. Debemos diseñar también la herramienta mediante la que se hará uso nuestro modelo, es decir, la herramienta que muestre al personal médico la información que se ha generado a través de los análisis de datos y que les ayude en la toma de decisiones. El diseño de la misma supone la creación de casos de uso y prototipado de la herramienta a través de diseños semi-funcionales.

- **Desarrollo y puesta en marcha**. El modelo que hemos creado se lleva a fase de implementación, siguiendo el diseño que hemos realizado a lo largo del proyecto. Será necesario también desarrollar la parte visible del proyecto, la herramienta de visualización de la información. En éste deberemos asegurarnos de que implementamos las funciones necesarias para cumplir con la protección de datos, así como también se satisfacen los requisitos propuestos y se pone en marcha un sistema totalmente funcional. Para su puesta en marcha lo ideal sería realizar un lanzamiento escalado, para poder poner en uso el sistema bajo un numero pequeño de centros sanitarios y así poder observar que aspectos son necesarios mejorar.

- **Adaptación y formación**. Debemos recoger todos los fallos, y proceder al mantenimiento del sistema para dar soporte a éste ante fallos, para el que se necesitará un departamento informático dedicado. A su vez se deberá formar al personal sanitario para la utilización de la herramienta a través de cursos, manuales y vídeos informativos. Cuando la herramienta esté totalmente pulida y el personal haya sido formado se podrá lanzar a nivel nacional.

Como otras iniciativas futuras tras la realización de un proyecto para el desarrollo e implementación del sistema que integra nuestro modelo, se proponen dos ítems que se desearía añadir para mejorar la interacción con el modelo a través de la herramienta creada.

- **App para pacientes**. Mediante la propuesta de dicha iniciativa se plantea la creación de una herramienta basada en la nube a través de la cual seremos capaces de ver la información que se recoge, se analiza y se genera sobre nosotros como pacientes. En ésta se habilitarían algunas funciones con el objetivo de proporcionar un mayor estado del bienestar a través de técnicas de prevención. Dichas técnicas variarían dependiendo del paciente, así como también de los datos generados a través de su dispositivo móvil inteligente, permitiendo recomendaciones en los cambios de sus hábitos de salud.

  Además ésta plantearía también una gestión familiar de los individuos del núcleo familiar, pudiendo visualizar y también gestionar los datos de los pacientes menores de edad. A través de dicha aplicación también podría ejercerse los derechos ARCO, y cancelar en cualquier momento la recogida y análisis de datos sobre un paciente.

  También se realizaría un estudio para comprobar la viabilidad de integrar un servicio asistencial telemático, con el objetivo de ayudar a las personas que por problemas sanitarios no pueden acudir al médico. Se vería reducido también la sobre frecuentación de servicios sanitarios por problemas banales.

- **Implementación de una IA**. Es sabido a día de hoy que la inteligencia artificial (IA) tiene cada vez un potencial más grande [43]. Se plantea como un añadido necesario a las tecnologías de la telemedicina, en la que mediante un dispositivo electrónico en casa con una pantalla podemos evitar visitas a nuestro centro sanitario mediante la resolución de algunas preguntas que podamos tener referentes con la salud. Dicho sistema se basaría en la implementación de la inteligencia artificial a través de un asistente que se encargase de comunicarse con los pacientes.

  Su uso junto a técnicas de análisis predictivo y *machine learning* ayudan a la identificación y predicción de la aparición de cáncer, un uso que en el ámbito de la oncología es algo que ya se está potenciando en los Estados Unidos [43]. Las pruebas de imágenes médicas también podrían verse afectadas positivamente para la identificación de masas, huesos rotos o cualquier tipo de problema identificable en las mismas.

  Todo ello llevaría a la reducción de costes debido a la reducción de visitas médicas y la temprana detección de enfermedades o problemas en los pacientes.

4.3 Objetivos Desarrollo de Sostenible

A principio del año 2000 la ONU, con la colaboración de multitud de países, lanza una iniciativa para conseguir alcanzar 8 objetivos para el año 2015. Es en este mismo año cuando se vuelven a evaluar los objetivos y se amplían hasta un total de 17, con un plazo de tiempo hasta 2030.

Con dichos objetivos se busca una mejora a nivel global en diversos aspectos para mejorar la vida de todos. Entre estos encontramos objetivos relacionados con la pobreza, la salud, la economía, la desigualdad o el cambio climático entre otros. Es responsabilidad de todos, y no sólo de

las grandes empresas, el compromiso del desarrollo de proyectos que cumplan los Objetivos de Desarrollo Sostenible. A continuación enunciamos detalladamente cuales son los objetivos que se incluyen en la propuesta de la ONU.[44]

- Objetivo 3: Garantizar una vida sana y promover el bienestar para todos en todas las edades.

El proyecto se presenta como una mejora para el sistema sanitario nacional, a pesar de ello este podría ser extrapolado a cualquier país con la infraestructura suficiente como para mantener la tecnología y explotar las fuentes de datos.

Con ello podremos ayudar a conseguir las metas de prevención y tratamiento del abuso de sustancias adictivas, epidemias de SIDA, tuberculosis, malaria, o las muertes evitables en recién nacidos e infantes. Así también queda cubierto de una forma más amplia la alerta temprana, reducción de riesgos y gestión de los riesgos para la salud, debido a las técnicas de prevención y de mejora de diagnósticos. El apoyo a las actividades de investigación y desarrollo de vacunas y medicamentos para las enfermedades transmisibles y no transmisibles también quedaría contemplado a través del uso de la medicina personalizada y el uso del análisis genómico.

Figura 4.11: Objetivos de Desarrollo Sostenible

- Objetivo 8: Promover el crecimiento económico sostenido, inclusivo y sostenible, el empleo pleno y productivo y el trabajo decente para todos.

Con el proyecto se promueven algunas de más metas que se proponen alcanzar como la de lograr niveles más elevados de productividad económica mediante la diversificación, la modernización tecnológica y la innovación. Relacionado estrechamente con nuestro modelo el cual supone una gran innovación y modernización tecnológica en el ámbito sanitario.

- Objetivo 9: Construir infraestructuras resilientes, promover la industrialización inclusiva y sostenible y fomentar la innovación.

En éste se busca una inversión en las tecnologías de la información, muy presentes en nuestro modelo, las cuales son capaces de promover un desarrollo sostenible y empoderar a las sociedades, como hemos comentado a lo largo del proyecto, para obtener una mayor estabilidad social.

Así mismo nuestro modelo se plantea como una solución para la mejora de recursos, en los que quedan implicados la mejora de la sostenibilidad del ámbito sanitario mediante la reducción de recursos naturales y humanos, y un aumento de eficiencia energética debido a la prevención y reducción de visitas médicas evitables, lo que repercute positivamente en el aprovechamiento de los recursos energéticos.

CAPÍTULO 5

Caso BDCAP

Se expondrá un caso de ejemplo a través de los datos que hemos conseguido del BDCAP, facilitados por el Ministerio de Sanidad, Consumo y Bienestar Social. Aquí observaremos, en pequeña medida, cómo seria la presencia de un sistema basado en un modelo como el que propone en el proyecto y sus efectos.

Se explicará el procedimiento seguido para la obtención de los datos necesarios para la elaboración de la herramienta, incluyendo los obstáculos encontrados a lo largo del proceso. También se detallará el proceso de extracción de muestras, así como el porqué del uso de éstas, para el análisis de los datos contenidas en las mismas.

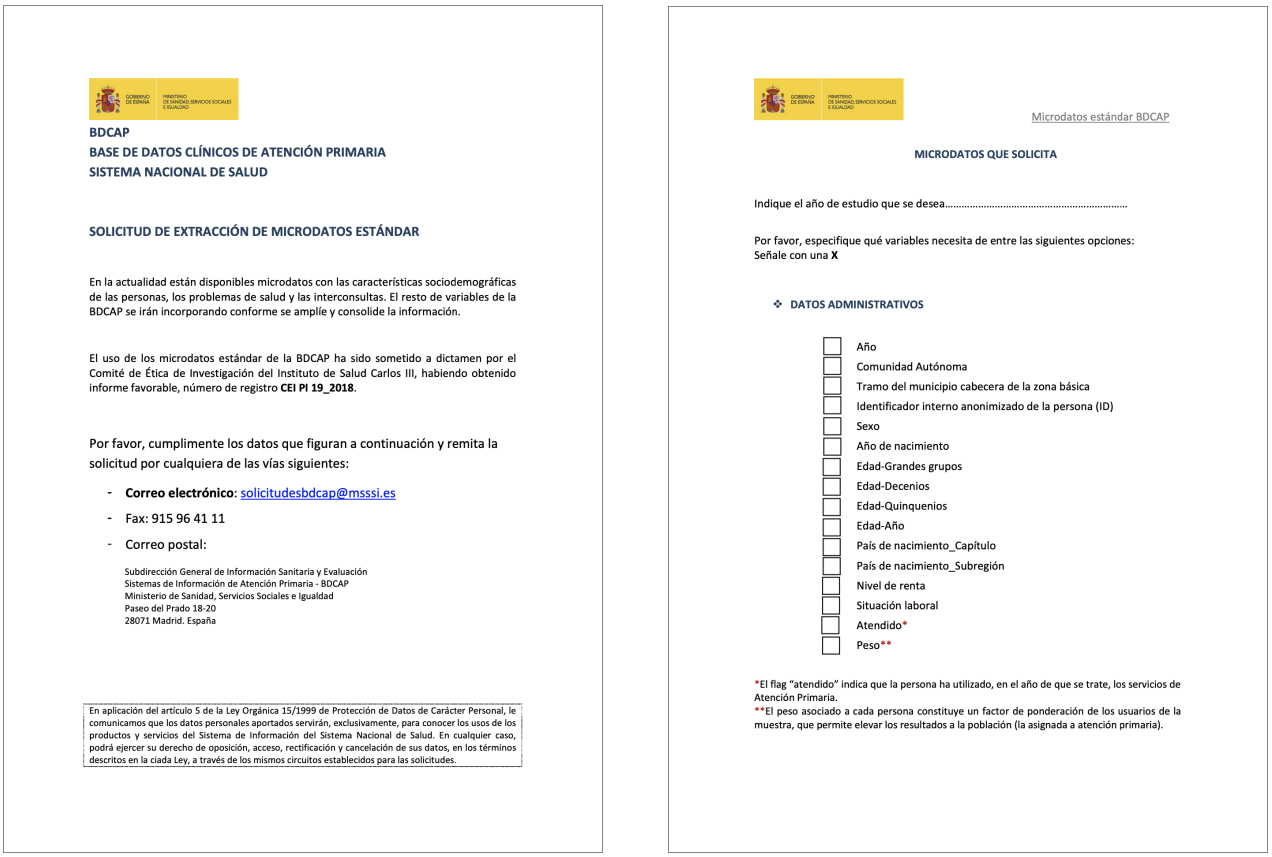
Para ello crearemos una herramienta, presentada como una pequeña parte del modelo que se quiere diseñar, que nos ayude a visualizar información a partir del procesamiento de los datos en tiempo real a través de gráficos o *dashboards* si fuera necesario, que aporten información clave para la toma de decisiones en situaciones del entorno sanitario, como pueden ser los diagnósticos a pacientes en consultas especializadas.

