El libro *Datos, información, tendencias: Tres miradas sobre un contexto cambiante* pretende dar una mirada a tres aspectos la vorágine técnica, política, económica y social relacionadas con la Cuarta Revolución Industrial, para contribuir a entenderla y a identificar formas nuevas de aprovechar, desde diferentes ángulos, las muchas oportunidades que surgen cada día.

Se analizan, la dimensión de la salud caracterizada por la complejidad de sus procesos, se exploran facetas de una particularidad de la actual sociedad hiperconectada y progresivamente tecnocéntrica y se penetra en un mundo más estructurado de información y datos, que se ha enriquecido con un subproducto de la masificación y abaratamiento de las tecnologías, los servicios y los dispositivos de comunicaciones.

EDICIONES UNIVERSIDAD SIMÓN BOLÍVAR

ISBN 978-958-5533-08-0

DATOS, INFORMACIÓN, TENDENCIAS

Vladimir Quintero - Diana Heredia Vizcaíno - Paul Sanmartín Mendoza - Ronald Romero Zúñiga - Jhonny Castillo Vélez

UNIVERSIDAD SIMÓN BOLÍVAR
BARRANQUILLA Y CÚCUTA - COLOMBIA | VIGILADA MINEDUCACIÓN
Res. 23095, del MEN

DATOS, INFORMACIÓN, TENDENCIAS,

tres miradas sobre un contexto cambiante

Vladimir Quintero - Diana Heredia Vizcaíno
Paul Sanmartín Mendoza - Ronald Romero Zúñiga
Jhonny Castillo Vélez

DATOS, INFORMACIÓN, TENDENCIAS,

tres miradas sobre un contexto cambiante

DATOS, INFORMACIÓN, TENDENCIAS: TRES MIRADAS SOBRE UN CONTEXTO CAMBIANTE

Proceso de arbitraje doble ciego
Recepción: Febrero de 2018
Evaluación de propuesta de obra: Abril de 2018
Evaluación de contenidos: Junio de 2018
Correcciones de autor: Agosto de 2018
Aprobación: Octubre de 2018

UNIVERSIDAD SIMÓN BOLÍVAR
BARRANQUILLA Y CÚCUTA - COLOMBIA | VIGILADA MINEDUCACIÓN
DE ALTA CALIDAD
Res. 23095. del MEN

DATOS, INFORMACIÓN, TENDENCIAS,

tres miradas sobre un contexto cambiante

Vladimir Quintero - Diana Heredia Vizcaíno
Paul Sanmartín Mendoza - Ronald Romero Zúñiga
Jhonny Castillo Vélez

EDICIONES UNIVERSIDAD SIMÓN BOLÍVAR

Datos, información, tendencias, tres miradas sobre un contexto cambiante/ Vladimir Quintero [y otros 4] -- Barranquilla: Ediciones Universidad Simón Bolívar, 2018.

95 páginas; gráficos a color

ISBN: 978-958-5533-08-0 (Pdf descargable)

1. Redes de información 2. Tecnología médica 3. Redes sociales online 4. Minería de datos I. Quintero, Vladimir II. Heredia Vizcaíno, Diana III. Sanmartín Mendoza, Paul IV. Romero, Ronald VI. Castillo Velez, Jhonny VII. Universidad Simón Bolívar. Grupo de investigación tecnológica y salud VIII. Universidad Simón Bolívar. Grupo de investigación ingebiocaribe IX. Título

006.333 D234 2018 Sistema de Clasificación Decimal Dewey 22ª. Edición

Universidad Simón Bolívar – Sistema de Bibliotecas

Impreso en Barranquilla, Colombia. Depósito legal según el Decreto 460 de 1995. El Fondo Editorial Ediciones Universidad Simón Bolívar se adhiere a la filosofía del acceso abierto y permite libremente la consulta, descarga, reproducción o enlace para uso de sus contenidos, bajo una licencia Creative Commons Atribución 4.0 Internacional. https://creativecommons.org/licenses/by/4.0/

©Ediciones Universidad Simón Bolívar
Carrera 54 No. 59-102
http://publicaciones.unisimonbolivar.edu.co/edicionesUSB/
dptopublicaciones@unisimonbolivar.edu.co
Barranquilla - Cúcuta

Producción Editorial
Editorial Mejoras
Calle 58 No. 70-30
info@editorialmejoras.co
www.editorialmejoras.co

Diciembre de 2018
Barranquilla

Made in Colombia

Cómo citar este libro:
Quintero, V., Heredia Vizcaíno, D., Sanmartín Mendoza, P., Romero Zúñiga, R. & Castillo Vélez, J. (2018). *Datos, Información, Tendencias: tres miradas sobre un contexto cambiante.* Barranquilla: Universidad Simón Bolívar.

Contenido

Introducción

El corto lapso entre la Tercera y la Cuarta Revolución Industrial fue suficiente para que el planeta volviera a cambiar radicalmente. El impacto de la informática y su convergencia con las industrias de la radiodifusión y las comunicaciones creó las condiciones para el nacimiento del mundo digital, que rápidamente reemplazó en forma significativa al mundo analógico que conocimos por algunos miles de años. En menos de tres décadas, la nueva convergencia de ciencias, tecnologías y nuevos materiales acelera la transformación, no solo de los modelos y herramientas de producción, sino de los mismos productos, propuestas de valor, modelos de negocios y finalmente, modelos de individuos y sociedades. Los datos, la información asociada y su intercambio permean todas las estructuras de este nuevo planeta-sociedad y generan retos y oportunidades a todas las jerarquías de sus habitantes digitales, de los nativos, los inmigrantes y hasta los refugiados digitales.

El presente libro pretende dar una mirada a tres aspectos de esa vorágine técnica, política, económica y social para contribuir a entenderla y a identificar formas nuevas de aprovechar, desde diferentes ángulos, las muchas oportunidades que cada día aparecen. Inicia desde la dimensión de la Salud, caracterizada por la siempre creciente complejidad de sus procesos y su

información, frente a la demanda de atención de una población también en aumento, a nivel mundial y que cada día necesita más atenciones como consecuencia de la inversión progresiva de la pirámide demográfica. Ha regresado con fuerza el antiguo modelo basado en la prevención, como una estrategia de incrementar la eficiencia y calidad de atención, y a la vez reducir costos de atención. Sin embargo, esos esfuerzos se pierden ante la ineficiencia del uso de la información, resultado de la falta de estandarización, tanto de los propios datos, como de su intercambio electrónico. Este problema, aparentemente solo técnico, es abordado desde las visiones complementarias de mercados y políticas, para identificar una muy interesante –pasajera– ventana de oportunidad para promover iniciativas colaborativas regionales en América Latina.

El segundo capítulo explora facetas de una particularidad de la actual sociedad hiperconectada y progresivamente tecnocéntrica. No es probable hallar en la historia de la humanidad una concentración tan elevada de iniciativas exitosas que han partido de la premisa: 'Tengo la solución, vamos a encontrarle un problema', como la que estamos presenciando. Este es precisamente el fenómeno tras las redes sociales. Hace menos de 15 años Mark Zuckerberg quería ayudar a intercambiar mensajes entre sus compañeros de la Universidad de Harvard. Difícilmente llegó a pensar que, en muy pocos meses, su aplicativo Facebook habría traspasado las fronteras de la Universidad y del país, o que su uso habría traspasado las fronteras del intercambio entre amigos, para convertirse en una miríada de negocios innovadores y rentables, que hacen uso de la misma 'solución'. La historia de Twitter no es muy diferente. Nacida un par de años después de Facebook como una plataforma de '*microblogging*' a partir de '*tweets*', que en el año de su lanzamiento fueron descritos como 'una corta ráfaga de información intrascendente'. Es posible que el concepto de la corta ráfaga se mantenga hoy en día (un poco menos corta al ampliarse a 280 caracteres en 2017), pero definitivamente está muy lejos de ser intrascendente. Twitter se ha convertido en una herramienta de alto impacto comercial, social y político. Los millones de

caracteres intercambiados diariamente forman un gigantesco volumen de datos con información escondida en sus relaciones, que puede y de hecho es utilizada en muchos sectores. El capítulo se centra en las aplicaciones de *marketing*, y analiza el potencial que este tiene para montar empresas viables que estimulen la economía.

Finalmente, el tercer capítulo penetra en un mundo más estructurado de información y datos, que se ha enriquecido con un subproducto de la masificación y abaratamiento de las tecnologías, los servicios y los dispositivos de comunicaciones. Se trata de un metadato que hasta hace menos de dos décadas era costoso y complejo de obtener, el geoposicionamiento de los datos. Prácticamente cualquier dispositivo inteligente cuenta hoy con la posibilidad de georreferenciar sus datos y archivos, sin que esto genere ningún costo adicional al usuario. Esta facilidad incrementada de contar con ese metadato adicional, permite ampliar el ámbito de las posibles investigaciones y aplicaciones del *Data Analysis* de grandes volúmenes de datos, o a repositorios de datos históricos, con mayor probabilidad de identificar patrones y proponer predicciones con un grado aceptable de confiabilidad. Las aplicaciones son incontables, desde el campo puramente científico, hasta temas de análisis sociales y de seguridad. Este último es explorado en más detalle y con resultados prometedores por la investigación realizada a partir de datos abiertos.

No cabe duda, los datos nos rodean y nosotros mismos estamos generando diariamente una gran cantidad. La política de datos abiertos adoptada en forma creciente por diversos países del mundo ofrece una mina cada vez mayor, más atractiva y más retadora para quienes acepten el reto de encontrar información, inteligencia y conocimiento en medio de ese aparente y amenazador caos. Es el camino de y hacia el futuro.

El documento a continuación contiene una descripción de los resultados de la investigación realizada como trabajo de grado en la Maestría de Ingeniería de Sistemas y Computación en la Universidad Simón Bolívar. Su objetivo era

aplicar las técnicas de minería de datos de *Clustering* y Árboles de decisión a datos georreferenciados de hurtos en la ciudad de Barranquilla, ocurridos entre los años 2016 y 2017, con el fin de obtener un patrón descriptivo y predictivo. Se obtuvieron dichos modelos a partir de datos abiertos (de uso libre para fines investigativos), con eficacia de clasificación de 51 % y 59,7 %, respectivamente, los cuales pueden considerarse no muy buenos como predictores, pero arrojan resultados interesantes en cuanto a manipulación de datos geográficos, atributos no numéricos con gran cantidad de valores.

CAPÍTULO I

Interoperabilidad: un reto que trasciende el intercambio de información y datos

Vladimir QUINTERO MÉNDEZ
Paul SANMARTÍN MENDOZA
Diana HEREDIA VIZCAÍNO

RESUMEN

La tecnología en salud ha experimentado un crecimiento exponencial en las últimas dos décadas como resultado, por una parte, de la penetración de las Tecnologías de Información y Comunicaciones en todos los sectores de la economía, de la multiplicación de las interacciones entre esas tecnologías producidas por las dos olas de convergencia que han dado nacimiento a la Tercera y Cuarta Revolución Industrial; por otra, por el impacto económico y social que a nivel mundial estos procesos han producido. La complejidad inherente a la información en salud, nacida de la multiplicidad de especializaciones, usuarios y propósitos de uso, se ha magnificado en correspondencia con esos cambios tecnológicos. En este contexto dinámico, con actores provenientes de múltiples sectores de la industria, la academia, el Estado y la sociedad en general, se crean nuevos y cambiantes retos tanto tecnológicos como económicos y de política pública. Este documento busca ofrecer una visión integral del proceso, desde su historia, los modelos próximos de intervención y sus resultados, el estado actual y sus tendencias, así como un análisis de las oportunidades y riesgos para los países de América Latina, particularmente para Colombia, en el contexto vigente del Modelo Integral de Atención de Salud.

Palabras clave: interoperabilidad, tecnología en salud, información en salud, intercambio de información, política de información en salud.

INTRODUCCIÓN

> La interoperabilidad en la Atención de la Salud no es algo en blanco o negro, sino que existe en diferentes tonos de gris, ubicados entre: "no ser capaz de intercambiar nada y fluidez de datos integral y ubícua". Además, no es algo que podamos conseguir en un solo paso gigantesco. (Doug Fridsma, ONC, 2013)

La interoperabilidad es una cualidad de cualquier proceso de intercambio de información (Tim Benson, 2016) y cada vez que ocurre y algo sucede como resultado, la interoperabilidad está presente... o no. Sin embargo, ella no es exclusiva de la información de salud, aunque dada la complejidad de las estructuras de información en salud, puede ser la más difícil de alcanzar. La interoperabilidad es básicamente una comprensión compartida entre el emisor y el receptor del mensaje y el flujo de trabajo resultante; en este proceso, el receptor y el remitente podrían ser un equipo de atención médico, una computadora, un sistema o servicio de información, un ser humano. Las características de este proceso, fundamentan la complejidad de la información en salud, que puede considerarse como un elemento 'fractal', esto es, cuanto más te fijas y detallas, más complejidad se le encuentra.

Los patrones de flujo de información y comunicación de salud involucran a muchas personas en un área geográfica amplia y en diversos temas (Giulio Gilia, 2015). El flujo de trabajo es variado, dependiendo de cuál es el problema con el paciente y en qué etapa se encuentra en el proceso; cambia de un paciente a otro y de un entorno de cuidado a otro. La medición de la presión sanguínea en un paciente hospitalizado y en recuperación de una cirugía no tiene la misma importancia que la misma medición en un paciente que llega a emergencia. Igualmente, ese parámetro puede ser medido con una periodicidad de horas en el primer paciente, mientras que debe ser monitoreado casi en tiempo real en el segundo caso.

La vida media de la información (durante cuánto tiempo tiene algún valor la información) difiere enormemente en diferentes contextos, como en las clínicas ambulatorias, en las áreas de hospitalización, en cuidados intensivos, o durante una operación en un quirófano. En el ejemplo del párrafo anterior, el parámetro de presión sanguínea puede tener una vida media de tres, o hasta seis meses en caso de un paciente ambulatorio, que es el período entre citas con su médico tratante; ese parámetro tiene una vida media de horas (6 a 12) en un paciente hospitalizado, y puede tener una vida media de segundos en el caso de una cirugía.

Adicionalmente, cada clase de clínico tiene sus propias necesidades, basadas en los procedimientos e información referente a su especialidad, y que frecuentemente aumenta en volumen, velocidad de producción o precisión, en correspondencia con los avances tecnológicos disponibles. Esta heterogeneidad clínica ayuda a explicar por qué muchos sistemas electrónicos exitosos de registro de pacientes se limitan a especialidades únicas, como la práctica general, la atención de maternidad o la diálisis renal, donde las necesidades son relativamente homogéneas y bien comprendidas por un número limitado de actores en un ambiente definido de procesos. Es fácil percibir el germen del problema cuando la información producida por esos diversos sistemas de información, en múltiples instituciones, debe ser intercambiada e integrada para una atención completa y oportuna al paciente.

ESTÁNDARES

Cualquier intercambio de información generalmente implica dos traducciones: primero desde el idioma nativo del remitente al formato de la conexión o transmisión; segundo, una vez se ha realizado exitosamente, desde el formato de la conexión hasta el idioma nativo del destinatario, que no necesariamente es el mismo del remitente (Oemig & Snelick, 2016). La elección del idioma de intercambio no es suficiente para garantizar la interoperabilidad. Cada transacción debe definirse con un detalle inequívoco, como parte de

un conjunto de especificaciones para esa transacción: que sea completo, coherente, y por supuesto, legible por computador, para asegurar la interoperabilidad y minimizar la posibilidad de error. El proceso de comunicación solo tiene éxito si se dan tres características: se reciben los datos, el significado de los datos recibidos es correcto, y se sigue el flujo de trabajo adecuado, utilizando los datos recibidos. La interoperabilidad solo aparece, cuando las tres características están presentes. La ausencia de cualquiera de ellas invalida el proceso. Esto incluye algunas ideas separadas, relacionadas con aspectos de la interoperabilidad: primero, el intercambio de información, que es la interoperabilidad técnica; segundo, la capacidad del destinatario para interpretar esa información, que es la interoperabilidad semántica, y el tercer concepto, relacionado con el uso real de la información, es la interoperabilidad de proceso. Cuanto más se comprende acerca de los tres tipos, es menos probable que se subestime el trabajo requerido para hacer que los sistemas de información de salud sean interoperables. Como ya se mencionó, estos tipos de interoperabilidad son interdependientes, y los tres son necesarios para brindar los beneficios esperados.

Las definiciones de interoperabilidad más ampliamente divulgadas son la del IEEE, desde el punto de vista técnico y la de HIMSS (*Health Information Management Systems Society*).

Según el IEEE, interoperabilidad es la habilidad de dos o más sistemas o componentes para intercambiar información y usar la que ha sido intercambiada.

De acuerdo al HIMSS, interoperabilidad es la habilidad de diferentes sistemas de tecnología de información y aplicaciones de *software* para comunicar e intercambiar datos, y usar la información que ha sido intercambiada (HIMSS, 2017).

Aparte de su similitud, es evidente que la definición del IEEE es más incluyente, puesto que debe poder aplicarse a un rango mucho mayor de tecnologías relacionadas con la Ingeniería. Por su parte, la de HIMSS se enfoca más al dominio del *software*, sin limitarse a aplicaciones del sector salud.

Dado el reconocimiento institucional de HIMMS, esta es la definición más ampliamente usada por los diferentes actores del sector a nivel mundial. Sin embargo, también puede ser la razón por la cual todas las iniciativas nacionales en busca de la interoperabilidad en las dos últimas décadas se han enfocado en aspectos de *software*, particularmente las Historias Clínicas Electrónicas, dejando de lado, y prácticamente al influjo de las Leyes de mercado, el componente de información de salud generada, cada vez con mayor cantidad y diversidad por los equipos médicos, desde las grandes instalaciones de radiodiagnósticos, hasta los sencillos y ubícuos dispositivos con IoT.

El objetivo integral de la interoperabilidad solo puede ser conseguido a través del uso de estándares en los tres componentes descritos: técnico, semántico y de proceso. El problema con los estándares no es que haya tantos para elegir, sino que no se han implementado los que existen y que, en general, no ha habido nadie con el poder para hacer que la implementación suceda (World Health Organization, 2012). Los estándares que no se implementan son una pérdida de tiempo y esfuerzo. Este reto ha sido enfrentado desde dos puntos de partida, relacionados con el origen de los datos y la información: *software* y hardware.

