UNIVERSIDAD DE VALLADOLID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

TRABAJO FIN DE MÁSTER

MÁSTER UNIVERSITARIO EN INVESTIGACIÓN

EN TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES

Análisis de fuentes y técnicas de Big Data en el sector sanitario

Autora:

Dña. Susel Góngora Alonso

Tutora:

Dr. Dña. Isabel de la Torre Díez

Valladolid, 07 de Septiembre de 2017

TÍTULO:	**Análisis de fuentes y técnicas de Big Data en el sector sanitario**
AUTOR:	**Dña. Susel Góngora Alonso**
TUTOR:	**Dr. Dña. Isabel de la Torre Díez**
DEPARTAMENTO:	

TRIBUNAL

PRESIDENTE:	**Dr. D. Miguel López - Coronado**
VOCAL:	**Dr. D. Mario Martínez Zarzuela**
SECRETARIO	**Dr. D. José Fernando Díez Higuera**
FECHA:	**13 de Septiembre de 2017**
CALIFICACIÓN:	

Resumen de TFM

En la última década, la recolección y análisis de datos ha aumentado enormemente en muchos campos de la sociedad. El análisis de Big Data ha empezado a desempeñar un papel fundamental en la evolución de las prácticas y la investigación sanitaria. Ha proporcionado herramientas para acumular, administrar, analizar y asimilar grandes volúmenes de datos dispares, estructurados y no estructurados producidos por los actuales sistemas de salud. La analítica de Big Data se ha aplicado recientemente para ayudar al proceso de la entrega del cuidado y de la exploración de la enfermedad. Con una gestión adecuada, las mejoras en la calidad, cantidad, almacenamiento y análisis de datos de salud podrían conducir a mejoras considerables en muchos de los resultados sanitarios. El objetivo principal de este trabajo es mostrar los resultados de una revisión bibliográfica de las fuentes y técnicas de Big Data empleadas en la sanidad, con el fin de conocer lo que existe e identificar las técnicas más utilizadas en el campo de las enfermedades crónicas. Además planteamos las áreas de investigación médica: imagen, señal, y genómica, las plataformas existentes para el análisis de los datos, las aplicaciones de Big Data en el sector de la salud y el trabajo futuro como posible tesis doctoral.

Palabras clave

Big Data, enfermedades crónicas, minería de datos, sector sanitario, fuentes, técnicas.

Abstract

In the last decade, the collection and analysis of data has increased enormously in many areas of society. The Big Data analysis has begun to play a key role in the evolution of health research and practice. It has provided tools to accumulate, manage, analyze and assimilate large volumes of disparate, structured and unstructured data produced by current health systems. Big Data's analytics have recently been implemented to aid the process of delivery, care and disease screening. With proper management, improvements in the quality, quantity, storage and analysis of health data could lead to considerable improvements in many health outcomes. The main aim of this work is to present a review of existing research in the literature, referring to sources and techniques of Big Data in the health sector and to identify which of these techniques are the most used in the prediction of chronic diseases. In addition we propose the medical research areas: image, signal, and genomics, existing platforms for analysis of the data, the applications of Big Data in the health sector and future work are also considered as possible doctoral thesis.

Keywords

Big Data, chronic diseases, data mining, healthcare, sources, techniques.

Agradecimientos

Me gustaría agradecer ante todo a la Universidad de Valladolid y al programa de Becas Santander por darme la oportunidad de venir desde Cuba para realizar la maestría y superarme profesionalmente.

A mi tutora y el claustro de profesores del Máster que me han brindado todo su apoyo en mi paso por la Universidad.

A mi madre y mi hermano Ulises, porque desde tan lejos me brindan la fuerza y la energía para seguir adelante.

A mi pareja por su apoyo y amor incondicional.

Y con un especial cariño a los amigos más cercanos, que he encontrado durante mi estancia en este país, que me han acogido como un miembro más de la familia.

Índice

Capítulo 1

Introducción

El término Big Data hace referencia a un conjunto de datos y de información tan grande y compleja que resulta complicada su procesamiento empleando herramientas de gestión de bases de datos de tipo convencional. Big Data son activos de información de gran volumen, alta velocidad y alta variedad que requiere nuevas formas de procesamiento para permitir una mejor toma de decisiones (Philip Chen & Zhang, 2014). En general son datos muy grandes, estructurados o no, estáticos o dinámicos, simples o complejos, que pueden ser capturados, almacenados, gestionados y analizados utilizando diferentes técnicas innovadoras (Philip Chen & Zhang, 2014).

El elevado volumen de datos (más de un *petabyte*) precisa nuevas técnicas de almacenamiento a gran escala y enfoques distintos para recuperar la información; la variedad de las fuentes de datos (texto, audio, vídeo, etc.) hace que las redes relacionales sencillas sean difícilmente aplicables; y por último, el incesante incremento con que se generan los datos, hace que la velocidad sea un parámetro clave en su manejo (Manuel & Sesmero, 2015).

La gran novedad de Big Data es el procesamiento de información no estructurada: lenguaje natural, redes sociales, telemedicina, sensores, etc. Vivimos en una sociedad donde los datos textuales en Internet están creciendo a un ritmo rápido y muchas empresas están tratando de usar este diluvio de datos para extraer las opiniones de la gente hacia sus productos (Cunha, Silva, & Antunes, 2015). Una gran fuente de información de texto no estructurada se incluye en las redes sociales, donde es imposible analizar manualmente tales cantidades de datos. Hay un gran número de sitios web de redes sociales que permiten a los usuarios contribuir, modificar y clasificar el contenido, así como expresar sus opiniones personales sobre temas específicos. Algunos ejemplos incluyen blogs, foros, sitios de revisiones de productos y redes sociales, como Twitter (http://twitter.com/). Estos ejemplos se presentan como la tendencia futura en los flujos de datos de la evolución de la minería (Cunha et al., 2015).

El análisis del Big Data ha abierto la puerta a una nueva era para la mejora en la prestación de servicios y solución de problemas en el ámbito de los sistemas sanitarios (Manuel & Sesmero, 2015). La gran mayoría de los agentes que participan en las estructuras de los servicios de salud reconocen que el análisis del Big Data puede ofrecer nuevas posibilidades en la elaboración de modelos predictivos, patrones de comportamiento, el descubrimiento de nuevas necesidades, reducir riesgos, así como proveer servicios más personalizados, todo ello en tiempo real y teniendo en cuenta toda la información relevante (Manuel & Sesmero, 2015).

Existen revisiones similares que basan su estudio en campos específicos del sector sanitario, como son el análisis de Big Data en informática biomédica (Merelli, Pérez-Sánchez, Gesing, & D'Agostino, 2014), el análisis de datos en tres áreas de investigación: procesamiento de imágenes, señales y genómica (Belle et al., 2015), el análisis de datos ómics en campos de la bioinformática, la

biomatemática y la bioestadística para desarrollar análisis de traslación de datos y aprovechar al máximo las tecnologías de alto rendimiento (Alyass, Turcotte, & Meyre, 2015), etc. En la revisión de la literatura no se encontró ningún estudio que abarcara las fuentes y técnicas de Big Data existentes de forma generalizada en los distintos campos de la medicina, por tanto se muestra la originalidad del Trabajo Fin de Máster (TFM). Además los estudios similares, encontrados en campos específicos de la salud hacen razonable pensar que nuestro trabajo es viable.

La era de los registros de salud electrónicos, la genómica y la mejora de los recursos de tecnología de la información (Trifiletti & Showalter, 2015)crean la oportunidad de aprovechar estos desarrollos para crear un sistema de salud de aprendizaje que puede entregar rápidamente evidencia clínica informativa. Al fusionar conceptos de investigación comparativa de efectividad con las herramientas y enfoques analíticos de Big Data, se espera que esta unión acelere el descubrimiento, mejore la evidencia para la toma de decisiones y aumente la disponibilidad de información altamente relevante y personalizada (Trifiletti & Showalter, 2015).

La expansión reciente de las últimas tendencias en Tecnologías de la Información (Redes Sociales, Movilidad, *Cloud Computing*, Big Data e Internet de las Cosas) supone importantes avances, pero también grandes desafíos, que exigen un considerable esfuerzo de adaptación por parte de las organizaciones, en especial el sector sanitario. Por tanto para que éste pueda beneficiarse de las posibles ventajas de su adopción es necesario previamente realizar una serie de cambios en la mayoría de los sistemas de información de los servicios sanitarios que existen en la actualidad. Estos cambios deben estar orientados, entre otros objetivos, a conseguir gestionar y analizar grandes volúmenes de datos procedentes de fuentes muy diversas y registrados en formatos muy heterogéneos. Las áreas de análisis de datos como historias clínicas electrónicas, análisis clínicos, gestión de centros de salud son algunas de las fuentes.

En este TFM se mostrarán los resultados de una revisión bibliográfica de las fuentes y técnicas de Big Data empleadas en la sanidad, con el fin de conocer lo que existe e identificar las técnicas más utilizadas en el campo de las enfermedades crónicas.

1.1 Objetivos principales

Este proyecto está orientado a una revisión de la literatura referente a las fuentes y técnicas de Big Data en el sector de la salud. Los objetivos principales que se plantean en este trabajo serán los siguientes:

1. Elaborar una descripción del estado del arte existente en Big Data en el sector de la sanidad.
2. Análisis de las fuentes de Big Data en la sanidad en este caso relacionadas con historias clínicas electrónicas, análisis clínicos, datos de compañías farmacéuticas, gestión de centros de salud, datos de ensayos clínicos, de genética e imágenes medicas
3. Análisis de las técnicas de Big Data en la salud, teniendo en cuenta las plataformas de Big Data y la analítica predictiva.
4. Identificar cuáles de las técnicas existentes son las más utilizadas en la

predicción de enfermedades crónicas tales como cáncer, diabetes, asma, enfermedades cardíacas, etc.

1.1 Estructura de la memoria

La estructura de la memoria sigue los requisitos de un trabajo de investigación, aborda en el Capítulo 1 una introducción que muestra la relevancia del tema, describe el contexto, los objetivos principales, y la estructura del trabajo, mientras que en el Capítulo 2 se muestra la metodología empleada en el desarrollo de la investigación.

En el Capítulo 3 se da una visión general de Big Data, en especial para lectores que tengan poco conocimiento del tema. Se describen las áreas de investigación médica, las principales fuentes e infraestructuras de Big Data así como las principales bases de datos existentes del sector sanitario.

En el Capítulo 4 se muestran las técnicas de Big Data en la salud encontradas en la literatura, las plataformas de Big Data, las técnicas de minería de datos existentes, el modelo predictivo de maquinas de aprendizaje y sus aplicaciones en el sector sanitario.

En el Capítulo 5 se identifican las técnicas de Big Data que más se utilizan en la predicción de enfermedades crónicas, haciendo una comparativa entre ellas. Por último en el Capítulo 6 se exponen las conclusiones del TFM y se proponen líneas futuras de investigación, para la mejora de este trabajo.

Capítulo 2

Metodología

En el desarrollo del proyecto, se aplica una metodología cualitativa para hacer un análisis de las fuentes y técnicas de Big Data en el campo de la salud. Para realizar una revisión sistemática de la literatura disponible se utiliza la metodología propuesta por (Brereton, Kitchenham, Budgen, Turner, & Khalil, 2007). De esta forma se definen las bases de datos a utilizar, los criterios de búsqueda, y los campos de búsqueda tal y como se muestran en la Tabla I.

Tabla I- Búsquedas realizadas en las diferentes bases de datos académicas. Fuente: Propia

Databases	Criterios de búsqueda	Campos de búsqueda
IEEE Xplore	"techniques" OR "sources" AND "Big Data" AND "medicine" OR "health" OR "chronic diseases"	"abstract, title, keywords"
Scopus	"techniques" OR "sources" AND "Big Data" AND "medicine" OR "health" OR "chronic diseases"	"abstract, title, keywords"
PubMed	"techniques" OR "sources" AND "Big Data" AND "medicine" OR "health" OR "chronic diseases"	"title/abstract"
Science Direct	"techniques" OR "sources" AND "Big Data" AND "medicine" OR "health" OR "chronic diseases"	"abstract, title, keywords"

Se desarrolló una revisión de los trabajos publicados relacionados con técnicas y fuentes de Big Data hasta diciembre del 2016. La revisión fue un estudio de la literatura donde se utilizaron diferentes sistemas y bases de datos académicas como son IEEE Xplore (http://ieeexplore.ieee.org/), Scopus (https://www.scopus.com/), PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) y Science Direct (http://www.sciencedirect.com/).

Las fechas de publicación a tener en cuenta para la revisión son a partir del 2006 hasta la actualidad. Se consideraron dos requisitos para la inclusión de un documento como pertinente. El documento debe estar redactado en inglés y debe incluir las fuentes y/o técnicas de Big Data en el sector sanitario. El proceso de selección de los artículos se llevó a cabo mediante la lectura de los títulos y resúmenes de los resultados obtenidos por uno de los autores, y los trabajos se clasificaron leyendo sus resúmenes así como el artículo completo cuando fuese necesario. La Figura 1.1 muestra los resultados obtenidos de las búsquedas, así como los artículos rechazados como duplicados o debido a un título que no está relacionado con nuestro tema.

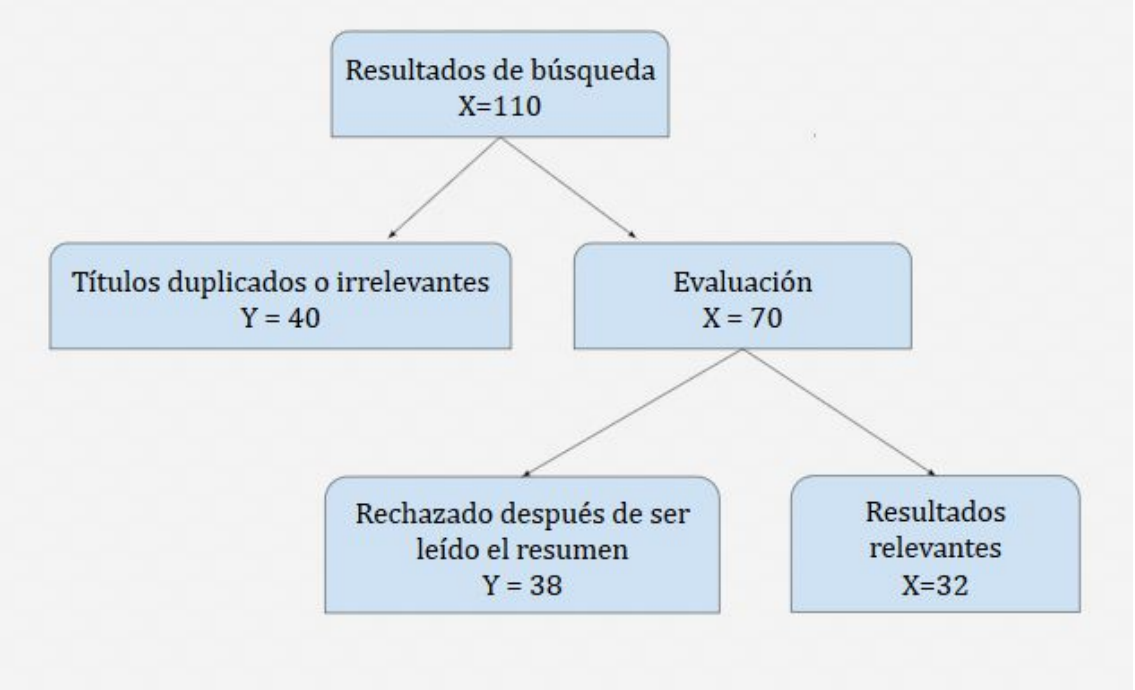

Figura 2.1: Diagrama de flujo de los pasos seguidos en la revisión de la literatura. Fuente: Propia

De las 110 publicaciones encontradas 40 fueron duplicadas o con un título irrelevante para esta investigación, los 70 restantes fueron leídos y analizados sus resúmenes para ver cuáles eran de interés, resultando 32 documentos los cuales dieron lugar a contribuciones pertinentes. A continuación, en los capítulos siguientes se muestran las obras más relevantes encontradas.