Además se detallará en profundidad la realización de la herramienta a través de los fragmentos de códigos en Python añadidos en el Apéndice A. También se abordará el prototipo de la herramienta con el objetivo de aportar una visión completa y funcional del caso que se desea exponer.

5.1 Solicitud de los datos

Al inicio de la propuesta del TFG a la actual tutora, se planteó la obtención de los datos a través del Hospital de la Fe de Valencia, pero finalmente dicha idea quedó descartada ante la incompatibilidad de agendas de la prestación de los datos con la fecha en la que se quería defender el proyecto. Es entonces cuando investigamos los datos a los que se podía acceder de manera pública en la sección de Sanidad en Datos de la pagina web del Ministerio de Sanidad [45], Consumo y Bienestar Social. Este caso ya explicado anteriormente nos lleva hasta datos generales de acceso libre cuyo objetivo estadístico dista bastante del perseguido en el proyecto.

En la misma página del ministerio podemos encontrar solicitudes para la extracción de datos de acceso mediante solicitud. Dentro de dicho apartado podemos encontrar dos opciones las cuales nos podrían servir, estas son las de 'Registro de Actividad de Atención Especializada (RAE-CMBD)' y 'Base de Datos Clínicos de Atención Primaria-BDCAP'. Rellenamos las solicitudes de ambos para una vez descargados los datos comprobar que ficheros serán los que mejor se ajustará a nuestras necesidades y con los que se podrá obtener un mejor resultado en el proyecto. En dichas solicitudes se nos solicita los datos de la persona que desea obtener los datos, también la finalidad y qué variables son las que solicitamos (figura 5.1). El año de estudio que se desea también es solicitado. En el caso de RAE-CMBD la solicitud es más complicada ya que abarca muchas mas variables y éstas no quedan muy claras a pesar de que están indicadas en un anexo a la solicitud. Ambas solicitudes son enviadas al mismo tiempo, y de ésta ultima no obtenemos respuesta de la obtención de los datos.



BDCAP
BASE DE DATOS CLÍNICOS DE ATENCIÓN PRIMARIA
SISTEMA NACIONAL DE SALUD

SOLICITUD DE EXTRACCIÓN DE MICRODATOS ESTÁNDAR

En la actualidad están disponibles microdatos con las características sociodemográficas de las personas, los problemas de salud y las interconsultas. El resto de variables de la BDCAP se irán incorporando conforme se amplíe y consolide la información.

El uso de los microdatos estándar de la BDCAP ha sido sometido a dictamen por el Comité de Ética de Investigación del Instituto de Salud Carlos III, habiendo obtenido informe favorable, número de registro **CEI PI 19_2018**.

Por favor, cumplimente los datos que figuran a continuación y remita la solicitud por cualquiera de las vías siguientes:

- **Correo electrónico**: solicitudesbdcap@msssi.es
- Fax: 915 96 41 11
- Correo postal:

Subdirección General de Información Sanitaria y Evaluación
Sistemas de Información de Atención Primaria - BDCAP
Ministerio de Sanidad, Servicios Sociales e Igualdad
Paseo del Prado 18-20
28071 Madrid. España

En aplicación del artículo 5 de la Ley Orgánica 15/1999 de Protección de Datos de Carácter Personal, le comunicamos que los datos personales aportados servirán, exclusivamente, para conocer los usos de los productos y servicios del Sistema de Información del Sistema Nacional de Salud. En cualquier caso, podrá ejercer su derecho de oposición, acceso, rectificación y cancelación de sus datos, en los términos descritos en la ciada Ley, a través de los mismos circuitos establecidos para las solicitudes.

Microdatos estándar BDCAP

MICRODATOS QUE SOLICITA

Indique el año de estudio que se desea.....

Por favor, especifique qué variables necesita de entre las siguientes opciones:
Señale con una **X**

❖ DATOS ADMINISTRATIVOS

- ☐ Año
- ☐ Comunidad Autónoma
- ☐ Tramo del municipio cabecera de la zona básica
- ☐ Identificador interno anonimizado de la persona (ID)
- ☐ Sexo
- ☐ Año de nacimiento
- ☐ Edad-Grandes grupos
- ☐ Edad-Decenios
- ☐ Edad-Quinquenios
- ☐ Edad-Año
- ☐ País de nacimiento_Capítulo
- ☐ País de nacimiento_Subregión
- ☐ Nivel de renta
- ☐ Situación laboral
- ☐ Atendido*
- ☐ Peso**

*El flag "atendido" indica que la persona ha utilizado, en el año de que se trate, los servicios de Atención Primaria.
**El peso asociado a cada persona constituye un factor de ponderación de los usuarios de la muestra, que permite elevar los resultados a la población (la asignada a atención primaria).

Figura 5.1: Solicitud BDCAP

Sobre la Base de Datos Clínicos de Atención Primaria-BDCAP se solicitan todas las variables (figura 5.1) para poder tener una elección más amplia a la hora del tratamiento de los datos y el año 2016 ya que es el más reciente del que se tiene datos. Este es un hecho que nos llama la

atención puesto que no pensábamos que hubiera un atraso de 3 años en el sistema de información o al menos en los datos que los ciudadanos pueden obtener para realizar estudios.

Al enviar la solicitud recibimos un mensaje de confirmación de dicha solicitud, y a poco menos de una semana recibimos la confirmación del acceso a los datos mediante el enlace a un servicio en la nube al cual tendremos acceso solamente durante unos días. En estos días debemos haber descargado los datos, los cuales vienen acompañados de un fichero en el que se nos explica en funcionamiento del servicio que los almacena, una descripción de la configuración de los ficheros txt en los cuales se incluye la información y las variables solicitadas. Dicha explicación viene en forma de documento Excel, en el que se nos indica que obtendremos tres documentos al realizar la descompresión del archivo descargado. Observamos la explicación de las variables así como el significado de cada una de las columnas de los archivos, también queda incluido el significado de multitud de códigos que son usados con la intención de compactar la información para ofrecer un mejor tratamiento de la misma.

5.2 Extracción de muestras

Tras la recepción de los archivos contenedores de los datos necesitados se procede a su exploración para concretar qué información será valiosa para analizar. Obtenemos tres archivos cuyo almacenamiento es distinto, lo que repercute directamente en el tamaño del archivo de texto. Los archivos, como observamos en la figura 5.3 son Interconsultas, Personas y Problemas. A lo largo del proyecto los tres archivos principales comentados son tratados como INTERCONSULTAS.txt, PROBLEMAS.txt y PERSONAS.txt para la simplificación de la tarea de lectura sobre los mismos.

Figura 5.2: Archivos

En cada uno de ellos quedan almacenadas variables distintas que nos aportan diversa información. El fichero Problemas es el más extenso, debido a que recoge todas las visitas de atención primaria, realizando una clasificación de diversos aspectos sobre las visitas. En concreto dicho archivo está compuesto por 33760415 lineas cuyo formato es el que se puede observar en la figura 5.2. Tras exportarlos a Excel somos capaces de obtener una visualización mejorada, en la que podemos transformar la separación por puntos y comas a una tabla en la que cada columna tiene su cabecera indicando el contenido.

Al darse cuenta de que algunos archivos están compuestos por decenas de millones de líneas y que nuestros recursos a nivel computacional son bastante limitados decidimos tomar muestras representativas de la población, realizando filtros sobre los datos para poder reducir aún más el campo y poder extraer el máximo partido de los datos. Esto nos facilitará realizar una simulación del análisis de datos en el ámbito sanitario a pequeña escala, demostrando el gran potencial de la explotación masiva de datos.

En la extracción de las muestras observamos que el componente común en los tres archivos, el identificador anonimizado del paciente. Este se compone por 20 dígitos entre los que quedan comprendidos números y el signo '-', para abarcar un mayor rango. Observamos que hay iden-

```
Año;Identificador del paciente recodificado;País - Capítulo;País -
Subregión;Comunidad Autónoma;Tramo municipio Agrupado;Sexo;Edad;Año de
nacimiento;Año de fallecimiento;Nivel de renta;Situación laboral;Atendido;Peso
2016;13672321191956581260;ESP;ESP;15;1;2;75;1941;;C;2;1;7,992559
2016;1367232110-1130426484;ESP;ESP;15;1;1;63;1953;;C;2;1;7,017634
2016;190395557849696705 2;ESP;ESP;15;4;2;57;1959;;B;1;1;6,730558
2016;-13172698991671372172;ESP;ESP;15;1;2;36;1980;;C;1;1;8,026749
2016;292933224-1902161012;ESP;ESP;15;1;2;55;1961;;C;2;1;6,730558
2016;-7816851341486864012;ESP;ESP;15;1;2;83;1933;;C;2;1;8,095169
2016;2137128689-2003840789;NS;OTRO;12;2;1;16;2000;;C;4;1;9,869990
2016;-98170715444019080;ESP;ESP;15;4;2;77;1939;;C;2;1;7,992559
2016;2256380032124404872;ESP;ESP;15;1;2;73;1943;;B;4;1;9,030808
2016;-1921960338-1835022712;ESP;ESP;15;4;2;78;1938;;B;5;1;7,992559
2016;-1899733027-323500605;NS;OTRO;12;1;2;16;2000;;C;4;0;10,060248
2016;-1183762823-1918908789;ESP;ESP;15;1;1;80;1936;;C;2;1;8,135000
2016;2037462646-207632757;ESP;ESP;15;2;2;52;1964;;C;1;1;6,054311
2016;1618136817951903915;NS;OTRO;12;2;1;16;2000;;C;4;1;9,869990
2016;-3776126891520418698;ESP;ESP;15;4;2;36;1980;;C;1;1;8,026749
2016;-914917770-1818233718;ESP;ESP;15;1;1;25;1991;;C;1;1;7,309681
2016;-5119450941033874829;ESP;ESP;15;4;1;49;1967;;C;1;1;7,050417
2016;-1586301314530569357;ESP;ESP;15;4;1;78;1938;;B;2;1;8,342180
2016;696129146-1382039668;ESP;ESP;15;2;1;39;1977;;B;1;1;7,863655
2016;21438369071051033095;ESP;ESP;15;4;2;57;1959;;B;4;1;6,730558
2016;1587011662 1805068;ESP;ESP;15;2;1;36;1980;;C;1;1;7,863655
2016;-915507604446643340;ESP;ESP;15;2;2;47;1969;;B;1;1;6,351368
2016;2932301931184865676;ESP;ESP;15;1;1;24;1992;;B;4;1;6,500275
2016;17007436659503142 55;ESP;ESP;15;1;2;83;1933;2016;C;4;1;8,095169
2016;-679008201150975666;ESP;ESP;15;2;2;59;1957;;B;1;1;6,730558
2016;-1796783163-477940531;ESP;ESP;15;4;1;55;1961;;C;3;1;6,472745
2016;-1812731273-1952809293;ESP;ESP;15;1;1;56;1960;;C;1;1;6,472745
2016;-1813206409-1952807757;ESP;ESP;15;1;1;80;1936;;C;2;1;8,135000
2016;19448899641419412659;ESP;ESP;15;1;2;76;1940;;C;2;1;7,992559
2016;10065373332114246 91;EXT;AMElat;15;4;1;36;1980;;C;1;1;7,583333
2016;-1342387598-1768290384;ESP;ESP;15;4;2;66;1950;;C;2;1;8,938108
2016;-4048436581807095093;ESP;ESP;15;1;2;48;1968;;C;1;1;6,351368
2016;1677723630-861134619;ESP;ESP;15;1;1;73;1943;;C;2;1;8,784182
2016;-1208997257345661622;ESP;ESP;15;4;1;71;1945;;C;2;1;8,784182
2016;1610402426-308672585;EXT;EURue;15;1;2;66;1950;;C;2;1;9,250000
```

Figura 5.3: Problemas

tificadores que no se componen de 20 dígitos, si no que constan de 19 o incluso 17 dígitos. Esto supone un problema ante la veracidad y la consistencia de la información que hemos obtenido y por ello debemos examinar en profundidad el resto de los datos, y realizar un filtro sobre los datos, desechando aquellos datos que no cumplen las especificaciones que se nos indican en las notas explicativas adjuntadas con los datos.