En el primer caso, los estándares se relacionan con los Sistemas de Información de Salud, y en particular con las Historias Clínicas Electrónicas. Las iniciativas existentes en busca de la interoperabilidad se basan en Modelos Nacionales apoyados y gradualmente reglamentados por los correspondientes órganos del Gobierno. La adopción se da a dos niveles: los desarrolladores, al ofrecer productos que incorporen estándares para interoperabilidad, y por los usuarios institucionales que demandan y adoptan esos productos, y ponen a prueba la validez o no de la interoperabilidad ofrecida.

El segundo caso está representado por la información y datos generados por equipos y dispositivos médicos, en un universo de posibilidades siempre creciente y un mercado muy competido. La responsabilidad de la adopción

de estándares que garanticen la interoperabilidad recae en los fabricantes, quienes ven una amenaza en el uso compartido de esos estándares y prefieren mantener sus 'silos' de información y clientes a través de la estrategia de vender '*suites*' de equipos que son totalmente interoperables, pero solo entre sí. La única excepción a este comportamiento fue propiciada por la Sociedad Americana de Radiología, que logró hacer presión para que el estándar DICOM fuese adoptado a nivel mundial (aunque con ligeras variaciones) por todos los fabricantes de equipos que generen imágenes radiológicas. Como subproducto de esta postura firme y clara ante la necesidad de uso amplio de los estándares, surgió la IHE (*Integrating the Healthcare Enterprise*), con presencia en más de 17 países, que ofrece desarrollos especializados de 'Perfiles de Interoperabilidad' para equipos y dispositivos en 11 especialidades clínicas.

La necesidad y urgencia de establecer y adoptar estándares de información e intercambio electrónico de datos aparece con la Tercera Revolución industrial, a finales del siglo pasado. La convergencia tecnológica de Informática, Comunicaciones y Radiodifusión, sumada a la caída del costo de producción de microprocesadores y memorias, y a la aparición del Computador Personal, cambiaron el mundo y crearon una nueva realidad: el mundo digital. Los elementos de texto, imágenes y multimedia, que hasta ese momento eran parte de un ambiente analógico, pudieron convertirse en series binarias, transmitirse electrónicamente y reconstruirse en su lugar de destino. Las características de interoperabilidad definidas eran aplicables también en ese contexto, pero su implementación era no solo más barata, sino también mucho más estable.

Veamos el ejemplo de un proceso de atención de salud que no ha cambiado en varias décadas: comunicar a Enfermería un nuevo medicamento para proporcionarlo al paciente. En el mundo digital esto representa los siguientes retos de interoperabilidad:

Técnica: un único documento físico que es intercambiado de una mano a otra.

Semántica: el único reto era, y aún es, entender la letra del médico.
De proceso: pasar la información escrita a Farmacia y de allí al siguiente turno de Enfermería.

Este esquema de interoperabilidad podrá no ser muy veloz, pero reduce considerablemente las fuentes de error.

El mismo proceso en el mundo digital ofrece estos retos de interoperabilidad:

Técnica: los datos deben estar completos y actualizados en la Historia Clínica Electrónica del paciente.
Técnica: debe existir compatibilidad con sistemas de Almacenamiento y Farmacia.
Semántica: debe contarse con el uso estandarizado de nombres y unidades de dosis.
De proceso: requiere una identificación única de pacientes.
Técnica: debe existir compatibilidad con el sistema de Administración y Facturación.

Sin duda, incorpora muchos más elementos de información, puede almacenarlos de manera segura y permanente, pero es significativamente más demandante técnicamente, e incrementa el riesgo de errores, a menos que se usen estándares claros y compartidos. La capacidad de manejar más datos es fundamental para responder a la complejidad de los sistemas de salud, resultante como ya se dijo, de la extrema heterogeneidad de sus procesos, datos y flujos de trabajo.

La interoperabilidad en los sistemas de salud requiere un mecanismo flexible para el mapeo de los procesos de negocios a un estándar (casos de uso). Por otro lado, requiere una comprensión profunda del modelo de interacción de los estándares y las lagunas creadas por implementaciones incompatibles. Aunque los beneficios de la interoperabilidad en la atención médica son considerables, tanto económicos, como en la calidad de la atención,

pueden ser difíciles de identificar, ya que se dispersan entre un gran número de partes interesadas, como proveedores, vendedores, responsables de políticas y los pacientes.

Se estima que, en Estados Unidos, por cada 100 pacientes, durante un año un médico de atención primaria, debe coordinar la prestación de servicios de salud con otros 229 médicos que trabajan en 117 instituciones (2017 Report, ONC). Esto genera una enorme complejidad para el intercambio de información entre todos esos actores.

Dentro de una Institución Prestadora de Salud, existen gran cantidad de puntos de origen, o modificación de la información:

- Los Profesionales de la Salud como tratantes de pacientes
- Las Unidades de Diagnóstico, Tratamiento y Monitoreo
- Los diferentes laboratorios
- Las Unidades Especializadas: Emergencias, Cirugía, etc.
- Los propios Sistemas de Información

Todos estos datos se relacionan de una u otra forma con la Historia del Paciente y pueden ser necesarios para intercambiarlos con otros actores del sistema dentro y fuera de la Institución:

- Instituciones de diferente nivel de complejidad
- Aseguradoras
- Farmacias
- Sistema de facturación.

ESTADO DEL ARTE

No son pocos los estándares disponibles y frecuentemente utilizados a nivel mundial, entre otros:

Imágenes: DICOM
Semántica: SNOMED, SNODENT

Laboratorio: LOINC
Comunicaciones: IEEE - ISO
Estructura de Datos: HL7 - CDA - FHIR - ISO
Estructura de Información: ISO
IoT: OneM2M
Identificación de Enfermedades: CID (10-11) reglamentado (OMS, 2016)
Información Administrativa y contable: generalmente son sistemas propietarios dentro de un contexto claramente reglamentado.

Sin embargo, "...los Hospitales en todo el país enfrentan el reto de no ser capaces de intercambiar eficientemente información de salud de sus pacientes con otras organizaciones, a pesar de la disponibilidad de estándares de datos a nivel global" (ONC, 2016).

Recientemente, HL7 ha dado un paso importante para reducir la complejidad del uso de estándares, con su propuesta de FHIR (*Fast Healthcare Interoperability Resources*). Es un marco de estándares de próxima generación creado por HL7. Combina las características de las líneas de productos v2, v3 y CDA de HL7, mientras aprovecha estándares web y aplica un enfoque estricto en la facilidad de implementación y en la curva de aprendizaje menos empinada.

Las soluciones FHIR se crean a partir de un conjunto de componentes modulares llamados "Recursos" que se pueden ensamblar fácilmente en sistemas operativos que resuelven problemas clínicos y administrativos del mundo real a una fracción del precio de las alternativas existentes. FHIR es adecuado para su uso en una amplia variedad de contextos: aplicaciones de teléfonos móviles, comunicaciones en la nube, intercambio de datos basado en EHR, comunicación de servidores en grandes proveedores de servicios de salud institucionales, etc.

Coexistiendo con esta diversidad de estándares individuales, se encuentran algunas propuestas de integración, entre las que cabe resaltar:

- **IHE.** Desarrolla perfiles que se integran en 11 dominios clínicos. Hace uso de todos los estándares mencionados, seleccionándolos de acuerdo con su pertinencia específica para cada caso de uso. Los resultados son ofrecidos libremente para ser usados por productores de *hardware* y *software*.
- **Continua.** (PCH Alliance-HIMSS). Ofrece guías de diseño basadas en estándares seleccionados, más los perfiles de IHE.

Estas dos propuestas de integración se enfocan mayoritariamente en la interoperabilidad de equipos y dispositivos médicos.

En el contexto del *software* resaltan los *Frameworks* propietarios de Epic, Cerner y Athena, que juntos representan cerca del 70 % del mercado de Estados Unidos en Historias Clínicas Electrónicas. Aunque no promocionan el uso de ningún estándar, sí ofrecen 'interoperabilidad' entre sus clientes, con diversos modelos de cobro.

Dentro de este escenario cabe resaltar IHE, particularmente por la estrategia de desarrollo que ha logrado implementar. Es una iniciativa internacional diseñada para estimular y promover la interoperabilidad entre equipos y dispositivos médicos, y sistemas de información que apoyan la operación de las modernas instituciones de salud. Su objetivo fundamental es asegurar que, en el cuidado de los pacientes, toda la información requerida para decisiones médicas sea correcta y esté disponible a los profesionales de la salud. La estrategia fundamental de desarrollo es la competencia colaborativa. Representantes (voluntarios) de los productores de equipos y dispositivos médicos trabajan en estrecha cooperación con representantes de los usuarios y compradores, miembros de la Academia y Centros de Investigación y representantes de Organismos Oficiales de todos los países participantes.

Esta es la descripción del Ciclo de Desarrollo de los Perfiles IHE. Los usuarios, generalmente Clínicas y Hospitales, reportan necesidades de intercambio

de información, que son priorizadas y convertidas en Casos de Uso por los miembros de los diferentes grupos técnicos de cada dominio (voluntarios).

Se convoca, entonces, a los principales productores de los equipos participantes en cada Caso de Uso, quienes aportan horas de sus ingenieros para conformar un Equipo de trabajo que comienza por identificar los estándares óptimos para el Caso de Uso y con ellos construyen colectivamente el perfil correspondiente de interoperabilidad y lo publican como *Technical Frameworks*, al alcance de cualquier interesado. Los productores que lo deseen pueden implementar estas especificaciones del nuevo perfil en sus productos y lo someten a prueba en eventos masivos de interoperabilidad llamados Connectathon. Una vez probada la eficiencia del perfil, es publicado en su versión final y se incorpora, para demostración pública en eventos de divulgación *Interoperability Showcase* que habitualmente se realizan durante las Conferencias anuales de HIMSS.

A la fecha, se dispone de *Technical Frameworks* con varias docenas de perfiles en estos dominios, todos de acceso público:

Anatomía Patológica Cardiología Odontología Oftalmología Infraestructura de TI Laboratorio Farmacia	Coordinación de Cuidado de Pacientes Dispositivos de Cuidado de Pacientes Calidad, Investigación y Salud Pública Radiación, Oncología Radiología - Medicina Nuclear Radiología - Mamografía

Se ha desarrollado *software* Especializado de Monitoreo y Control (Gazelle) de los perfiles, que es utilizado en cada Connectathon. Igualmente, se cuenta con *software* de simulación y pruebas desarrollado por el National Institute for Standards and Technology (NIST) de Estados Unidos, adscrito al US Department of Commerce.

Esta es una lista de los 15 productores de equipos y *software* de salud que más frecuentemente han participado en Connectathones en América, Europa,

o Asia desde 2001, y la cantidad de veces que han probado perfiles de interoperabilidad en sus productos:

Participantes más frecuentes a Connectathon

VENDEDOR	NO.
GE Healthcare	40
Philips Medical Systems	35
Siemens Healthcare	35
AGFA Healthcare	32
Fujifilm	32
INFINITT	31
Carestream Health (Kodak)	24
Toshiba Medical Systems	24
Cerner Corporation	23
Hitachi Medical Corporation	22
International Business Machines	20
InterSystems Corporation	20
ITH icoserve technology for healthcare	20
Tiani-Spirit Gmbh	18
Epic Systems Corporation	16

Fuente: IHE.net

No cabe duda que los grandes productores tanto de *software* como de *hardware* en salud se han venido preparando, a lo largo de más de 15 años, para incorporar estos perfiles de interoperabilidad en sus productos, en un proceso con un costo prácticamente marginal y en muy corto tiempo. Dicho de otra forma, están preparados para dar el salto a la interoperabilidad completa, en el momento en que las demandas del mercado o reglamentarias se los exijan. Esto crea una gran oportunidad para los usuarios y compradores de los productos, que aunque totalmente indefinida en el tiempo, también

representa una gran amenaza para los productores en los países latinoamericanos, particularmente de *software*, que ni siquiera han comenzado a pensar en incorporar estándares internacionales en sus productos. En las condiciones actuales, podrían quedar fuera del mercado en cuestión de días, por obsolescencia de sus productos que súbitamente serían los únicos no interoperables del mercado.

POLÍTICAS PÚBLICAS

Con diferente grado de éxito en su implementación, los países latinoamericanos han iniciado esfuerzos por aproximarse a una solución al problema de la falta de interoperabilidad y a los costos y riesgos que esto conlleva. El estudio de la OPS/OMS sobre estándares de interoperabilidad para la eSalud en Latinoamérica (OMS, 2016) identificó la utilización parcial y dispersa de estándares internacionales, y encontró un número alarmantemente bajo de publicaciones científicas en la región relacionadas con el tema.

Estándar Identificado	No.
DICOM	12
Health Level 7 HL7 R2.0	4
HL7 Comm. Std	3
ICD-10	3
International Class. Primary Care	2
ISO	2
Nursing Intervention Classification	2
SNOMED	2
IHE	1
UMLS	1

Figura 1.1. Número y distribución de publicaciones relacionadas con estándares en salud

Bases consultadas: PubMed – LILACS – IEEE – CINAHL - Scopus
Fuente: Revisión de Estándares de Interoperabilidad para eSalud en L.A. OPS-2016

Brasil definió en 2014 su estrategia para incorporar la información electrónica en salud a nivel nacional (Ministerio Da Saúde, 2014). Es el primer documento oficial de estrategia o política que claramente identifica y recomienda el uso amplio de un estándar, en este caso los perfiles de interoperabilidad de

IHE. Sin embargo, aunque en su revisión de 2016 (Ministerio da Saúde, 2016) permanece la misma recomendación, los organismos especializados están desarrollando versiones básicas de estándares utilizando XML.

Argentina adoptó una estrategia mixta: "Marco Argentino de Interoperabilidad en Salud" (MAIS- http://www.mais.org.ar/) en la que actores del sector privado y público, principalmente prestadores de salud conforman equipos que desarrollan estándares de interoperabilidad basados en HL7 Ver. 3 y CDA.

Chile incorporó el tema de la información en salud en su "Agenda Digital 2020, Chile Digital para todos" con dos metas complementarias: [...] 6. Apoyar las políticas sectoriales del Estado a través del uso de tecnologías, y [...] 26. Informatización de la red de atención médica: historia clínica electrónica, que son implementadas a través de la estrategia Salud+Desarrollo con su Programa Nacional Estratégico de Tecnologías y Servicios de Salud y el Proyecto de "Cuenta médica interoperable".

En forma semejante, Uruguay había incorporado el concepto de la información en salud en su "Agenda digital Uruguay: 15 objetivos para 2015". El proyecto central HCEN: Historia Clínica Electrónica Nacional con el apoyo del Ministerio de Salud, consiguió realizar en diciembre de 2016 la primera Conectatón de Salud en América Latina, con participación de 21 Proveedores Privados de Servicios de Salud, 7 Proveedores Públicos y 14 Vendedores de Historias Clínicas. Es un ejemplo interesante de articulación público-privado, con participación de quienes pueden modificar con facilidad sus productos.

México ha avanzado en el tema con normativas específicas que guían el desarrollo de Sistemas de Información de registro electrónico para el intercambio de información en salud. Evidencia la posibilidad de utilizar, entre otros, estándares de IHE, o HL7".

Las Directrices y Formatos se basan en la Arquitectura de Referencia y los procedimientos emitidos por la Secretaría a través de la Dirección General de Información de Salud. Esta Arquitectura de Referencia puede considerar

los estándares y lineamientos publicados internacionalmente por IHE, HL7 o aquellos determinados por la propia Secretaría a través de la DGIS de acuerdo con los avances tecnológicos y la situación del Sistema Nacional de Salud.

Colombia ha incursionado en el tema de la interoperabilidad desde hace casi dos décadas cuando publicó su Marco de Interoperabilidad (MinTIC, 2010). Este documento definió los principios que debían aplicarse a la información de la estrategia Gobierno en Línea y fue complementado con la Guía de Uso del Marco de Interoperabilidad (MinTIC, 2011). Aunque estos dos documentos trajeron el tema de la interoperabilidad al centro de la discusión pública, no avanzaron en el tema específico de la información en salud. Los problemas inherentes a la falta de estandarización en la estructura de la información, en los datos y en los procesos de intercambio, se agudizaron con la adopción de la Política de Atención Integral en Salud, que genera nuevas necesidades de intercambiar e integrar información para procesar los pagos por los servicios prestados (MinSalud, 2016):

> El sistema de información se basa en la integración en interoperabilidad en la información entre los diferentes agentes. Se deben adoptar estándares (semánticos y sintácticos) integrados de interoperabilidad y gestión de información entre entidades públicas de diferentes sectores y operadores públicos y privados del sistema de Salud, que permitan la interacción entre repositorios e integración de información.

Se encuentra en estudio un borrador de Estándares para la Interoperabilidad en Salud enfocado específicamente en la Interoperabilidad semántica. Es claro que el concepto de estándares no es ajeno al sector Salud en Colombia, donde resalta la utilización de los Estándares ISO 9000, la Administración de Calidad Total y el enfoque de mejora de calidad Six Sigma (Hugo, 2016). Sin embargo, más allá de la misma tecnología y las políticas públicas relacionadas,

un componente que dificulta la situación en Colombia es la inestabilidad de los trabajadores de la salud, tanto en su capacitación, como en su seguridad laboral, que puede reducir el impacto de cualquier esfuerzo por estandarizar el intercambio de información en salud, al depender de personal con altas tasas de rotación, salarios poco atractivos y una capacitación poco enfocada en este aspecto de la tecnología (Piteres, Cabarcas & Gaspar, 2018).

MODELOS Y ESTRATEGIAS DE INTEROPERABILIDAD

A pesar del dinamismo observado en diferentes países latinoamericanos, no son frecuentes iniciativas que promuevan desde el Estado, la incorporación de estándares de interoperabilidad en sus sistemas de información y dispositivos médicos (OECD Health Policy Studies, 2016).

El estudio de la OECD se enfoca en dos dimensiones clave: a) el desarrollo y vinculación de datos de salud y atención médica, b) el desarrollo y uso de sistemas electrónicos de registros de salud. Un rol para la OECD en los próximos años es continuar apoyando a los países a alcanzar el objetivo de fortalecer la infraestructura de información de salud, a través del monitoreo continuo del desarrollo de la infraestructura de información de salud para ayudar a promover el aprendizaje compartido sobre los avances y desafíos en el adelanto y uso de datos de salud.

Otro paso clave será ayudar a los países a reducir los obstáculos innecesarios al uso de datos que pueden surgir a raíz de las diferencias en las legislaciones sobre la protección de la privacidad de la información de salud. En este contexto, trabajar juntos es particularmente importante, ya que hay reformas legislativas en el horizonte en muchos países, incluidas las propuestas a la Directiva de Protección de Datos que se traducirán en legislación dentro de los Estados miembros.