Capítulo 3

Fuentes de Big Data en el sector sanitario

3.1 Visión General de Big Data en la Salud

Los datos masivos proceden de múltiples fuentes de información, derivados de diferentes contextos, tales como los financieros, la informática de negocio, el ocio, las redes sociales y las redes laborales, las ciencias ambientales y también la salud (Pérez, 2016). En este último ámbito la industria de la salud ha llegado a un consenso sobre el valor Big Data como una herramienta transformadora. Las principales fuentes de datos incluyen años de investigación y desarrollo como son: datos de compañías farmacéuticas, *Electronic Medical Records (EMR)*, *Electronic Health Records (EHRs)*, de proveedores de atención médica, y de ensayos clínicos. Estos datos provienen de múltiples fuentes de información derivada de la medicina asistencial, la genómica, la biología molecular, la clínica, la epidemiología y la salud pública, entre otras. Los avances tecnológicos recientes en el hardware y el software están haciendo que sea más fácil la recolección, transferencia, almacenamiento, y análisis de datos de múltiples fuentes (Nambiar, Bhardwaj, Sethi, & Vargheese, 2013).

Existe ya una gran base de conocimiento sobre los resultados de la aplicación de Big Data en salud y biomedicina, incluso en salud pública en particular. Conforme aumenta su aplicación se identifican nuevos retos a los que enfrentarse, así como nuevas oportunidades que acrecientan el interés por el desarrollo de la investigación en este dominio. Los métodos y herramientas de Big Data se caracterizan por el volumen, la complejidad y la velocidad de la información que manejan. Las propiedades, los retos y los asuntos relevantes que caracterizan la aplicación de los mismos en biomedicina son la gran variedad en la naturaleza de los datos y la alta velocidad de proceso requerida; retos relacionados con la veracidad de los datos, con los flujo de trabajo, con los métodos computacionales, con la extracción de información significativa, con el intercambio de datos y con la necesidad de expertos en el uso de estas tecnologías (Luis & Calderón, 2016).

3.2 Áreas de investigación médica

Las nuevas tecnologías permiten capturar grandes cantidades de información sobre cada paciente individual en una gran escala de tiempo. Según (Belle et al., 2015) existen tres áreas populares de investigación donde los conceptos de análisis de Big Data se están aplicando actualmente, las mismas son: el procesamiento de imágenes, de señales y la genómica.

3.2.1 Imágenes Médicas

Las imágenes médicas son una fuente importante de datos utilizados con frecuencia para el diagnóstico, la evaluación de la terapia y la planificación (Belle et al., 2015). La tomografía computarizada, la resonancia magnética, la radiografía, la imagen molecular, el ultrasonido, la imagen fotoacústica, el fluoroscopio, la

tomografía computarizada por emisión de positrones y la mamografía son algunos de los ejemplos de técnicas de imagen.

La integración del análisis informático con la atención adecuada tiene potencial para ayudar a los médicos a mejorar la precisión diagnóstica. La integración de imágenes médicas con otros tipos de datos de EHRs y datos genómicos, también puede mejorar la precisión y reducir el tiempo necesario para un diagnóstico.

El objetivo del análisis de imágenes médicas es mejorar la interpretabilidad de los contenidos representados (Belle et al., 2015). Muchos métodos y marcos se han desarrollado para el procesamiento de imágenes médicas. Sin embargo, estos métodos no son necesariamente aplicables para grandes aplicaciones de datos.

3.2.2 Señales Médicas

Similar a las imágenes médicas, las señales médicas también plantean obstáculos de volumen y velocidad, especialmente durante la adquisición y almacenamiento continuos de alta resolución desde una multitud de monitores conectados a cada paciente (Belle et al., 2015). El análisis de las señales fisiológicas suele ser más significativo cuando se presenta junto con la conciencia contextual y situacional, que debe integrarse en el desarrollo de monitoreo continuo y sistemas predictivos para asegurar su efectividad y robustez.

La telemetría y los dispositivos de monitorización de señales fisiológicas son omnipresentes. Sin embargo, los datos continuos generados a partir de estos monitores no se han almacenado típicamente durante más de un breve período de tiempo, con lo que se descuida la extensa investigación en los datos generados.

El análisis de datos en flujo continuo en la asistencia sanitaria puede definirse como un uso sistemático de la forma de onda continua (señal que varía en función del tiempo), y la información relacionada de registros médicos desarrollada a través de disciplinas analíticas aplicadas (por ejemplo estadística, cuantitativa, contextual, cognitiva y predictiva) en atención al paciente (Belle et al., 2015). Se requiere una plataforma para la adquisición y la ingestión de datos de flujo continuo, que tiene el ancho de banda para manejar múltiples formas de onda a diferentes fidelidades. La integración de estos datos dinámicos en forma de onda con datos estáticos del EHRs, es un componente clave para proporcionar una conciencia situacional y contextual para el motor de análisis.

3.2.3 Genómica

En la actualidad se están llevando a cabo iniciativas a lo largo de los años para integrar los datos clínicos desde el nivel genómico al nivel fisiológico de un ser humano (Belle et al., 2015). Estas iniciativas ayudarán a brindar atención personalizada a cada paciente. La entrega de recomendaciones en un entorno clínico requiere un análisis rápido de Big Data de genoma de una manera fiable.

El análisis genómico que utiliza microarrays ha sido exitoso en el análisis de rasgos a través de una población y contribuyó con éxito en tratamientos de enfermedades complejas como la enfermedad de Crohn y la degeneración muscular relacionada con la edad (Belle et al., 2015).

3.3 Principales fuentes de Big Data

Los datos a gran escala se pueden recopilar de muchas fuentes diferentes. A continuación, en la Figura 3.1 mostramos cuatro fuentes populares en la generación de datos a gran escala (Huang et al., 2015).

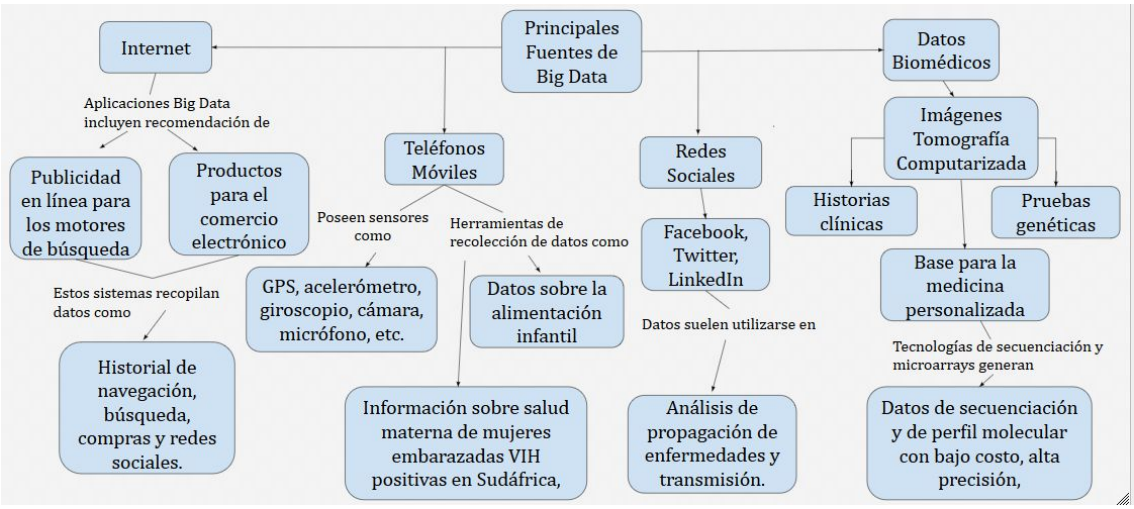

Figura 3.1: Principales fuentes de Big Data. Fuente: (Huang et al., 2015)

El Internet, los teléfonos móviles, las redes sociales, los sensores y los datos biomédicos son las principales fuentes existentes de Big Data (Huang et al., 2015). Las primeras aplicaciones de Big Data incluyen la recomendación de publicidad en línea para los motores de búsqueda y la de productos para el comercio electrónico. Todos estos sistemas recopilan datos de Internet tales como, historial de navegación, historial de búsqueda, historial de compras y redes sociales.

En el sector de la salud Google proporciona una herramienta llamada *Google Flu Trends* para la vigilancia en tiempo real de brotes de influenza (Huang et al., 2015). Dicha herramienta basa su suposición en que, cuando el número de personas tiene síntomas de influenza, las búsquedas de temas relacionados con la enfermedad aumentarán. Por tanto, basándose en búsquedas en Internet, se puede estimar el número de personas con síntomas gripales.

El uso de teléfonos móviles reflejado en la Figura 3.1 ha crecido excepcionalmente en la última década, estos dispositivos tienen incluido una serie de sensores como GPS, acelerómetro, giroscopio, cámara, micrófono, etc. (Huang et al., 2015). La ubicuidad del teléfono móvil y una gran cantidad de datos generados a partir de sensores incorporados ofrecen nuevas oportunidades para caracterizar y entender el comportamiento de vida (por ejemplo, movilidad humana, patrones de interacción, etc.) de un usuario.

Otra de las fuentes de Big Data reflejada en la Figura 3.1 son las redes sociales. Facebook, Twitter y LinkedIn han ganado notable atención en los últimos años y son muy ricas en contenido. La enorme cantidad de datos disponibles en las mismas ofrece oportunidades sin precedentes para el análisis de datos en múltiples áreas (Huang et al., 2015). En el ámbito de la salud pública, estos datos suelen utilizarse para el análisis de la propagación/transmisión de enfermedades.

El uso de estas nuevas tecnologías como son los smartphones, las aplicaciones que contiene, así como las redes sociales, ofrecen la promesa de poder utilizar los datos que recolectan para desarrollar nuevos métodos de vigilancia y epidemiología de una de las enfermedades que por más de tres décadas sigue siendo un tremendo reto para la salud pública, el VIH. En el estudio (Young, 2015) se demostró que los datos recogidos a través de redes sociales tienen el potencial de proporcionar una alternativa más rentable y en tiempo real para el monitoreo remoto y la vigilancia del VIH. Estos datos también han ayudado a los investigadores en los esfuerzos de prevención del VIH, tales como la capacidad de distribuir kits de pruebas de VIH en el hogar a los necesitados.

Los datos biomédicos (Nambiar et al., 2013) de la comunidad hospitalaria y científica son otra fuente importante; incluyen datos de años de investigación y desarrollo de las compañías farmacéuticas, los EMR y EHR (Nambiar et al., 2013). Estos registros están conformados por notas de médicos, informes de laboratorio, informes de rayos X, historias clínicas, régimen de dieta, pruebas genéticas, lista de médicos y enfermeras en un hospital en particular, imágenes de tomografía computarizada, datos del registro nacional de salud, medicina e instrumentos quirúrgicos, etc., los cuales son la base para la medicina personalizada (Huang et al., 2015).

En el área biomédica, con el rápido desarrollo de las tecnologías de secuenciación y microarrays, se generan toneladas de datos de secuenciación y datos de perfil molecular con bajo costo, alta precisión, velocidad rápida y requerimiento mínimo de muestras (Huang et al., 2015). Estos conjuntos de datos proporcionan puntos de referencia para el desarrollo del método y la elevación del rendimiento del análisis Big Data.

Las tecnologías de habilitación que van desde nano y microelectrónica, materiales avanzados, computación portátil/móvil y sistemas de telecomunicaciones, así como sistemas de teledetección y de información geográfica han hecho posible que la información sobre la salud se recopile de forma penetrante y discreta.

3.4 Big Data integrado en la Bioinformática Médica

El análisis e interpretación eficientes de Big Data abre nuevas vías para explorar la biología molecular, se necesitan nuevos paradigmas para almacenar y acceder a los datos, para su anotación e integración, y finalmente para inferir el conocimiento y ponerlo a disposición de los investigadores (Merelli et al., 2014). La bioinformática puede ser vista como la unión de todos estos procesos. Un claro conocimiento de las actuales soluciones de computación de alto rendimiento en bioinformática, los paradigmas de análisis Big Data para la biología computacional y las cuestiones que aún están abiertas en los campos biomédico y de la salud, representan el punto de partida para ganar este desafío.

El desarrollo de metadatos para la información biológica sobre la base de los estándares de la web semántica puede ser visto como un enfoque prometedor para una integración semántica de la información biológica. La web semántica es un movimiento colaborativo, que promovió el estándar para la anotación y la integración de datos (Merelli et al., 2014). Al fomentar la inclusión del contenido

semántico en los datos accesibles a través de Internet, se pretende convertir la web actual, dominada por documentos no estructurados y semiestructurados, en una red de datos. Se trata de publicar información en lenguajes diseñados específicamente para los datos: Resource Description Framework (RDF), Web Ontology Language (OWL) y SPARQL (lenguaje de consulta de grafos RDF).

En (Merelli et al., 2014) citan de ejemplo en el campo de la biomedicina, el proyecto *Open Biomedical Ontologies (OBO)*, como un esfuerzo para crear vocabularios controlados para uso compartido en diferentes dominios biológicos y médicos. OBO pertenece a los recursos del *National Center for Biomedical Ontology (NCBO)*, donde formará un elemento central del BioPortal de la NCBO. Otro ejemplo que se referencia es el intento de publicar datos vinculados realizado por el proyecto Bio2RDF (Merelli et al., 2014). El objetivo del proyecto es crear una red de datos coherentemente vinculados a través de las bases de datos biológicas. Como parte del proyecto se ha construido un almacén bioinformático integrado en la web semántica. Este almacén RDF tiene más de 70 millones de triples que describen los genomas humanos y de ratón.

EL mayor impacto de RDF y OWL en bioinformática (Merelli et al., 2014) es ayudar a integrar todos los formatos de datos y estandarizar las ontologías existentes. Si los identificadores únicos se convierten a referencias URI, las ontologías se pueden expresar en OWL y los datos pueden anotarse a través de estos recursos basados en RDF.

3.5 Infraestructuras de Big Data en Salud

Como líder mundial en redes, *Cisco* está bien posicionada para ayudar a mejorar el futuro de la asistencia sanitaria a través de las tecnologías de red que transforma la manera de las personas al conectarse, acceder, colaborar y compartir información. Según (Nambiar et al., 2013) las soluciones *Cisco Connected Health (CCH)* permiten la atención en equipo, colaboración y la eficiencia del negocio. Ayuda a simplificar las comunicaciones de la salud que utilizan una red de tecnologías interoperables que conectan mejor a los pacientes con los proveedores médicos, pagadores y organizaciones de ciencias de la vida. Las soluciones de Cisco conectan información crítica, las personas, y el conocimiento para ayudar a mejorar la experiencia de la atención médica.

La arquitectura CCH (Nambiar et al., 2013) fue desarrollado específicamente para profesionales de la salud para ayudar a construir una infraestructura robusta, altamente segura y escalable para la entrega de servicios de salud, mediante la utilización de las directrices y las mejores prácticas de la experiencia en el cuidado de la salud, que se centra en los datos del hospital, espacio de trabajo unificado, y la interoperabilidad de la asistencia sanitaria.

En el dominio de enfermedades cardiovasculares, se han iniciado esfuerzos para establecer infraestructuras para el intercambio de datos, proporcionando exámenes anónimos de referencia, datos de imagen y otros cálculos derivados, como son los modelos anatómicos. Una gran cantidad de datos ya está disponible en exámenes de imágenes cardíacas no invasivas. Se han establecido estudios (Suinesiaputra, Medrano-Gracia, Cowan, & Young, 2015) prospectivos longitudinales derivados de los datos de imagen a gran escala, para investigar la

patogénesis de las enfermedades cardíacas. Los estudios longitudinales, que siguen a los pacientes a través del tiempo, permiten a los científicos entender la evolución de la enfermedad cardíaca de las manifestaciones subclínicas a síntomas clínicos. Según (Suinesiaputra et al., 2015) la Tabla II resume las infraestructuras existentes de intercambio de datos médicos, que también incluyen los datos cardiovasculares.

Tabla II- Principales estudios epidemiológicos cardiovasculares con la inclusión de datos de imagen. Fuente: (Suinesiaputra et al., 2015)

Estudios	País	Imágenes	Inicio	Tamaño	Grupo	Años
MESA	USA	ECG, CT, MRI	2000	6814	Hispanic, Chinese, White, African-American	45–84
JHS	USA	ECG, CT, MRI	2000	5302	African-American	21–84
UK Biobank	UK	MRI, ECG, DEXA	2006	100 000	Multi	40–69
CPTP	Canadá	MRI	2009	10 000	Multi	35–69
Iceland MI	Islandia	MRI	2004	936	European	67–93
Framingham Offspring	USA	MRI	2002	1707	Multi	NA
EuroCMR	Europa	MRI	2012	27 781	Multi	47–70

Establecido en 2010, el Cardiac *Atlas Project (CAP)* es un consorcio mundial para albergar Big Data en imagen cardíaca, con modelos de elementos finitos derivados del corazón e información de diagnóstico asociado (Suinesiaputra et al., 2015). Más de 3000 casos de resonancia magnética cardíaca han contribuido a la base de datos, que está siendo utilizada por más de 20 grupos de investigación en todo el mundo para diversas actividades de investigación. CAP ha desarrollado métodos para combinar los datos de diferentes fuentes en una forma normalizada, y para corregir el sesgo derivado de protocolo de formación de imágenes o análisis.