Para la extracción de las muestras aplicamos los siguientes filtros sobre la muestra personas:

- Sexo = 2 (Hombre)
- Edad >= 40 años
- CCAA = 10 (Comunidad Valenciana)
- País = ESP (España)
- Longitud de ID = 20 dígitos

Dichos filtros se ejecutan a través de un *script* de Python con el objetivo de obtener una muestra de 1000 personas. Se selecciona dicho número de personas tras haber realizado pruebas con diferentes muestras, decidiendo finalmente que éste número es acorde a nuestros propósitos y también nos permite trabajar sobre las muestras con agilidad. Tras este paso procedemos a cruzar los archivos Interconsultas y la muestra de Personas con el objetivo de almacenar todas las interconsultas relacionadas con los individuos almacenados en nuestra muestra.

Para ello programaremos otro *script* en Python, y haremos uso de la libreria *multiprocessing* con el objetivo de resolver la tarea con más rapidez. Dicha librería nos permite utilizar diversos

procesos y buscar todos los *matches* de ID en interconsultas para posteriormente almacenarlos. Lo que hacemos es buscar dónde estarán los *matches* a lo largo de las millones de lineas que tiene el archivo de Interconsultas, y posteriormente analizamos dicho rango donde se encuentran las coincidencias y seleccionamos los valores coincidentes para almacenarlos en una lista que será escita en nuestro fichero 'interconsultas_data.txt' . (Apéndice A.1)

La última tarea que realizamos en nuestra extracción de muestras es el análisis del archivo donde se encuentran todos los problemas de atención primaria (PROBLEMAS.txt). Como en los *scripts* anteriores utilizaremos la libreria *multiprocessing* para facilitar la búsqueda de las coincidencias de los ID almacenados en el archivo 'personas_data.txt' y el archivo cedido 'PROBLEMAS.txt'. En este *script* también hemos analizado dónde se encuentran los *matches* a lo largo de todo el archivo para luego ejecutar la salvaguarda de los mismos a través del rango de registros que hemos obtenido.

5.3 Herramienta

Una vez creadas las muestras con las que trabajaremos, se propone la implementación de una herramienta a modo de ejemplo de la utilidad del análisis de datos en una consulta médica. La herramienta que desarrollamos viene creada nuevamente a través de un *script* de Python. Dicha herramienta es desarrollada con el objetivo de ser usada por los profesionales médicos, pero desarrollaremos posteriormente un prototipo de diseño en el que dicha interacción del profesional con la herramienta será más cómoda.

Cuando se lanza la herramienta se le pregunta al personal sanitario por el identificador del paciente, el cual nosotros tratamos como un id de manera anonimizada pero que en un caso real sería el DNI del paciente. Dicha información podría también ser extraída de forma más cómoda a través de la lectura de la tarjeta sanitaria sin la necesidad de tener que teclear el número de identificación de cada paciente.

A través del DNI, nuestra herramienta busca todos los datos relacionados con el paciente que están almacenados en las muestras ya preparadas, y a partir de estos datos se produce un análisis de los datos para la extracción de información y conocimiento que podríamos dividir en tres partes:

- Observación de datos generales.
- Análisis y creación de conocimiento.
- Visualización de datos y gráficas.

En la primera parte se muestra un contenido general de los episodios sanitarios del paciente a lo largo del año de estudio, 2016. Ésta información sirve de precedente para el facultativo al recibir a dicho paciente en su consulta, aportando información que podría ser extraída del Historial Clínico Electrónico, pero que necesitaría un mayor tiempo de lectura y atención.

Dichos datos se agrupan por capítulos, los cuales recogen los diferentes problemas de salud relacionados por un conjunto de órganos o por temática (figura 5.4). Para ello hemos creado anteriormente listas con los diferentes Códigos de Capítulo BDCAP que existen y otra con los Capítulos BDCAP, las cuales nos servirán para usarlas como un diccionario que con ayuda de la función *index* nos permitirá obtener el nombre del capítulo teniendo sólo su código, que en éste caso es una letra en mayúsculas. (Apéndice A.2)

```
Paciente encontrado, problemas registrados en 2016:

A. Problemas generales e inespecificos : 4
F. Ojo y anejos : 2
K. Aparato circulatorio : 7
L. Aparato locomotor : 3
P. Problemas psicologicos : 1
R. Aparato respiratorio : 2
S. Piel y faneras : 2
T. Aparato endocrino, metabolismo y nutricion : 2
U. Aparato urinario : 2
```

Figura 5.4: Problemas por Capítulo BDCAP

En el segundo apartado analizamos las fechas de apertura y de cierre de los problemas que nos figuran sobre el paciente, así como también de los mismos problemas que tenemos almacenados en las muestras, con el fin de obtener una media estadística por capítulos.

La fecha de apertura nos proporciona la fecha en la que dicho problema se abre debido a la asistencia de un paciente a un centro sanitario en el que se le diagnóstica dicho problema. Pueden abrirse varios problemas al día y también varios del mismo tipo, como por ejemplo un paciente puede tener diversas artrosis en las extremidades. En algunos de los casos las fecha de apertura coincide con la de cierre, la cual establece el fin de la enfermedad debido a que dicho paciente ha sido curado o bien se trata de problemas como por ejemplo embarazos, los cuales tienen la misma fecha de apertura y de cierre.

```
Fechas de apertura : (y/n) y

Capitulo A. Problemas generales e inespecificos. //  Abierto el 23/06/2011
 - Problema recurrente : 4 veces.  ALERTA.(Posible enfermedad crónica)
 - Tiempo medio de recuperación: 1.477 años.
 - Tiempo medio transcurrido para el paciente : 3.7 años

Capitulo F. Ojo y anejos. //  Abierto el 24/04/2008
 - Problema recurrente : 0 veces.
 - Tiempo medio de recuperación: 3.068 años.
 - Tiempo medio transcurrido para el paciente : 8.5 años

Capitulo K. Aparato circulatorio. //  Abierto el 07/12/2007
 - Problema recurrente : 3 veces.  ALERTA.(Posible enfermedad crónica)
 - Tiempo medio de recuperación: 0.0 años. No existen datos disponibles de
pacientes curados en 2016
 - Tiempo medio transcurrido para el paciente : 5.3 años
```

Figura 5.5: Generación de información

A través de los datos que disponemos de nuestra muestra creamos medias estadísticas de los tiempos de recuperación de cada uno de los capítulos, tanto para los pacientes como para todos los problemas de nuestra muestra, con el objetivo de poder comparar los tiempos de recuperación del paciente y el tiempo medio de recuperación de todos los pacientes. Ésto nos permite saber si el paciente está teniendo una correcta evolución de las enfermedades agrupadas en capítulos, lo que permite al profesional sanitario explorar nuevos tipos de fármacos o tratamientos que se ajusten más a las necesidades del paciente, así como realizar observaciones en mayor profundidad para analizar cuales son las causas de dicho desajuste en comparación con la media de los otros pacientes. Para extraer dicha información transformamos las fechas almacenas como *strings*, almacenadas en el archivo 'problemas_data.txt', a objetos de tipo *date* para poder operar con ellos y realizar restas entre fechas (apéndice A.3). Posteriormente establecemos una fecha de referencia

para analizar problemas que pueden estar relacionados y ser recurrentes. Esta fecha se guarda en la variable b1.

A pesar de que nosotros establecemos la fecha a nuestro criterio, el correcto uso sería tener un estudio en el que pudiéramos extraer la información necesaria para concluir a partir de ella cuantas veces y en cuanto tiempo un problema puede tratarse como una enfermedad crónica. Dichos estudios no han sido encontrados de forma pública y desconocemos si actualmente se está realizando algún tipo de estudio similar pero consideramos que puede tratarse de una característica que sirva de gran ayuda para la mejora de diagnósticos en servicios asistenciales sanitarios.

Cuando nuestro script detecta que hay más de 3 problemas que han sido recurrentes en los últimos X años se genera un mensaje de alerta para que el profesional sanitario sepa que puede haber un problema relativo a dicho Capítulo BDACP y evalúe los datos según su criterio médico (apéndice A.4). Se muestra una alerta por cada Capítulo BDCAP del que se cumple la condición comentada anteriormente. El objetivo de dicha alerta es que se le trate al paciente con una mayor agilidad y eficacia sobre el problema relativo. Para la implementación de la alerta utilizamos la librería *pymsgbox*, la cual muestra la alerta en el centro de la pantalla y debe ser atendida para poder seguir visualizando el resto de información. De éste modo nos aseguramos de que el profesional sanitario visualice dicha información de vital importancia.