El Foro de la Organización Mundial de la Salud sobre Estandarización de Datos e Interoperabilidad (World Health Organization, 2012) convocó a los Ministros

de Salud de las naciones pertenecientes a la ONU en 2012 en Ginebra, para discutir el estado actual y futuro del tema. Los participantes del Foro incluyeron representantes de los Estados miembros de la OMS, organizaciones de desarrollo de normas sanitarias, instituciones académicas y de investigación, socios implementadores, la comunidad de donantes y expertos en la materia relacionados con el desarrollo, adopción e implementación de estándares de datos de salud en los campos nacional y sub-nacional.

Las discusiones trataron temas como: conceptos esenciales de datos de salud para la entrega de la atención en salud; las perspectivas de los países sobre la implementación de estándares de datos en salud; el acceso a estándares de datos de salud; los mecanismos de política y gobernanza existentes, maneras innovadoras de financiar estas iniciativas y talento humano para la implementación y mantenimiento. La barrera identificada más importante como inhibidor del proceso fue la falta de políticas para la adopción de los estándares. La segunda barrera se relaciona con la falta de enfoques nacionales hacia la implementación de estándares y la consiguiente falta de financiación para apoyar proyectos relacionados. La tercera barrera identificada fue el compromiso insuficiente de los países con el proceso de desarrollo e implementación de estándares, así como la falta de compromiso en general de los diferentes interesados en el proceso. Finalmente, el Foro identificó la necesidad de establecer recursos educativos y de información para habilitar a los interesados a navegar en los aspectos técnicos de la Información y Tecnologías en Salud (HIT).

El Foro acordó reunirse periódicamente y recomendó la participación más comprometida de los Gobiernos en estos procesos.

En 2014 se convocó el segundo Foro interministerial sobre el tema (World Health Organization, 2014). Se discutieron básicamente los mismos temas del primer Foro, con poco o ningún avance evidenciado por los países en vías de desarrollo. Se reconoció la urgencia de adoptar políticas coherentes que

permitan afrontar esta situación y se incluyó en el diagnóstico a los productores y comercializadores de equipos y *software*. Se sugirió la conveniencia de incluir en los términos de referencia de las adquisiciones de equipos y *software*, el requisito de incorporar estándares de interoperabilidad. Fue evidente que muchos de los retos para la implementación de los estándares son comunes a la mayoría de los Estados Miembros, y en consecuencia, la colaboración efectiva entre actores de diferentes países es crucial para el desarrollo de políticas habilitantes. Al finalizar el Foro no se acordó repetirlo en el futuro.

En febrero de 2018, la Organización Mundial de la Salud creó la asociación Global Digital Health (DGH- https://www.gdhp.org/) que reúne delegados de 23 países, agencias de gobierno, multinacionales y la WHO. El propósito es discutir, identificar y promover iniciativas en los temas de Ciberseguridad, Interoperabilidad, Evidencia y Evaluación, Entornos de Políticas, y Relaciones de Consumidor y Proveedor de Salud. A la fecha, de América Latina se han suscrito Brasil, Argentina y Uruguay. Sin que se haga mención de ello, esta asociación está creando condiciones adecuadas para promover las iniciativas transnacionales que fueron identificadas como cruciales en el Segundo Foro sobre Estandarización de Datos e Interoperabilidad.

En octubre de 2014, la Unión Europea, a través de su plataforma TIC, evaluó 27 perfiles de Interoperabilidad e IHE, los encontró adecuados, los reconoció como 'EU- ICT Technical Specifications' que pueden ser referenciados directamente en cualquier proceso de adquisición de equipos o *software* médico en los países de la Unión. Esta decisión fue publicada el 25 de julio de 2015 (European Union, 2015) en el Diario Oficial de la Unión Europea. Esta aceptación oficial crea un valioso precedente para dar fuerza a las demandas de los compradores, sobre la existencia de estándares de interoperabilidad en los productos que compren, como fuera sugerido en el Foro de 2014. Sin duda, es un ejemplo que podría ser seguido por los países latinoamericanos.

El modelo de Canadá para implementación de interoperabilidad

En el año 2001 el Gobierno de Canadá decidió invertir 500 millones de dólares como un 'fondo semilla' para apoyar iniciativas, cofinanciadas con el sector privado, orientadas a promover el uso de la información digital en salud. Este fondo sirvió para crear el organismo sin ánimo de lucro InfoWay Canada que comenzó a implementar el Plan Estratégico *Opportunitites for Action: a Pan-Canadian Digital Health Strategic Plan* (Inforoute Canada, 2001). La propuesta del plan fue crear una estructura de 'Habilitadores Clave' que permitieran construir cinco metas de largo plazo. Los habilitadores identificados son:

- Gobernanza y Liderazgo
- Política Pública y Legislación
- Capacidad y Competencias de Recursos. Cultura
- Financiación
- Privacidad y Seguridad
- Soluciones Digitales de Salud Interoperables
- Desarrollo de Casos de Negocio y Beneficios
- Cambio de Prácticas y Procesos

Estos ocho habilitadores han servido como base para avanzar hacia las cinco metas:

- Llevar la atención más cerca de casa
- Proveer acceso fácil 24/7
- Apoyar nuevos modelos de atención
- Mejorar la seguridad del paciente
- Habilitar un Sistema de Información de Alto Desempeño.

A partir de su creación, InfoWay ha solicitado un aporte del Gobierno de aproximadamente $300 millones anuales, para una inversión pública total hasta 2018 cercana a los $2.800 millones. Este es un resumen de los resultados luego de poco más de 15 años de acción de InfoWay Canada (Inforoute Canada, 2018):

- » $26.000 millones en ahorro y mejora de eficiencia desde 2007 (de $19.200 en el año anterior).
- » La red de información en salud produce al año un estimado de 18 millones de horas ahorradas a los pacientes y 5,9 millones de horas ahorradas a los prestadores de salud.
- » Cerca de 191.000 médicos usan la Historia Clínica Electrónica. Un incremento superior al 300 % en cinco años.
- » Los datos de la Historia Clínica Electrónica de más del 96 % de los canadienses son accesibles para los proveedores de salud, así como para los individuos.
- » El 85 % de los médicos de atención primaria usan la Historia Médica Electrónica.
- » Cada $1 invertido en programas de atención domiciliaria a distancia ACCESS (activo desde 2010) genera $4 en valor al sistema de salud en el primer año.
- » Más de 31.500 pacientes se han beneficiado del programa de atención domiciliaria a distancia. Esto evitó 45.000 visitas a emergencia y más de 26.000 hospitalizaciones, representando 188.000 días de hospitalización.
- » PrescribeIT (teleprescripción) está activo en 6 provincias y trabaja con 27 cadenas de farmacias y 3.400 tiendas, a los tres meses de su inicio.

No hay duda de que se trata de indicadores muy positivos tanto desde el punto de vista de la atención a la salud y la obtención de las metas estratégicas, como desde la dimensión del negocio.

En abril de 2018 se hizo el lanzamiento de InfoWay 2.0 con una visión empresarial aún más ágil y creativa. Esto incluye una nueva estructura organizativa que permite un mayor enfoque en ACCESS y PrescribeIT (prescripción electrónica).

El modelo de Estados Unidos para implementación de interoperabilidad

En 1996, es emitida la HIPAA Act (*Health Insurance Portability and Accountability Act*), que requiere estándares para las transacciones electrónicas en la atención de salud. Los estándares se orientan a mejorar la efectividad y eficiencia del sistema por medio de:

- Identificación única de pacientes (UPI)
- Privacidad de la información del paciente
- Seguridad de la información de salud.

Como dato curioso, pero de gran impacto, en 1999 el Congreso prohíbe el uso del Identificador Único de Pacientes (UPI) por razones de privacidad. Hasta la fecha se han realizado cuatro intentos de derogar esta prohibición y aún no se logra, a pesar de que los errores por falsa identificación del paciente son numerosos y están debidamente registrados.

En 2004 se crea la Oficina del Coordinador Nacional de Tecnología de Información en Salud (ONC), que lidera los esfuerzos nacionales de TI de la salud, y está a cargo de la principal entidad federal para coordinar los esfuerzos a nivel nacional para implementar y utilizar la tecnología de información de salud más avanzada y el intercambio electrónico de información de salud.

El Gobierno Federal promulga en 2015 el Plan Estratégico de Tecnología en Salud 2015-2020 (HHS Department, 2015) que da los lineamientos generales para la implementación de TIC en salud a nivel nacional. Casi simultáneamente, la ONC publica la propuesta de implementación de ese Plan Estratégico enfocada en la interoperabilidad (ONC, 2015). La propuesta de implementación busca generar un ecosistema de TIC en Salud como un Sistema de Salud que aprende: "... un ecosistema en el que todos los interesados pueden en forma segura, eficaz y eficiente contribuir, compartir y analizar datos y crear nuevo conocimiento que puede ser usado por una amplia variedad de

sistemas electrónicos de información de salud para apoyar la toma efectiva de decisiones para mejorar los resultados de salud".

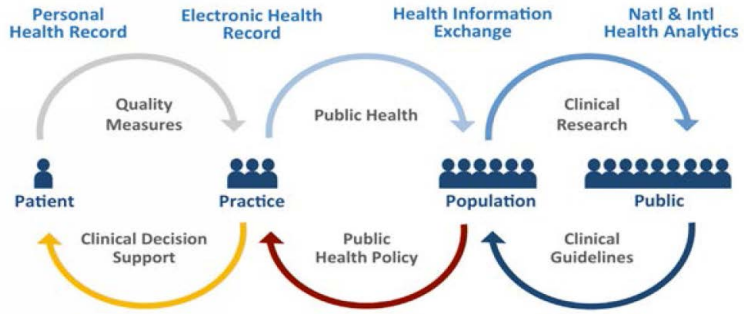

Figura 1.2. Representación gráfica del modelo de Sistema de Salud que aprende
Fuente: ONC

El sistema visiona una creciente complejidad en herramientas, actores, procesos e interacciones entre los integrantes.

Así, a nivel de paciente se maneja la Historia Clínica Personal sobre la cual se aplican procesos de mejoramiento de calidad para integrar las Historias Clínicas Electrónicas que son usadas por los profesionales de la salud para atender al paciente y generar el soporte de sus decisiones clínicas. La información acumulada, a nivel de población, a través de los sistemas de Intercambio de Información en Salud determinan las Políticas de Salud Pública. Esta información sirve para realizar análisis clínicos a nivel nacional e internacional, generando nuevas o mejoradas guías clínicas. Es un sistema altamente dinámico, con participación y responsabilidades distribuidas en todos los niveles.

El Plan propone diez principios de interoperabilidad, a saber:

- Construir sobre la infraestructura de salud existente
- Considerar el entorno actual y apoyar niveles múltiples de avance
- Proteger la seguridad y privacidad en todos los aspectos de interoperabilidad
- Mantener la modularidad

- Empoderar al individuo
- Apalancar el mercado
- Acceso universal y escalable
- Una talla no le sirve a todos
- Simplificar
- Foco en el valor.

Adicional a estos principios transversales, el Plan define un conjunto básico común de datos para las Historias Clínicas Electrónicas.

Simultáneamente con la creación del ONC, es aprobada la Ley HITECH: *Health Information Technology for Economic and Clinical Health*, que requiere el uso de Historias Clínicas Electrónicas, el acceso de los pacientes a sus datos y la interoperabilidad. ONC es el encargado de construir un sistema de información de salud interoperable, privado y seguro a nivel nacional y apoyar el uso generalizado y significativo de la tecnología de la información de salud.

En cumplimiento de este mandato, en 2009 se reglamenta HITECH e inicia implementación de la estrategia Meaningful Use (MU) a partir de incentivos (reembolsos) a quienes cumplan los requisitos dentro de períodos establecidos, y penalizaciones luego de esos períodos. Esta es la característica predominante del modelo de implementación en Estados Unidos: inversión del Estado (zanahoria) para estimular la adopción de los requisitos establecidos, y garrote (multas) para quien no cumpla en el plazo establecido.

La estrategia de Uso Significativo (MU) se divide en tres etapas con estos criterios y resultados esperados:

- MU1- desde 2011 Captura/comparte datos. Funcionalidad básica de Historias Clínicas Electrónicas (Conjunto básico de datos del Plan).
- MU2- desde 2013 Procesos avanzados de atención con apoyo a toma de decisiones. Certificación de Historias Clínicas Electrónicas. Participación de Pacientes. Intercambio de información dentro de las instituciones.
- MU3- desde 2015 hasta 2017 Resultados mejorados.

Los requisitos de MU3 son:

- Información de salud protegida (PHI)
- Prescripción electrónica
- Soporte a decisiones clínicas (CDS)
- Ingreso automatizado de órdenes (CPOE) para laboratorio, medicamentos y diagnóstico por imágenes
- Acceso electrónico de pacientes a su Historia Clínica
- Coordinación de la atención mediante la participación del paciente
- Intercambio de Información de Salud (HIE). Intercambio de Historias en la referencia de pacientes, o en cuidado continuo. Transferencia de historia a médico tratante por primera vez. Prescripción electrónica
- Reportes de salud pública y datos de registros clínicos

Al finalizar 2013 estos fueron los resultados de MU1:

- % de Hospitales que se acogen al beneficio de MU-1

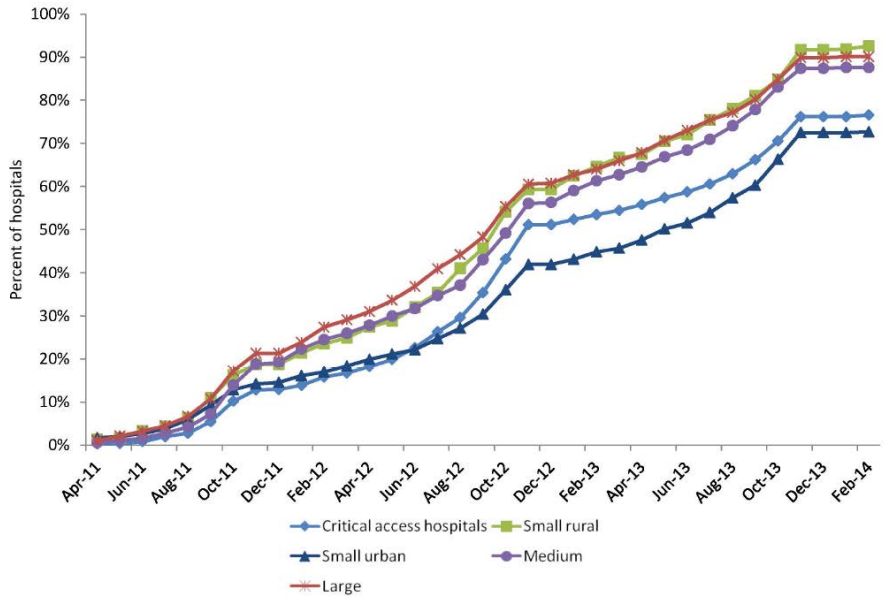

Figura 1.3. % de Hospitales que se acogen al beneficio de MU-1.
Fuente: ONC Report to Congress

Es evidente una detención abrupta del ritmo de adopción una vez se agotaron los fondos de esa etapa, independiente del tipo de usuario, reflejando un interés en el reembolso, por encima de un compromiso con la adopción de la tecnología para mejorar el servicio.

En su informe al Congreso en 2016, la ONC reporta los siguientes resultados:

- 96 % de los Hospitales y 78 % de los médicos particulares cuentan con Historias Clínicas Electrónicas certificadas.
- El 82 % de los Hospitales de atención aguda han intercambiado información con proveedores de atención ambulatoria u Hospitales fuera de su organización, aunque esto no constituye una práctica común y establecida.
- 95 % de los Hospitales de atención aguda le dan a sus pacientes la opción de ver su información de salud; 87 % les permiten descargarlos; 71 % los habilitan para transmitirlos y 69 % pueden hacer los tres procesos: ver, descargar y transmitir la información de salud.
- 51 % de los Hospitales pueden realizar búsquedas electrónicas de información de salud crítica desde lugares externos, como una sala de emergencias o un consultorio particular.

Los cuatro principales obstáculos a la interoperabilidad identificados por la ONC son:

a. La información de salud no está suficientemente estandarizada;
b. Falta alineación con el pago de incentivos;
c. Errores de interpretación y diferencias en las leyes existentes sobre privacidad;
d. Falta de confianza.

Las cinco principales barreras a la interoperabilidad, identificadas por los Hospitales usuarios de los sistemas, son:

a. El destinatario no tiene Historia Clínica Electrónica, u otro sistema para recibir los datos, 59 % de los Hospitales;
b. El sistema del receptor no tiene la capacidad para recibir los datos, 58 % de los Hospitales;
c. Dificultad para encontrar la dirección (electrónica) del proveedor, 45 % de los Hospitales;
d. Proceso difícil e incómodo para enviar información desde un sistema de Historia Clínica Electrónica, 30 % de los Hospitales.
e. Muchos receptores reportan que los resúmenes de la Historia Clínica enviados no son útiles.

La inversión en incentivos en los ocho años de vigencia de la estrategia 'Meaningful Use' se calcula en $30.000 millones.

Una comparación de los resultados reportados por las correspondientes entidades responsables en Canadá y Estados Unidos, Inforoute y ONC permite extraer algunas conclusiones iniciales:

- El modelo de Canadá es significativamente más eficiente económicamente. Ha demostrado ahorros financieros que representan nueve veces la cantidad invertida.
- La propuesta de cofinanciación de iniciativas en busca de las metas estratégicas generó un entorno de apropiación colectiva público-privada que le da coherencia y sostenibilidad al proceso.
- El énfasis en la Telemedicina (por razones coyunturales de dispersión de la población) contribuyó a darle mayor importancia a la transferencia confiable de datos clínicos.
- La propuesta de Inforoute, estimula la identificación y financiación de iniciativas priorizadas por los mismos usuarios, a partir de sus expectativas tanto económicas como técnicas y de servicio, pero todas con un propósito común: las metas estratégicas.
- Esta dinámica de colaboración puede explicar el que se haya alcanzado el indicador sorprendente del 96 % de Historias Clínicas Electrónicas

remotamente accesibles por diversos proveedores y por los individuos, sin que se haga evidente ningún conflicto entre los proveedores de los sistemas de información para llegar a ese nivel de interoperabilidad.