Anatomical Models Database (AMDB) es un marco de acceso web para compartir y reutilizar modelos cardiovasculares (Suinesiaputra et al., 2015). AMDB almacena varios modelos de geometría cardíacos, accesible para cualquier investigador para llevar a cabo una simulación o un estudio de referencia en la electrofisiología cardíaca y la mecánica. Emplea una herramienta de servicio web que personaliza un modelo de corazón geométrico dado imágenes binarias del corazón y los parámetros cardíacos.

Integrated Data analysis, Anonymization and Sharing (iDASH) (Suinesiaputra et al., 2015) es un marco general para proporcionar una herramienta escalable para compartir y acceder a los datos médicos, incluyendo las enfermedades cardíacas. iDASH almacena y comparte datos heterogéneos de múltiples dominios.

CardioVascular Research Grid (CVRG) proporciona una (Suinesiaputra et al., 2015) infraestructura para acceder de forma segura a complejos datos de estudios cardiovasculares, incluyendo un etiquetado automatizado ontológico de las diferencias anatómicas. CVRG está creando una plataforma de acceso fácil con las herramientas basadas en la nube y en el navegador sin necesidad de instalar software complejo.

Una infraestructura específica que comparte solamente señales de *Electrocardiogram (ECG)* (Suinesiaputra et al., 2015) es proporcionado por *PhysioNet*. En la actualidad, las tiendas *PhysioNet* comparten a gran escala señales fisiológicas desprovistas de identificación, series de tiempo y otros datos biomédicos relacionados.

Virtual Imaging Platform (VIP) (Suinesiaputra et al., 2015)es otra plataforma de intercambio de datos en línea de libre acceso, que se centra más en extensos procesos computacionales de simulación biomédica.

La historia médica de los pacientes con cáncer añade volumen y complejidad adicional, especialmente en los grandes hospitales y centros médicos universitarios donde gran cantidad de datos adicionales son generados por el personal científico. En (Seebode, Ort, Regenbrecht, & Peuker, 2013)se plantea una plataforma que integra y procesa grandes conjuntos de datos basado en la explotación semántica de los datos clínicos recogidos de diversos recursos, como los sistemas de información hospitalarios, relatos clínicos, la literatura científica y las plataformas de colaboración.

Esta plataforma se desarrolla dentro de un proyecto conjunto con Charité y Vivantes, proveedores de salud líderes en Europa. El principal motor para el desarrollo es el apoyo en la identificación de cohortes en los ensayos clínicos para los estudios de factibilidad y de reclutamiento. Es compatible con la explotación flexible y los datos clínicos de diversas fuentes, incluyendo los textos clínicos. El conocimiento está representado por ontologías y diseñado por los propios expertos del dominio de una manera directa en un banco de trabajo semántico (Seebode et al., 2013). La explotación semántica de datos clínicos a partir de registros estructurados de *Health Information Systems (HIS)* y textos clínicos, proporcionan una base de evidencia actualizada para la investigación médica y la prestación de asistencia sanitaria. Los médicos y expertos son capaces de diseñar ontologías y adaptar los conocimientos necesarios para un caso de aplicación o uso. Las ontologías siempre pueden ser validadas y probadas contra los datos reales que desembocan en el sistema de fuentes estructuradas o textos clínicos.

En (Brinkmann, Bower, Stengel, Worrell, & Stead, 2009) se describe una plataforma de electrofisiología escalable y de última generación diseñada para la adquisición, compresión, cifrado y almacenamiento de datos a gran escala. Los datos se almacenan en un formato de archivo que incorpora la compresión de datos sin pérdidas, utilizando diferencias de rango de codificación, una suma de comprobación de 32 bits cíclicamente redundante para garantizar la integridad de los datos, y cifrado de 128 bits para la protección de la información del paciente.

En el centro de enfoque está un sistema de adquisición escalable (hasta 1024 canales), un *Storage Area Network (SAN)* a gran escala y un nuevo *Multiscale Electrophysiology Format (MEF)* (Brinkmann et al., 2009), tal y como se muestra en la Figura 3.2.

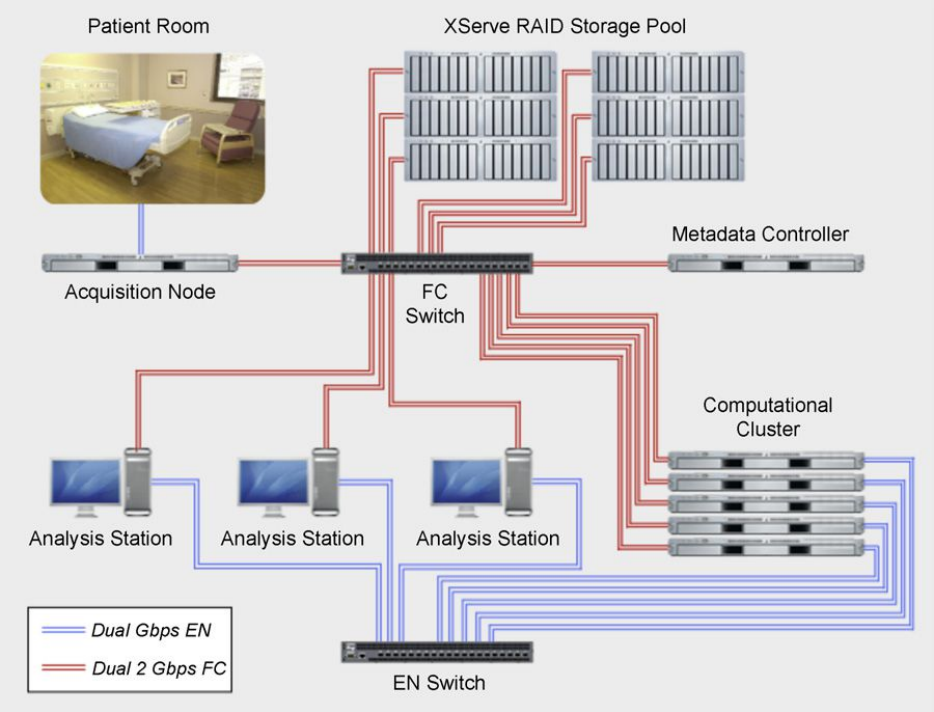

Figura 3.2: Plataforma de electrofisiología escalable. Fuente: (Brinkmann et al., 2009)

La capacidad de recolectar, almacenar y extraer el ancho de banda de electrofisiología a lo largo de múltiples escalas espaciales, fue desarrollada originalmente para investigar la estructura fina del cerebro epiléptico humano (Brinkmann et al., 2009). Una plataforma de adquisición escalable (32-320 canales) capaz de grabar continuamente a largo plazo, fue desarrollada en colaboración con *Neuralynx Inc*. El sistema *Digital-Lynx* es el único que utiliza un convertidor A / D individual de alta resolución de 24 bits por canal para digitalizar directamente la señal del electrodo, utilizando un solo amplificador diferencial de bajo acoplamiento de DC y filtro anti-aliasing (paso bajo de 9 kHz). Todos los canales se muestrean simultáneamente a 32 kHz con un ancho de banda de señal DC a 8 kHz. Este diseño de alta resolución proporciona un rango de entrada dinámica de ± 132mV con resolución de 1µV (18th bit). Todos los datos muestreados se empaquetan y se transfieren a un PC a través de un enlace de datos de fibra óptica a 600Mbits/s.

El cifrado de la información de identificación (Brinkmann et al., 2009)del paciente dentro del archivo con un algoritmo apropiado representa una solución elegante para mantener la confidencialidad del paciente, evitando la necesidad de protocolos de transferencia especializados, y reduciendo el potencial para perder información relevante o causar errores de registro en los datos de investigación. El protocolo de investigación implica la sustitución de electrodos clínicos estándar con electrodos híbridos personalizados.

El MEF (Brinkmann et al., 2009) consta de tres partes principales: (1) una cabecera de 1024 bytes de longitud fija, que contiene información del paciente e información técnica sobre la grabación, (2) una sección de datos, que consta de una serie de bloques de datos codificados y (3) una sección de índice de tiempo, que consta de tres bloques, de elementos de 8 bytes que mantienen el tiempo de inicio del bloque, el desplazamiento del archivo y los valores del índice de muestra para facilitar el acceso aleatorio rápido a los datos.

La tecnología de grabación actual junto con el formato de archivo MEF desacopla la adquisición de datos de las restricciones de almacenamiento y análisis, permitiendo a los neurobiólogos de sistemas adquirir, almacenar y manipular todos los datos fisiológicamente relevantes (Brinkmann et al., 2009). La estructura de bloques de los datos hace que el archivo resista daños menores durante el almacenamiento o la transmisión, ya que sólo se perderán los bloques dañados, mientras que los bloques de datos restantes no se verán afectados.

En (Archenaa & Anita, 2015) se plantea el análisis de datos referentes a notas médicas, informes de laboratorio, informes de rayos X, historial de casos, régimen de dieta, lista de médicos y enfermeras de un hospital en particular, datos del registro nacional de salud, identificación de la fecha de vencimiento de los instrumentos médicos y quirúrgicos basada en datos *RFID*, etc. Para ello se utiliza *MapReduce* y *Hive*, mediante la implementación de algoritmos de aprendizaje automático que ayudan a analizar un patrón similar de datos. Esto ayuda a predecir el riesgo de condición de salud del paciente en las primeras etapas.

En (Vito, Casagrande, Bianchi, & Costantino, 2015) se desarrolla una infraestructura de datos común que involucra cuatro centros de diálisis clínicas entre Lombardía y Suiza. La plataforma ha sido construida para almacenar gran cantidad de datos clínicos relacionados con las 716 sesiones de diálisis de 70 pacientes de diferentes hospitales. La plataforma está compuesta por una combinación de una base de datos *MySQL* y una biblioteca de minería de datos basada en *MATLAB*.

Dicha base de datos está organizada para almacenar información heterogénea procedente tanto de sistemas de información hospitalarios como de adquisición en tiempo real de máquinas de diálisis (Vito et al., 2015). El esquema resultante recoge datos de registro para cada sujeto, y prescripción clínica, estado de hidratación, datos de adquisición de la máquina y datos hemato-químicos para cada sesión de tratamiento.

3.6 Bases de datos del sector sanitario

National Database for Autism Research (NDAR) (Payakachat, Tilford, & Ungar, 2016) es un repositorio de datos de investigación financiado por los *National Institutes of Health (NIH)* de los Estados Unidos, creado mediante la integración de datos heterogéneos, a través de acuerdos de intercambio de datos, entre investigadores del autismo y los NIH. NDAR es considerado el mayor repositorio de datos de neurociencia y genómica para la investigación del autismo. Además de los datos biomédicos, contiene una gran colección de evaluaciones clínicas, comportamiento y resultados de salud de las nuevas intervenciones.

Tiene dos componentes clave que sirven de base para un portal de bioinformática: *global unique identifier (GUID)* para sujetos de investigación y un diccionario de definición de datos armonizado definido por el investigador para las descripciones de los experimentos (Payakachat et al., 2016). El GUID sirve como el eslabón esencial para agrupar los datos clínicos y de investigación biomédica entre los laboratorios, los proyectos de investigación y los repositorios de datos sin vulnerar la información de identificación personal.

Otro componente clave del NDAR es la biblioteca de extensas definiciones de datos, que abarca 800 medidas de autismo de investigación clínica, de imágenes y genómica en el autismo. Este componente ayuda a estandarizar los datos entre diferentes laboratorios y repositorios (Payakachat et al., 2016). NDAR proporciona una herramienta de validación de datos para los investigadores de autismo para confirmar si sus datos son compatibles con las definiciones existentes antes de la presentación. También integra recursos de computación en nube dentro de los datos compartidos disponibles, permitiendo a los investigadores mover datos de imágenes de alta dimensión directamente en un software, diseñado para procesamiento y análisis altamente eficiente.

Los EMRs se archivan digitalmente y posteriormente se pueden extraer y analizar. Entre 2011 y 2019, se espera que la prevalencia de EMR crezca de un 34% a un 90% entre las prácticas de oficina, y la mayoría de los hospitales han reemplazado o están en proceso de reemplazar los sistemas de papel por EMRs (Payakachat et al., 2016). Varias organizaciones y centros médicos académicos han comenzado a aprovechar el potencial de Big Data a través de su aplicación tanto en clínicas como en campos de investigación.

En (Moskowitz, McSparron, Stone, & Celi, 2015) citan algunos ejemplos tales como: la implementación del software en *Mayo Clinic* que se desarrolló utilizando datos clínicos, incluyendo *Ambient Warning and Response Evaluation (AWARE)* que apoya las mejores prácticas en la UCI y la sala de operaciones; *Syndromic Surveillance*, que proporciona "*sniffers*" para detectar sepsis; y *YES Board*, una herramienta de gestión multipaciente que ofrece un conocimiento de la situación en tiempo real para el Departamento de Emergencias.

En *Cleveland Clinic*, se han desarrollado calculadoras médicas que tienen en cuenta la demografía de los pacientes, así como detalles sobre la condición médica con el fin de guiar a los clínicos y pacientes en la toma de decisiones con respecto a las pruebas y tratamientos (Moskowitz et al., 2015). Por último, se están realizando esfuerzos para crear bases de datos clínicos internacionales. Con el financiamiento de los NIH, el Laboratorio de Fisiología Computacional de la División de Ciencias y Tecnología de Salud de Harvard-MIT está liderando una iniciativa para crear un repositorio de acceso abierto de datos de EHR (Moskowitz et al., 2015) de las UCI de países socios, incluyendo Estados Unidos, Bélgica, Estados Unidos Reino Unido y Francia. Financiado por la Comisión Europea, el *Brain Monitoring with Information Technology (BrainIT)* ha creado un conjunto de datos básicos recogidos de 20 centros de atención neurointensivos de 11 países de Europa.

Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC) (Moskowitz et al., 2015) de acceso público, contiene datos clínicos de más de

60.000 estancias en las UCI, en *Beth Israel Deaconess Medical Center*. En dicha base de datos participan activamente clínicos en todos los niveles de capacitación, incluyendo estudiantes de medicina, residentes, becarios y profesores. Los datos médicos están asociados con científicos del Instituto de Tecnología de Massachusetts y la Escuela de Salud Pública de Harvard. El equipo de científicos bajo la supervisión de un experto en el campo de la informática clínica, extrae los datos de MIMIC y forma los análisis necesarios para responder a las preguntas que surgen durante las rondas.

La colaboración investigativa para la base de datos de estructura de proteínas bioinformáticas (Andreu-Perez, Poon, Merrifield, Wong, & Yang, 2015), es un archivo mundial de datos estructurales de macromoléculas biológicas, que proporciona acceso a las estructuras tridimensionales de las mismas, así como la integración con externos biológicos, como las bases de datos de genes y fármacos (Rose et al., 2011). Paralelamente a esta evolución, la base de datos de metabolitos humanos (Wishart et al., 2013) consta con más de 40 000 anotaciones de metabolitos. Proporciona datos experimentales de concentración de metabolitos y análisis, mediante espectrometría de masas y resonancia magnética nuclear (Wishart et al., 2013). Las bases de datos como tal facilitan en gran medida la traducción de la información al conocimiento para transformar la práctica clínica, en particular para las enfermedades metabólicas, como la diabetes y las enfermedades de las arterias coronarias.

El proyecto internacional 1000 Genomas es una iniciativa gubernamental lanzada en 2008 (Costa, 2014) que tiene como objetivo secuenciar el genoma completo de miles de personas, y continúa creciendo como el mayor conjunto de datos en todo el mundo sobre la variación genética humana. Además, los datos de este proyecto se combinan con genotipos de datos para crear un gran repositorio de datos en biomedicina (Buchanan, Torstenson, Bush, & Ritchie, 2012).

Otro proyecto en curso de Big Data en biomedicina es *ENCODE* (Maher, 2012). Su objetivo principal es cartografiar y caracterizar cómo funciona todo el genoma humano. Los datos generados por *ENCODE* destacan las funciones bioquímicas de aproximadamente el 80% del genoma humano, con un enfoque particular en las regiones de codificación de proteínas del ADN (Costa, 2014).