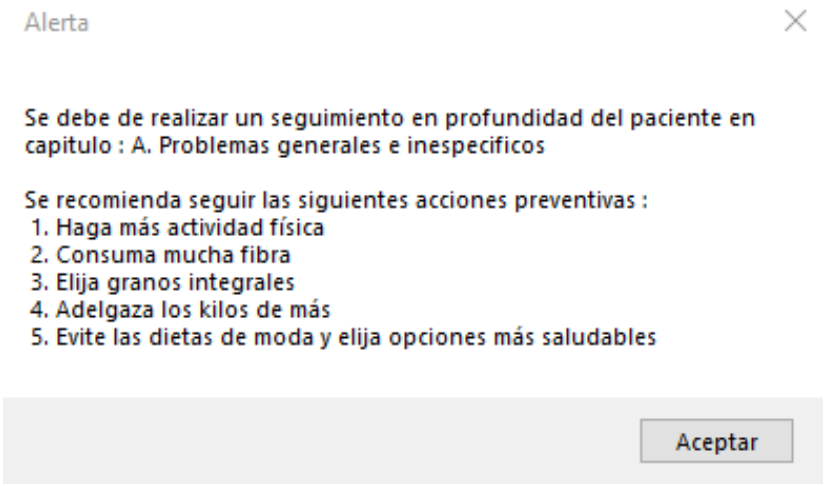


Figura 5.6: Generación de información

Tras la aparición de la alerta debido a la 'detección' de enfermedad crónica aparecerá un listado con medidas preventivas para el paciente que puedan ayudar a cambiar sus hábitos diarios y tomar un estilo de vida más saludable que prevenga y disminuya los episodios de dicha enfermedad. Actualmente la lista de acciones preventivas que se muestran es simplemente una lista orientativa de prevención ante la diabetes, ésta nos sirve para visualizar y comprender cómo el sistema actuaría ante la detección de una enfermedad crónica. (Figura 5.6)

Como última parte de nuestra herramienta se ofrece la posibilidad de observar en un gráfico la comparativa de tiempo entre los Capítulos BDCAP del paciente en cuestión y del resto de pacientes (figura 5.7). Al mostrar este gráfico hemos tenido que eliminar el Capítulo BDCAP "K. Aparato circulatorio"del cual no se tienen fechas de cierre en nuestra muestra y por lo que su inclusión en el gráfico no tiene ninguna finalidad ante la imposibilidad de establecer una comparativa entre ambos tiempos de recuperación.

Para la implementación de los gráficos utilizamos la librería *matplotlib*.

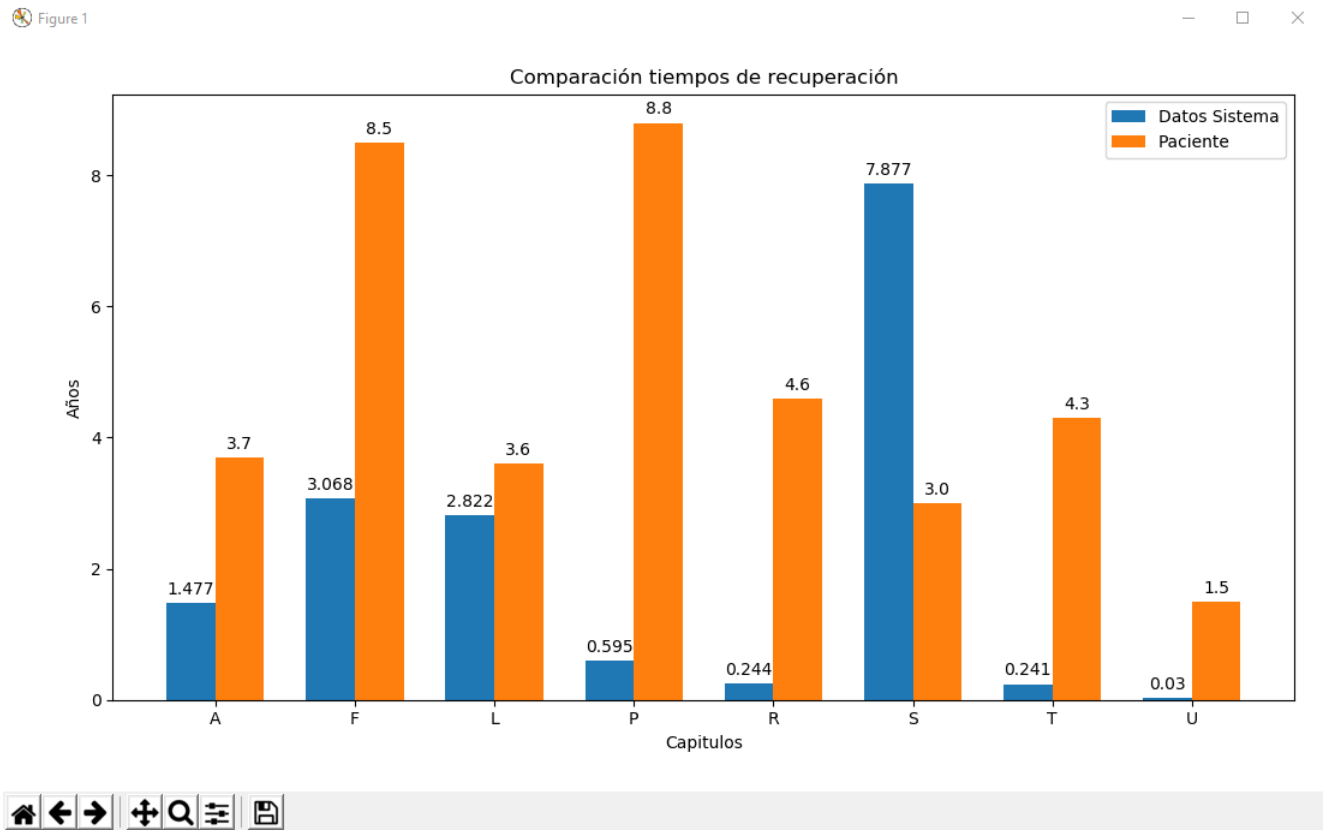


Figura 5.7: Gráfico tiempos recuperación

Como observamos en el apéndice A.5 analizamos en que rango de edad se encuentra el paciente para posteriormente concluir que puede que presente algún episodio de la enfermedad más común de su rango si no lo ha presentado ya. Éste es un ejemplo básico de análisis con objetivo predictivo que puede alertar al profesional médico de posibles enfermedades que el paciente pueda presentar.

```
Gráfico : (y/n) y

Primera enfermedad más común de los 40 a los 59 años : L. Aparato locomotor
Segunda enfermedad más común de los 40 a los 59 años : S. Piel y faneras
Tercera enfermedad más común de los 40 a los 59 años : P. Problemas psicologicos

Primera enfermedad más común de los 60 a los 79 años : L. Aparato locomotor
Segunda enfermedad más común de los 60 a los 79 años : K. Aparato circulatorio
Tercera enfermedad más común de los 60 a los 79 años : P. Problemas psicologicos

Primera enfermedad más común a partir de los 80 años : K. Aparato circulatorio
Segunda enfermedad más común a partir de los 80 años : L. Aparato locomotor
Tercera enfermedad más común a partir de los 80 años : T. Aparato endocrino, metabolismo
y nutricion

Edad del paciente : 94

El paciente se encuentra en un rango de edad en el que podría desarrollar algún problema
relacionado con K. Aparato circulatorio
Observar el gráfico "Enfermadades más comunes" para más información
```

Figura 5.8: Enfermedades más comunes

Posteriormente estudiamos cuales son las tres enfermedades que más aparecen en cada rango de edad de la muestra con el objetivo de comparar como afectan a distintos grupos de edad (figura 5.8). Para ello almacenamos en una lista el número de veces que aparecen y posteriormente creamos tres variables para cada rango de edad en las que guardamos las tres enfermedades cuyo número de apariciones es el más alto. Se da el caso de que las enfermedades más comunes en cada rango son distintas. Después obtenemos el porcentaje de aparición frente al total de enfermedades por rango de edad, ya que en nuestra muestra existen mayores registros del rango que va desde los 40 hasta los 60 años de edad. Estos datos son los que nos servirán para desarrollar el gráfico de la figura 5.9. Los rangos de edad van desde los 40 a los 59, los 60 a los 79 y finalmente mayores de 80 años.

Como podemos observar cada rango de edad se caracteriza por unas enfermedades comunes en concreto, siendo el Capitulo BDCAP Aparato Locomotor el único que está presente en cada uno de los grupos.

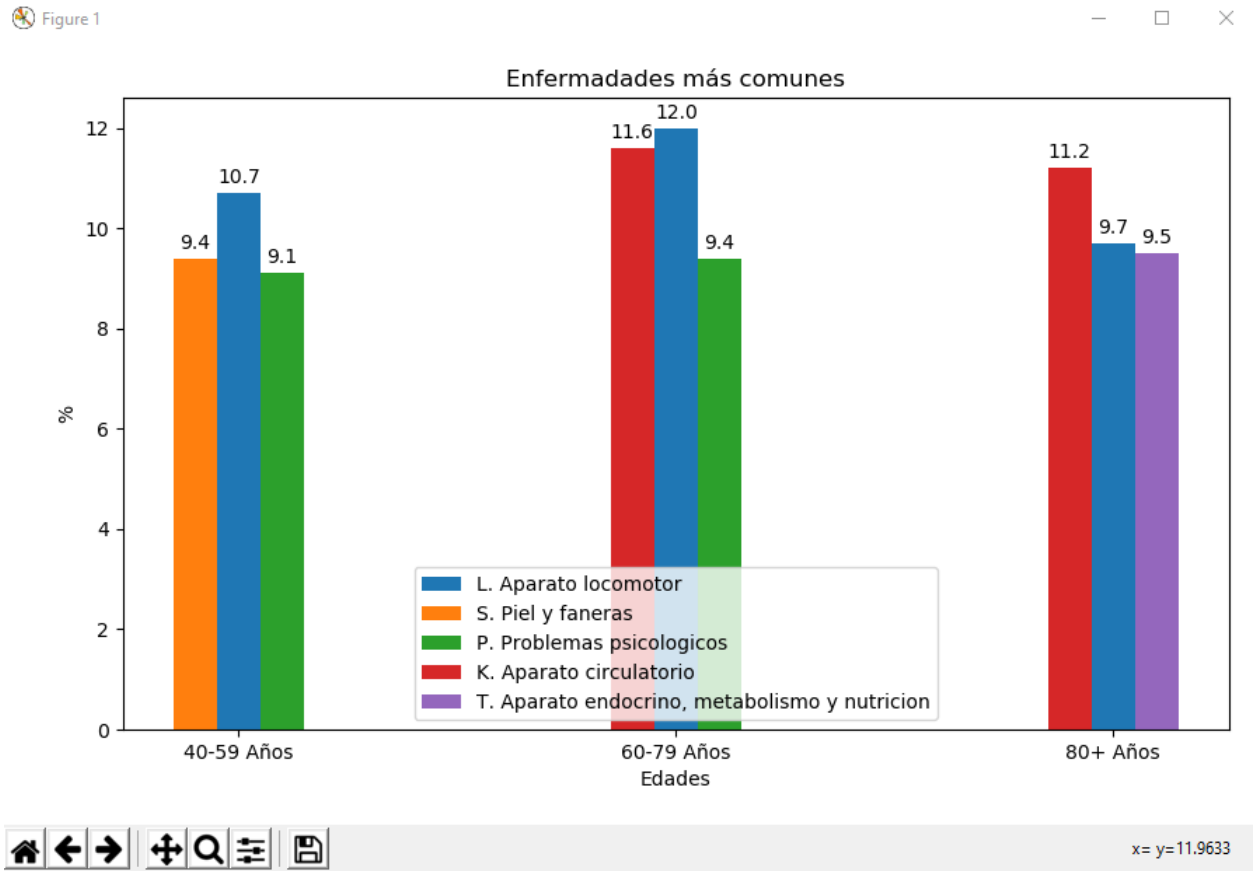


Figura 5.9: Gráfico enfermedades más comunes

Cabe señalar que para la toma de capturas sobre la información y gráficos extraídos por la herramienta no se ha utilizado un único paciente, si no que hemos hecho uso de varios identificadores para poder ofrecer una mejor representación de los objetivos.