- A pesar de demostrar su sostenibilidad en forma temprana, el Gobierno canadiense ha continuado aportando fondos a Inforoute para nuevas iniciativas prioritarias, con un monto razonablemente bajo para el tamaño de los resultados (aproximadamente $300 millones por año).
- Por su parte, el modelo de Estados Unidos ofrece algunos indicadores comparables en su valor absoluto (95 % de los Hospitales permiten a sus pacientes ver su información de salud), pero cuestionables en cuanto a su eficiencia si se analizan las barreras identificadas por esos mismos usuarios.
- El modelo es un barril sin fondo de inversiones, a fondo perdido, en el que el primer beneficiado es el prestador de salud que recibe reembolsos como premio por adoptar determinada tecnología; el beneficio para el paciente, sea en calidad, o tiempo de atención es secundario al proceso.
- El modelo genera competencia (carrera) por el premio en dinero, pero falla estrepitosamente al no generar una apropiación del proceso.
- Todas las iniciativas han estado definidas (no abiertamente) por la influencia de los grandes productores de *Software* Clínico, que siguen luchando por su propuesta de monopolio informático, que llaman interoperabilidad.

En medio de este contexto dinámico y poco definido del mercado de la información de salud en Estados Unidos, la ONC lanza en marzo de 2016 una iniciativa que puede ser el inicio de un punto de coyuntura: se trata de la 'Interoperability Pledge Strategy', en español ´El compromiso por la Interoperabilidad'. En forma voluntaria, las empresas que proporcionan el 90 % de las historias clínicas electrónicas utilizadas por los hospitales de todo el país, junto a Asociaciones Profesionales y a las mayores Corporaciones de

Prestación de Servicios de Salud se han comprometido en 47 Estados con estos principios fundamentales:

- Acceso del consumidor: para ayudar a los consumidores a acceder de manera fácil y segura a su información de salud electrónica.
- Sin bloqueo/transparencia: para ayudar a los proveedores a compartir la información de salud de las personas con otros proveedores y sus pacientes.
- Estándares: implementar normas, políticas, guías y prácticas de interoperabilidad nacional reconocidas a nivel federal para la información de salud electrónica.

Aunque las condiciones de competencia predatoria entre estas empresas no han cambiado, la firma de este acuerdo ha creado un contexto temporal para que se desarrollen algunas interesantes iniciativas que mejoran las perspectivas de interoperabilidad en el futuro inmediato, como se verá en las siguientes secciones de este documento.

Finalmente, en diciembre de 2016, terminando su mandato, el presidente Obama aprobó la Ley "21st Century Cures Act" que hace obligatoria la interoperabilidad de las Historias Clínicas Electrónicas a finales de 2017. Conflictos con los primeros decretos del presidente Trump sobre reglamentación, han obligado a retrasar, al menos en un año, las metas de esta nueva Ley, que plantea profundos cambios a la estructura del modelo de prestación de salud, entre ellos, el intercambio electrónico de información. Los elementos más resaltantes de la Cures Act en esta materia son:

- Incluye mandatos para mejorar la TI de la atención médica, sobre todo en relación con la interoperabilidad a nivel nacional y el bloqueo de la información. De repente, esas "promesas de interoperabilidad" que los proveedores de EHR firmaron a principios de año no serán solo expresiones de buena voluntad.

- Ciertas secciones se centran en "mejorar la calidad de la atención para los pacientes" en el área de la tecnología de la información, con la interoperabilidad como la preocupación principal y central. HHS recibirá $15 millones en fondos para cambiar el proceso de certificación de ONC, para ayudar a impulsar la interoperabilidad y luchar contra el bloqueo de información por parte de los proveedores de Historias Clínicas Electrónicas.
- Para certificarse, los proveedores deben desarrollar interfaces de programación de aplicaciones (API) u otras tecnologías para permitir el acceso, el intercambio y el uso de la aplicación sin un esfuerzo especial. También deben haber probado con éxito el "uso real de la tecnología para la interoperabilidad". Los proveedores que estén bloqueando la información están sujetos a sanciones de hasta 1 millón por infracción.
- El objetivo será establecer alianzas público-privadas para generar consenso y desarrollar un "marco de intercambio de confianza, que incluya un acuerdo común entre las redes de información de salud a nivel nacional".

Dimensión tecnológica

Dentro del ambiente de gran competitividad en el sector de las tecnologías e información en salud, destacan las siguientes tendencias en el futuro próximo:

- Blockchain
- Inteligencia Artificial: seguridad, diagnóstico y reconocimiento de imágenes
- Data AnalyTIC en información de salud
- Salud Digital.

¿Cómo se puede usar blockchain en el cuidado de la salud?

- Rastreabilidad de medicamentos: cada transacción entre los fabricantes de medicamentos, mayoristas, farmacéuticos y pacientes puede rastrearse.
- Mejora y autenticación de registros y protocolos de salud en el intercambio de registros.
- Contratos inteligentes con métodos basados en reglas para el acceso de datos del paciente. Se pueden otorgar permisos a organizaciones seleccionadas.
- Ensayos clínicos: erradica la alteración o modificación fraudulenta de datos de ensayos clínicos.
- Medicina de precisión: pacientes, investigadores y proveedores pueden colaborar para desarrollar cuidados individualizados.
- Investigación genómica: acceso a datos genéticos asegurados en blockchain.
- Historia Clínica Electrónica.
- Interoperabilidad a nivel nacional.

Blockchain podría resolver y agilizar la mayoría de las preocupaciones relacionadas con la conectividad, la privacidad y el intercambio de registros de pacientes. La interoperabilidad de TI en Salud habilitada por blockchain tiene la capacidad de transferir datos relevantes del paciente de un proveedor a otro, sin importar la ubicación o la Historia Clínica Electrónica particular de los proveedores. La falta de estándares para esta tecnología, aún inmadura, está causando incertidumbre regulatoria, mientras que la industria anticipa orientaciones de las reglas federales en algún momento en 2018.

ALGUNAS APLICACIONES ESPECÍFICAS DE BLOCKCHAIN A LAS HISTORIAS CLÍNICAS ELECTRÓNICAS:

Blockchain puede ayudar a mejorar tres características principales de los sistemas de HCE que son:

- Inmutabilidad a través de Integridad de archivo donde cada evento en el blockchain tiene un hash único correspondiente al contenido de un registro. Esto significa que los usuarios pueden verificar si el contenido ha sido modificado o no.
- Ciberseguridad a través de Gestión de Acceso a Datos, donde cada hash puede contener permisos de usuario particulares para médicos, pacientes, enfermeras o cualquier usuario o dispositivo autorizado. Por lo tanto, solo el personal acreditado puede acceder a la información del registro.
- Interoperabilidad a través de Control Colaborativo de Versión donde cada parte tiene un registro vinculado al original que está guardado en blockchain. De esta forma, todos los que tengan el rol y la responsabilidad apropiados, pueden agregar información al registro evitando problemas como inconsistencias o duplicados.

El Instituto ECRI (https://www.ecri.org) publica anualmente la lista de los Top Ten Technology Hazards y en 2018 identificó como el más alto riesgo: Ramsomware y otras amenazas de ciberseguridad a la entrega de la atención de salud que pueden poner en peligro a los pacientes. En su reporte de Top Ten Technology Hazards for 2019, nuevamente el primer lugar es ocupado por Hackers que pueden explotar el acceso remoto a los sistemas, entorpeciendo las operaciones de la atención de salud.

La ONC apoya el uso de blockchain como herramienta para promover el intercambio en condiciones de seguridad. En agosto de 2016 lanzó la campaña "The Use of Blockchain in Health IT and Health-related research".

En el dominio de la Inteligencia Artificial y diagnóstico resaltan aplicaciones que ya se vienen desarrollando hace algún tiempo, pero que pueden cobrar un ritmo de crecimiento acelerado con el apoyo de las TIC:

- Reconocimiento de imágenes y análisis de datos para apoyo en toma de decisiones clínicas.

- Uso de Big Data para gestión de Salud Pública, Medicina de precisión y Diagnóstico Preventivo.
- Un mejor uso de la computación en la nube y una mayor exploración de la inteligencia artificial y blockchain reforzará la protección de los datos del paciente.

Data AnalyTIC sigue creando tendencia en todos los sectores, como puede verse en los otros capítulos de este libro y en el campo de la Tecnología en Salud, con los sistemas de registro electrónico de salud (EHR) en plena vigencia en toda la industria médica; 2019 será el año en que los proveedores de EHR explorarán más sólidamente el análisis de datos de pacientes. Las ventajas de hacerlo son claras, como mejorar los esfuerzos de salud de la población y proporcionar informes en tiempo real entre varias especialidades de atención médica. La mayoría de los médicos utilizan algunas capacidades de análisis ya integradas en sus sistemas de EHR, según encuestas realizadas en 2017. El potencial de los gigantes de EHR, como Epic Systems y Cerner para agregar datos de pacientes dentro de sus productos es una nueva línea de negocios que está floreciendo. Cerner ya ha iniciado su participación en la gestión de la salud de la población, que requiere cálculos numéricos intensivos para medir los resultados de los pacientes.

Finalmente, en el campo de la salud digital, los dispositivos móviles se están volviendo cada vez más frecuentes en la atención médica, con la proliferación de teléfonos inteligentes y tabletas. La Administración de Alimentos y Medicamentos (FDA) también emitió una nueva guía que relajó sus regulaciones para ciertas tecnologías de salud móviles como rastreadores de actividad física y aplicaciones de salud. La salud digital permite a los ancianos envejecer y vivir mejor en sus propios hogares, usando tecnología como monitores de detección de caídas.

La telesalud y la telemedicina también formarán parte de la transformación de la salud digital. En Colombia es justo reconocer los avances que ha obtenido el Sistema de Salud de Antioquia en este campo, al concentrar esfuerzos de

investigación y desarrollo alrededor de la telemedicina. Un área que se está ampliando es la de los servicios de telesalud mental y conductual.

En la primera mitad de 2017 se invirtieron 3.500 millones de dólares en 188 compañías digitales de salud, y se proyecta que la cantidad de dispositivos portátiles (*wearables*) llegará a 34 millones para el año 2022.

El Internet de las Cosas (IoT) ha impactado considerablemente en todos los sectores de la economía. A nivel de la atención de la salud, estos son algunos ejemplos en que IoT puede ser utilizado en el cuidado de la salud:

- 60 % de las organizaciones de atención médica han introducido dispositivos IoT en sus instalaciones.
- 73 % de las organizaciones de atención médica usan IoT para monitoreo y mantenimiento.
- El uso más común de la tecnología IoT en la industria de la salud son monitores de pacientes (60 %), seguidos de contadores de energía (56 %).
- En la industria de la salud, la mayoría de los encuestados (80 %) citan el aumento de la innovación como el principal beneficio de IoT.
- 89 % de las organizaciones de atención médica ha sufrido una infracción de seguridad relacionada con IoT.
- 87 % de las organizaciones de salud planea implementar la tecnología IoT para 2019. Un poco más que el 85 % de las empresas en todas las industrias que planean hacerlo.

Datos tomados del informe The Internet of Things: Today and Tomorrow (2017), encuestó a 3.100 tomadores de decisiones de TI y de negocios en 20 países para evaluar el impacto del IoT en diferentes industrias. Este crecimiento acelerado del número de dispositivos, cada día más económicos, que pueden captar, procesar y transmitir datos e información de salud tienen el potencial de complicar, aún más el contexto de la interoperabilidad, pues si bien son diseñados en cumplimiento de estándares internacionales como IEEE,

ISO, FCC, sin embargo en muy pocos casos tienen en cuenta los estándares propios de la información en salud que se han planteado en este documento (Jimeno, De la Hoz, & Wilches, 2014).

A nivel de estándares e interoperabilidad, la Figura central en el escenario del crecimiento acelerado de usuarios es FHIR, *Fast Healthcare Interoperability Resources*, una especificación de HL7 que describe cómo soportar una API con el fin de intercambiar datos entre sistemas HIT. La promesa de FHIR es que hará que los elementos de datos más finos y granulares sean accesibles y transferibles a través de interfaces API. En lugar de tener que intercambiar documentos completos o conjuntos de datos más grandes, FHIR utiliza el concepto de "Recursos" para delinear conjuntos muy básicos de datos estructurados. Un recurso, por ejemplo, podría definirse como una lista de medicamentos o resultados de laboratorio.

La empresa McKesson Health Solutions está promoviendo su 'Intelligence Hub' que ofrece interoperabilidad impulsada por API basadas en FHIR. El Hub es diseñado utilizando estándares de comunicación abiertos de la industria, como HL7 FHIR. Intelligence Hub se conecta con desarrolladores de aplicaciones y sistemas de datos de salud para construir servicios digitales basados en API y FHIR.

Por su parte, IHE ha fortalecido su alianza con HLT y creado el grupo de trabajo IHE-FHIR para adaptar sus perfiles de interoperabilidad a desarrollos con FHIR. Las colecciones de perfiles IHE reutilizables pueden emplearse en diferentes configuraciones para poner en funcionamiento múltiples flujos de trabajos de cuidados basados en directrices, por ejemplo, atención prenatal, inmunización, VIH. Cada Perfil IHE describe con precisión cómo se aprovecha una pila de estándares básicos subyacentes (especificaciones de contenido, codificación, protocolos de comunicación, etc.) para cumplir con un determinado nivel de funcionalidad reutilizable, por ejemplo, consulta demográfica del paciente (PDQ).

La tendencia a migrar y gestionar la información a servicios en la nube –*CLOUD Computing*– (Gutiérrez, Almeida & Romero, 2018), aunque no resuelve en sí misma el problema de la interoperabilidad, sí puede crear condiciones más favorables a la utilización de estándares para el intercambio de la información en salud.

DIMENSIÓN DEL MERCADO

Como ha quedado evidente en las secciones previas, el mundo de la información en salud representa un mercado en constante ebullición, afectado por aspectos no solo de tecnología, sino de políticas y claro, del mismo mercado. En este contexto solo es posible identificar tendencias para el futuro muy próximo:

- Los fabricantes de equipos médicos buscarán asociarse con proveedores e integradores de EHR, para encontrar modos de monetización basados en el valor.
- La atención de la salud se está moviendo hacia una interoperabilidad en toda la industria que no puede emerger completamente hasta que la tecnología, para permitirla, haya madurado.
- Hay una buena posibilidad que a mediados de 2019 se comenzará a ver un beneficio real en los indicadores en Estados Unidos.

Al observar el comportamiento de los dos segmentos del mercado, oferta y demanda, es posible reconocer algunas tendencias:

Tendencias del Mercado (oferta) 2015-2019

- Aproximación de actores principales alrededor de la interoperabilidad
- Posicionamiento de las propuestas de HIE (*Health Information Exchange*)
- Foco de los servicios en la integración de sistemas
- Primeros pasos de aproximación de productores de equipos
- Experimentación con tecnologías emergentes
- Ingreso/diversificación de nuevos actores
- Interacción de actores público-privado.

Tendencias del Mercado (demanda) 2017-2019

- Incremento de generación de datos desde los usuarios
- Movilidad de usuarios según perfil de oferta (información)
- Organizaciones buscando sistemas de integración.

La primera tendencia de la oferta de mercado se basa en una iniciativa que comenzó en octubre de 2015, cuando 12 de los más grandes productores de Historias Clínicas Electrónicas, representando el 65 % del mercado de Estados Unidos se reúnen para definir métricas de interoperabilidad. Como resultado de esa reunión, en septiembre de 2016 publicaron un listado de parámetros que utilizarán para hacer un ejercicio de *Benchmarking* sobre interoperabilidad. Los participantes de esta alianza voluntaria son:

- AllScripts
- athenaHealth
- Cerner
- eClinicalWorks
- Epic
- GE Healthcare
- HCIT Greenway
- Healthland
- McKesson
- MEDITECH
- MEDHOST
- NextGen

Los cuatro parámetros sobre los que proponen realizar el *Benchmarking* son:

1. Disponibilidad de la información necesaria
2. Facilidad para localizar registros
3. Capacidad de ver registros externos dentro del flujo de trabajo clínico
4. Impacto en el cuidado del paciente.

El inicio del estudio se programó para 2017; todos los participantes firmaron el Interoperability Pledge de la ONC, pero a la fecha no se conocen los resultados del análisis. Sin embargo, algunas iniciativas de negocio han surgido recientemente, que integran en el mismo negocio, competidores hasta entonces acérrimos, como Carequality, un proyecto de intercambio de información entre Historias Clínicas Electrónicas de diferentes productores, apoyado por ONC bajo su Proyecto Sequoia. La iniciativa reúne a casi la totalidad de los firmantes del acuerdo de *Benchmarking*, más organizaciones prestadoras de salud, más asociaciones profesionales, más desarrolladores de estándares y otros actores del sector. Da la impresión que la movilización de los agentes económicos, alrededor de un potencial próspero negocio, con visos de monopolio ha logrado más resultados en menos tiempo que la larga y costosa iniciativa oficial de casi 15 años. A la fecha se reportan más de 600.000 médicos particulares, 35.000 clínicas y 1.250 Hospitales asociados al Carequality, incluyendo gigantes del sector como Kaiser Permanente.

En febrero de 2018, el CTO/CIO del Proyecto Sequoia (Carequality) se une al grupo de trabajo central de datos de ONC para Interoperabilidad.

Junto a estas mega-iniciativas, muchas otras de menor escala se han registrado en los últimos meses. Estas son algunas en orden cronológico:

- Diciembre 2017. Philips adquirió Forcare, una solución de *software* de interoperabilidad basada en estándares abiertos para flujos de datos rápidos e impecables entre sistemas médicos y fuentes de información a nivel departamental y empresarial, así como intercambios de información médica (HIE) en todos los sistemas de salud. Forcare y su equipo proporcionarán capacidades altamente complementarias para Philips, y permitirán a Philips ofrecer soluciones informáticas integradas y más efectivas que mejoran el flujo de trabajo clínico, mejoran la atención del paciente y optimizan la gestión empresarial. Philips es el segundo participante más frecuente a las Connecthaton de IHE.