Clinical Practice Research Datalink (CPRD) (Lu & Keech, 2015) es el nuevo servicio de datos de observación y de investigación intervencionista financiado conjuntamente por el *National Health Service (NHS)* y *Medicines and Healthcare Products Regulatory Agency (MHRA)*. Desde que se originó a mediados de los años ochenta, el CPRD es una de las bases de datos de atención primaria más validadas y utilizadas con mayor intensidad del mundo. Comprende aproximadamente 14 millones de pacientes anónimos registrados con 660 prácticas de atención primaria repartidas en todo el Reino Unido. Se utiliza en todo el mundo para la investigación entre las organizaciones clínicas y las principales instituciones académicas.

3.7 Resultados

Los resultados obtenidos en la revisión de la literatura referente a las principales fuentes de Big Data existentes en el sector de la salud y las bases de datos encontradas se resumen en la Tabla III y IV respectivamente.

Tabla III- Fuentes de Big Data. Fuente: Propia

Fuentes de Big Data	Publicación	Año	Descripción de la fuente
Internet	Huang et al.	2015	Aporta datos tales como, historial de navegación, historial de búsqueda e historial de compras, etc.
Teléfonos móviles	Huang et al.	2015	Aportan datos a través de sensores como GPS, acelerómetro, giroscopio, cámara, micrófono, etc. Aplicaciones móviles de salud permiten recolectar gran cantidad de datos del paciente.
Redes sociales	Young	2014	Facebook, Twitter y LinkedIn proporcionan una enorme cantidad de datos en el ámbito de la salud pública, los cuales se utilizan para el análisis de la propagación/transmisión de enfermedades.
Datos biomédicos	Nambiar et al.	2013	Aporta datos de años de investigación y desarrollo de compañías farmacéuticas, los registros médicos electrónicos (EMR) y registros electrónicos de salud (EHR).

Tabla IV- Bases de Datos en Salud. Fuente: Propia

Bases de Datos en Salud	Publicación	Año	Descripción de las Bases de Datos
Base de Datos estructura de proteínas bioinformática	Andreu-Pérez et al.	2015	Archivo mundial de datos estructurales de macromoléculas biológicas. Proporciona acceso a las estructuras tridimensionales de las mismas, así como la integración con externos biológicos, como las bases de datos de genes y fármacos.
Proyecto Internacional 1000 Genomas	Costa	2014	Continúa creciendo como el mayor conjunto de datos en todo el mundo sobre la variación genética humana. Los datos de este proyecto se combinan con genotipos de datos para crear un gran repositorio en biomedicina.
CPRD	Lu & Keech	2015	Base de datos de atención primaria más validada y utilizada con mayor intensidad del mundo. Se utiliza para la investigación entre las organizaciones clínicas y las principales instituciones académicas.
ENCODE	Maher	2012	Los datos generados destacan las funciones bioquímicas de aproximadamente el 80% del genoma humano, con un enfoque particular en las regiones de codificación de proteínas del ADN.
MIMIC	Moskowitz et al.	2015	Contiene datos clínicos de más de 60.000 estancias en las unidades de cuidados intensivos (UCI), en el Centro Médico Beth Israel Deaconess. Fomenta la colaboración en línea de estudiantes de medicina, residentes, becarios y profesores.
NDAR	Payakachat et al.	2016	Mayor repositorio de datos de neurociencia y genómica para la investigación del autismo. Contienen una gran colección de evaluaciones clínicas, comportamiento y resultados de salud de nuevas intervenciones.
Base de Datos de metabolitos humanos	Wishart et al.	2012	Proporciona datos experimentales de concentración de metabolitos y análisis, mediante espectrometría de masas y Resonancia Magnética Nuclear.

Capítulo 4

Técnicas de Big Data en el sector sanitario

En este capítulo se definen las plataformas de Big Data y la analítica predictiva. Se presentan las técnicas de minería de datos, las máquinas de aprendizaje así como las aplicaciones del mismo en el entorno de la salud.

4.1 Plataformas de Big Data

El uso de los datos masivos requiere la utilización de nuevas herramientas tecnológicas para su captura desde las diferentes fuentes y sistemas, así como su transformación, almacenamiento, análisis, visualización, etc. La plataforma de código abierto *Apache Hadoop* (Merelli et al., 2014) es la que ha liderado desde un principio los distintos proyectos de software especializado en Big Data. Ha sido adoptada tanto por la comunidad de desarrolladores de aplicaciones de software libre como por los principales proveedores de software propietario de bases de datos (*Oracle, IBM y Microsoft*).

4.1.1 Hadoop

La plataforma *Hadoop* (Cunha et al., 2015) fue diseñada para manejar grandes cantidades de datos. Su tecnología utiliza una metodología de división y conquista para el procesamiento, al manejar complejos datos no estructurados que usualmente no encajan en tablas relacionales. Hadoop se compone de dos componentes principales, *MapReduce* y *Hadoop Distributed File System* (HDFS).

MapReduce es la solución de Google para el procesamiento de Big Data (O'Driscoll, Daugelaite, & Sleator, 2013) y se desarrolló como gran proveedor de motores de búsqueda de Internet, indexando miles de millones de páginas web de una manera rápida y significativa. La implementación de *MapReduce* se basa en modelos de programación para procesar grandes datos o conjuntos de datos dividiéndolos en pequeños bloques de tareas (Saravana Kumar, Eswari, Sampath, & Lavanya, 2015). Utiliza algoritmos distribuidos, en un grupo de equipos de un clúster, para procesar grandes conjuntos de datos. Consta de dos funciones:

- La función *Map* que reside en el nodo maestro y luego divide los datos o tareas de entrada en subtareas más pequeñas, distribuyendo a los nodos que procesan las tareas más pequeñas y pasan las respuestas al nodo maestro. Las subtareas se ejecutan en paralelo en varios equipos.

- La función *Reduce* recopila los resultados de todas las subtareas y las combina para producir un resultado final agregado, que devuelve como respuesta a la consulta original.

HDFS (Huang et al., 2015) es un tipo de sistema de archivos de clúster que está diseñado para almacenar de forma fiable una gran cantidad de datos entre máquinas. Separa los metadatos del sistema de archivos y los datos de la aplicación. Todos los nodos contenidos en un clúster Hadoop están completamente conectados y se comunican entre sí mediante protocolos basados en TCP. En comparación con los sistemas de archivos distribuidos tradicionales, HDFS tiene

dos ventajas importantes (Huang et al., 2015): Altamente tolerante a fallos y datos a gran escala. Utiliza una arquitectura maestro-esclavo con cada grupo, que consiste en un *NameNode* que gestiona las operaciones del sistema de archivos y el *DataNode*, que gestiona el almacenamiento de datos en nodos de computación individuales, por lo que es la opción más útil en la gestión de Big Data y Twitter. La Figura 4.1 representa la arquitectura general de *Apache Hadoop* (Cunha et al., 2015) que se utiliza para procesar Big Data en salud.

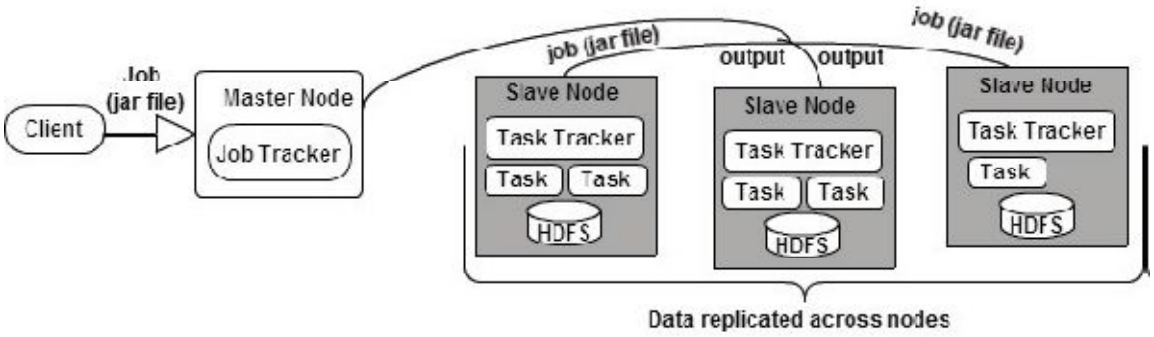

Figura 4.1: Arquitectura general de *Apache Hadoop*. Fuente: (Cunha et al., 2015)

Los datos se almacenan en una arquitectura HDFS, en la que cada archivo de datos se distribuye por varios nodos, conectados a través de una red de alta velocidad. El modelo tiene un *Master Node* que regula la distribución de la información manejada a los *Slave Node* (Cunha et al., 2015). Los nodos esclavos son responsables de las operaciones de tareas en el sistema de archivos (lecturas / escrituras) tales como la creación de bloques, la eliminación, la replicación de datos y la comprobación de integridad de datos. Existen varios *Task Trackers* que informan y ayudan al progreso de la tarea al *Job Tracker*, haciendo que el sistema sea tolerante a fallos y reduzca la pérdida de datos en las operaciones de tareas. Las operaciones de *MapReduce*, se aplican de forma distribuida a los datos y los resultados se fusionan y reducen.

Los clústeres *Hadoop* se pueden ejecutar en infraestructura privada, sin embargo las ofertas públicas como el servicio *Amazon Elastic MapReduce* están demostrando popularidad, permitiendo a los usuarios procesar de forma fácil y rentable grandes conjuntos de datos, y aplicar técnicas analíticas adicionales, aprendizaje automático así como análisis estadístico (O'Driscoll et al., 2013).

Un clúster de *Hadoop* divide los datos en partes pequeñas y los distribuye entre los distintos servidores/nodos. Los datos de un cluster *Hadoop* se descomponen en bloques más pequeños (denominados bloques) y se distribuyen por todo el cluster (Cunha et al., 2015). De esta manera, las funciones *map* y *reduce* se pueden ejecutar en subconjuntos más pequeños de conjuntos de datos más grandes, proporcionando la escalabilidad que se necesita para el procesamiento de Big Data.

4.1.2 Hive

Hadoop fue la base para otras soluciones de alto nivel como *Apache Hive* (Merelli et al., 2014), una plataforma de código abierto que proporciona un resumen de datos, consulta y análisis. *Apache Hive* soporta el análisis de grandes conjuntos de datos almacenados en HDFS y sistemas de archivos compatibles como el sistema de archivos Amazon S3. Proporciona un lenguaje semejante a SQL conocido como *HiveQL* (Merelli et al., 2014). Los usuarios pueden escribir fácilmente consultas *HiveQL* para recopilar y analizar datos con diversos propósitos, tales como inteligencia empresarial, resumen de datos o minería de datos interactiva. *Hive* traduce estas consultas a trabajos de *MapReduce* y las envía a *Hadoop* para su ejecución (Grover et al., 2015). También permite *user defined functions (UDFs)* y *aggregation functions (UDAFs)* escritas en Java para realizar operaciones que no son compatibles con *HiveQL*.

4.1.3 Mongo DB

MongoDB (Cunha et al., 2015) es una base de datos *NoSQL* diseñada para reemplazar a las SQL tradicionales, que soporta formatos de texto, no relacional. Está especialmente diseñado para proporcionar soluciones de almacenamiento de datos escalables y de alto rendimiento, y utiliza un lenguaje de consulta versátil con una sintaxis algo similar a los lenguajes de consulta orientados a objetos. Dados los datos no estructurados de Twitter, las soluciones suelen requerir sistemas de gestión "*NoSQL*", que modelan datos sin usar relaciones tabulares, contrastando bases de datos relacionales.

4.1.4 Spark

Spark proporciona una plataforma escalable de análisis de datos con computación en memoria. Se ha demostrado que la computación en memoria proporciona un acceso más rápido a los datos eliminando la sobrecarga (Patel & Sharma, 2014). Está diseñado para aplicaciones explícitas como algoritmos de aprendizaje automático y procesamiento del lenguaje natural.

Spark se ejecuta en *Apache Mesos*, un gestor de clústeres en el que las aplicaciones *Spark* coexisten con *Hadoop*. Utiliza dos tipos de operaciones Acción y Transformación. La acción es similar a reducir mientras que la transformación es similar al mapeo y la operación de caché. Se desarrolla y soporta en el lenguaje de programación funcional *Scala* (Patel & Sharma, 2014), utilizado para proporcionar un entorno distribuido e iterativo.

Además el sistema *Spark* contiene una de las herramientas de análisis avanzadas más fiables, llamada *MLlib* (biblioteca de aprendizaje automático) y *SparkR*, que es una nueva estructura de modelado para manejar Big Data (Elsebakhi et al., 2015).

4.1.5 Storm, HPCC, SAP-HANA y Pig

Storm es lanzado por Twitter como una fuente abierta en septiembre de 2011. Se implementa en el lenguaje *Clojure* para apoyar el entorno de aprendizaje de máquinas (Patel & Sharma, 2014). *Ruby* y *Python* están soportados para hacer aplicaciones en *Storm*. La idea clave es que se utiliza para los procesos de

transmisión. *Storm* no usa ningún concepto de almacenamiento, sino que utiliza datos semi-estructurados, no estructurados y estructurados en conjunto.

High Performance Computing Cluster (HPCC) (Patel & Sharma, 2014) utiliza *MapReduce* para el análisis de Big Data y funciona con *Enterprise Control Language (ECL)*, un lenguaje de programación declarativo. ECL proporciona un paradigma de programación completo en el que se logra un alto paralelismo. Las dos principales ventajas proporcionadas por HPCC sobre *Hadoop* son la escalabilidad y la velocidad. En la memoria de computación para Big Data *SAP* ideó una nueva herramienta *HANA*, que procesa los datos en bloque mediante el uso de la arquitectura paralela avanzada y algoritmos, para así obtener una mayor velocidad (Patel & Sharma, 2014).

Pig (Chennamsetty, Chalasani, & Riley, 2015), es una de las herramientas comerciales que se han desarrollado para analizar grandes conjuntos de datos. Proporciona un lenguaje llamado *PigLatin* así como también un modelo de datos y los operadores correspondientes, lo que permite a los usuarios describir las consultas como programas de *PigLatin* que pueden ser compilados automáticamente en los trabajos de *Hadoop*. Existen en las literaturas muchas herramientas que no fueron mencionadas y que se utilizan para el análisis de Big Data.

4.2 Minería de Datos

En la minería de datos, hay muchas iniciativas de código abierto y dos de las más citadas son: aprendizaje de máquina escalable basado en *Apache Mahout* y minería de datos de código abierto basado en *Hadoop* (Lu & Keech, 2015).

La minería de datos en *Cloud Computing* permite a las organizaciones centralizar la gestión de software y almacenamiento de datos, lo que lleva a servicios adecuados, mejorados, fiables y seguros para los usuarios.

Las técnicas de minería de datos pueden aplicarse al conjunto de datos de asistencia sanitaria para una toma de decisiones eficaz. La metodología es comenzar con la definición del problema, la recopilación de datos y finalmente la minería de datos. Se pueden aplicar las principales técnicas de agrupación, clasificación y asociación (Chauhan & Kumar, 2013), un ejemplo de ellas es *K-means* que se utiliza para explorar la clasificación del cáncer de mama con datos genómicos. Con el aumento exponencial de los datos de atención médica, las técnicas de minería de datos no pueden ofrecer un enfoque holístico para el análisis de datos en la salud debido a su gran volumen.

Los enfoques tradicionales en el descubrimiento de patrones no son suficientes para determinar la naturaleza temporal de las enfermedades. Estas técnicas no consideran el tiempo transcurrido entre dos eventos y, por lo tanto, no pueden producir información valiosa sobre enfermedades temporalmente frecuentes, ya que no toman la dimensión temporal como una variable en su estructura (Lu & Keech, 2015).

La minería de *Time-Annotated Sequences (TAS)* es un paradigma novedoso que tiene como objetivo resolver este problema. TAS son patrones secuenciales

donde cada transición entre dos eventos se anota con un típico tiempo de transición que se encuentra frecuentemente en los datos (Lu & Keech, 2015). En particular, muestra cómo la lógica difusa y el agrupamiento jerárquico divisivo pueden usarse para extraer frecuentes patrones secuenciales con intervalos de tiempo de una serie de diagnósticos de cáncer de mama.

4.2.1 Técnicas de minería de datos sobre WSN

La amplia adopción de *Wireless Senor Networks (WSNs)* (Fouad, Oweis, Gaber, Ahmed, & Snasel, 2015) en la actualidad ha aumentado la cantidad de datos sensoriales complejizando grandes volúmenes de datos. WSN es una red inalámbrica que consiste en un gran número de nodos (pequeños sensores) densamente desplegados. La mayoría de los datos, generados a partir de las WSN, son originalmente datos de transmisión que representan mediciones o eventos que ocurren a lo largo de intervalos de tiempo.