5.4 Prototipo de una interfaz

A pesar de haber creado una herramienta funcional para mostrar el funcionamiento de nuestro modelo, dicha herramienta, en un entorno real, estaría recogida en una herramienta con una interfaz gráfica y no solo de comandos por terminal.

Por ello se propone realizar un prototipo de una interfaz en la que insertar nuestra herramienta. De esta forma quedará un concepto mucho más claro de la utilización de un sistema basado en nuestro modelo de creación de conocimiento y toma de decisiones. Se hará uso de la herramienta Adobe XD, la cual es usada para el desarrollo profesional de interfaces en dispositivos digitales. Dicha plataforma nos permitirá realizar un prototipo avanzado de como podría ser la interfaz final del sistema.

Además añadiremos elementos en la interfaz que harán referencia a funciones, que aunque no están desarrolladas en el apartado anterior debido a la imposibilidad de dicha tarea por causa de la naturaleza de los datos, ayudarán a comprender el funcionamiento de nuestro modelo en una consulta médica y la modernización de los servicios asistenciales que ello supone.

Figura 5.10: Pantalla Inicial

Como observamos en la figura 5.10 la pantalla inicial del prototipo que hemos diseñado el profesional sanitario, en este caso un médico tiene la opción de consultar los datos del paciente cuando éste llega a la consulta. Inmediatamente después de seleccionar dicha opción, y como ya se había introducido en nuestra herramienta, el médico deberá insertar el identificador del paciente (figura 5.11) para poder acceder a los datos del mismo. Como se comentó en la implementación de la herramienta, la inserción del número identificador del paciente a través de la lectura de la tarjeta sanitaria del paciente es una opción más cómoda y rápida.

Una vez validado el número identificador se cambiará de pantalla permitiendo al profesional sanitario la visualización de multitud de informaciones relevantes sobre el paciente. Por defecto, se realizará un análisis del año actual y se mostrará los problemas de dicho paciente a lo largo del trascurso del año. Esta opción en en la herramienta basada en un *script* de Python se mostraba también automáticamente al ingresar el identificador. Como opción adicional hemos añadido la opción de cambiar el año de estudio, lo que facilita al facultativo realizar consultas sobre los problemas de salud del paciente en años pasados.

El resto de informaciones como fechas de apertura y gráficos se mostrarán también de forma automática, con la opción de esconder dichas informaciones y visualizaciones de gráficos. Dicha información se sitúa en el centro de la pantalla dónde podemos realizar *scroll down* para acceder

Figura 5.11: Identificador de paciente

a toda la información. Además se ha intentado mantener al máximo el formato de la información ya presentada en la herramienta, para poder asociar mejor la inclusión de dicha herramienta en el prototipo del sistema de información que un médico puede usar en su consulta. Podemos observar que en la parte izquierda se muestran algunos datos sobre el paciente para facilitar la interacción humana con el mismo por parte del facultativo. Como en la herramienta desarrollada, también hemos añadido la aparición de mensajes de alerta y de predicciones para aportar información relevante al médico que pueda ayudar en el diagnóstico. Ésta información puede verse reflejada en la parte inferior izquierda de las interfaces del prototipo desarrollado, que a diferencia de la herramienta en Python, las alertas no irrumpen la tarea del profesional sanitario y quedan agrupadas en una zona visible de la pantalla en la que la información es estática a diferencia del bloque central.

Algunas opciones no implementadas en nuestra herramienta, pero propias de un sistema de información de consulta primaria son también mostradas. En este caso la opción 'Historial Clínico Electrónico' nos permite ver en profundidad toda la información referente al paciente, dónde podemos encontrar los tratamientos actuales y datos médicos más detallados sobre consultas pasadas. A su vez se ofrece la opción de 'Importar datos médicos' una acción que se considera necesaria ante la creación masiva de información relevante para la salud por dispositivos móviles o wearables, así como también los tradicionales marcapasos o inyectores de insulina automáticos. Toda ésta información podría ser de gran utilidad, como ya vimos en nuestro modelo, a la hora de realizar posibles diagnósticos y contrastar tratamientos.

Como novedad, hemos introducido el apartado de diagnosis en el que se pueden observar algunos conceptos y técnicas de *big data* en salud de las cuales ya se ha hablado en capítulos anteriores. A través de introducción en el sistema de los síntomas del paciente en la opción 'Sintomas del Paciente', como se suele hacer en el HCE, el sistema es capaz de contrastar a través de toda la información de nuestro *data lake* los diagnósticos ya realizados de pacientes cuyos síntomas han sido similares o los mismos. De esta forma se ofrece al profesional médico en el apartado inferior 'Posibles Diagnósticos' la estimación de un diagnóstico basado en los datos, el cual puede ser

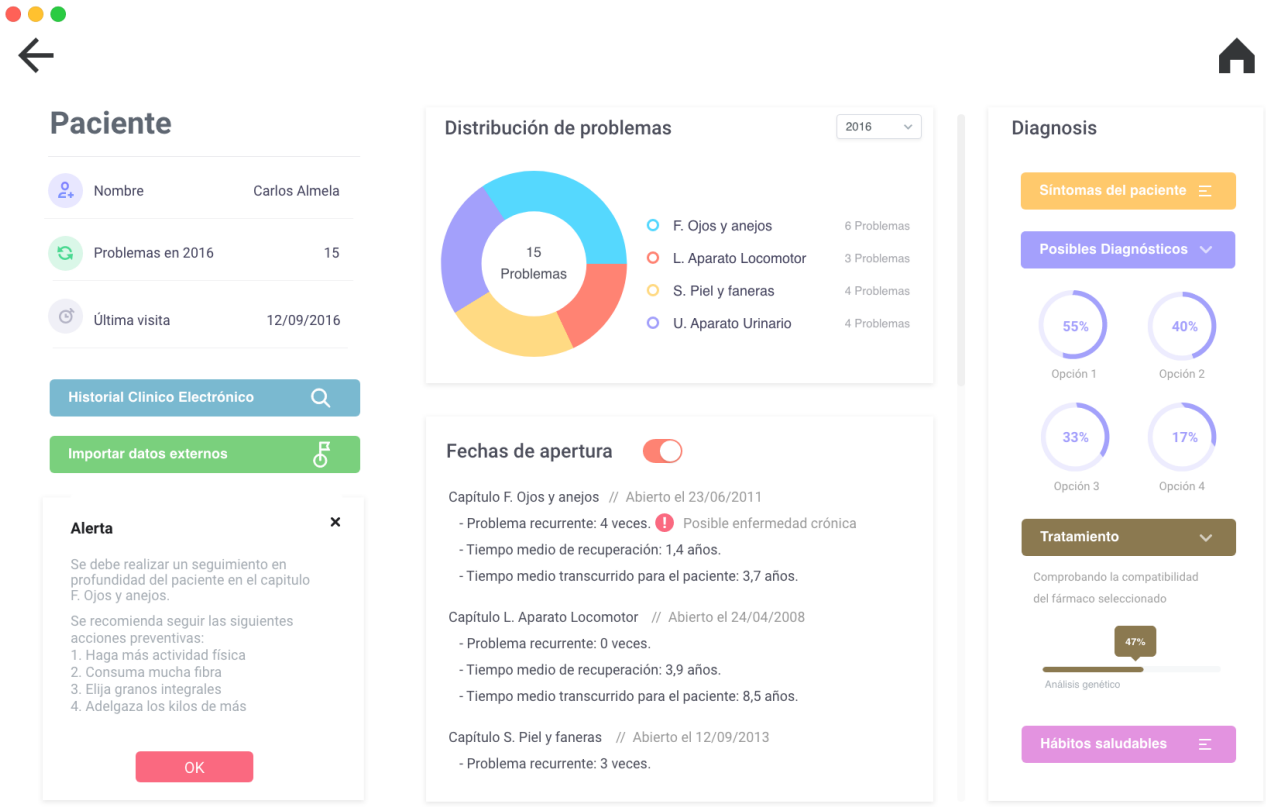


Figura 5.12: Información 1

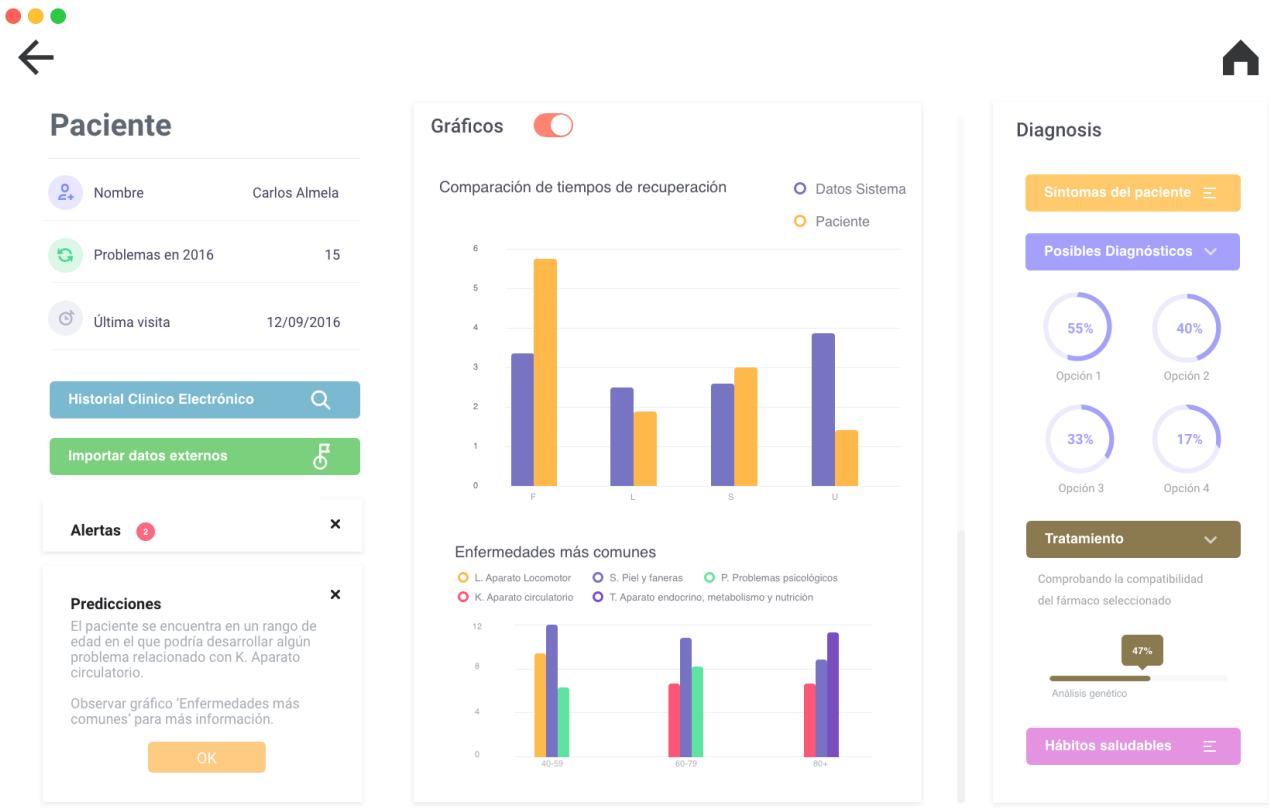


Figura 5.13: Información 2

decisivo en el diagnóstico final, mostrando por orden de probabilidad las opciones que son más previsibles de acertar con el diagnóstico del paciente.