- Enero 2018. Por iniciativa de ONC, los gigantes de EHR Epic y Cerner coordinan junto con el Departamento de Salud y Servicios Humanos Oficina del Coordinador Nacional de TI de Salud para crear interoperabilidad. Los proveedores de EHR están trabajando con aplicaciones de salud móviles de terceros para que los pacientes puedan acceder a su información de salud; en este caso, información de medicamentos, por medio de estas aplicaciones móviles. Las dos empresas que participan en la iniciativa y las aplicaciones móviles de terceros pudieron conectarse y compartir información a través del estándar de recursos de interoperabilidad de Fast Healthcare (FHIR).
- Enero 2018. Epic Systems, anunció el lanzamiento de su nueva iniciativa 'One Virtual System Worldwide' para garantizar que los médicos de todas las organizaciones que usan Epic no solo puedan intercambiar más datos, sino que también puedan interactuar entre sí. La iniciativa One Virtual System Worldwide consta de tres partes:
    1. Come Together: Epic busca registros de pacientes y reúne sus datos de organizaciones Epic, organizaciones que usan otros EHR, Gobiernos y redes.
    2. Happy Together: el portal para pacientes MyChart les permite a los pacientes tener una vista combinada de su registro de atención médica sin tener que iniciar sesión o salir a diferentes portales.
    3. Working Together: realice acciones a través de las organizaciones usando imágenes, haga reservas, envíe mensajes, busque en cualquier lugar, tenga telesalud en cualquier lugar, "Estamos tomando la interoperabilidad de poder 'ver más' para poder 'hacer más'.
- Enero 2018. Durante la Conferencia HIMSS18, Google presentó su nueva API de Cloud Healthcare, que aborda los importantes desafíos de interoperabilidad en los datos de atención médica. La nueva API proporcionará una solución de infraestructura robusta y escalable para ingerir y gestionar tipos de datos clave de atención médica, incluidos HL7, FHIR y DICOM.

Los estándares abiertos son fundamentales para la interoperabilidad de la atención médica, así como para permitir la investigación biomédica. Hemos estado usando la API de Google Cloud Genomics durante mucho tiempo y estamos muy emocionados de ver a Google Cloud expandir sus ofertas para incluir la nueva API de Cloud Healthcare. La capacidad de combinar la interoperabilidad con el análisis escalable de Google Cloud tendrá un impacto transformador en nuestra comunidad de investigación.

VISIÓN INTEGRAL

El panorama de la Interoperabilidad descrito hasta ahora, a partir de hechos y documentos, es complejo y confuso. Está compuesto por actores del Gobierno, de la industria y por los pacientes. Estos actores están presentes en todas las dimensiones desde la personal hasta la internacional y participan en una o más de cuatro dimensiones que se intersectan: Política Pública, Tecnología, Demanda de Mercado y Oferta de Mercado. Tal vez la forma más intuitiva de ver la imagen completa sea a través de un gráfico de tiempo que represente los hitos de cada una de las dimensiones identificadas en este documento, así:

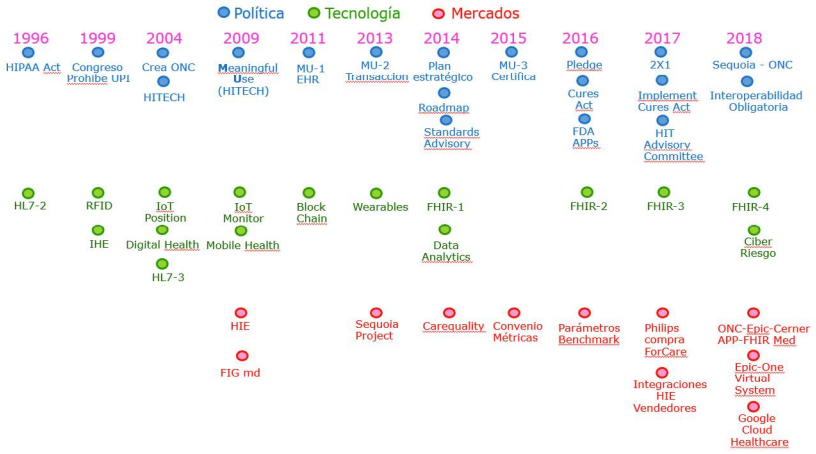

Figura 1.4. *Timeline* de Hitos relacionados con las tres dimensiones de interoperabilidad.
Fuente: Construcción propia

La observación de esta gráfica permite reconocer algunos hallazgos en correspondencia con el análisis individual de los hechos, pero enriquecidos desde la visión integral de sus interacciones:

- Es evidente la continuidad y fortalecimiento de la Política Pública en 20+ años.
- La tecnología puede presionar la política, atender tiempos, ritmos.
- La adopción voluntaria de estándares/interoperabilidad es poco realista.
- La integración de vendedores de *software* con las empresas de intercambio de información en salud, HIE, crea nuevo modelo de negocio.
- Este modelo requiere vendedores+usuarios con apoyo de sector público.
- Esta HIE 'recargada' usa estándares abiertos ==> facilita intercambio.
- Estrategia de 'concentrar interfases' facilita interoperabilidad y competitividad.
- Incremento de la presencia de actores 'Grandes Ligas' en el mercado HIE.
- Mayor articulación de los esfuerzos público-privados.
- Los actores pesados continúan tratando de mantener 'feudos'.
- Productores de equipos se aproximan más a apoyar el intercambio de datos.

CONCLUSIONES

Los países latinoamericanos se encuentran en una encrucijada técnico-política-económica en lo que se refiere a la interoperabilidad de equipos y sistemas de información en salud. Existe más que suficiente información, referencias y ejemplos en el mundo, con niveles variables de éxito, costo y duración, que pueden y deben ser utilizados para obtener un 'leapfrogging' que nos permita avanzar rápida y eficazmente, a un costo bajo. El panorama

de alternativas (tonos de gris, según el ONC Doug Fridsma) estaría conformado por dos opciones iniciales:

a. No hacer nada y esperar que las fuerzas del mercado definan el futuro (seguidores) con la amenaza de actores externos, ejemplo: Infor Coverleaf Health Information Exchange, que están atentos a tomarse los relativamente pequeños pero rentables mercados de Latinoamérica.
b. Identificar fortalezas, necesidades y oportunidades a nivel local y regional, y ACTUAR.

Dentro de las fortalezas identificadas pueden resaltarse:

- Todos los países de la región tienen algún avance en cada una de las dimensiones de interoperabilidad previamente identificadas.
- Existen grupos localizados que trabajan en interoperabilidad, principalmente en Universidades y Hospitales.
- Algunas empresas de desarrollo de *software* han estado exportando durante años.
- Existen redes activas para llegar a los profesionales de la región.
- América Latina representa un mercado potencial de 600 millones de personas.

En diferentes Foros se ha llegado a una conclusión común: la solución será mucho más fácil, rápida y eficiente si se trabaja articuladamente en la región. El problema y las restricciones son comunes y comparables, pero la fuerza negociadora como un bloque puede representar una gran diferencia. Esta articulación debe darse horizontalmente a través de los seis componentes básicos de la ecuación de interoperabilidad:

1. Talento Humano
2. Política Pública
3. *Software*
4. *Hardware*

5. Estándares
6. Demanda

También puede y debe darse como articulación vertical a través de las dimensiones local, país y región.

Las herramientas están disponibles, las debilidades y fortalezas son conocidas, la necesidad es crítica. El momento es ahora... la interoperabilidad existe en diferentes tonos de gris, ubicados entre: "no ser capaz de intercambiar nada" y "fluidez integral de datos".

Los países de América Latina tienen la oportunidad de hacer *Leapfrogging*.

Que tan alto y lejos, dependerá de nuestra habilidad para sumar esfuerzos.

> *Vivimos en un mundo donde hay más y más información, y menos y menos significado*.
> Jean Baudillard

REFERENCIAS BIBLIOGRÁFICAS

European Union (2015). Legislation. *Official Journal of the European Union - L199*, 44.

Giulio Gilia, V. M. (2015). Improving Interoperability of clinical documents. In *6th IESM Conference*. Seville, Spain.

Gutiérrez, C., Almeida, R. & Romero, W. (2018). Diseño de un modelo de migración a *cloud computing* para entidades públicas de salud. *Investigación e Innovación en Ingenierías, 6*(1), 10-26.

HHS Department (2015). *Federal Health IT Strategic Plan 2015-2020*. Washington: HHS.

HIMSS (2017). *Dictionary of Health Information Technology Terms, Acronyms and Organizations*, Fourth Edition. New York: CRC Press.

Hugo, H. (2016). Factores críticos para promover la calidad en el sector salud del departamento del Atlántico. *Investigación e innovación en Ingenierías, 4*(2), 94-103.

Inforoute Canada (2001). *Opportunities for Action: A Pan-Canadian Digital Health Strategic Plan*. Quebec: Inforoute Canada.

Inforoute Canada (2018). *Year in Review 2017-2018*. Quebec: Inforoute Canada.

Jimeno, M., De la Hoz, Y. & Wilches, J. (2014). Wireless ECG and PCG portable telemedicine kit for rural areas in Colombia. *Investigación e Innovación en Ingenierías, 1*(2), 1-9.

Ministerio Da Saúde (2014). *Estrategia e-Saúde para o Brasil*. Brasilia: Ministerio da Saúde.

Ministerio da Saúde (2016). *Estrategia e-Saúde para o Brasil.* Brasilia: Ministerio da Saúde.

MinSalud (2016). *Política de Atención Integral en Salud*. Bogotá, Colombia: MinSalud.

MinTIC (2010). *Marco de Interoperabilidad para Gobierno en Línea*. Bogotá, Colombia: MinTIC.

MinTIC (2011). *Guía de Uso del Marco de Interoperabilidad*. Bogotá, Colombia: MinTIC.

OECD Health Policy Studies (2016). *Strengthening Health Information Infrastructure for Care Quality Governance*. Washington: OECD.

Oemig, F. & Snelick, R. (2016). *Healthcare Interoperability Standards Compliance Handbook*. Geneva: Springler.

OMS (2016). *Revisión de Estándares de Interoperabilidad para la e-Salud en Latinoamérica y el Caribe*. Washington DC: OPS/OMS.

ONC (2015). *Connecting Health and Care for the nation: A shared nationwide Interoperability Roadmap*. Washington: ONC.

Piteres, R., Cabarcas, M. & Gaspar, H. (2018). El Recurso Humano, factor de competitividad en el sector salud. *Investigación e innovación en Ingenierías, 6*(1), 93-101.

Tim Benson, G. G. (2016). *Principles of Health Interoperability*. London: Springler.

World Health Organization (2012). *WHO Forum on Health Data Standardization and Interoperability*. Geneva: WHO.

World Health Organization (2014). *Joint-Inter-Ministerial Policy Dialogue on eHealth Standardization*. Geneva: WHO.

Cómo citar este capítulo:
Quintero Méndez, V., Sanmartín Mendoza, P., & Heredia Vizcaíno, D. (2018). Interoperabilidad: Un reto que trasciende el intercambio de información y datos. En V. Quintero, D. Heredia Vizcaíno, P. Sanmartín Mendoza, R. Romero, & J. Castillo Vélez, *Datos, Información, Tendencias, tres miradas sobre un contexto cambiante* (pp.11-53). Barranquilla: Ediciones Universidad Simón Bolívar.

CAPÍTULO II

Los Datos y las Redes Sociales: Caso Twitter

Paul SANMARTÍN MENDOZA
Ronald ROMERO ZÚÑIGA
Diana HEREDIA VIZCAÍNO
Vladimir QUINTERO MÉNDEZ

RESUMEN

Las redes sociales se han convertido en un medio de comunicación muy amplio que nos ha permitido obtener información incluso antes que cualquier otro medio tradicional. Por otro lado, las empresas están utilizando cada vez más estos datos que le brindan las redes sociales para conocer futuros clientes y las tendencias de los que ya poseen.

Twitter es una de las tantas redes sociales que existen y son utilizadas para acceder a información en tiempo real, ya que la gran parte de su contenido es accesible de forma pública. Además, las herramientas y API para extraer su contenido son de libre acceso. Actualmente existen diversas aplicaciones que permiten analizar los datos de las diferentes redes, la mayoría de carácter pago, mientras que las gratuitas brindan información muy limitada.

El propósito de este capítulo es poder contribuir al *marketing* que realizan las diferentes empresas a través de un modelo que, aplicado a una red social, permita identificar patrones que relacionen eventos, productos y usuarios a través de filtros por palabras que vislumbren el mercado de estudio en el cual la empresa desea enfocarse.

Palabras clave: red social, Twitter, minería de datos.

INTRODUCCIÓN

Actualmente, las redes sociales se han convertido en una importante fuente de comunicación a nivel mundial. Las interacciones que se producen entre sus usuarios proporcionan una inmensa cantidad de información sobre sus preferencias, lo cual resulta verdaderamente útil para la detección de tendencias. Algunas de las redes sociales más relevantes y populares son Facebook y Twitter, en las que millones de usuarios comparten sus comentarios y opiniones a través de mensajes de texto cortos. Estos mensajes pueden ser extraídos y a continuación tratados mediante diferentes técnicas de minería de datos, para así poder distinguir grupos de tendencias de opiniones sobre un tema (twitter.com, 2017; (Real-time Twitter Trending Hashtags and Topics - Trendsmap, 2017).

Las redes sociales han sido de gran provecho en muchas áreas de la industria incluyendo el mercadeo y ventas, educación (López, Sanmartín & Méndez, 2014), entre otras, razón por la cual son temas de estudio en las Ciencias Sociales y en muchas ramas de la academia y posgrados a nivel mundial.

Hoy en día mucha de la información que circula en las redes sociales es tomada como base para la creación de bancos de datos para luego realizar un análisis dependiendo del campo de interés de quien realiza esta intervención.

Este capítulo muestra cómo se puede manipular la información de las redes sociales, en especial Twitter, para su recolección y almacenamiento, con el propósito de que sea útil para quien la necesite analizar. En primer lugar, se expone una breve introducción de las redes sociales y luego de Twitter, explicando qué es y cómo funciona, acompañada de un glosario de términos comunes que se repetirán a lo largo de todo el documento, luego se analiza la información que podemos tratar desde esta red social.

REDES SOCIALES

Se definen como un conjunto finito de actores (individuos, grupos, organizaciones, comunidades, sociedades, etc.) vinculados unos a otros a través

de una relación o un conjunto de relaciones sociales (Menéndez, 2003). Son el resultado de un proceso evolutivo de formas de organización social, en las que se conectan grupos de individuos para poder coordinarse y actuar en conjunto (Sandoval-Almazan & Saucedo-Leyva, 2010). Algunas de las características intrínsecas de los grupos como el mandar mensajes a todos los miembros, y la factibilidad de segmentación, han desarrollado un ambiente idóneo para la actividad del *marketing* (Vásquez, Escamilla & Vera, 2017). En otras palabras, las redes sociales usan programas de medios sociales basados en Internet para hacer conexiones con amigos, familiares, compañeros de clase y clientes; también pueden usarse con fines sociales, comerciales o ambos a través de sitios como Facebook, *Twitter*, Instagram, *WhatsApp*, LinkedIn, Classmates.com y Yelp (Degenne & Forsé, 1999). Las redes sociales también son un área objetivo importante para los profesionales del *marketing* que buscan atraer a los usuarios. Los profesionales del *marketing* utilizan las redes sociales para aumentar el reconocimiento y la lealtad de marca; hacen que la compañía sea más accesible para los clientes nuevos y más reconocibles para los clientes existentes, y ayudan a promover la voz y el contenido de una marca (Carballar Falcón, 2014). Por ejemplo, un usuario frecuente de Twitter puede escuchar hablar de una empresa por primera vez a través de un servicio de noticias y decide comprar el producto o servicio. Mientras más personas expuestas están a la marca de una compañía, mayores son sus posibilidades de encontrar y retener nuevos clientes.

TWITTER

Es un servicio gratuito de *microblogging* de redes sociales que permite a los miembros registrados transmitir publicaciones cortas llamadas *tweets* (Goergen, Migliosi, Gurbani, State & Engel, 2014). Sus miembros pueden transmitir y seguir *tweets* de otros usuarios mediante el uso de múltiples plataformas y dispositivos. Los *tweets* y sus respuestas se pueden enviar por mensaje de texto del teléfono celular o publicándolos en Twitter.com.

La configuración predeterminada para Twitter es pública. A diferencia de Facebook o LinkedIn, donde los miembros deben aprobar las conexiones sociales, cualquiera puede seguir a cualquiera. Para tejer *tweets* en un hilo de conversación o conectarlos a un tema general, los miembros pueden agregar "hashtags" a una palabra clave en su publicación. Incluso, es considerado una metaetiqueta y se hace referencia a él a través de un numeral y seguido de una palabra clave, como por ejemplo, #keyword (Efron).

Los *tweets* pueden incluir hipervínculos. Estos a veces llegan en tiempo real, y algunos usuarios que inician su actividad en esta red social pueden creer que estos mensajes instantáneos fueron enviados a ellos sin saber que cualquiera puede verlos sin necesidad de ser un seguidor de la persona en la red social (Gerbaudo, 2012). Los mensajes pueden desaparecer en el momento que el usuario cierra la aplicación; pero los *tweets* también se publican en el sitio web de Twitter y son permanentes, se pueden buscar y son públicos, donde cualquiera los puede buscar, ya sean miembros o no.

Twitter utiliza un marco web de código abierto llamado Ruby on Rails (RoR) (Bächle & Kirchberg, 2007). La API está abierta y disponible para los desarrolladores de aplicaciones. Además de su relativa novedad, su gran atractivo es su rapidez y facilidad de exploración: puede rastrear cientos de interesantes *Twitters* y leer su contenido con solo echar un vistazo (Ghanem, Magdy, Musleh, Ghani & Mokbel, 2014).

Twitter emplea una restricción de tamaño de mensaje útil para mantener las cosas fáciles de escanear: cada '*tweet*' está limitado a 280 caracteres o menos. Este límite de tamaño promueve el uso enfocado e inteligente del lenguaje, lo que hace que los *tweets* sean muy fáciles de escanear, pero también es un gran reto escribir bien. Esta restricción de tamaño realmente ha convertido a Twitter en una herramienta social popular.

BIG DATA ANALYTIC

Es el proceso de examinar conjuntos de datos grandes y variados, es decir, *Big Data*, para descubrir patrones ocultos, correlaciones desconocidas,

tendencias del mercado, preferencias del cliente y otra información útil que puede ayudar a las organizaciones a tomar decisiones comerciales más informadas (Russom, 2011).

Impulsado por sistemas analíticos especializados y *software*, el análisis de Big Data puede señalar el camino hacia diversos beneficios comerciales, incluyendo nuevas oportunidades de ingresos, *marketing* más efectivo, mejor servicio al cliente, mejor eficiencia operativa y ventajas competitivas sobre sus rivales (Zikopoulos & Eaton, 2011).