La necesidad de extraer conocimiento de los datos de sensores, se ha convertido en una cuestión importante en los sistemas de decisión en tiempo real. Sin embargo, los recursos restringidos de estas redes, junto con sus características de transmisión continua de datos detectados, hacen que las técnicas tradicionales de minería de datos no sean aplicables.

Los algoritmos de minería de datos podrían clasificarse generalmente en un procesamiento de datos centralizado o distribuido, tal y como lo muestra la Figura 4.2.

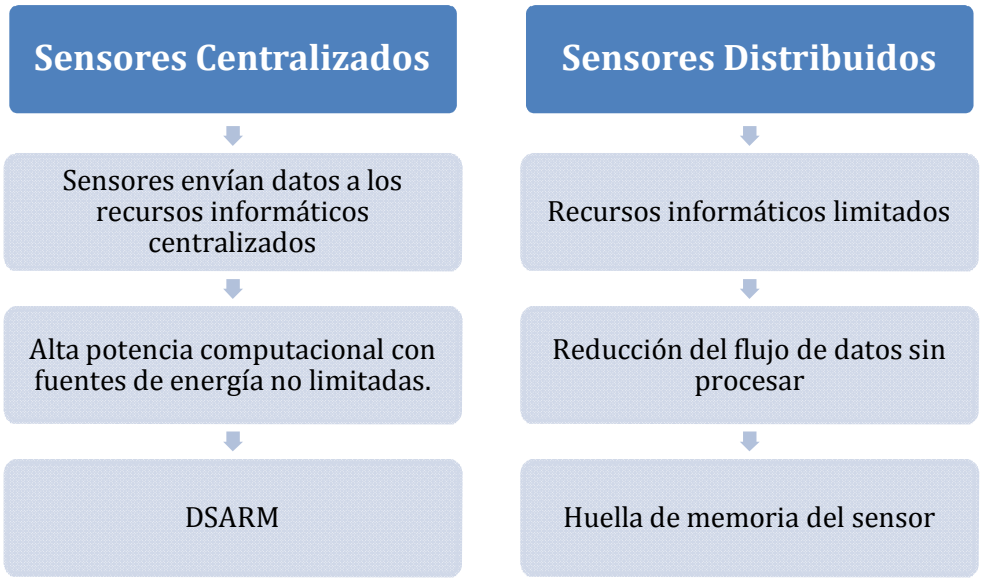

Figura 4.2: Sensores centralizados y distribuidos en el procesamiento de Minería de Datos. Fuente: Propia

En los sensores centralizados *Data Stream Association Rule Mining (DSARM)* hace uso de *Adaptive Multiple Regression (AMR)* (Fouad et al., 2015) que refleja la correlación espacial de los datos del sensor en la estimación de los que faltan. Dado que los datos detectados pueden cambiar dinámicamente, el AMR puede estimar el nivel de confianza de los datos faltantes capturando la correlación dinámica de los

ya detectados, y seleccionar la muestra óptima para regresar los coeficientes apropiados de la función de estimación.

En el enfoque de procesamiento de datos de sensores distribuidos, cada nodo utiliza sus recursos informáticos limitados para realizar el proceso de minería. La principal ventaja es la reducción de flujos de datos sin procesar que se van a entregar al nodo sink. Sin embargo, puede agotar los recursos de la red en términos de huella de memoria y consumos de energía. En cuanto a la huella de memoria del sensor, el número creciente de patrones frecuentes requiere un gran espacio de la memoria como se indica en la referencia (Fouad et al., 2015).

4.2.2 Técnicas de clasificación

La minería de datos es el proceso de extraer conocimiento de Big Data. Con el tiempo se almacenan enormes datos y podemos aplicar algunas técnicas de inteligencia sobre estos para apoyar el proceso de toma de decisiones. Algunas de las tareas de análisis de datos tienen que encontrar valores continuos o valores ordenados para la variable, este modelo se denomina predictor y las tareas se denominan predicción numérica (Al-Janabi, Patel, Fatlawi, Kalajdzic, & Al Shourbaji, 2015). Hay muchos métodos de aprendizaje de máquina que se pueden utilizar con técnicas de predicción, en la Figura 4.3 mostramos las principales.

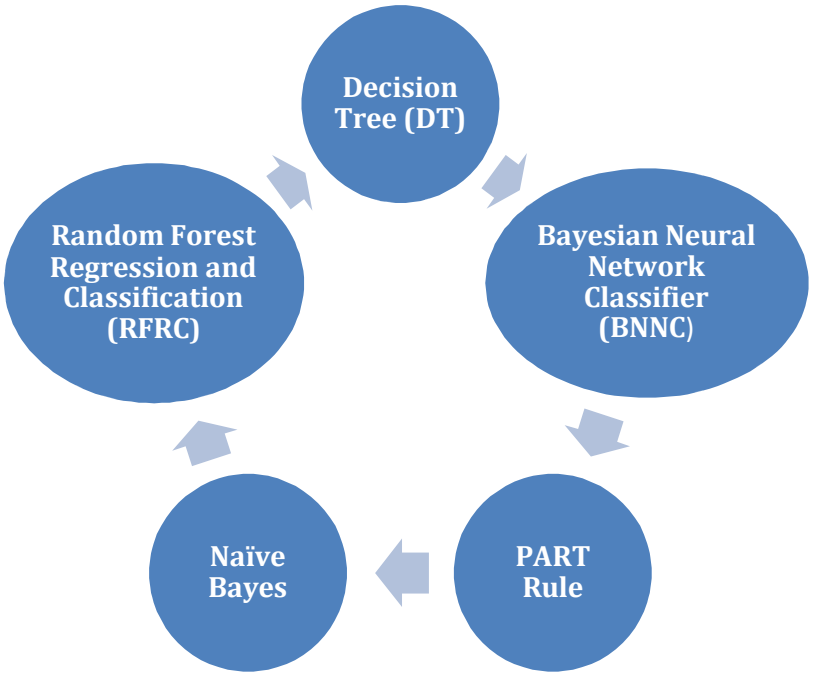

Figura 4.3: Técnicas de predicción. Fuente: Propia

La técnica *DT* (Al-Janabi et al., 2015) se utiliza para clasificar los datos de una manera más fácil y comprensible. Para clasificar el problema, el valor de la variable de destino debe ser modificado usando alguna variable interesante. Realiza recursivamente un procedimiento de división de arriba hacia abajo tratando de construir un árbol, cada rama representa una pregunta sobre el valor de una de las variables. Hay muchas implementaciones de *DT* tales como ID3, J48 y CART.

RFRC tiene una precisión tan buena como *Adaboost*. Proporciona estimaciones internas de error, fuerza, correlación y variables. Es simple y fácilmente paralelizado(Al-Janabi et al., 2015). Cuando se utiliza *Random Forest*, su tiempo de entrenamiento computacional requiere de horas a días, especialmente para conjuntos más grandes y depende de la aleatoriedad.

BNNC es uno de los métodos de clasificación que utilizó la teoría de probabilidades y de gráficos. El gráfico se utiliza como una parte cualitativa con el nodo que representa las variables y los arcos que representan la dependencia entre ellos. Un conjunto de parámetros se utiliza como parte cuantitativa para representar las distribuciones de probabilidad condicional (Al-Janabi et al., 2015). BNNC puede manejar datos incompletos y debido a que se combinan de forma probabilística con el conocimiento previo, el método es bastante robusto para modelar el ajuste excesivo.

La técnica *Naïve Bayes* se basa en el denominado teorema bayesiano y es particularmente adecuada cuando la dimensionalidad de las entradas es alta (Al-Janabi et al., 2015). Es uno de los métodos de clasificación que utiliza la teoría de probabilidades para resolver el problema de la relación no determinista entre las variables establecidas y la clase objetivo. A pesar de su simplicidad, a menudo puede superar métodos de clasificación más sofisticados. Se compone de estructuras simples en forma de estrella, por lo que la estimación de parámetros es su núcleo de aprendizaje con datos completos.

PART Rule es una combinación entre el algoritmo C4.5 de *DT* y el algoritmo de *RIPPER* para generar reglas, creando repetidamente *DT* parciales. El conjunto de reglas de este método tiene el mismo tamaño y exactitud de las reglas que se genera por C4.5 y es mejor que *RIPPER*, además, es más rápido porque no necesita post-procesamiento. La principal ventaja de *PART* es generar buenos conjuntos de reglas sin optimización global (Al-Janabi et al., 2015).

4.3 Máquinas de aprendizaje

El aprendizaje automático es un subcampo de la inteligencia artificial (Kanevsky et al., 2016). En el ámbito de la medicina, el sistema de computación cognitiva *International Business Machines Watson Health (IBM)* ha utilizado la tecnología de aprendizaje de máquinas para crear un sistema de apoyo en la toma de decisiones de los médicos que tratan a pacientes con cáncer, con la intención de mejorar la precisión diagnóstica y reducir los costos.

Los modelos predictivos de aprendizaje automático pertenecen al dominio del aprendizaje supervisado, donde el algoritmo ha sido entrenado usando ejemplos de entradas y salidas deseadas, permitiendo el mapeo de entradas futuras a salidas. El objetivo de este proceso es un modelo matemático único capaz de predecir valores deseados a partir de nuevos datos (Kanevsky et al., 2016).

Este tipo de aprendizaje automático se ha aplicado en el campo de la genética molecular y la genómica para organizar e interpretar grandes cantidades de información genética, así como también en muchos otros campos de la medicina.

4.3.1 Redes funcionales

En los últimos años, las redes funcionales se convirtieron en algoritmos computacionales de aprendizaje de máquinas (Elsebakhi et al., 2015). Es una estructura útil para resolver una amplia gama de problemas en probabilidades, estadísticas, procesamiento de señales, reconocimiento de patrones, aproximaciones de funciones, previsión de inundaciones en tiempo real, ciencia, bioinformática, medicina, ingeniería de estructuras y otras aplicaciones de negocios e ingeniería.

Para tratar con las redes funcionales en la identificación de resultados continuos o categóricos se requiere una comprensión específica del problema, la arquitectura basada en datos, las comunicaciones intermedias y los mecanismos de flujo de datos hacia el objetivo deseado. Los procesos de capacitación utilizan dos tipos de aprendizaje: aprendizaje estructural y aprendizaje paramétrico. En el aprendizaje estructural, la topología inicial se elige sobre la base de algunas propiedades disponibles para el diseñador, y la arquitectura final se modifica simplemente mediante la ecuación funcional (Elsebakhi et al., 2015). En el aprendizaje paramétrico, usualmente las funciones de activación (neuronas) se eligen considerando la combinación de funciones "base", y luego se estiman usando distintos criterios de optimización, a saber, mínimos cuadrados, pendiente más pronunciada, gradiente conjugado y mínimos -máximos.

Los datos biomédicos reales se utilizan para investigar el desempeño del aprendizaje de máquinas a gran escala, basado en redes funcionales; utilizando la técnica de optimización de *Newton-Raphson* para identificar individuos de quimioterapia y optimizar el correcto curso de quimioterapia contra el cáncer.

4.3.2 Aplicaciones de Máquinas de Aprendizaje

4.3.2.1 Cirugía Plástica

Existen varias aplicaciones del aprendizaje mecánico relacionadas con la cirugía plástica, en la Figura 4.4 se muestran algunas de ellas. En la cirugía de quemaduras se desarrolló un método para determinar con precisión el tiempo de cicatrización en la lesión por quemaduras (Kanevsky et al., 2016). Usando la espectrometría de reflectancia y una red neuronal artificial, los investigadores desarrollaron un modelo para predecir si una quemadura requeriría aproximadamente 14 días para sanar, sirviendo en última instancia como una aproximación para la evaluación de la profundidad de la quemadura para la planificación quirúrgica.

Durante el entrenamiento, las *Artificial Neural Network (ANN)*, de manera similar a las neuronas biológicas, toman parte en un proceso llamado retropropagación, en el que el peso de las conexiones entre nodos se ajusta en función de la diferencia entre los valores de salida de *Artificial Neural Network* y los valores objetivos conocidos (Kanevsky et al., 2016). Este proceso asegura que la salida de *Artificial Neural Network* esté lo más cerca posible de los valores objetivos deseados.

Figura 4.4: Aprendizaje automático aplicado a los tipos de Cirugía Plástica. Fuente: Propia

En la actualidad, se está estudiando el aprendizaje automático para facilitar el diagnóstico automatizado de la craneosinostosis no sindrómico (Kanevsky et al., 2016). Se realizó un algoritmo de análisis discriminante lineal, regularizado para diagnosticar la craneosinostosis y distinguir entre diferentes tipos, utilizando un índice de fusión de sutura craneal junto con deformación y curvatura discrepante. Otra potencial aplicación de aprendizaje mecánico para la cirugía craneofacial implica la identificación de genes candidatos en casos no sindrómicos de fisura labial y palatina. El aprendizaje automático tiene el potencial de descubrir genes candidatos y secuencias reguladoras, previamente desconocidas para el labio leporino y palatino no sindrómico, permitiendo una mejor comprensión de la patogénesis de esta condición.

El aprendizaje automático también tiene aplicaciones potenciales en áreas más subjetivas de la cirugía plástica, como la estética. Los algoritmos *DT* evalúan un conjunto de atributos descriptivos que, la investigación incluyó diferentes relaciones faciales e intentó determinar características faciales atractivas, más estrechamente relacionadas con las variables objetivo postoperatorio (Kanevsky et al., 2016). Al someter el conjunto de imágenes de prueba, se demostró que el clasificador automatizado tenía una alta precisión, por tanto puede servir como una herramienta predictiva para estimar la belleza percibida de un paciente después de la cirugía estética, proporcionando una medida cuantitativa para establecer expectativas y posiblemente disuadir a los pacientes de someterse a procedimientos que ofrezcan mejoras marginales.

El espectro de aplicaciones potenciales del aprendizaje de la máquina para la microcirugía no se limita solamente al monitoreo postoperatorio. El aprendizaje mecánico también podría beneficiar la consulta preoperatoria y la planificación quirúrgica de la microcirugía (Kanevsky et al., 2016).

4.3.2.2 Plataforma Neuroinformática

En la referencia(Melethadathil, Chellaiah, Nair, & Diwakar, 2015)se plantea el desarrollo de una plataforma de neuro-informática llamada *Neuroinformatics Natural Language Processing (NeuroNLP)*, con una estructura de aprendizaje de máquina que incluye clústeres de documentos para mejorar la eficacia de la categorización del proceso de búsqueda basado en términos de artículos de investigación minados por texto.

El estudio incluyó la clasificación de los datos de entrenamiento generados usando dos términos de búsqueda en la herramienta *NeuroNLP*: "ataxia de la función cerebelosa" y "ataxia de la fisiología del cerebelo". Durante el procesamiento de la consulta de búsqueda *NeuroNLP* creó una tabla de base de datos que contiene los detalles sobre la consulta de búsqueda, la cual se exportó y se utilizó como conjunto de datos para el procesamiento previo (Melethadathil et al., 2015). Estos puntos de datos relacionados con la consulta se combinaron y se utilizaron como datos de formación en oncología y neurociencia.

La Figura 4.5 muestra algunos de los algoritmos utilizados en el desarrollo de la plataforma.