Posteriormente el profesional médico mediante el sistema planteado en la interfaz es capaz de realizar un análisis de compatibilidad de tratamientos a través del apartado 'Tratamiento', en los que se tendrá en cuenta la información generada por nuestra herramienta, analizando así los factores genéticos y contrastándolos con tratamientos ofrecidos a pacientes con mismos diagnósticos y también los tratamientos que ha tenido a lo largo de los años con problemas relacionados, teniendo en cuenta si dicho problema es recurrente. Por último, tenemos de nuevo, un ejemplo de medicina preventiva bajo el apartado Hábitos Saludables', en el cual se ofrece al paciente una serie de pautas que seguir a parte del tratamiento ya asignado. A través de estas pautas el paciente podrá mejorar su calidad de vida respecto a la patología diagnosticada y potenciar a través de dichas medidas preventivas una curación más efectiva.

Dicho apartado, al igual que el bloque izquierdo de la interfaz, es estático en todo momento y solo varía cuando se introduce información para realizar la tarea de diagnosis.

5.4.1. Normas

Para la realización de las interfaces se han seguido algunas normas que se han visto en la asignatura Interfaces Persona Computador del grado. Éstas son las Leyes Gestalt (Ref ppt) y hacen referencia a las percepciones que que recibe el usuario ante una interfaz, dichas leyes son las siguientes:

- **Proximidad**: Objetos cercanos parece que forman grupos, en vez de una sucesión aleatoria de elementos.
- **Similitud**: Elementos de la misma forma o color parecen formar grupos.
- **Cierre**: Los huecos de una figura incompleta y áreas cerradas parece que forman un objeto.
- **Continuidad**: Elementos alineados parecen formar lineas.
- **Simetría**: Zonas delimitadas por bordes simétricos son considerados un todo.
- **Separación figura-fondo**: Cuando existen bordes o diferencias de color, así como textura o brillo, separamos un objeto y su fondo.

Dichas leyes, específicamente Proximidad, Similitud, Continuidad y Simetría pueden observarse en el bloque de diagnosis, en el que los elementos siguen un diseño unificado por apartados con colores llamativos que facilitan la atención de usuario. Dichos elementos conforman un grupo debido también a su forma y aportan la sensación de continuidad a lo largo del bloque, formando un flujo de información que viene marcado por los pasos de la realización del diagnóstico.

Podemos observar la ley de Separación figura-fondo en los gráficos representados en la figura 5.13, en la que las diferencias de color marcadas por las barras del gráfico marcan la distinción del objeto con el fondo.

Además se ha procurado seguir algunas de las reglas de Ben Shneiderman's en 'Designing the User Interface' y también Jakob Nielsen's en 'Ten Usability Heuristics' [**?**], entre las que encontramos la consistencia en el diseño de la interfaz, la cual se puede observar pues hemos seguido los mismos patrones para diseñarla en su totalidad, incluyendo colores llamativos que destacan los apartados a los que se puede acceder para realizar acciones. La inclusión de diálogos y comentarios informativos también queda recogida en las reglas de Scheiderman's, las cuales se ven

reflejadas en los diálogos de alerta y predicciones mostrados a la izquierda de la interfaz principal, así como de comentarios informativos relacionados con el porcentaje del análisis genético realizado o la posibilidad de introducir el identificador del paciente mediante el paso de la tarjeta sanitaria. Como queda incluido en la regla octava del manual de Nielsen se ha tratado de diseñar una interfaz minimalista y estética que recogiera toda la información importante sin mostrar información innecesaria.

5.5 Relación con el modelo

La herramienta creada, como ya se introdujo al principio del capítulo, tiene una relación directa y representa la utilidad del modelo diseñado en este proyecto. En la siguiente imagen observamos cómo en la representación del modelo sobre el que se construye la herramienta implementada se consigue introducir elementos representativos de nuestro modelo a excepción de la parte del*big data* por razones ya comentadas.

A diferencia de nuestro modelo no debemos realizar tareas de extracción de datos ya que estos son cedidos y ya han sido preprocesados con anterioridad, por lo que tras la recepción de los mismos las tareas ETL son bastante simples. A través de técnicas ETL y *data mining* conseguimos obtener nuestras muestras con las que trabajaremos en la implementación de una herramienta capaz de incluir análisis utilizados en *big data*. Estos análisis son como se pueden ver en la figura 5.14 *data mining*, *statistical analysis* y *data visualization*, todo ello a través de la ejecución de *scripts* en el lenguaje de programación Python.

Mediante de estas técnicas de análisis obtenemos métodos para la obtención de conocimiento en el área sanitaria como *medical alerts*, *evidence based on data* y algunas opciones más que están incluidas en la figura. Todo este conocimiento sirve de ayuda al profesional sanitario para la toma de decisiones y la realización de mejores diagnósticos.

Dichas mejoras sanitarias que aparecen en la imagen se observan en nuestro caso a través de la oferta de información y conocimiento como soporte, todo ello basado en una evidencia de datos provenientes del SNS. Los nuevos enfoques de la medicina quedan plasmados en la predicción de posibles episodios a través del estudio de las enfermedades más comunes en el rango de edad del paciente, así como la medicina preventiva mediante el cambio de los hábitos de salud del paciente propuestos al detectar posibles enfermedades crónicas.

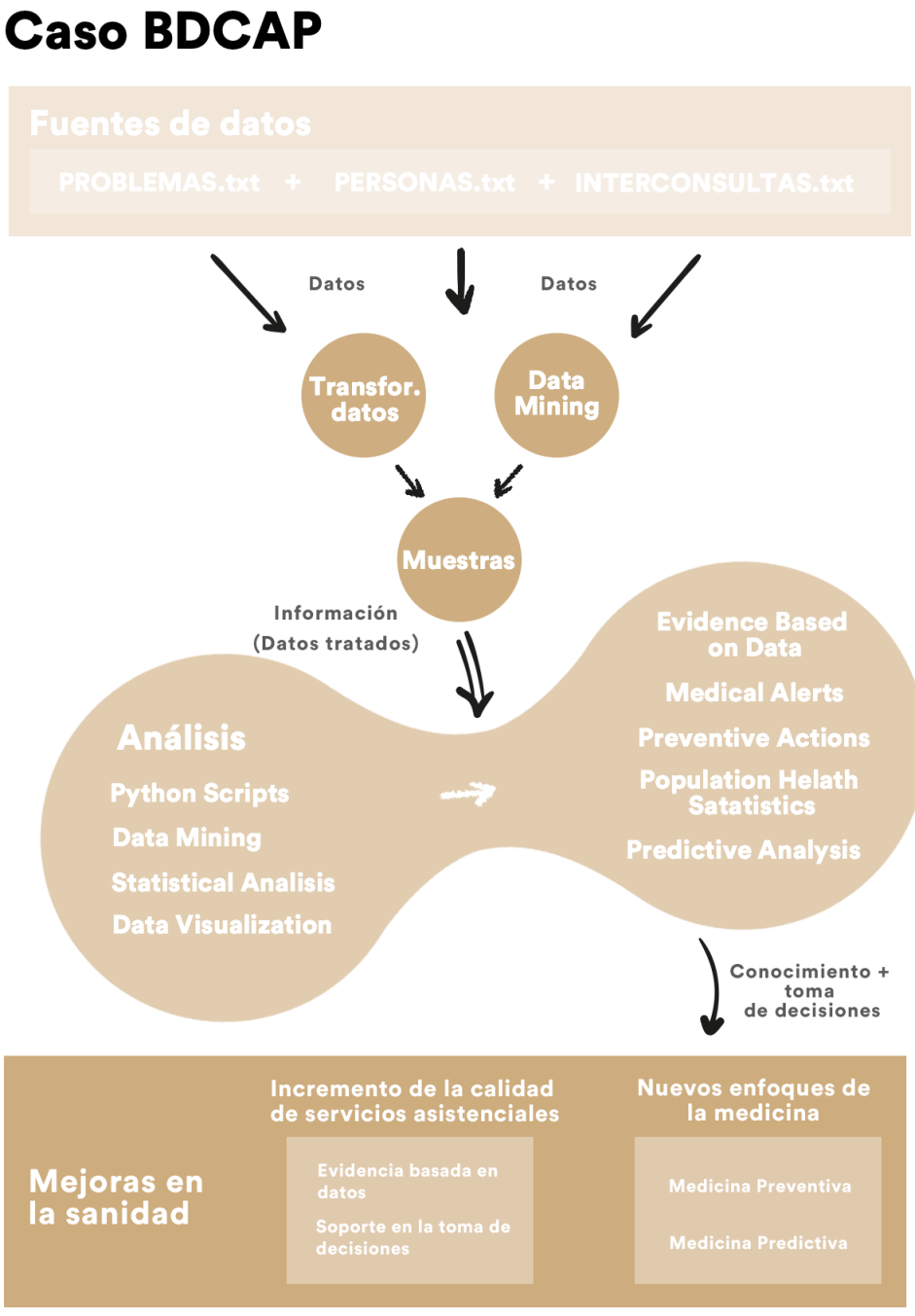


Figura 5.14: Modelo del Caso BDCAP. **Fuente:** Elaboración propia

CAPÍTULO 6

Conclusiones

Ante la inevitable aplicación de las técnicas de *big data* al ámbito sanitario y la observación de la falta de un modelo que englobe el funcionamiento de un sistema como tal en sanidad, se decide la realización de este proyecto para cubrir un nicho no cubierto, a pesar de existir algún estudio que se limita a la exposición y explicación de conceptos sobre el tema sin desarrollar la mecánica de dicha sinergia en profundidad.

Los grandes beneficios que aportan las técnicas de análisis masivos en la salud han quedado expuestos y argumentados, así como la importancia de la necesidad de una modernización de la sanidad que pasa por la inclusión del *big data*.

Se han cumplido los objetivos propuestos de plasmar la importancia de un sistema de información de las características explicadas en nuestro modelo para la mejora de la calidad asistencial, nuevos enfoques médicos y la eficiencia económica. Todo ello a través del diseño de un modelo que engloba el funcionamiento del sistema de información y el flujo de datos desde la extracción de los mismos hasta la toma de decisiones a partir del conocimiento generado por el sistema, y también de la implementación de una breve herramienta a modo de demostración sobre el valor de la aplicación de nuestro modelo en un entorno real como es una consulta médica.

Cabe indicar que la colaboración del Ministerio de Sanidad, Consumo y Bienestar Social ha sido clave para la obtención de datos con la que desarrollar la herramienta vista en el proyecto.