Las aplicaciones de análisis de datos grandes permiten a los científicos de datos, modeladores predictivos, estadísticos y otros profesionales de análisis analizar volúmenes crecientes de datos de transacciones estructuradas, además de otras formas de datos que a menudo quedan sin explotar en los programas de inteligencia de negocios (BI) y análisis convencionales (Chen, Chiang & Storey, 2012). Esto incluye una mezcla de datos semiestructurados y no estructurados, por ejemplo, datos de *clickstream* de Internet, registros de servidor web, contenido de redes sociales, mensajes de correo electrónico de clientes y respuestas de encuestas, registros detallados de llamadas de teléfonos móviles y datos de máquinas capturados por sensores conectados a Internet de las cosas a gran escala; las tecnologías y técnicas de análisis de datos proporcionan un medio para analizar conjuntos de datos y sacar conclusiones sobre ellos para ayudar a las organizaciones a tomar decisiones comerciales informadas. Las consultas de BI responden preguntas básicas sobre las operaciones comerciales y el rendimiento. El análisis de datos grandes es una forma de análisis avanzado, que involucra aplicaciones complejas con elementos tales como modelos predictivos, algoritmos estadísticos y análisis hipotéticos basados en sistemas de análisis de alto rendimiento (Vitt, Misner & Gallardo, 2002).

ESTADO DEL ARTE

Dentro del estado actual encontramos a nivel nacional algunos trabajos en los que se implementan técnicas de extracción y almacenamiento de la información de Twitter. Por ejemplo en la Universidad Católica de Colombia se

realizó el proyecto "Desarrollo y aplicación de una herramienta de extracción y almacenamiento de datos de Twitter a un contexto social de violencia política" (Barriga Mariño, 2017).

Por otro lado, vemos que en el campo internacional en la Universidad Politécnica de Valencia en España la tesis de maestría "Aplicación de técnicas de minería de datos en redes sociales/web" de la autora Elena Asensio Blasco, aplican minería de datos para extraer y analizar la información de Twitter (Asensio Blasco, 2015). También encontramos a Mauricio Iturralde (dir) y Christian Alexis Álvarez Espín, quienes desarrollaron el trabajo de grado "Minería de datos en redes sociales por medio de un correlacionador de datos" en donde buscan ofrecer una herramienta que permita analizar tendencias sobre cualquier tema publicado en Twitter (Iturralde & Álvarez Espín, 2017)

Manuel Artero Anguita realizó el trabajo "Extracción, análisis y visualización de información social desde Twitter" (Artero Anguita & Marcos Lorenzo, 2014) en el cual se aplican diferentes técnicas de extracción, análisis que le permiten clasificar la información y obtener usuarios por géneros, categorización de *tweets* por contenido y el posicionamiento de *tweets* por áreas geográficas.

TWITTER Y LA MINERÍA DE DATOS

Al igual que muchas otras redes sociales, Twitter es una gran fuente de datos. Por ello, gran multitud de estudios enmarcados en el campo de la minería de datos se han centrado en esta plataforma social.

Sin duda, estudios basados en *marketing* se destacan sobre muchos otros trabajos y proyectos desarrollados en torno a Twitter (Jansen, Zhang, Sobel & Chowdury, 2009). Las marcas aprovechan toda la información que los usuarios generan al comunicarse con familiares y amigos, explotándola y descubriendo nuestros intereses, además de conocer qué se está comentando sobre una marca en concreto. También envían *spam* a los usuarios de Twitter, y este estudio recoge que esta red social es muy susceptible a ese tipo de prácticas. Se analizan marcas de distintos sectores: del sector

textil, de la industria automovilística, del sector hotelero, etc. Además, se trata de detectar qué porcentaje de los *tweets* corresponde a opiniones de las marcas. El resultado muestra que un 20 % de las menciones pertenecen a esa categoría.

Recientemente han aparecido varios trabajos que muestran el potencial de Twitter para poder hacer seguimiento y predicción de brotes de Enfermedades. Los usuarios pueden publicar comentarios acerca de enfermedades y sus síntomas, y realizando su geolocalización se podrían utilizar para conocer anticipadamente zonas de posibles Brotes de Enfermedades. Basándose en esta idea, la Universidad de Northwestern ha desarrollado un sistema para detectar brotes de gripe (Lee, Agrawal & Choudhary). El sistema extrae continuamente textos relacionados con enfermedades y utiliza técnicas de minería de datos basados en texto para mostrar gráficas y mapas de la enfermedad.

Por último, también se han llevado a cabo trabajos que estudian el impacto de Twitter en otros medios de comunicación, en concreto, la televisión (Highfield, Harrington & Bruns, 2013). Este estudio se centró en el análisis de las audiencias, poniendo de ejemplo el Festival de Eurovisión, aunque no solo tuvo en cuenta la comunidad europea para desarrollar este trabajo. Una cadena televisiva en Australia retransmite este evento con cierto retraso, por lo que anima a los televidentes a seguir el concurso de forma más dinámica e interactiva a través de hashtags en Twitter. Se muestra cómo las cadenas televisivas utilizan este sistema para conocer sus audiencias.

Algunos trabajos de investigación importantes que estudian el análisis de polaridad a nivel de redes sociales utilizan para ello críticas escritas por los usuarios en Twitter (Wilson, Wiebe & Hoffmann), las analizan centrándose en cuatro dominios diferentes (automóviles, bancos, películas y destinos vacacionales). Presentan un algoritmo de aprendizaje no supervisado para clasificar dichas opiniones como recomendadas (thumbs up) o no recomendadas (*thumbs down*). Estas clasificaciones las predice calculando la media de la Orientación Semántica (OS) de las frases del texto que contienen adjetivos

o adverbios. Esta OS consiste en estimar, haciendo uso de buscadores de páginas web, la Información Mutua Puntual (PMI) entre los términos a analizar y dos palabras que representan inequívocamente una orientación semántica positiva y negativa (en este caso escoge las palabras "*excellent*" y "*poor*"). La PMI mide estadísticamente la posible aparición de una palabra a partir de la inserción de otra.

En los últimos años, debido a la importancia de las redes sociales en el campo de la minería de opiniones, muchos investigadores se han dedicado a su estudio. Existen interesantes sondeos realizados sobre la información recogida de los comentarios escritos en Twitter (Bollen, Mao & Pepe, 2010), han estudiado la posibilidad de predecir los resultados del mercado de valores a partir de los sentimientos expresados en esta red. Un artículo interesante que examina el funcionamiento de los clasificadores en la minería de opinión de *tweets* en español es el de Grigori Sidorov (2013). Explora diferentes configuraciones para ver cómo cada una afecta a la precisión de los algoritmos de aprendizaje automático. Experimenta con los algoritmos de Naive Bayes, Decision Tree y SVM, dado que estos ya han presentado buenos resultados para el idioma inglés. En sus configuraciones tienen en cuenta diferentes tamaños *n*-gram, la longitud del corpus, el número de clases de sentimientos, corpus balanceado vs. corpus no balanceado y diferentes dominios para entrenar y testear (teléfonos móviles y política). En sus conclusiones determinan que la mejor configuración corresponde al uso de unigramas como *features*, un número tan pequeño como se pueda de clases (positivo y negativo), un tamaño de al menos 3.000 *tweets* en el conjunto de entrenamiento (un tamaño superior no incrementa la precisión significativamente), un corpus no balanceado muestra una ligera mejoría en los resultados y el clasificador con más precisión es el SVM. Además, concluyen que el hecho de entrenar el sistema con *tweets* de un dominio diferente al que posteriormente se utilizará, empeora significativamente la precisión de los resultados, llegando a bajar del 85,8 % al 28,0 % en la prueba realizada con SVM.

En España, gracias a la SPLN y al TASS, diversos grupos de investigación españoles han presentado sus algoritmos sobre el análisis de sentimientos en Twitter en idioma castellano, dando así un impulso a esta línea de investigación. Saralegi y San Vicente (2013), consiguieron en este taller los mejores resultados en la tarea de análisis de sentimientos a nivel global de *tweet*. El método de aprendizaje supervisado que presentan usa un clasificador SVM que construyen con la herramienta WEKA. Esta solución incluye un procesamiento basado en conocimiento lingüístico para preparar los *features* que utilizará el clasificador. Dicho procesamiento incorpora *lematización* y etiquetado POS realizado con Freeling, etiquetado de polaridad para el que construye su propio léxico, tratamiento de emoticonos y de negación. Además, para aumentar su precisión realizan un pre-procesamiento del texto de los *tweets* en el que hacen correcciones ortográficas.

DISEÑO DE EXTRACCIÓN DE INFORMACIÓN EN LAS REDES SOCIALES

Como caso de estudio se ha tomado la red social Twitter para el diseño de extracción y análisis de información en una red social; la librería TweetInvi y la base de datos NoSql MongoDb han sido tomadas como referencia. Sin embargo, se debe dejar claro que existen diversas herramientas relacionadas en numerosos artículos de investigación, tesis e Internet en general, entre ellas encontramos a Hootsuite, utilizada para hacer seguimientos a una cuenta, además, permite administrar varias redes sociales para ver los mensajes y enviarlos o publicarlos en el muro de cada red social. Una de las cosas más interesantes es poder programar los mensajes al momento de publicarlos por fecha en el muro de la red social (H. M. Inc., 2017), Trendsmap muestra las últimas tendencias de Twitter, para cualquier parte del mundo. Trendsmap Plus permite explorar tendencias ahora y desde la última semana, desde cualquier parte del mundo. Trendsmap Pro proporciona la capacidad de analizar, visualizar y recibir alertas sobre cualquier tema en Twitter desde cualquier lugar del mundo. Trendsmap Premium tiene toda la funcionalidad de los marcos de tiempo extendidos Pro plus (5 semanas en lugar de 2), más alertas y la integración de Slack (Real-time Twitter Trending Hashtags and

Topics - Trendsmap, 2017; Sprout Social, 2017). Otra herramienta conocida se llama Trendinalia, los hashtags y frases que –según Twitter– fueron Temas del Momento (Trending Topics) en Worldwide (por día) (@trendinalia, 2017). Existe otra herramienta llamada Sproutsocial, muy parecida a Trendsmap, ya que permite administrar varias redes sociales como *Twitter, Facebook, Instagram, Linkedin, Google+* (Sprout Social, 2017).

EXTRACCIÓN Y CATEGORIZACIÓN DE LA INFORMACIÓN DE LA RED SOCIAL TWITTER

Para poder hacer extracción de la información almacenada debemos conocer las herramientas que brinda Twitter para acceder a los datos. En este caso de estudio se utilizaron API Streaming y API REST, que para su funcionamiento necesitan obtener un token que puede ser por medio de autenticación de usuario o por aplicación. Para esto se debe crear una aplicación en Twitter con una cuenta registrada, para poder obtener las credenciales de autenticación necesarias.

API Streaming

Proporciona a los desarrolladores un acceso de baja latencia al flujo global de Twitter de datos de Twitter. Un cliente de transmisión continuará recibiendo mensajes que indican que se han producido *tweets* y otros eventos, sin ninguna de las sobrecargas asociadas al sondeo de un punto final REST (Wang, Callan & Zheng, 2015).

API REST

Esta interfaz proporciona acceso programático para leer y escribir datos de Twitter. Crear un nuevo *tweet*, leer el perfil de usuario y datos de seguimiento, y más. La API REST identifica las aplicaciones de Twitter y los usuarios que utilizan OAuth; las respuestas están en formato JSON (Pol & Holubov, 2015).

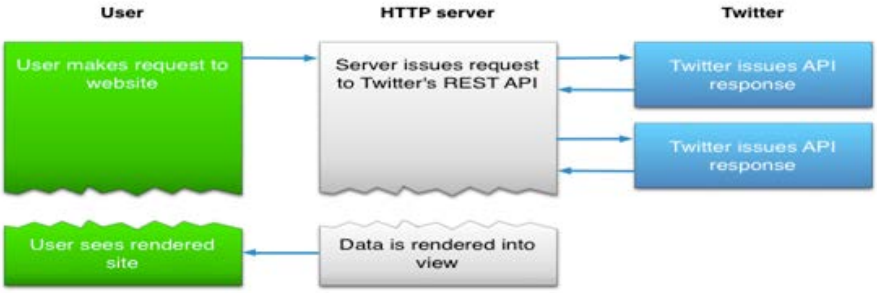

Figura 2.1. Esquema del funcionamiento de la API REST de Twitter
Fuente: Twitter.com, 2017

Diferencias entre API Streaming y API REST: La conexión a la API Streaming requiere mantener abierta una conexión HTTP persistente. En muchos casos esto implica pensar en su aplicación de manera diferente a como si estuviera interactuando con la API REST. Para un ejemplo, considere una aplicación web que acepte solicitudes de usuario, haga una o más solicitudes a la API de Twitter, luego formatee e imprima el resultado al usuario, como respuesta a la solicitud inicial, la aplicación que se conecta a la API Streaming no podrá establecer una conexión en respuesta a una solicitud de usuario. En su lugar, el código para mantener la conexión de transmisión se ejecuta normalmente en un proceso separado del proceso que gestiona las solicitudes HTTP:

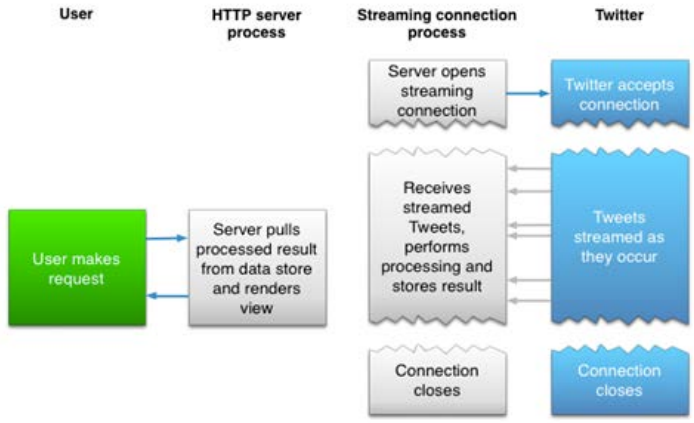

Figura 2.2. Esquema del funcionamiento de la API Streaming de Twitter
Fuente: Twitter.com, 2017

EXTRACCIÓN DE LA INFORMACIÓN

En la siguiente Figura se muestra el proceso de extracción que debe implementarse.

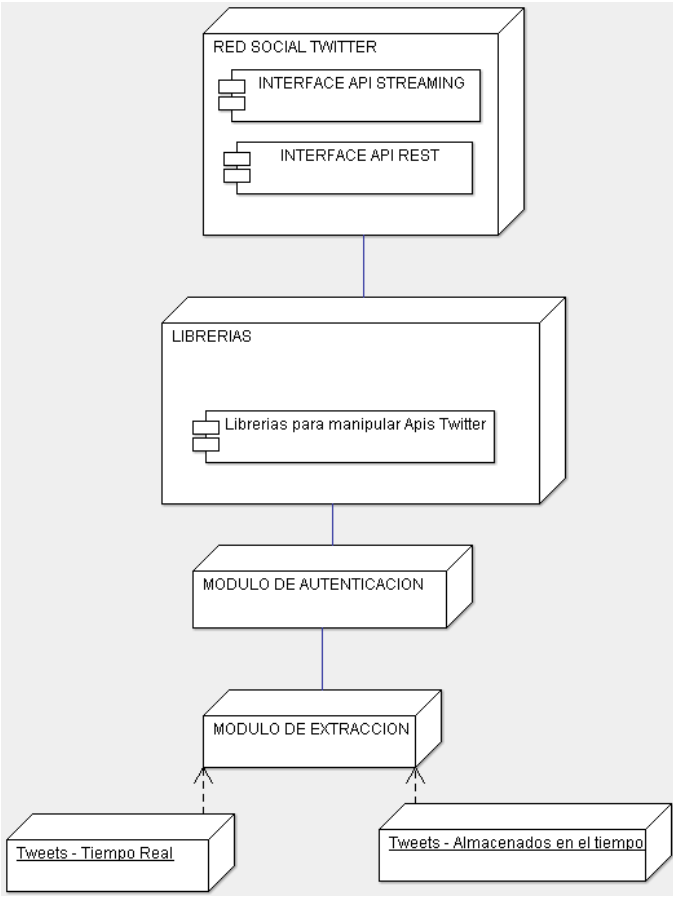

Figura 2.3. Extracción de la información de Twitter

Como podemos observar en la Figura 2.3, en este proceso es posible implementar un módulo de autenticación que nos permite acceder de forma segura a los datos de Twitter. Luego vemos las librerías externas (como TweetInvi para el lenguaje de programación C# y Twython para el lenguaje de programación

Python) para manipular las opciones que nos brindan las API de twitter para acceder a los campos propiamente de esta red social.

Por último, el módulo de extracción que implementa los métodos ofrecidos por el API Streaming y API REST.

En este proceso se pueden aplicar pre-filtros a los datos por medio de los campos idioma y localización ya establecidos en esta red social.

Un ejemplo a implementar con algunas herramientas es ilustrado en la Figura 2.4.

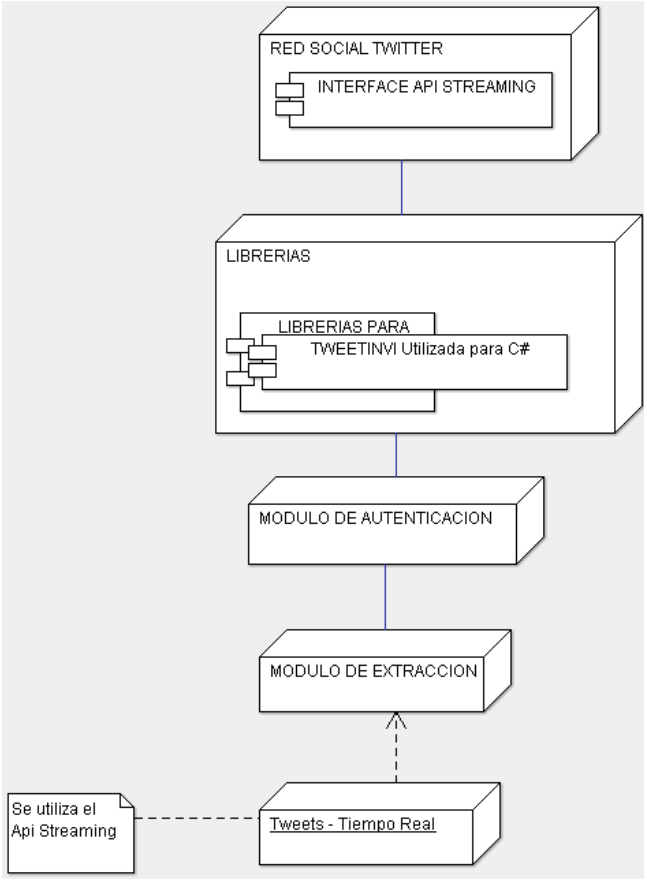

Figura 2.4. Ejemplo para extracción de la información de Twitter

ALMACENAMIENTO

Para este proceso se contemplan el módulo de extracción, citado en el anterior apartado, y la elección del sistema de base de datos. En esta investigación, se recomienda los motores de base de datos NoSql y explícitamente MongoDb por su afinidad en la estructura de los datos que se asemejan al formato JSON, obtenidos por las API de Twitter.