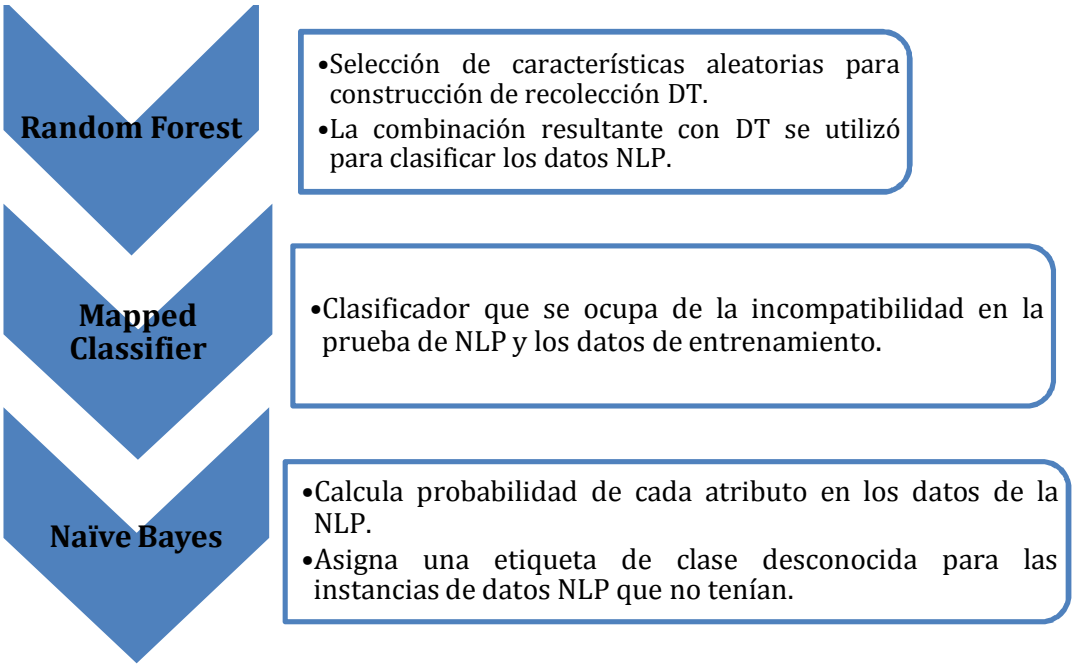

Figura 4.5: Algoritmos de la plataforma *NeuroNLP*. Fuente: Propia

4.4 Otras Aplicaciones de Técnicas de Big Data en la Salud

4.4.1 Modelo predictivo usando RBFNN

En (Pombo, Garcia, Felizardo, & Bousson, 2014) se plantea un estudio de caso basado en un modelo predictivo usando *Radial Basis Function Neural Network (RBF NN)* en combinación con una técnica de filtrado que apunta a la estimación de la forma de onda del *Electrocardiogram (ECG)*. El método propuesto reveló la

idoneidad para apoyar a profesionales de la salud en la toma de decisiones y prácticas clínicas.

El modelo predictivo se basa en el RBFNN combinado con el filtro *Savitzky-Golay* (Pombo et al., 2014) para el suavizado y la diferenciación que realizan automáticamente el ajuste polinomial de mínimos cuadrados en ejecución, cuando la señal de entrada se convoluciona con los coeficientes del filtro.

El modelo propuesto reveló su idoneidad para predecir la forma de onda ECG, basada en una muestra reducida de datos y por lo tanto representa una herramienta útil para profesionales de la salud en la toma de decisiones clínicas (Pombo et al., 2014). El modelo abarca una RBFNN y una técnica de filtrado, que asume gran importancia porque la señal de ECG es a menudo perturbada por el ruido de banda ancha, compuesto principalmente de interferencias de alta frecuencia debido al electromagnetismo y la conexión a tierra. Por lo tanto, es importante minimizar la distorsión en las formas de onda de característica para mantener aquellas que serían de mayor interés en términos de análisis, mientras que al mismo tiempo elimina el ruido.

Este modelo reveló ser preciso y adecuado cuando se aplica en el contexto de salud y bienestar. Estudios adicionales deben ser dirigidos con el objetivo de evaluar la combinación de varios parámetros como ECG y temperatura de la piel, en relación con la predicción de las condiciones de salud de los pacientes.

4.4.2 Técnicas de imágenes

La imagen molecular es una técnica no invasiva de los eventos celulares y subcelulares (Belle et al., 2015), tiene el potencial para el diagnóstico clínico de estados de enfermedad como el cáncer. Sin embargo, para que sea clínicamente aplicable a los pacientes, la interacción de la radiología, la medicina nuclear y la biología es crucial, lo que podría complicar su análisis automatizado.

La integración de imágenes de diferentes modalidades y otra información clínica y fisiológica podría mejorar la precisión del diagnóstico y la predicción de resultados de la enfermedad. En (Gessner, Frederick, Foster, & Dayton, 2013) se exploran los avances en la imagen neurovascular y el papel de la tomografía computarizada multimodal o la resonancia magnética, incluyendo la angiografía y la perfusión de imágenes en la evaluación del trastorno cerebro vascular y el logro de la medicina de precisión. Se ha utilizado una resonancia magnética tardía para la evaluación exacta de la cicatriz del infarto de miocardio (Belle et al., 2015). Para este tipo de enfermedad, *Electroanatomic Mapping (EAM)* puede ayudar a identificar la extensión subendocárdica del infarto, la evaluación simultánea de todas las técnicas de imagen disponibles es una necesidad insatisfecha.

Se ha desarrollado un sistema de soporte de decisión asistido por computadora (Belle et al., 2015) que podría ayudar a los médicos a proporcionar una planificación precisa del tratamiento para los pacientes que sufren de *Traumatic Brain Injury (TBI)*. La tecnología propuesta está diseñada para ayudar en la detección temprana del cáncer mediante la integración de información molecular y fisiológica con información anatómica. Usando esta técnica de imagen para pacientes con cáncer de ovario avanzado, la precisión del predictor de

respuesta a un tratamiento especial se ha incrementado en comparación con otros criterios clínicos o histopatológicos.

4.4.3 Grandes aplicaciones de datos en genómica

El análisis genómico que utiliza microarrays ha sido exitoso en el análisis de rasgos a través de una población contribuyendo con éxito en tratamientos de enfermedades complejas como la enfermedad de Crohn y la degeneración muscular (Belle et al., 2015).

En la actualidad se están llevando a cabo iniciativas a lo largo de los años para integrar los datos clínicos desde el nivel genómico al nivel fisiológico de un ser humano (Belle et al., 2015). Estas iniciativas ayudarán a brindar atención personalizada a cada paciente.

Un único genoma humano completo obtenido por *Next-Generation Sequencing (NGS)* es típicamente de 3 GB (Andreu-Perez et al., 2015). Dependiendo de la profundidad media de cobertura, esto puede variar hasta 200 GB, por lo que es una fuente clara de Big Data para la salud.

La secuenciación del genoma entero por NGS es importante para el estudio de enfermedades complejas como el cáncer. En el tratamiento del cáncer ha sido un problema que los fármacos a menudo tienen respuestas de tratamiento heterogéneo incluso para el mismo tipo de cáncer, y algunos fármacos sólo muestran una sensibilidad profunda en un pequeño número de pacientes (Andreu-Perez et al., 2015). Actualmente, se han generado conjuntos de datos de genómica personal y farmacogenómica a gran escala para descubrir patrones de señalización únicos de pacientes individuales y descubrir fármacos que apuntan a estos patrones únicos.

Hay múltiples enfoques para el análisis escalar de datos del genoma utilizando un sistema dinámico. Las aplicaciones desarrolladas para la inferencia de redes en biología de sistemas para grandes aplicaciones de datos pueden dividirse en dos grandes categorías que consisten en la reconstrucción de redes metabólicas y redes reguladoras de genes (Belle et al., 2015).

4.5 Resultados

Al término de ver realizado una revisión de la literatura se resumen en la Tabla V las principales técnicas de Big Data empleadas en el sector sanitario así como en la Tabla VI las plataformas existentes para el desarrollo de Big Data tanto en la salud como en otro sector de la sociedad.

Tabla V- Técnicas de Big Data. Fuente: Propia

Técnicas de Big Data	Publicación	Año	Descripción de la técnica
Decision Tree (DT)	Al-Janabi et al.	2014	Clasificación de datos.
Bayesian Neural Network Classifier (BNNC)	Al-Janabi et al.	2014	Clasificación de datos; utiliza la teoría de probabilidades y de gráficos.
Naïve Bayes	Al-Janabi et al.	2014	Clasificación de datos, que crea una guía paso a paso para determinar la salida de una nueva instancia de datos.
Part Rule	Al-Janabi et al.	2014	Combinación entre el algoritmo C4.5 de DT y el algoritmo de RIPPER, para generar reglas creando repetidamente DT parciales.
Random Forest Regression and Classification (RFRC)	Al-Janabi et al.	2014	Proporciona estimaciones internas de error, fuerza, correlación y variables.
K-means	Chauhan & Kumar	2013	Clasificación del cáncer de mama con datos genómicos.
Newton-Raphson	Elsebakhi et al.	2015	Identifica y optimiza la quimioterapia contra el cáncer.
Adaptive Multiple Regression (AMR)	Fouad et al.	2015	Estima el nivel de confianza de los datos del sensor faltantes capturando la correlación dinámica de los ya detectados. Selecciona la muestra óptima para regresar los coeficientes apropiados de la función de estimación.
Mapped Classifier	Melethadathil et al.	2015	Clasificación de datos.

Tabla VI- Plataformas de Big Data. Fuente: Propia

Plataformas de Big Data	Publicación	Año	Descripción de la técnica
MongoDB	Cunha et al.	2015	Proporciona soluciones de almacenamiento de datos escalables y de alto rendimiento
Pig	Chennamsetty et al.	2015	Analiza grandes conjuntos de datos.
Hadoop Distributed File System (HDFS)	Huang et al.	2015	Diseñado para almacenar de forma fiable una gran cantidad de datos entre máquinas. Separa los metadatos del sistema de archivos y los datos de la aplicación.
Hive	Merelli et al.	2014	Plataforma de código abierto que proporciona un resumen de datos, consulta y análisis. Analiza grandes conjuntos de datos almacenados en HDFS y sistemas de archivos compatibles como el sistema de archivos Amazon S3.
Spark	Patel & Sharma	2014	Proporciona una plataforma escalable de análisis de datos con computación en memoria. Diseñado para aplicaciones explícitas como algoritmos de aprendizaje automático y procesamiento del lenguaje natural.
Storm	Patel & Sharma	2014	Apoya el entorno de aprendizaje de la máquina. Utiliza datos semi-estructurados, no estructurados y estructurados en conjunto.
HPCC	Patel & Sharma	2015	Analiza grandes conjuntos de datos. Proporciona un paradigma de programación completo en el que se logra un alto paralelismo.
SAP - HANA	Patel & Sharma	2015	Procesa los datos en bloque mediante el uso de la arquitectura paralela avanzada y algoritmos, para así obtener una mayor velocidad.
Hadoop MapReduce	Saravana Kumar et al.	2015	Procesamiento de complejos datos no estructurados. Se desarrolló como gran proveedor de motores de búsqueda de Internet, indexando miles de millones de páginas web de una manera rápida y significativa.

Capítulo 5

Técnicas de Big Data en la predicción de enfermedades crónicas. Aplicaciones

Big Data es una de las tecnologías actuales que tienen el potencial de cambiar radicalmente la forma en que las organizaciones utilizan la información, para mejorar la experiencia del cliente y transformar sus modelos de negocio (Koppad & Kumar, 2016). El sector de la salud ha estado manejando grandes cantidades de datos y es impulsado en gran medida por el cumplimiento, los requisitos reglamentarios, el mantenimiento de registros y aspectos similares de la atención al paciente. El objetivo es presentar a los analistas y profesionales de la salud los avances en el campo de la computación para manejar de forma efectiva los datos y hacer inferencias a partir de Big Data.

La mayoría de las personas en el mundo se ven afectadas por enfermedades crónicas. En países extranjeros como Estados Unidos la mayor parte de las muertes ocurren debido a enfermedades crónicas. Algunas de ellas son alergia, cáncer, asma, enfermedades del corazón, glaucoma, obesidad, enfermedades virales como Hepatitis C y VIH/SIDA; de todas ellas la diabetes y las enfermedades coronarias son las más peligrosas (Ramkumar, Prakash, & Sangeetha, 2016).

En este capítulo se plantea identificar de las técnicas existentes en Big Data, cuáles son las más utilizadas en la predicción de las enfermedades crónicas.

5.1 Minería de Datos relacionada con enfermedades crónicas

La minería de datos es una de las áreas de investigación más alentadoras con el propósito de encontrar información útil de voluminosos conjuntos de datos. Sus aplicaciones incluyen la detección de la anomalía, análisis de datos financieros, análisis de datos médicos, análisis de redes sociales, análisis de mercado etc. (Kunwar, Chandel, Sabitha, & Bansal, 2016).

Es particularmente útil en el campo médico cuando no se encuentra evidencia disponible que favorezca una opción de tratamiento en particular. La industria de la salud está generando una gran cantidad de datos sobre pacientes, enfermedades, hospitales, equipos médicos, costos de tratamiento, etc., que requieren procesamiento y análisis para la extracción del conocimiento(Kunwar et al., 2016).

Combina el análisis estadístico, el aprendizaje automático y la tecnología de bases de datos para extraer patrones ocultos y relaciones de grandes bases de datos (Palaniappan & Awang, 2008). En (Palaniappan & Awang, 2008) la definen como "un proceso de selección, exploración y modelización de Big Data, para descubrir regularidades o relaciones que son desconocidas en un primer momento, con el objetivo de obtener resultados claros y útiles para el propietario de la base de datos".

La minería de datos viene con un conjunto de herramientas y técnicas como la clasificación, el agrupamiento, la regresión y la asociación, que han sido utilizadas en el campo médico para detectar y predecir la progresión de la enfermedad, y para tomar decisiones con respecto al tratamiento del paciente (Kunwar et al., 2016).

Cada técnica tiene un propósito diferente dependiendo del objetivo de modelado. Los dos objetivos de modelado más comunes son la clasificación y la predicción. Los modelos de clasificación predicen etiquetas categóricas (discretas, no ordenadas), mientras que los modelos de predicción predicen funciones de valor continuo (Palaniappan & Awang, 2008). La clasificación es un método de aprendizaje supervisado, es el proceso que clasifica los objetos o datos en grupos, cuyos miembros tienen una o más características en común. Las técnicas de clasificación son *Support Vector Machines (SVM)*, *Decision Tree*, *Naïve Bayes*, *Artificial Neural Network* etc. (Kunwar et al., 2016). Para la asociación, las redes neuronales radiales son enfoques bien conocidos, el clúster y la optimización de búsqueda también se aplican como estrategias de minería de datos. Para la regresión, pueden usarse máquinas de análisis de componentes principales o de vector de soporte, y en cuanto al mapa de organización propia, la cuantificación vectorial y el algoritmo genético también se aplican como técnicas de agrupamiento (Sankaranarayanan & Perumal, 2014).

Mediante la aplicación de las técnicas de minería de datos, el conocimiento valioso se puede extraer del sistema de atención de salud, un ejemplo de ello es la enfermedad cardíaca (Sivagowry, Durairaj, & Persia, 2013). Se han realizado investigaciones con muchas técnicas híbridas para diagnosticar enfermedades del corazón.

La actividad de minería de datos en los sistemas de atención de salud en concreto tiene vital importancia, sin ella es difícil obtener el potencial completo de los datos recolectados a través de diversas fuentes. A continuación planteamos un enfoque de dichas técnicas aplicadas en la predicción de enfermedades crónicas.

5.1.1 Técnicas de Minería de Datos en el campo de la enfermedades crónicas

Haciendo una revisión de la literatura (Sivagowry et al., 2013), (Palaniappan & Awang, 2008), (Alfisahrin & Mantoro, 2013), (Kunwar et al., 2016), (Sankaranarayanan & Perumal, 2014) y (Koppad & Kumar, 2016) las técnicas de minería de datos que más se utilizan en la predicción de enfermedades crónicas como el corazón, cáncer de pulmón, de hígado, diabetes, enfermedad de Parkinson, etc. son *Decision Tree*, *Naïve Bayes* y *Artificial Neural Network*.

En (Sivagowry et al., 2013) se investiga la aplicación de técnicas de minería de datos en la predicción de un ataque cardíaco. La enfermedad cardíaca afecta la estructura y las funciones del corazón y es una de las principales causa de muerte en el mundo en los últimos años. El diagnóstico inicial de un ataque al corazón se realiza mediante una combinación de síntomas clínicos y cambios característicos del *Electrocardiogram (ECG)*, un registro de la actividad eléctrica del corazón, que se reflejan en la historia clínica del paciente, la cual es una fuente de datos muy utilizada para diagnosticar una enfermedad en particular. Los investigadores en el campo médico identifican y predicen la enfermedad con la ayuda de técnicas de

minería de datos (Sivagowry et al., 2013), de ahí que el objetivo de dicha investigación es examinar el uso potencial de dichas técnicas basadas en clasificación, como *Naïve Bayes*, *DT* y *Artificial Neural Network*.

Desarrollaron un prototipo llamado *Intelligent Heart Disease Prediction System (IHDPS)* (Sivagowry et al., 2013) utilizando las técnicas de minería de datos ya planteadas; el conjunto de datos proviene del repositorio de minería de datos de la *University of California, Irvine (UCI)*, contiene 76 atributos en total de los cuales solo se utilizan 14 atributos. Según el estudio (Sivagowry et al., 2013) realizado *Naïve Bayes* es la más efectiva en la predicción de la enfermedad cardíaca, ya que tiene el mayor porcentaje de predicciones correctas (86,53%). Se utiliza para crear modelos con capacidad predictiva y proporciona nuevas formas de explorar y comprender los datos, obteniendo resultados eficientes.