A lo largo del proyecto hemos observado la falta de estudios y publicaciones relacionado con el tema tratado en este proyecto, posiblemente debido a la complejidad del ámbito sanitario y su carácter público, ya que se han encontrado multitud de artículos anglosajones cuyos temas principales quedan comprendidos en el proyecto. Ha sido imposible encontrar estudios clínicos que tengan en cuenta la inclusión de un sistema de información, y ello a mermado la creación del modelo en algunos aspectos, así como la implementación de la herramienta correspondiente.

Algunos errores cometidos a lo largo del proyecto han estado relacionados ante la suposición de la estructura del proyecto antes de la puesta en marcha del mismo, dependiendo estrechamente de los datos que han sido cedidos para el Caso BDCAP. Éste, al principio del proyecto, se planteó como elemento principal del proyecto. Por ello es necesaria una planificación marcada ante la premisa de que dichas tareas podrán ser llevadas a cabo en su totalidad.

Tras la realización del proyecto se ha obtenido un creciente interés en las tecnologías y sistemas de información aplicados a la sanidad, cuyo objetivo es la mejora del bienestar de los pacientes. Se ha incrementado los conocimientos base sobre las tecnologías y conceptos que abarcan el *big data*, teniendo que realizar un estudio intenso ante la existencia de un gran volumen de fuentes en referencia al *big data*, así como multitud de soluciones y configuraciones dependiendo de las necesidades de cada sistema de información. Se han adquirido multitud de conocimientos sobre

dichas plataformas que conforman las soluciones *big data*, despertando un especial interés por la creación de conocimiento que nos proporciona los análisis de información procesada por ecosistemas de administración de datos como Hadoop.

Dicho trabajo abre las puertas a la motivación de la realización del Máster de Gestión de Información de la escuela, planteándose éste proyecto como un posible comienzo del desarrollo de un proyecto de final de máster o paso previo de la inserción al mundo laboral del análisis y gestión de los datos.

CAPÍTULO 7

Futuros trabajos

En el planteamiento inicial del proyecto se deseaba implementar una herramienta a partir de la cual poder extraer diversas informaciones tras el tratamiento de los datos, con la intención de descubrir patrones y tendencias en los datos que nos permitiese proporcionar conocimiento para la prevención y detección de enfermedades. El desarrollo de la herramienta suponía en tiempo un gran exceso de las horas curriculares del proyecto, y por ello se plantea el uso de los datos obtenidos por el Ministerio de Sanidad, Consumo y Bienestar Social como un caso de aplicación de algunas pinceladas que podrían plasmarse en la creación de una futura herramienta. De este modo se ha decidido profundizar y ampliar el diseño del modelo en el que se basaría la herramienta comentada.

Debido al gran interés que me surge como alumno de la escuela por el análisis de los datos y el mundo de la tecnología en la sanidad y la positiva repercusión de la unión de ambos conceptos. A lo largo del proyecto se plantea la idea de la implementación de dicha herramienta para la concepción de un sistema real en el que nuestro modelo queda plasmado y puede ser utilizado en un ámbito sanitario. Este es un plan que se deja plasmado debido al gran aporte que podría hacer al ámbito sanitario, ofreciéndose como una herramienta de modernización de la sanidad y de ayuda de toma de decisiones en los diagnósticos de los profesionales médicos. Es probable que en un futuro no muy lejano algún plan parecido llegue a ser integrado en sanidad debido a la inevitable aplicación del *big data* en casi todos los campos de la sociedad, presentándose en la sanidad un especial interés debido a su relevancia en el bienestar social.

Este plan de propuesta que se realiza queda suspendido en el aire para futuros trabajos que puedan ser realizados a través de proyectos de Máster. Se deja abierta la posible modificación del modelo, ya que a pesar de haberse usado tecnologías y conceptos actuales para la elaboración del proyecto, puede que por el entorno cambiante en el ámbito de las tecnologías y sistemas de la información, hayan surgido nuevos estándares, herramientas o técnicas que supongan una mejora sustancial del modelo de creación de conocimiento y toma de decisiones que se plantea.

Como iniciativas futuras han quedado planteadas el deseo de implementar una inteligencia artificial para mejorar el diagnóstico en imágenes médicas así como el servicio telemático asistencial. La aplicación para el control de los datos y la prevención de visitas evitables en la sanidad se esboza como un futuro más cercano y tangible. A pesar de ello y como se comenta, ya hemos visto esbozos de las mismas en algunos sistemas de EEUU.

Queda abierto también el planteamiento de un concepto de ámbito sanitario en el que se trate en profundidad el cambio que sufrirían las relaciones médico-paciente así como el empoderamiento del paciente debido a los grandes volúmenes de información que dispone. Se aconseja si se va a realizar dicha tarea en el futuro, trabajar minuciosamente el campo del *big data* y las herramientas que este concepto comprende, ya que al tratarse de un concepto que se ha moder-

nizado masivamente existen multitud de contenidos de consulta que pueden aportar puntos de vista distintos, haciendo tediosa la investigación sobre el mismo causada por la inexistencia de un estándar de lo que éste significa y la configuración del mismo como herramienta tecnológica.

Así mismo supone un impedimento la falta de estudios sanitarios sobre la implementación de la tecnología existente y sobre técnicas de análisis de datos. Ello supone un gran campo de estudio que queda abierto para una mayor exploración del mismo y no nos permite avanzar tanto como se quisiera, retrasando el desarrollo de aplicaciones o herramientas para el sector sanitario que puedan suponer una modernización de la sanidad.

Bibliografía

[1] Amir Gandomi, Murtaza Haider. International Journal of Information Management. *Beyond the hype: Big data concepts, methods, and analytics*, December 2015.

[2] Philip Russom. TWDI Best Practices Report. *Big Data Analytics*, fourth quarter, 2011.

[3] Paul C. Zikopoulos, Chris Eaton, Dirk deRoos, Thomas Deutsch, George Lapis. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. The McGraw-Hill Companies, 2012.

[4] The Four V's of Big Data IBM Infographics Animations https://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg.

[5] Big Data: Qué es y por qué es importante SAS https://www.sas.com/es_es/insights/big-data/what-is-big-data.html.

[6] An Oracle White Paper. *Oracle: Big Data for the Enterprise*, June 2013.

[7] Kapow Software Infographic https://www.columnfivemedia.com/work-items/infographic-intelligence-by-variety.

[8] Difference between Structured, Semi-structured and Unstructured data. https://www.geeksforgeeks.org/difference-between-structured-semi-structured-and-unstructured-data/.

[9] Ralph Kimball, Joe Caserta. *The Data Warehouse ETL Toolkit Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Wiley Publishing, 2004

[10] Data warehouse overview https://en.wikipedia.org/wiki/Data_warehouse#/media/File:Data_warehouse_overview.JPG.

[11] Bill Schmarzo. *Big Data: Understanding How Data Powers Big Business*. John Wiley Sons, 2013.

[12] Hrushikesha Mohanty, Prachet Bhuyan, Deepak Chenthati. *Big Data: A Primer*. Springer India 2015, Studies in Big Data.

[13] Jeffrey Dean and Sanjay Ghemawat. Google, Inc. *MapReduce: Simplified Data Processing on Large Clusters*.

[14] Divya D. Patel, Kavita R. Singh . International Conference on Innovative Mechanisms for Industry Applications, 2017. *Genome Sequencing using MapReduce and Hadood - A Technical Review*.

[15] Philip Russom. TWDI Best Practices Report *Data Lakes Purposes, Practices, Patterns, and Platforms*, first quarter, 2017.

[16] Gert H. Laursen, Jesper N Thorlund. *Business analytics for managers taking business intelligence beyond reporting*. Wiley and SAS business series, John Wiley Sons, 2017.

[17] Top 6 tools to master and get started with Big Data. Mythili Devi. Jan 3, 2017. https://bigdata-madesimple.com/top-6-big-data-tools-to-master-in-2017/.

[18] Ramesh Sharda, Dursun Delen, Efraim Turban. *Business Intelligence, Analytics and Data Science: A Managerial Perspective*. Pearson Education Inc, Fourth edition, 2018.

[19] Watson IoT IBM 2017. *Descriptive, predictive, prescriptive: Transforming asset and facilities management with analytics*.

[20] Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey. *Business Intelligence and Analytics From Big Data to Big Impact*, MIS Quarterly Vol. 36 No. 4/December 2012.

[21] Salud (Según la OMS) https://concepto.de/salud-segun-la-oms/.

[22] Historia clínica electrónica. Javier Cabo Salvador. https://www.gestion-sanitaria.com/1-historia-clinica-electronica.html.

[23] Historia clínica electrónica HCE en España y en América Latina y el Caribe https://www.e-global.es/software/10-programas-opensource-software-de-historia-clinica-electronica-hce-ece-emr-o-hcue.html.

[24] HL7 normalizando la comunicacion en salud. Febrero 2011. http://informatica-medica.blogspot.com/2011/02/hl7-normalizando-la-comunicacion-en.html.

[25] Las aseguradoras introducen el médico online en sus pólizas Ciencia y Salud, El País. https://cincodias.elpais.com/cincodias/2018/01/25/companias/1516879236_403610.html.

[26] La asistencia médica remota salva vidas y ahorra costes. Salud, La razón. https://www.larazon.es/atusalud/salud/la-asistencia-medica-remota-salva-vidas-y-ahorra-costes-GG11533036.

[27] Ernestina Menasalvas, Consuelo Gonzalo, Alejandro Rodríguez González. International Universidad Politécnica de Madrid. *Big Data en Salud: Retos y oportunidades*.

[28] Alberto Urueña López, María Pilar Ballestero Alemán, Eva Prieto Morais, José María San Segundo Encinar, Iván Soler. International Fundación Vodafone España. *Big Data en salud digital*, 2018.

[29] Cognitive Computing and the Future of Health Care. Mohamed Nooman Ahmed, Andeep S. Toor, Kelsey O'Neil and Dawson Friedland. May 17, 2017. https://pulse.embs.org/may-2017/cognitive-computing-and-the-future-of-health-care/.

[30] Krisa Tailor *The Patient Revolution: How Big Data and Analytics Are Transforming the Healthcare Experience*. 2016 John Wiley Sons, Inc., Hoboken, New Jersey, 2016.

[31] Wullianallur Raghupathi and Viju Raghupathi. Health Information Science and Systems 2014 2:3. *Big data analytics in healthcare: promise and potential*.

[32] Itziar de Lecuona. Departamento de Medicina, Facultad de Medicina; Observatorio de Bioética y Derecho de la Universidad de Barcelona; Comité de Ética de la Investigación del Hospital Clínic de Barcelona, Barcelona, España. *Evaluación de los aspectos metodológicos, éticos, legales y sociales de proyectos de investigación en salud con datos masivos (big data)*, Barcelona, 31 de mayo de 2018.

[33] Leonardo Pucheta. El Derecho: Diario de Doctrina y Jurisprudencia. *Big Data y su impacto en el ámbito de la salud*, Buenos Aires, 4 noviembre de 2017.