En la Figura 2.5 podemos ver que el modelo también puede ser implementado para motores de base de datos relacional, pero solo a manera de ilustración.

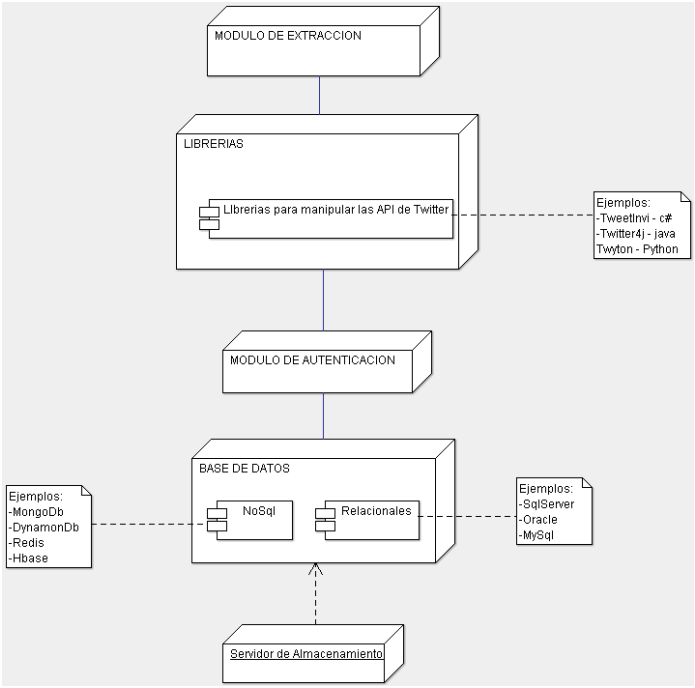

Figura 2.5. Almacenamiento de la información de Twitter

En la Figura 2.6**,** se muestra explícitamente que para el almacenamiento se necesita haber ejecutado el proceso de extracción para luego almacenar la información pre-procesada en la base de datos MongoDb a través de la librería TweetInIv.

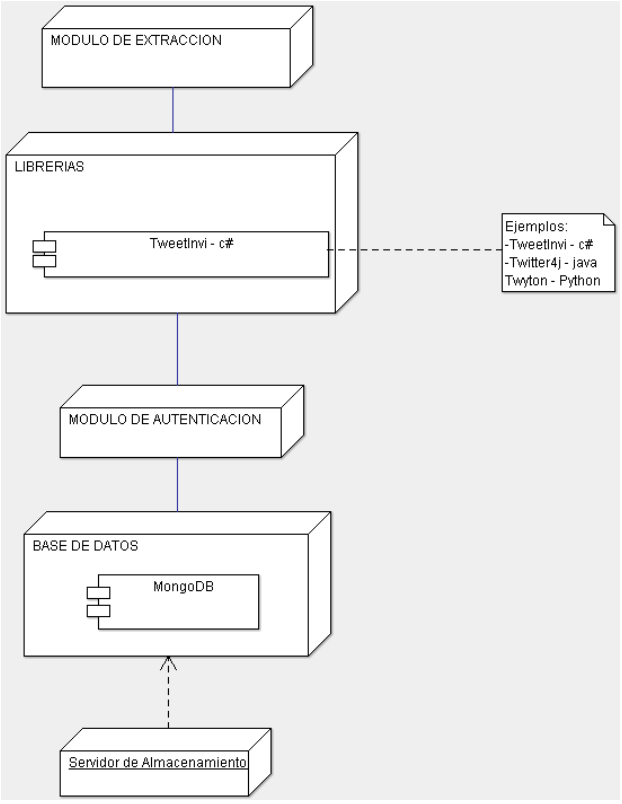

Figura 2.6. Ejemplo para almacenamiento de la información de Twitter

El proceso de transmisión recibe los *tweets* de entrada y realiza cualquier análisis, filtrado y/o agregación necesarios antes de almacenar el resultado en un almacén de datos. El proceso de procesamiento HTTP consulta el almacén de datos para obtener resultados en respuesta a las solicitudes de los usuarios. Si bien este modelo es más complejo que el primer ejemplo, los beneficios de tener un flujo en tiempo real de datos Tweet hacen que la integración valga la pena para muchos tipos de aplicaciones.

Para el almacenamiento de los datos, en esta investigación se propone utilizar el motor de base de datos MongoDB, que es un tipo NoSql y maneja documentos en vez de tablas relacionadas. Es de código abierto y proporciona alto rendimiento, alta disponibilidad y escalado automático (Li & Yang, 2014).

Un registro en MongoDB es un documento, una estructura de datos compuesta por pares de campo y valor. Los documentos MongoDB son similares a los objetos JSON. Los valores de los campos pueden incluir otros documentos, matrices y matrices de documentos.

IDE Visual studio community 2017: Un completo IDE extensible y gratuito para crear aplicaciones modernas para Windows, Android e iOS, además de aplicaciones web y servicios en la nube (Cook, Jones, Kent & Wills, 2007).

Microsoft.Net Core 1.0.1 VS 2015 Toolig Preview 2: .NET Core es una plataforma de desarrollo de uso general cuyo mantenimiento se encargan Microsoft y la comunidad .NET en GitHub. Es multiplataforma, admite Windows, macOS y Linux y puede usarse en escenarios de dispositivo, nube, IoT e incrustados.

Las siguientes características definen mejor a .NET Core:

- Implementación flexible: se puede incluir en la aplicación o se puede instalar de forma paralela a nivel de usuario o de equipo.
- Multiplataforma: se ejecuta en Windows, macOS y Linux; se puede portar a otros sistemas operativos. Los sistemas operativos (SO), CPU y escenarios de aplicaciones admitidos proporcionados por Microsoft, otras empresas e individuos, crecerán con el tiempo.
- Herramientas de línea de comandos: todos los escenarios de productos pueden ejecutarse en la línea de comandos.
- Compatible: .NET Core es compatible con .NET *Framework*, Xamarin y Mono, mediante la biblioteca estándar .NET.
- Código abierto: la plataforma .NET Core es de código abierto, con licencias de MIT y Apache 2. Documentación con licencia de CC-BY. .NET Core es un proyecto de .NET Foundation.
- Compatible con Microsoft: .NET Core incluye compatibilidad con Microsoft, como se indica en .NET Core Support (Compatibilidad de .NET Core) (Al-Omari, Keivanloo, Roy, & Rilling).

FILTROS

Los filtros se pueden aplicar de dos formas:

1. Directamente sobre los campos de un *tweet*:

Para ello es necesario utilizar las interfaces API Streaming o Api REST de Twitter que nos proporcionan el modelo de los datos contenidos en un *tweet*. Las librerías externas como TweetInvi interactúan con las interfaces realizando las peticiones de consultas por las cuales se desea filtrar.

En la Figura 2.7 observamos los filtros aplicados directamente a los *tweets*.

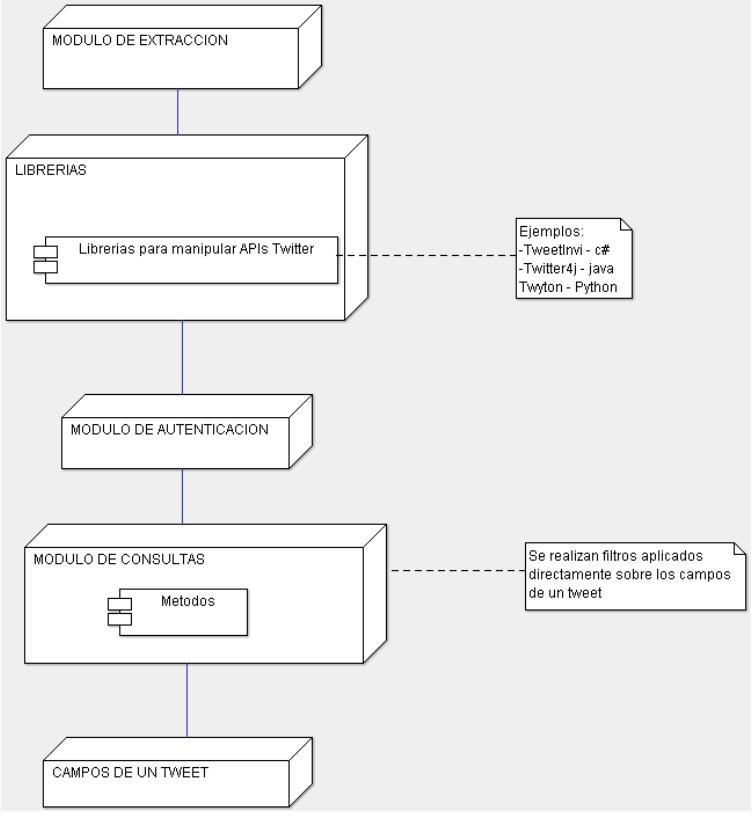

Figura 2.7. Filtros aplicados directamente sobre los *tweets*

2. Información almacenada en el modelo o base de datos:

Para aplicar filtros sobre los datos almacenados en el modelo es decir, en la base de datos:

Se implementan sentencias o comandos propiamente establecidos por el motor de base de datos escogido. La Figura 2.8 visualiza la manera de hacer los filtros a la base de datos implementada en el modelo.

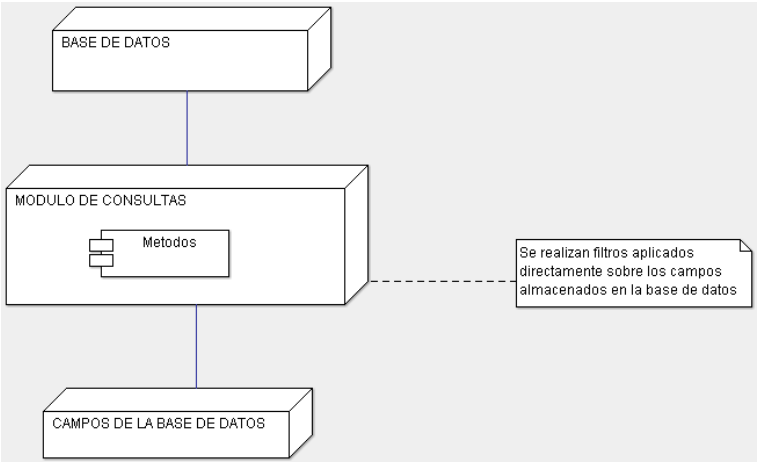

Figura 2.8. Filtros aplicados a los datos almacenados en la base de datos

CONCLUSIONES

Para la construcción de un modelo de patrones basado en las redes sociales que permita la relación que pueda existir entre los eventos, productos y usuarios fue necesario tomar como caso de estudio la red social Twitter y generar un marco teórico de la información que está disponible en ella.

En el apartado teórico se ha estudiado qué trabajo previo existía relacionado con su uso como fuente de información; se ha puesto sobre la mesa a qué información es posible acceder exactamente y se ha emprendido la labor de clasificarla a través de filtros y métodos que apliquen técnicas de minería de datos, análisis de sentimientos y Big Data para el almacenamiento.

Posteriormente se propone el modelo de patrones a través de dos procesos:

1. Extracción y Almacenamiento.
2. Análisis de los datos.

Finalmente, se crea una herramienta que permita validar el modelo propuesto y así establecer los principios básicos para las aplicaciones que necesiten extraer, almacenar y clasificar información desde las redes sociales y de esta manera las empresas puedan beneficiarse con el uso de aplicaciones de este tipo para ayudar al *marketing* digital y fidelizar a más clientes y mantener a los que ya tienen.

REFERENCIAS BIBLIOGRÁFICAS

@trendinalia (2017). *Trendinalia Worldwide. Available: http://www.trendinalia.com/twitter-trending-*topics/globales/globales-171125.html

Al-Omari, F., Keivanloo, I., Roy, C. K. & Rilling, J. *Detecting clones across microsoft. net programming languages*. IEEE.

Artero Anguita, M. & Marcos Lorenzo, R. (2014). *Extracción, análisis y visualización de información social desde Twitter*.

Asensio Blasco, E. (2015). *Aplicación de técnicas de minería de datos en redes sociales/web* (Tesis de máster). Universitat Politècnica de Valèncía.

Barriga Mariño, J. C. (2017). *Desarrollo y aplicación de una herramienta de extracción y almacenamiento de datos de Twitter a un contexto social de violencia política*.

Bollen, J., Mao, H. & Pepe, A. (2010). *Determining the Public Mood State by Analysis of Microblogging Posts*.

Bächle, M. & Kirchberg, P. (2007). Ruby on rails. *IEEE software, 24*(6).

Carballar Falcón, J. A. (2014). Twitter*: marketing personal y profesional*. Madrid, España: RC Libros. Eskup. http://eskup.elpais.com/Estaticas/ayuda/quees.html.

Chen, H., Chiang, R. H. L. & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4).

Cook, S., Jones, G., Kent, S. & Wills, A. C. (2007). *Domain-specific development with Visual Studio DSL tools*. Pearson Education.

Degenne, A. & Forsé, M. (1999). *Introducing social networks*. Sage.

Efron, M. *Hashtag retrieval in a microblogging environment*. ACM.

García Miranda, I. *Plataforma escalable para la captura y gestión de datos en* Twitter *(PARSE4U).*

Gerbaudo, P. (2012). *Tweets and the streets: Social media and contemporary activism*. Pluto Press.

Ghanem, T. M., Magdy, A., Musleh, M., Ghani, S. & Mok*bel, M. F. (2014). VisCAT: spatio-temporal visualization and aggregation of categorical attributes in Twitter data.* Presented at the Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Dallas, Texas.

Goergen, D., Migliosi, A., Gurbani, V. K., State, R. & Engel, T. (2014). *Spatial and temporal analysis of* Twitter*: a tale of two countries.* Presented at the Proceedings of the Conference on Principles, Systems and Applications of IP Telecommunications. Chicago, Illinois.

H. M. Inc. (2017). *Hootsuite - Su plataforma para interactuar, escuchar y compartir en redes sociales*. Available: https://hootsuite.com/es/

Highfield, T., Harrington, S. & Bruns, A. (2013). Twitter as a technology for audiencing and fandom: The# Eurovision phenomenon. *Information, Communication & Society, 16*(3), 315-339.

Iturralde, M. D. & Álvarez Espín, C. A. (2017). *Minería de datos en redes sociales por medio de un correlacionador de datos, (bachelor Thesis).* USFQ, Quito.

Jansen, B. J., Zhang, M., Sobel, K. & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the Association for Information Science and Technology, 60*(11), 2169-2188.

Lee, K., Agrawal, A. & Choudhary, A. *Real-time disease surveillance using* twitter *data: demonstration on flu and cancer.* ACM.

Li, C. & Yang, W. (2014). The distributed storage strategy research of remote sensing image based on Mongo DB. In *2014 Third International Workshop on Earth Observation and Remote Sensing Applications (EORSA)*, 101-104.

López, R., Sanmartín, P. & Méndez, F. (2014). Revisión de las evaluaciones adaptativas computarizadas (CAT). *Educación y Humanismo, 16*(26), 27-40. https://doi.org/10.17081/eduhum.16.26.2345

Menéndez, L. S. (2003). Análisis de redes sociales: o cómo representar las estructuras sociales subyacentes. *Documento de Trabajo, 3*, 07.

Pol, M. & Holubov, I. (2015). *Advanced REST API Management and Evolution Using MDA*. Presented at the Proceedings of the 3rd International Workshop on (Document) Changes: modeling, detection, storage and visualization. Lausanne, Switzerland.

Real-time Twitter *Trending Hashtags and Topics – Trendsmap (2017).* Available: https://www.trendsmap.com/

Russom, P. (2011). Big data analytics. *TDWI best practices report, fourth quarter, 19*, 40.

Sandoval-Almazan, R. & Saucedo-Leyva, N. K. (2010). Grupos de interés en las redes sociales: el caso de Hi5 y Facebook en México. *Tecnociencia Chihuahua, 4*(3), 132-141.

Sidorov, G. (2013). Syntactic dependency based n-grams in rule based automatic English as second language grammar correction. *International Journal of Computational Linguistics and Applications, 4*(2), 169-188.

Sprout Social (2017). *Software de gestión de redes sociales | Sprout Social.* Available: https://es.sproutsocial.com/

Twitter.com (2017). Publish and analyze Tweets, optimize ads, and create unique customer experiences. Available: https://developer.twitter.com/en.h*tml*

Twitter.com (2017). *Tweet data dictionaries.* Available: https://developer.twitter.com/en/docs/*tweets*/data-dictionary/overview/*tweet*-object

Vásquez, G. A. N., Escamilla, E. M. & Vera, J. V. A. (2017). Estudio exploratorio sobre el uso de las redes sociales en las mipymes de la Zona Metropolitana de Guadalajara, México. *Revista de Investigación en Ciencias y Administración, 8*(15), 379-403.

Vitt, E., Misner, M. & Gallardo, S. R. (2002). *Business intelligence: técnicas de análisis para la toma de decisiones estratégicas.* McGraw-Hill.

Wang, Y., Callan, J. & Zheng, B. (2015). Should We Use the Sample? Analyzing Datasets Sampled from Twitter's Stream API. *ACM Trans. Web, 9*(3), 1-23.

Wilson, T., Wiebe, J. & Hoffmann, P. *Recognizing contextual polarity in phrase-level sentiment analysis.* Association for Computational Linguistics.

Zikopoulos, P. & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data.* McGraw-Hill Osborne Media.

Cómo citar este capítulo:
Sanmartín Mendoza, P., Romero Zúñiga, R., Heredia Vizcaíno, D. & Quintero Méndez, V. (2018). Los datos y las redes sociales: Caso Twitter. En V. Quintero, D. Heredia Vizcaíno, P. Sanmartín Mendoza, R. Romero, & J. Castillo Vélez, *Datos, Información, Tendencias, tres miradas sobre un contexto cambiante* (pp.55-75). Barranquilla: Ediciones Universidad Simón Bolívar.