En (Palaniappan & Awang, 2008) la investigación está centrada en predecir la probabilidad de que los pacientes tengan una enfermedad cardíaca utilizando perfiles médicos como edad, sexo, presión arterial y azúcar en la sangre. Para ello usan el sistema IHDPS (Sivagowry et al., 2013) haciendo una comparativa entre las técnicas de minería de datos, tales como *DT*, *Naïve Bayes* y *Neural Network*. El sistema extrae el conocimiento de la base de datos *Cleveland Heart Disease*, específicamente un total de 909 registros con 15 atributos médicos (factores). Los registros se dividieron por igual en dos conjuntos de datos: datos de formación (455 registros) y datos de pruebas (454 registros) (Palaniappan & Awang, 2008). Para evitar sesgos, los registros de cada conjunto fueron seleccionados al azar; los modelos son entrenados y validados contra un conjunto de datos de prueba.

La tabla de elevación sin valor predecible y los métodos de clasificación de matriz se utilizan para evaluar la efectividad de los modelos. La comparativa tiene en cuenta el 50 % de la población y la totalidad de la misma para determinar el porcentaje de predicciones correctas. En la Tabla VII se muestran los resultados obtenidos.

Tabla VII. Comparación de Técnicas de Minería de Datos. Fuente: Propia

Técnica / %	DT	Naïve Bayes	Neural Network
50 % de la población			
Porcentaje de predicciones correctas	41.85%	47.58%	49.34%
100 % de la población			
Porcentaje de predicciones correctas	80.4%	86.12%	85.68%

La Tabla VII muestra que al ser procesada el 50% de la población, *Neural Network* da el mayor porcentaje de predicciones correctas (49.34%), seguido de *Naïve Bayes* (47.58%) y *DT* (41.85%). Si se procesa a toda la población, el modelo *Naïve Bayes* parece tener un mejor desempeño que los otros dos, ya que da el mayor número de predicciones correctas (86.12%), seguida de *Neural Network*

(85.68%) y *DT* (80.4%).Procesar menos del 50% de la población hace que las líneas de elevación para *Neural Network* y *Naïve Bayes* siempre sean más altas que las de *DT*, lo que indica que son mejores al hacer un alto porcentaje de predicciones correctas (Palaniappan & Awang, 2008). Para algunos grupos de población, *Neural Network* parece ser mejor que *Naïve Bayes* y viceversa.

En (Palaniappan & Awang, 2008) se definen cinco metas de minería basadas en inteligencia de negocios y exploración de datos. Los objetivos se evalúan con los modelos entrenados y dichos modelos podrían responder a consultas complejas, cada una con su propia fuerza con respecto a la facilidad de interpretación del modelo, el acceso a la información detallada y la precisión. Los resultados arribaron que *Naïve Bayes* responde a cuatro de los cinco objetivos; *DT* a tres; y la *Neural Network* a dos; por tanto el modelo más eficaz para predecir a los pacientes con enfermedad cardíaca en base a este estudio es *Naïve Bayes*.

En (Alfisahrin & Mantoro, 2013) la investigación tiene como objetivo identificar si los pacientes tienen una enfermedad hepática, basada en una serie de párametros propias de la enfermedad, tales como edad, sexo, *total bilirubin (TB)*, *direct bilirubin (DB)*, *total proteins (TP)*, *albumin (ALB)*, etc. Este estudio utilizó un conjunto de datos del repositorio de la *University of California Irvine (UCI).* El conjunto de datos contiene 583 pacientes, donde 416 pacientes fueron afectados positivamente por la enfermedad hepática y 167 pacientes no sufren de enfermedad hepática. Al clasificar a los pacientes que sufren de enfermedad hepática con pacientes que no, aplicaron dos técnicas de clasificación: *Naïve Bayes* y *DT* (C4.5). El resultado muestra que la precisión del algoritmo C4.5 es mayor que el algoritmo de *Naïve Bayes*.

Naïve Bayes tiene una gran precisión en las bases de datos pequeñas, pero para las grandes bases de datos, la precisión del algoritmo *DT* es mejor (Alfisahrin & Mantoro, 2013). Para mejorar la precisión de la clasificación del algoritmo de *Naïve Bayes*, se combina con el algoritmo *DT*, llamado entonces *NBTree*. Aunque *NBTree* fue declarado como de mayor precisión este algoritmo se utiliza muy poco.

En el estudio (Alfisahrin & Mantoro, 2013) se comparan entonces las técnicas de clasificación de minería de datos, *DT* (C4.5), *Naïve Bayes* y *NBTree* para determinar qué algoritmo es más preciso y óptimo para detectar la enfermedad hepática. El algoritmo más preciso se puede aplicar a las instituciones médicas para la detección temprana de la enfermedad del hígado y se puede utilizar para ayudar a los médicos a proporcionar tratamiento adecuado a los pacientes. Al comparar los tres algoritmos, basado en el valor de precisión y el tiempo computacional requerido se obtienen los siguientes resultados que se muestran en la Tabla VIII.

De este estudio, concluimos entonces que: El algoritmo *NBTree* tiene la más alta precisión, seguido por *DT* y *Naïve Bayes*. El algoritmo *Naïve Bayes* da el tiempo de computación más rápido seguido por *DT* y *NBTree*. Además el número de reglas de clasificación del algoritmo *NBTree* es más simple que el número de reglas de clasificación producidas por el algoritmo *DT*.

Tabla VIII. Comparación de técnicas en cuanto a precisión y tiempo computacional. Fuente: (Alfisahrin & Mantoro, 2013)

Técnicas	Precisión	Tiempo computacional (s)
DT	66.14%	0.45 s
Naïve Bayes	56.14%	0.04 s
NBTree	67.01%	2.51 s

En la investigación (Kunwar et al., 2016) el objetivo es predecir la enfermedad renal crónica utilizando técnicas de clasificación como *Naïve Bayes* y *Artificial Neural Network*. La comparación experimental de *Naïve Bayes* y *ANN* se realiza sobre la base de los vectores de rendimiento. Se tomaron datos clínicos de 400 registros provenientes de *UCI Machine Learning Repository*, una vez que se han limpiado y eliminado los valores faltantes quedan 220 registros. Algunos de los factores considerados fueron la edad, la diabetes, la presión arterial, el conteo de eritrocitos, creatinina sérica, sodio, potasio, hemoglobina, glóbulos blancos, glóbulos rojos, etc.

Dentro de los párametros a tener en cuenta para la comparación están el análisis de rendimiento. El valor de rendimiento (Kunwar et al., 2016) incluye la precisión que se refiere al número de predicciones correctas y la variable Kappa tiene en cuenta las predicciones correctas que ocurren por casualidad. Da una medida cuantitativa de la magnitud del acuerdo entre los observadores. Se encuentra en el rango de -1 a 1, donde 1 es el acuerdo perfecto, 0 es el acuerdo al azar, y los valores negativos indican el desacuerdo entre los observadores.

Los resultados obtenidos se muestran en la Tabla IX demostrando que en el estudio realizado *Naïve Bayes* es el clasificador más preciso con un 100% en comparación con la *ANN* con un 72,73% de precisión.

La diabetes es otra de las enfermedades crónicas del metabolismo corporal, caracterizada por la incapacidad de producir suficiente insulina para procesar carbohidratos, grasas y proteínas de manera eficiente. El estudio (Sankaranarayanan & Perumal, 2014) aborda la aplicación de técnicas de minería de datos en la investigación de la diabetes, ofrececiendo una perspectiva racional para modelar los patrones que pueden pronosticar la incidencia de la diabetes mellitus en la raza humana. Los registros clínicos de pacientes y los informes de pruebas patológicas representan inherentemente conjuntos de datos que pueden aplicarse para la investigación sobre la diabetes.

Tabla IX. Comparación de técnicas de clasificación. Fuente: (Kunwar et al., 2016)

Párametros a comparar	Naïve Bayes	Neural Network Artificial
Análisis de rendimiento:		
-Precisión	100%	72.73 %
-Valor Kappa	1	0.455

En este caso *DT* es la técnica más común de clasificación para la toma de decisiones y la representación del conocimiento (Sankaranarayanan & Perumal, 2014). El conjunto de datos utilizados para el estudio proviene de *UCL ML Data Repository (https://kdd.ics.uci.edu/)*. En este caso contiene tres atributos predictores: edad, sexo, valor de hemoglobina glicosilada (HbA1C) y un atributo de decisión que determinan si una persona está o no teniendo diabetes mellitus. Desarrollando el algoritmo *DT* (Sankaranarayanan & Perumal, 2014) que se muestra en la Figura 5.1 se puede determinar si la persona durante los últimos 3 meses es diabética, teniendo en cuenta el control de azúcar glicémico en la sangre.

La técnica *DT* es también utilizada para el diagnóstico de la enfermedad pulmonar obstructiva crónica (Koppad & Kumar, 2016), que es una de las principales causas de morbilidad y mortalidad en todo el mundo. Para el desarrollo de la investigación se recopilan datos a partir de diferentes fuentes como son notas médicas, documentos en papel y EMR.

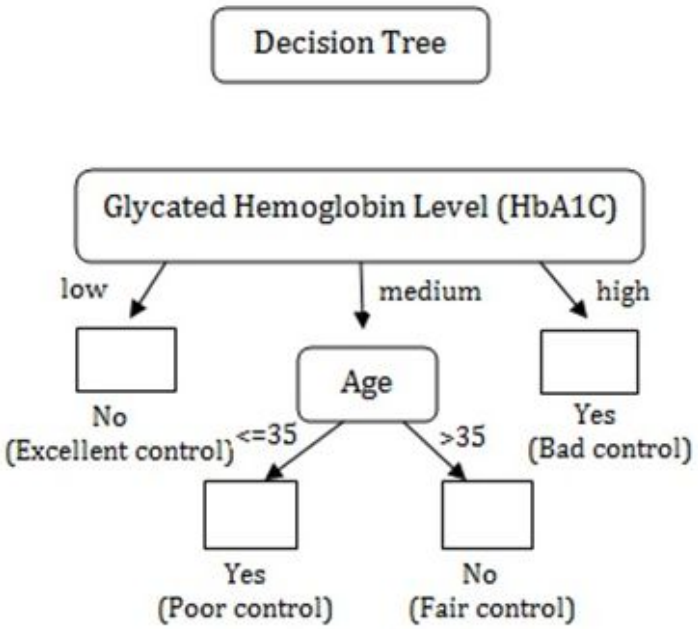

Figura 5.1: Algoritmo *DT* para determinar diabetes mellitus. Fuente: (Sankaranarayanan & Perumal, 2014)

Posteriormente, los datos se analizan y se utiliza el algoritmo J48 (Koppad & Kumar, 2016) para predecir el modelo de aprendizaje automático y clasificar su precisión. Después de ejecutar este algoritmo la salida es analizada por el clasificador, dando varias estadísticas basadas en la validación cruzada, y de esta forma hacer una predicción de cada instancia del conjunto de datos.

5.2 Aplicaciones de Técnicas de Big Data en las enfermedades crónicas.

5.2.1 Análisis predictivo para la diabetes

En (Saravana Kumar et al., 2015) se plantea el análisis predictivo del tratamiento diabético usando técnicas de minería de datos basadas en regresión, usando el algoritmo de aprendizaje supervisado SVM, que da solución a problemas de clasificación y regresión, cumpliendo la función en este caso de identificar el mejor modo de tratamiento para la diabetes en diferentes edades.

El modelo de predicción basado en *soft computing* fue desarrollado para encontrar los riesgos acumulados por los pacientes diabéticos. Se ha experimentado con datos clínicos en tiempo real utilizando algoritmo genético (Saravana Kumar et al., 2015), los resultados obtenidos están relacionados con el nivel de riesgo propenso a un ataque cardíaco o un accidente cerebrovascular.

Para el tratamiento diabético es necesario probar patrones como: concentración de glucosa en plasma, insulina, presión arterial, *Body Mass Index (BMI)*, edad, número de veces embarazada (Saravana Kumar et al., 2015), etc. En la Figura 5.2 se muestra el ciclo de análisis predictivo para el descubrimiento de patrones en la diabetes.

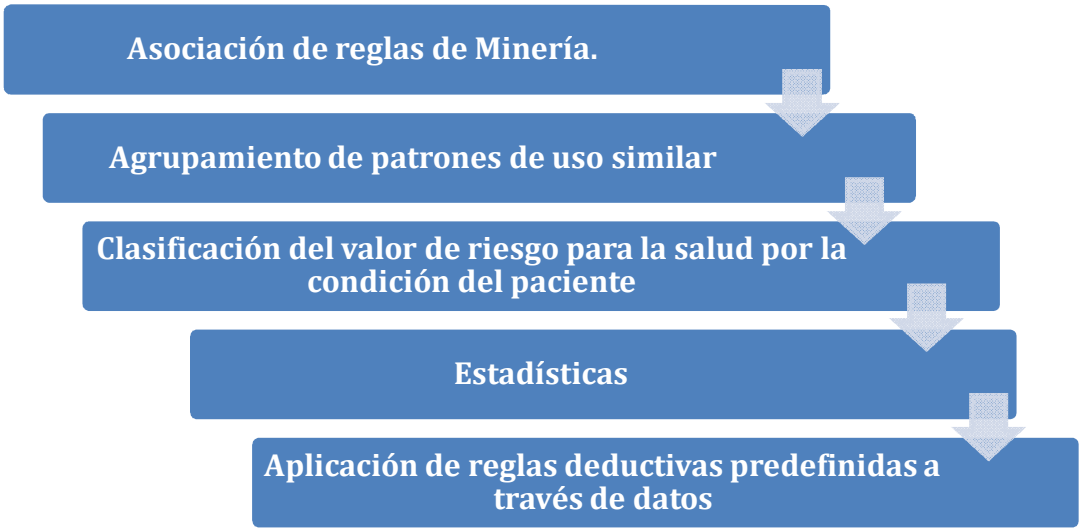

Figura 5.2: Ciclo de análisis predictivo para el descubrimiento de patrones en la diabetes. Fuente: Propia

La diabetes puede asociarse con enfermedades graves como ataques cardíacos, derrames cerebrales, enfermedades oculares y enfermedades renales, etc. Analizar el valor del riesgo por el nivel de condición de salud del paciente

utilizando los resultados anteriores puede ser utilizado por los médicos en lugares remotos a bajo costo.

5.3 Resultados

Las revisiones realizadas no dan la idea de cuáles son las técnicas más utilizadas en la predicción de enfermedades crónicas, pero las técnicas tienen mayor o menor efectividad en dependencia del tipo de enfermedad crónica y de la predicción que se pretenda obtener.

Según los resultados obtenidos que se muestran en la Tabla X la técnica *Naïve Bayes* es la más recomendable para la predicción de enfermedades cardíacas, mientras que *DT* es la técnica más común para la clasificación de datos en la diabetes y la enfermedad pulmonar obstructiva.

Tabla X. Comparativa de técnicas de minería de datos en la predicción de enfermedades crónicas. Fuente: Propia

Publicación	Técnicas	Enfermedades crónicas	Bases de datos	Resultados
Palaniappan & Awang, 2008	Naïve Bayes, DT and Neural Network	Enfermedad Cardíaca	Cleveland Heart Disease	Con 50% de población : Red Neuronal: (49.34%) Naïve Bayes: (47.58%) DT: (41.85%) Con 100% de población: Naïve Bayes: (86.12%) Neural Network: (85.68%) DT: (80.4%)
Kunwar et al., 2016	Naïve Bayes and ANN	Enfermedad Renal Crónica	UCI Machine Learning Repository	Rendimiento: Precisión/Kappa Naïve Bayes: (100%/1) ANN: (72,73% /0.455)
Sivagowry et al., 2013	Naïve Bayes, DT and ANN	Enfermedad Cardíaca	Repository University of California, Irvine (UCI)	Naïve Bayes: 86.53%.
Alfisahrin &Mantoro,2013	Naïve Bayes, DT (C4.5) and NBTree	Enfermedad Hepática	Repository University of California, Irvine (UCI)	Precisión/tiempo computacional: DT: (66.14%/0.45s) Naïve Bayes:(56.14%/0.04s) NBTree: (67.01% /2.51 s)

La revisión ha demostrado que a pesar de que *Naïve Bayes*, *DT* y *ANN*, son las más utilizadas, es necesario ampliar el estudio con un mayor número de investigaciones para corroborar los datos obtenidos.