[34] SMUFIN. Localizador de mutaciones tumorales Revista de la Universitat de Barcelona sobre salut i benestar http://www.ub.edu/senesciencia/noticia/smufin-nuevo-metodo-analisis-genomica/.

[35] Christopher Burton, Maria Klara Wolters, Antoni Serrano Blanco International Journal of Integrated Care, Volume 12, 15 June 2012 *Help4Mood: avatar-based support for treating people with major depression in the community*.

[36] Comité del Sistema Estadístico Europeo 16 de noviembre de 2017 *Código de Buenas Prácticas de las Estadísticas Europeas*.

[37] Banco de Datos del Ministerio de Sanidad, Consumo y Bienestar Social https://www.mscbs.gob.es/estadEstudios/estadisticas/bancoDatos.htm.

[38] Ministerio de Sanidad, Consumo y Bienestar Social Sanidad 2019 *Marco Estratégico para la Atención Primaria y Comunitaria*.

[39] Guía para el tratamiento de datos personales en el ámbito sanitario Grupo Ático34. 17 agosto, 2017 https://protecciondatos-lopd.com/empresas/guia-centros-sanitarios/.

[40] Cumplimiento de las obligaciones https://www.aepd.es/reglamento/cumplimiento/.

[41] Esquema Interoperabilidad Servicios web del Sistema Nacional de Salud - Ministerio de Sanidad, Consumo y Bienestar Social https://www.mscbs.gob.es/organizacion/sns/servWebSNS/EsqInterope/home.htm.

[42] Informe sobre percepción ciudadana en la prestación de los servicios públicos Ministerio de Hacienda y Función Pública, Junio 2018 https://uxplanet.org/golden-rules-of-user-interface-design-19282aeb06b.

[43] Jiang F, Jiang Y, Zhi H. Stroke and Vascular Neurology 2017. *Artificial intelligence in healthcare: past, present and future*.

[44] Objetivos y metas de desarrollo sostenible https://www.un.org/sustainabledevelopment/es/sustainable-development-goals/.

[45] Sanidad en Datos del Ministerio de Sanidad, Consumo y Bienestar Social https://www.mscbs.gob.es/estadEstudios/sanidadDatos/home.htm.

[46] Golden Rules of User Interface Design Nick Babich https://uxplanet.org/golden-rules-of-user-interface-design-19282aeb06b.

APÉNDICE A

Códigos caso BDCAP

A.1 Obtención de muestras

```
# Imprimimos el numero de matches para cada tramo del archivo INTERCONSULTAS.txt, con un escalon de 1.000.000 de lineas por cada proceso
def func1(arg):
    c=0
    x, y = arg
    for i in range(x,y):
        col2 = line2[i].split(";")[1]
        if col2 in id :
            c+=1
            print(c)
    return(c)

a=1
b=100000 #step

# Estudiamos en que parte del archivo se encuentran las coincidencias entre el archivo PERSONAS.txt e INTERCONSULTAS.txt
p = multiprocessing.Pool(processes=12)
table = p.map(func1, [(a,a+b), (a+b,a+2*b), (a+2*b,a+3*b), (a+3*b,a+4*b), (a+4*b,a+5*b),
                      (a+5*b,a+6*b),(a+6*b,a+7*b),(a+7*b,a+8*b),(a+8*b,a+9*b),(a+9*b,a
                          +10*b)
                     ,(a+10*b,a+11*b),(a+10*b,n2-1)])
print(table)
p.close()

# Ejecutamos de forma independiente al fragmento anterior una vez ya sabemos el rango donde se encuentran las coincidencias
for i in range (a+3*b,a+5*b):
    col2 = line2[i].split(";")[1]
    if col2 in id :
        newdata2.append(line2[i])
    print(len(newdata2))
```

A.2 Observación de datos generales.

```
# Guarda del identificador del paciente en la variable DNI
DNI = input('ID del paciente: ')

# Para dicho DNI obtencion de cada C digo de Capitulo BDCAP del paciente, guardado en la lista cap
for i in range(0,n3-1):
    c=0
    col = line3[i].split(";")
    if col[1] == DNI:
        fechas.append(col[2])
        newdata3.append(col[6])
        cierre.append(col[3])
        if col[6] not in cap:
            cap.append(col[6])

# Impresion a traves de un alfabeto generado con los nombres de cada Capitulo BDCAP, obteniendo la posicion del nombre del Capitulo BDCAP a trav s de la funci n indice sobre cada elemento de la lista cap
if len(cap) == 0:
    print('De este paciente no se tiene ningun registro')
else :
    print('Paciente encontrado, problemas registrados en 2016: \n')
    cap.sort()
    for j in range(0, len(cap)):
        count = newdata3.count(cap[j])
        index = alphabet.index(cap[j])
        if index!=-1:
            print(capitulo[index] + ' : ' + str(count))
```

A.3 Análisis de fechas

```
# Transformacion de las fechas de apertura de string a tipo date
for j in range(len(cap)):
    list=[]
    r=0
    count = 0
    for i in range(len(newdata3)):
        if newdata3[i] == cap[j]:
            fecha = datetime.datetime.strptime(fechas[i],'%d%m%Y').strftime('%d/%m/%Y')
            list.append(fecha)
            list.sort(key = lambda x:time.mktime(time.strptime(x,'%d/%m/%Y')))
            primero = list[0]
            total = 0


```

```
# Calculo de la media del tiempo transcurrido a traves de las fechas de apertura de los problemas de cada capitulo del paciente
    total = 0
    for i in range(len(list)):
        d2, m2, y2 = [int(x) for x in list[i].split('/')]
        b2 = date(y2, m2, d2)
        if b2 > b1:
            r+=1

        dif = (b3 - b2).days
        total = dif + total

    media = total/len(list)
    media = round(float(media/365),1)
    avg_pat.append(media)

# Calculo de la media del tiempo transcurrido a trav s de las fechas de apertura de los problemas de cada capitulo de todos los individuos del archivo 'problemas_data.txt'
    for i in range(0, n3-1):
        col = line3[i].split(";")
        if cap[j] == col[6] and col[3] != '':
            count = 0
            total = 0
            s1 = datetime.datetime.strptime(col[2],'%d %m %Y').strftime('%d/%m/%Y')
            d1, m1, y1 = [int(x) for x in s1.split('/')]
            b1 = date(y1, m1, d1)

            s2 = datetime.datetime.strptime(col[3],'%d %m %Y').strftime('%d/%m/%Y')
            d2, m2, y2 = [int(x) for x in s2.split('/')]
            b2 = date(y2, m2, d2)

            dif = (b2 - b1).days
            total = dif + total
            count+=1

    avg = round(float((total/count)/365),3)
    avg_data.append(avg)

# Excepcion para capitulos de los que no se tienen problemas de pacientes con fecha de cierre
    con = ''
    if avg == 0:
        con = 'No existen datos disponibles de pacientes curados en 2016'

# Impresion de resultados
    index = alphabet.index(cap[j])
    print('Capitulo ' + capitulo[index] +'. //  Abierto el ' + str(primero) + '\n - Problema recurrente : ' + str(r) + ' veces. ' + str(alerta) + '\n - Tiempo medio de recuperaci n: ' + str(avg) + ' a os. ' + str(con))
```

```
59      print('- Tiempo medio transcurrido para el paciente : ' + str(media) + ' a os' + '
            \n' )
```

A.4 Alerta y medidas preventivas

```
1  # Se crea una ventana de alerta ante la detecci n de 3 o m s problemas en cada
       Capitulo BDCAP
2  if r >= 3:
3      alerta = ' ALERTA.(Posible enfermedad cr nica)'
4      index = alphabet.index(cap[j])
5      alert(text='Se debe de realizar un seguimiento en profundidad del paciente en
           capitulo : '
6      + capitulo[index] + '\n \n' + 'Se recomienda seguir las siguientes acciones
           preventivas : \n '
7      + '1. Haga m s actividad f sica \n 2. Consuma mucha fibra \n 3. Elija granos
           integrales \n'
8      + ' 4. Adelgaza los kilos de m s \n 5. Evite las dietas de moda y elija opciones
           m s saludables \n'
9      , title=' Alerta', button='OK')
```

A.5 Datos para gráficos

```
1  # Numero de veces que aparece cada enfermedad por paciente
2  for i in range(0, n3-1):
3      col3 = line3[i].split(";")
4      for j in range(0, n-1):
5          col = line[j].split(";")
6          if col[1]==col3[1] and col[1]+col3[6] not in id:
7              if col[7] >= '40' and col[7] < '60':
8                  index40 = alphabet.index(col3[6])
9                  c40[index40]+=1
10             if col[7] >= '60' and col[7] < '80':
11                 index60 = alphabet.index(col3[6])
12                 c60[index60]+=1
13             if col[7] >= '80' :
14                 index80 = alphabet.index(col3[6])
15                 c80[index80]+=1
16             id.append(col3[1]+col3[6])
17
18 mostcomun_40 = c40.index(max(c40))
19 max1_40 = alphabet[mostcomun_40]
20 print("Primera enfermedad m s com n de los 40 a los 59 a os : " + str(capitulo[
       mostcomun_40]))
21 c40[mostcomun_40] = 0
22 mostcomun2_40 = c40.index(max(c40))
23 max2_40 = alphabet[mostcomun2_40]
```

```
print("Segunda enfermedad m s com n de los 40 a los 59 a os : " + str(capitulo[
    mostcomun2_40]))
c40[mostcomun2_40] = 0
mostcomun3_40 = c40.index(max(c40))
max3_40 = alphabet[mostcomun3_40]
print("Tercera enfermedad m s com n de los 40 a los 59 a os : " + str(capitulo[
    mostcomun3_40])+ '\n')

# Evaluamos si el paciente podr a tener alg n episodio de la enfermedad m s comun en
    su rango de edad
indice = 0
for i in range(0, n-1):
    columna = line[i].split(";")
    if columna[1] == DNI:
        if max1 not in cap and columna[7] >= '60' and columna[7] < '80' :
            indice = mostcomun
            break
        if max2 not in cap and columna[7] >= '40' and columna[7] < '60':
            indice = mostcomun2
            break
        if max2 not in cap and columna[7] >= '80':
            indice = mostcomun2
            break

if indice != 0:
    print('El paciente se encuentra en un rango de edad en el que podr a desarrollar
        alg n problema relacionado con ' + str(capitulo[indice]))

# Contamos el numero de enfermedades por rango de edad para mostrar el grafico
id=[]
for i in range(0, n3-1):
    col3 = line3[i].split(";")
    for j in range(0, n-1):
        col = line[j].split(";")
        if col[1]==col3[1] and col3[1]+col3[6] not in id:
            if col[7] >= '40' and col[7] < '60':
                enf40.append(col3[6])
                count40 += 1
            if col[7] >= '60' and col[7] < '80':
                enf60.append(col3[6])
                count60 += 1
            if col[7] >= '80':
                enf80.append(col3[6])
                count80 += 1
            id.append(col3[1]+col3[6])
```