Capítulo 6

Conclusiones

Big Data surge como respuesta a la creciente necesidad de un número cada vez mayor de organizaciones de disponer de nuevas herramientas, capaces de procesar de forma eficiente un enorme volumen de datos de diversa tipología, procedentes de fuentes muy heterogéneas. También ofrece la oportunidad de permitir una medicina eficaz y de precisión mediante la estratificación del paciente. Esto es de hecho una tarea clave hacia la atención sanitaria personalizada. Un mejor uso de los recursos médicos por medio de la personalización puede conducir a servicios de salud bien gestionados que pueden superar los retos de una población en rápido crecimiento y envejecimiento. Por lo tanto, los avances en el procesamiento de Big Data para la bioinformática, la detección de enfermedades crónicas, la genómica y la biomedicina tendrán un gran impacto en la investigación clínica futura.

Otro factor importante a considerar es la rápida adquisición de datos en la salud, lo que contribuirá al éxito de Big Data en medicina. La frecuencia recopilación de datos de este tipo todavía implica un proceso lento y complejo, que requiere la participación de personal y laboratorios especiales de salud. En este contexto el uso de sensores significa la capacidad de cubrir grandes períodos de monitoreo continuo sin necesidad de realizar exámenes esporádicos, lo cual puede representar solamente una imagen estrecha del desarrollo de una enfermedad.

A diferencia de otras soluciones tradicionales, Big Data permite gestionar grandes volúmenes de datos de todo tipo, especialmente no estructurados, a una velocidad muy superior y con un consumo de recursos mucho menor. Otra de las grandes ventajas que ofrece esta tecnología es la posibilidad de utilizar técnicas de analítica avanzada, como la analítica predictiva, con el objetivo de operar sobre datos masivos y construir modelos predictivos de calidad que sirvan de soporte a la toma de decisiones.

En este TFM se han planteado las principales fuentes de Big Data y las tres áreas de investigación donde los conceptos de análisis de Big Data se están aplicando actualmente: el procesamiento de imágenes, procesamiento de señales y la genómica. El crecimiento exponencial del volumen de imágenes médicas obliga a los científicos a encontrar soluciones innovadoras para procesar este gran volumen de datos en escalas de tiempo manejables. La tendencia de la adopción de sistemas computacionales para el procesamiento de la señal fisiológica de la investigación y la práctica de profesionales médicos, está creciendo constantemente con el desarrollo de algunos sistemas que ayudan a salvar vidas.

Los nuevos avances tecnológicos han dado lugar a una mayor resolución, dimensión y disponibilidad de imágenes multimodales, que conducen al aumento de la precisión del diagnóstico y la mejora del tratamiento. Sin embargo, la integración de imágenes médicas con diferentes modalidades o con otros datos médicos es una oportunidad potencial. Se requieren nuevos marcos analíticos y métodos para analizar estos datos en un entorno clínico.

También se han mostrado en la investigación realizada las distintas infraestructuras que se usan para el intercambio de datos, tanto en el dominio de las enfermedades cardiovasculares como en electrofisiología, esta última desarrollada originalmente para investigar la estructura fina del cerebro epiléptico humano; así como en otras áreas de salud como las enfermedades renales, ejemplificando la infraestructura que incluye cuatro centros de diálisis clínicas entre Lombardía y Suiza, construida para almacenar gran cantidad de datos clínicos, relacionados con las 716 sesiones de diálisis de 70 pacientes de diferentes hospitales.

Los campos científicos y médicos necesitarán implementar el mismo tipo de estructura escalable para manejar los volúmenes de datos generados por las diferentes tecnologías e información de salud. La biomedicina tendrá que adaptarse a los avances de la informática para abordar con éxito los grandes problemas de datos que se enfrentarán en el futuro, especialmente en los programas de medicina personalizada, para mejorar significativamente la atención al paciente.

Como resultado de la revisión bibliográfica realizada se muestran además las bases de datos y técnicas utilizadas por Big Data en el sector sanitario. Se identifican en las publicaciones existentes cuales son las técnicas de minería de datos que se utilizan en la predicción de enfermedades crónicas, obteniendo como resultado las tres de mayor utilidad: *Decision Tree, Naïve Bayes y Artificial Neural Network.*

Por último, se plantean estudios que hacen comparativas entre dichas técnicas, buscando cual es la que proporciona el porcentaje más alto en la predicción de enfermedades crónicas, pero los resultados arrojan que varía en dependencia del tipo de enfermedad. Aún así la cantidad de estudios encontrados es un factor clave para demostrar la veracidad de los resultados alcanzados, por tanto en este caso planteamos como investigación futura ampliar el número de artículos a revisar y obtener la técnica más factible para la predicción de enfermedades crónicas.

En otro enfoque, la arquitectura puede ser escalable para incluir datos de otros medios, como centros de agregados hospitalarios para aplicaciones móviles. El futuro tiene acceso a importantes fuentes de datos, tales como relojes inteligentes y dispositivos portátiles para ayudar a nuestra investigación en el sector de la salud. Tenemos información importante para combatir la enfermedad y permitir el seguimiento y la predicción de la evolución de las epidemias y brotes de enfermedades.

Por último, pero no menos importante, la política y la regulación gubernamentales son necesarias para garantizar la privacidad durante la transmisión y el almacenamiento de datos, así como durante las tareas posteriores de análisis de datos.

El desarrollo de este TFM permite la continuidad de la investigación para una posible Tesis Doctoral, proyectando su objetivo al campo de las enfermedades crónicas y actualizando la bibliografía utilizada, por las nuevas referencias que van publicando en las bases de datos.

Referencias Bibliográficas

Al-Janabi, S., Patel, A., Fatlawi, H., Kalajdzic, K., & Al Shourbaji, I. (2015). Empirical rapid and accurate prediction model for data mining tasks in cloud computing environments. *2014 International Congress on Technology, Communication and Knowledge, ICTCK 2014*, 26–27. https://doi.org/10.1109/ICTCK.2014.7033495

Alfisahrin, S. N. N., & Mantoro, T. (2013). Data Mining Techniques for Optimization of Liver Disease Classification. *2013 International Conference on Advanced Computer Science Applications and Technologies*, 379–384. https://doi.org/10.1109/ACSAT.2013.81

Alyass, A., Turcotte, M., & Meyre, D. (2015). From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Medical Genomics*, *8*(1), 33. https://doi.org/10.1186/s12920-015-0108-y

Andreu-Perez, J., Poon, C. C. Y., Merrifield, R. D., Wong, S. T. C., & Yang, G. Z. (2015). Big Data for Health. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1193–1208. https://doi.org/10.1109/JBHI.2015.2450362

Archenaa, J., & Anita, E. A. M. (2015). A survey of big data analytics in healthcare and government. *Procedia Computer Science*, *50*, 408–413. https://doi.org/10.1016/j.procs.2015.04.02

Belle, A., Thiagarajan, R., Soroushmehr, S. M. R., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big Data Analytics in Healthcare. *Hindawi Publishing Corporation*, 1–16. https://doi.org/10.1155/2015/370194

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, *80*(4), 571–583. https://doi.org/10.1016/j.jss.2006.07.009

Brinkmann, B. H., Bower, M. R., Stengel, K. A., Worrell, G. A., & Stead, M. (2009). Large-scale electrophysiology: Acquisition, compression, encryption, and storage of big data. *Journal of Neuroscience Methods*, *180*(1), 185–192. https://doi.org/10.1016/j.jneumeth.2009.03.022

Buchanan, C. C., Torstenson, E. S., Bush, W. S., & Ritchie, M. D. (2012). A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *Journal of the American Medical Informatics Association*, *19*(2), 289–294. https://doi.org/10.1136/amiajnl-2011-000652

Chauhan, R., & Kumar, A. (2013). Cloud computing for improved healthcare: Techniques, potential and challenges. *2013 E-Health and Bioengineering Conference, EHB 2013*. https://doi.org/10.1109/EHB.2013.6707234

Chennamsetty, H., Chalasani, S., & Riley, D. (2015). Predictive analytics on Electronic Health Records (EHRs) using Hadoop and Hive. *Proceedings of 2015 IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT 2015*, 1–5. https://doi.org/10.1109/ICECCT.2015.7226129

Costa, F. F. (2014). Big data in biomedicine. *Drug Discovery Today*, *19*(4), 433–440. https://doi.org/10.1016/j.drudis.2013.10.012

Cunha, J., Silva, C., & Antunes, M. (2015). Health Twitter Big Bata Management with

Hadoop Framework. *Procedia Computer Science*, *64*, 425–431. https://doi.org/10.1016/j.procs.2015.08.536

Elsebakhi, E. ., Lee, F. ., Schendel, E. ., Haque, A. ., Kathireason, N. ., Pathare, T. ., … Al-Ali, R. . (2015). Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms. *Journal of Computational Science*, *11*, 69–81. https://doi.org/10.1016/j.jocs.2015.09.008

Fouad, M. M., Oweis, N. E., Gaber, T., Ahmed, M., & Snasel, V. (2015). Data Mining and Fusion Techniques for WSNs as a Source of the Big Data. *Procedia Computer Science*, *65*, 778–786. https://doi.org/10.1016/j.procs.2015.09.023

Gessner, R. C., Frederick, C. B., Foster, F. S., & Dayton, P. A. (2013). Acoustic angiography: A new imaging modality for assessing microvasculature architecture. *International Journal of Biomedical Imaging*, *2013*. https://doi.org/10.1155/2013/936593

Grover, A., Gholap, J., Janeja, V. P., Yesha, Y., Chintalapati, R., Marwaha, H., & Modi, K. (2015). SQL-like big data environments: Case study in clinical trial analytics. *2015 IEEE International Conference on Big Data (Big Data)*, 2680–2689. https://doi.org/10.1109/BigData.2015.7364068

Huang, T., Lan, L., Fang, X., An, P., Min, J., & Wang, F. (2015). Promises and Challenges of Big Data Computing in Health Sciences. *Big Data Research*, *2*(1), 2–11. https://doi.org/10.1016/j.bdr.2015.02.002

Kanevsky, J., Corban, J., Gaster, R., Kanevsky, A., Lin, S., & Gilardino, M. (2016). Big Data and Machine Learning in Plastic Surgery. *Plastic and Reconstructive Surgery*, *137*(5), 890e–897e. https://doi.org/10.1097/PRS.0000000000002088

Koppad, S. H., & Kumar, A. (2016). Application of Big Data Analytics in Healthcare System to Predict COPD. *Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference . IEEE*, 1–5.

Kunwar, V., Chandel, K., Sabitha, A. S., & Bansal, A. (2016). Chronic Kidney Disease Analysis Using Data Mining Classification. *Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference. IEEE*, 300–305. https://doi.org/10.1109/CONFLUENCE.2016.7508132

Lu, J., & Keech, M. (2015). Emerging Technologies for Health Data Analytics Research: A Conceptual Architecture. *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*, 225–229. https://doi.org/10.1109/DEXA.2015.58

Luis, C., & Calderón, P. (2016). Big data en sanidad en Espã na: la oportunidad de una estrategia nacional Big data in health in Spain: now is the time for a national strategy. *Gac Sanit*, *30*(1), 63–65. https://doi.org/10.1016/j.gaceta.2015.10.005

Maher, B. (2012). ENCODE: The human encyclopaedia. *Nature*, *489*(7414), 46–48. https://doi.org/10.1038/489046a

Manuel, J., & Sesmero, M. (2015). "Big Data"; aplicación y utilidad para el sistema sanitario. *Farm Hosp*, *39*(2), 69–70. https://doi.org/10.7399/fh.2015.39.2.8835

Melethadathil, N., Chellaiah, P., Nair, B., & Diwakar, S. (2015). Classification and clustering for neuroinformatics: Assessing the efficacy on reverse-mapped NeuroNLP data using standard ML techniques. *2015 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2015*, 1065–1070.

https://doi.org/10.1109/ICACCI.2015.7275751

Merelli, I., Pérez-Sánchez, H., Gesing, S., & D'Agostino, D. (2014). Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives. *BioMed Research International*. https://doi.org/10.1155/2014/134023

Moskowitz, A., McSparron, J., Stone, D. J., & Celi, L. A. (2015). Preparing a New Generation of Clinicians for the Era of Big Data. *Harvard Medical Student Review*, *2*(1), 24–27.

Nambiar, R., Bhardwaj, R., Sethi, A., & Vargheese, R. (2013). A look at challenges and opportunities of Big Data analytics in healthcare. *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*, 17–22. https://doi.org/10.1109/BigData.2013.6691753

O'Driscoll, A., Daugelaite, J., & Sleator, R. D. (2013). "Big data", Hadoop and cloud computing in genomics. *Journal of Biomedical Informatics*, *46*(5), 774–781. https://doi.org/10.1016/j.jbi.2013.07.001

Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. *2008 IEEE/ACS International Conference on Computer Systems and Applications*, 108–115. https://doi.org/10.1109/AICCSA.2008.4493524

Patel, J. A., & Sharma, P. (2014). Big data for Better Health Planning. *Advances in Engineering and Technology Research (ICAETR), 2014 International Conference. IEEE.*, 0–4.

Payakachat, N., Tilford, J. M., & Ungar, W. J. (2016). National Database for Autism Research (NDAR): Big Data Opportunities for Health Services Research and Health Technology Assessment. *PharmacoEconomics*, *34*(2), 127–138. https://doi.org/10.1007/s40273-015-0331-6

Pérez, G. (2016). Peligros del uso de los big data en la investigación en salud pública y en epidemiología Risks of the use of big data in research in public health and epidemiology, *30*(1), 66–68.

Philip Chen, C. L., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314–347. https://doi.org/10.1016/j.ins.2014.01.015

Pombo, N., Garcia, N., Felizardo, V., & Bousson, K. (2014). Big data reduction using RBFNN: A predictive model for ECG waveform for eHealth platform integration. *2014 IEEE 16th International Conference on E-Health Networking, Applications and Services, Healthcom 2014*, (Ssh), 66–70. https://doi.org/10.1109/HealthCom.2014.7001815

Ramkumar, N., Prakash, S., & Sangeetha, K. (2016). Data Analysis for Chronic disease – Diabetes using Map Reduce Technique.

Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., … Bourne, P. E. (2011). The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Research*, *39*, 392–241. https://doi.org/10.1093/nar/gkq1021

Sankaranarayanan, S., & Perumal, T. P. (2014). A Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies. *2014 World Congress on Computing and Communication Technologies*, 231–233. https://doi.org/10.1109/WCCCT.2014.65

Saravana Kumar, N. M., Eswari, T., Sampath, P., & Lavanya, S. (2015). Predictive methodology for diabetic data analysis in big data. *Procedia Computer Science*, *50*, 203–208. https://doi.org/10.1016/j.procs.2015.04.069

Seebode, C., Ort, M., Regenbrecht, C., & Peuker, M. (2013). BIG DATA infrastructures for pharmaceutical research. *2013 IEEE International Conference on Big Data*, 59–63. https://doi.org/10.1109/BigData.2013.6691759

Sivagowry, S., Durairaj, M., & Persia, a. (2013). An empirical study on applying data mining techniques for the analysis and prediction of heart disease. *2013 International Conference on Information Communication and Embedded Systems (ICICES)*, 265–270. https://doi.org/10.1109/ICICES.2013.6508204

Suinesiaputra, A., Medrano-Gracia, P., Cowan, B. R., & Young, A. A. (2015). Big Heart Data: Advancing Health Informatics Through Data Sharing in Cardiovascular Imaging. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1283–1290. https://doi.org/10.1109/JBHI.2014.2370952

Trifiletti, D. M., & Showalter, T. N. (2015). Big Data and Comparative Effectiveness Research in Radiation Oncology: Synergy and Accelerated Discovery. *Frontiers in Oncology*, *5*, 5–9. https://doi.org/10.3389/fonc.2015.00274

Vito, D., Casagrande, G., Bianchi, C., & Costantino, M. L. (2015). How to extract clinically useful information from large amount of dialysis related stored data. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, *2015–Novem*, 6812–6815. https://doi.org/10.1109/EMBC.2015.7319958

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., … Scalbert, A. (2013). HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Research*, *41*(D1), D801–D807. https://doi.org/10.1093/nar/gks1065

Young, S. D. (2015). A " big data " approach to HIV epidemiology and prevention. *Preventive Medicine*, *70*, 17–18. https://doi.org/10.1016/j.ypmed.2014.11.002