CAPÍTULO III

Aplicación de técnicas de minería de datos sobre datos georreferenciados para obtener un modelo predictivo: Caso de Estudio Hurtos en la Ciudad de Barranquilla

Diana HEREDIA VIZCAÍNO
Jhonny CASTILLO VÉLEZ
Paul SANMARTÍN MENDOZA
Vladimir QUINTERO MÉNDEZ

RESUMEN

El texto contiene una descripción de los resultados de la investigación realizada como trabajo de grado en la Maestría de Ingeniería de Sistemas y Computación en la Universidad Simón Bolívar. Su objetivo era aplicar las técnicas de minería de datos de *Clustering* y Árboles de decisión a datos georreferenciados de hurtos en la ciudad de Barranquilla, ocurridos entre los años de 2016 y 2017, con el fin de obtener un patrón descriptivo y predictivo. Se obtuvieron dichos modelos a partir de datos abiertos (de uso libre para fines investigativos), con eficacia de clasificación de 51 % y 59,7 %, respectivamente, los cuales pueden considerarse no muy buenos como predictores, pero arrojan resultados interesantes en cuanto a manipulación de datos geográficos, atributos no numéricos con gran cantidad de valores.

Palabras clave: minería de datos, datos georreferenciados, *clustering*, árboles de decisión.

INTRODUCCIÓN

La minería de datos es un conjunto de técnicas de análisis de datos, usada desde hace varios años para extraer información y conocimiento valioso de gran cantidad de datos provenientes de muy diversas fuentes, tales como sistemas de información y aplicaciones web y móviles, documentos físicos, encuestas, entre otros. La proliferación actual de dispositivos de comunicación móviles con disponibilidad de sistemas de posición geográfica (GPS) al alcance de las personas ha permitido que la información se enriquezca con este tipo de datos.

Surge entonces la pregunta: ¿Cómo aprovechar los datos de posición geográfica para lograr predecir la ocurrencia de cierto tipo de hechos en lugares específicos y, con base en ello, tomar decisiones? A partir de esta pregunta, se decidió realizar una investigación acerca de la aplicación de técnicas de minería de datos a datos georreferenciados (coordenadas geográficas) para obtener un modelo predictivo de hechos delictivos, para lo cual se tomó como caso de estudio la ocurrencia de hurtos en la ciudad de Barranquilla.

En la primera sección se presenta un breve marco teórico sobre minería de datos y las técnicas utilizadas en el estudio. En la siguiente sección, el estado del arte, se describen algunos estudios previos. A continuación, la descripción del trabajo realizado y los resultados obtenidos. Finalmente, se presentan las conclusiones.

MINERÍA DE DATOS

Puede definirse inicialmente como un proceso en el que se utilizan herramientas y técnicas de descubrimiento de nuevas y significativas relaciones, patrones y tendencias de grandes cantidades de datos almacenados en los repositorios, dando como resultado la extracción de conocimiento" (Han, Pei & Kamber, 2011); (Larose, 2014).

Según su tarea de descubrimiento, las principales técnicas de minería de datos se clasifican en: Agrupamiento o *Clustering*, Clasificación y Asociación.

Agrupación o clustering

Agrupar un conjunto de datos basándose en la similitud que tengan en los valores de sus atributos se le denomina *clustering* o agrupación. Este tipo de técnicas construye grupos, que son la identificación de las regiones más densamente pobladas, de acuerdo a alguna media distancia previamente establecida (Chen, Han & Yu, 1996).

La técnica de *clustering* ha sido utilizada en varios campos de la ciencia, como la medicina en "The detection of disease *clustering* and a generalized regression approach - Cancer research" (Mantel, 1967), "Medical image segmentation using k-means *clustering* and improved watershed algorithm" (Ng, Ong, Foong, Goh & Nowinski, 2006) y "Broad patterns of gene expression revealed by *clustering* analysis of tumor and normal colon tissues probed by oligonucleotide arrays" (Alon, Barkai, Notterman, Gish, Ybarra, Mack & Levine, 1999), en el cual se utiliza el *clustering* para la detección de enfermedades; en la estadística, en el artículo "*Clustering* of fast-food restaurants around schools: a novel application of spatial statisTIC to the study of food enviroment" (Austin, Melly, Patel, Buka & Gortmaker, 2005), además de la minería de datos en "Uncertain data mining: An example in *clustering* location data" (Chau, Cheng, Kao & Ng, 2006) y "Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification" (Ester, Kriegel & Xu, 1995).

Dos de los algoritmos más importantes y más utilizados para el *clustering* son: Self Organizing Maps (SOM) y K-means. Creado principalmente por Teuvo Kohonen en 1982, SOM es un modelo de redes neuronales artificiales, y uno de los más utilizados para el aprendizaje no supervisado, que de igual manera ha sido empleado en investigaciones de otras ramas de la ciencia como en el artículo "Network Anomaly Detection with Bayesian Self-Organizing

Maps" (De la Hoz Franco, Ortiz García, Ortega Lopera, De la Hoz Correa & Prieto Espinosa, 2013).

K-means fue ideado originalmente por Steinhaus en 1957, y su algoritmo fue propuesto en el mismo año por Lloyd (aunque se publicó en el año 1982), pero fue en el año 1967, cuando James MacQueen utiliza por primera vez el término k-means (Jain, 2010). Este algoritmo de agrupamiento es descrito en detalle en 1975 por (Hartigan, 1979). A pesar de todo el tiempo que ha transcurrido, k-means sigue siendo uno de los algoritmos más utilizados, debido a su facilidad de implementación, simplicidad y eficiencia. El algoritmo de k-means tiene como objetivo dividir los datos en N números de subconjuntos no vacíos en M clúster, luego calcula un punto medio asignado a cada grupo cuyo punto medio es más cercano (Hartigan, 1979).

Clasificación

La clasificación encuentra los atributos comunes entre un conjunto de objetos y los divide en diferentes categorías (subpoblaciones), de acuerdo a un modelo. Para su realización, se utiliza un conjunto de entrenamiento. El objetivo de la clasificación es analizar los datos de entrenamiento y desarrollar una descripción para cada clase empleando las características disponibles en los datos.

La clasificación es una instancia de aprendizaje supervisado, es decir, que para el conjunto de entrenamiento, se conoce la clase de pertenencia y se le indica al modelo si la división que realiza es correcta o no (Chen, Han & Yu, 1996); (Swian & Sarangi, 2013). Un algoritmo que implementa la clasificación, se le denomina clasificador, que se refiere a la función matemática, implementada por un algoritmo de clasificación que mapea datos de entrada a una categoría (Swian & Sarangi, 2013).

El algoritmo que más se destaca es el árbol de decisión ID3, es de aprendizaje, y fue desarrollado por Ross Quinlan (Peng, Chen & Zhou, 2009); es famoso por su alta velocidad, clasificación fácil, gran capacidad de aprendizaje y fácil

construcción. El ID3 construye un árbol de decisión realizando una búsqueda minuciosa de arriba hacia abajo a través de los conjuntos para probar todos los atributos de cada nodo, y debido a esta búsqueda minuciosa, existe la posibilidad de que pueda inclinarse a escoger atributos que tienen muchos valores, lo que afecta su practicidad (Peng, Chen & Zhou, 2009); (Yuxun & Niuniu, 2010). A pesar de este pequeño inconveniente, es utilizado en muchos campos como la medicina y la estadística.

PRE-PROCESAMIENTO DE DATOS

Todas las bases de datos de hoy en día están expuestas a datos ruidosos, perdidos e incoherentes, debido a su gran tamaño, que normalmente son gigabytes o más, y su origen probable se derive de múltiples fuentes heterogéneas.

El pre-procesamiento se encarga de hacer menos ruidosos, perdidos e incoherentes los datos en la base de datos (Han, Pei & Kamber, 2011).

El preprocesamiento cuenta con los siguientes pasos:

- Resumen de datos (*Data Summarization*): Para que el procesamiento de datos tenga éxito, es primordial una imagen general de sus datos. Las técnicas de resumen de datos descriptivos pueden ser utilizadas para identificar los atributos de los datos y resaltar qué valores de datos deben ser tratados como ruidos, perdidos o incoherentes (Han, Pei & Kamber, 2011). Para realizar el resumen de datos se tienen tres pasos: primero se transfiere el archivo a una base de datos relacional. A continuación, aplica la generalización a los datos y se calculan los datos agregados para visitas y sesiones de usuario (Tanasa & Trousse, 2004).
- Limpieza de datos (*data cleaning*): Este paso consiste en eliminar repeticiones inútiles de registros. Generalmente este proceso elimina solicitudes o recursos no analizados como imágenes y archivos multi-

media. La limpieza de datos también identifica robots web y elimina sus peticiones (Han, Pei & Kamber, 2011); (Tanasa & Trousse, 2004). Se realiza un análisis profundo de los datos, para luego poder detectar qué tipos de errores e inconsistencias deben eliminarse (Rahm & Do, 2000).

- Integración y transformación de datos (*Data Integration and Transformation*): Cuando se realiza la tarea de análisis de datos se debe incluir la integración de datos, donde se combinan con datos de múltiples fuentes en un almacén de datos, ya sea una base de datos, *data warehouse*, o cualquier lugar donde sea posible almacenar datos. Estas fuentes pueden incluir ya sea una o múltiples bases de datos, cubos de datos o archivos planos (Han, Pei & Kamber, 2011). El proceso de transformación básicamente consiste en varios pasos, en los que cada uno realiza una serie de transformaciones relacionadas con el esquema y con las asignaciones (Rahm & Do, 2000).
- Reducción de datos (*Data Reduction*): Se aplicarán técnicas de reducción de datos para conseguir un pequeño volumen que será la representación de un conjunto de datos, pero manteniendo la integridad de los datos originales (Han, Pei & Kamber, 2011). Las estrategias para la reducción de datos incluyen lo siguiente: agregación de cubos de datos, selección de subconjuntos de atributos, reducción dimensional, reducción de la numerosidad y discretización y generación de jerarquías de conceptos.

ESTADO DEL ARTE

La minería de datos espaciales no es nueva, se ha venido utilizando en diversos tipos de estudios, en muchas áreas del conocimiento humano y la vida cotidiana. A continuación se mostrarán algunos estudios similares al propuesto.

Denisse Cangrejo y Juan Agudelo en el artículo "Minería de datos espaciales", muestran la importancia de la minería de datos espaciales, su vinculación con la cartografía analizando y gestionando datos espaciales, representando dicha información en mapas y símbolos donde se pueden utilizar los métodos de modelamiento predictivo, resumen de datos espaciales y *clustering*. Por medio de los cuales se puede realizar una predicción de algún evento en un espacio geográfico específico e identificar las relaciones que tiene un grupo de datos georreferenciados (Cangrejo Aljure & Agudelo, 2011).

En 2004 Deren Lluna y Shuliang Wang, realizaron el artículo "Concepts Principles and applications of spatial data mining and knowledge discovery" (Wang & Li, 2004), este documento presenta los principios de SDMKD (minería de datos espaciales y descubrimiento del conocimiento), el cual propone tres nuevas técnicas y su aplicabilidad. SDMKD es un proceso de descubrir una forma de reglas, en el que se visualizan ángulos con diferentes umbrales, perforación, corte en datos, base de datos, generalizar, caracterizar, clasificar entidades, resumir características de datos, que describe las reglas y predice las tendencias futuras. Después de exponer en este documento el concepto y el principio de la minería de datos espaciales y descubrimiento de conocimiento (SDMKD), se proponen tres técnicas de extracción, clasificación de imágenes basadas en SDMKD, modelo de nube y el campo de datos junto con ejemplos de aplicación, es decir, la clasificación de imágenes de teledetección, deslizamientos.

En 2007 (Perversi & Fernández, 2007) presentan el documento "Aplicación de minería de datos para la exploración y detección de patrones delictivos en Argentina", que explica la importancia del análisis de registros criminales en la prevención del delito con datos georreferenciados. En esta investigación se utiliza el *dataset* de la base de datos del SAT (Sistema de Alerta Temprana) del Ministerio de Justicia de Argentina, al cual se le aplica el algoritmo de K-means para agrupar los hechos según su similitud, interpretar y convalidar los resultados obtenidos con los usuarios, utilizando la herramienta Weka, y

luego el algoritmo de inducción C4.5 para identificar reglas de pertenencia de los clústers, y por último la interpretación e implementación de resultados (Perversi & Fernández, 2007).

En 2009 María Ximena Dueñas Reyes realiza el artículo "Minería de datos espaciales en búsqueda de la verdadera información", en el que pretende abordar la importancia de la inteligencia de negocios, la cual surge como un nuevo concepto orientado al tratamiento inmediato de los datos, la información y el conocimiento, con el fin de mejorar los procesos de cualquier organización frente a las exigencias del mercado. Debido a la creciente información que actualmente manejan las empresas, fue necesario crear una nueva herramienta que ofrezca un manejo ágil, eficaz y eficiente, llamado Spatial Data Warehouse, que combina las bases de datos espaciales y las tecnologías del Data Warehouse, que maneja un esquema multidimensional que permite elementos espaciales y no espaciales (Dueñas-Reyes, 2009).

En 2011 el autor D. Rajesh realizó el artículo "Application of Spatial Data Mining for Agriculture". En este trabajo se trata de conseguir la temperatura y la precipitación como datos especiales iniciales y así mejorar el rendimiento de los cultivos y la reducción de pérdidas en ellos; basados en el algoritmo de K-means (Rajesh, 2011).

En 2014 Juan Manuel Suárez realizó la investigación sobre "Técnicas de agrupamiento de minería de datos espaciales para la caracterización y análisis de los hurtos que afectan Bogotá"; donde analizó el comportamiento de este delito en la ciudad capital. Se utilizaron las bases de datos del Instituto Nacional de Medicina Legal, el Centro de Investigaciones Criminológicas de la Policía Metropolitana de Bogotá y la Infraestructura de Datos Espaciales del Distrito Capital (IDECA). Propone implementar técnicas de minería de datos espaciales que permitan analizar, clasificar y predecir tendencias de delito (Suárez, 2016).

En 2013 el autor José Javier Ruiz presenta el estudio "Análisis espacial y supervisado de datos criminalísticos utilizando un sistema de información

geográfica". Este ofrece un análisis espacial y reconocimiento de patrones para determinar el comportamiento de la criminalidad en un caso de estudio particular, enfocándose en un área específica, implementando una aplicación de escritorio, con la finalidad de proyectar actividades criminales a través de los datos históricos; de esta manera se pretende realizar un estudio comparativo de zonas con base en sus incidencias (Ruiz Gómez, 2016).

En 2007 los autores Diego Orlando Abril Frade y José Nelson Pérez Castillo realizaron el artículo "Estado actual de las tecnologías de bodega de datos y OLAP aplicadas a bases de datos espaciales"; plantean que las organizaciones necesitan de una información oportuna, dinámica, centralizada y que sea de fácil acceso, para poder analizar y tomar decisiones acertadas y correctas en el momento preciso. La centralización se logra con la tecnología de almacenes de datos. El análisis lo pueden proporcionar los diferentes sistemas de procesamiento analítico en línea, OLAP (*On Line Analytical Processing*). Después serían muy útiles los Sistemas de Información Geográfica, SIG, que son diseñados especialmente para ubicar la información y representarla por medio de mapas. Las bodegas de datos generalmente se implementan con el modelo multidimensional para facilitar los análisis con OLAP (Abril Frade & Pérez Castillo, 2007).

Descripción del Trabajo Realizado

El primer paso dentro del trabajo es obtener un conjunto de datos de libre acceso, que contuviera datos de georreferenciación, es decir, la posición geográfica de un punto en términos de longitud y latitud. Para este estudio, se buscaron datos de hurtos en la ciudad de Barranquilla. En el sitio web de Datos Abiertos de Colombia (www.datos.gov.co), se obtuvo un dataset con estas características; inicialmente tenía más de 23.000 registros, correspondientes a los años 2016 y 2017.

Muchos de los registros contenían datos en blanco, incompletos o errados; estos fueron eliminados para reducir la probabilidad de sesgos o errores

durante la ejecución de los procesos de minería. Igualmente, durante esta etapa de pre-procesamiento, se redujo la cantidad de los valores en varios de los atributos alfanuméricos. Por ejemplo, en el caso del atributo Caracterización, el cual presentaba originalmente 12 valores diferentes (correspondientes a las distintas clases de hurtos).

También se aplicó discretización a aquellos campos que lo requirieron, probando la cantidad de bins, hasta optar por usar únicamente dos en todos los casos. Esta discretización, en el caso de atributos numéricos, se hace con respecto a su media aritmética. Cabe anotar que durante la realización de las distintas pruebas con las técnicas de minería elegidas, fue necesario aplicar diversas formas de reducir los valores.

Al finalizar el pre-procesamiento de datos, se obtuvieron ocho atributos definitivos, asentado en un archivo de tipo Excel cvs, formato de amplio uso en herramientas de análisis de datos:

- Caracterización: Tipos de hurto (por ejemplo: Hurto a personas, a establecimientos, automóviles, residencias, comercio, entidades financieras, entre otros). Es un dato alfanumérico.
- Modalidad: Las circunstancias en que se produjo el hurto (por ejemplo: atraco, halado, fleteo, entre otros). Es un dato de tipo alfanumérico.
- Longitud: Coordenada geográfica; los valores negativos corresponden al hemisferio occidental. Se asume y manipula como un dato numérico real.
- Latitud: Coordenada geográfica; los valores positivos corresponden al hemisferio norte. Se asume y manipula como un dato numérico real.
- Año: De ocurrencia del hecho delictivo.
- Mes: Cuando ocurrió el delito.
- DíaMes: Valor numérico, entre 1 y 31, que indica qué día del mes ocurrió el hurto.
- DíaSemana: Valor alfanumérico, indica el día de la semana en el que se presentó el delito (lunes a domingo).

PRUEBAS REALIZADAS CON LAS TÉCNICAS DE MINERÍA DE DATOS

Una vez se logró el dataset antes descrito, se procedió a realizar pruebas de las técnicas de minería de datos. Para ello se utilizó la herramienta de *software* Weka[1], versión 3.8, la cual implementa muchos de los algoritmos ampliamente reconocidos y utilizados.

Se eligieron las técnicas de *clustering* k-means y árboles de decisión J48 (versión de los árboles ID3), dada su facilidad de uso y obtención de resultados fáciles de interpretar. El atributo Caracterización es el clasificador (Class), dado que es la variable dependiente y la que se pretende evaluar.

Las pruebas se realizaron aplicando en cada ocasión las dos técnicas al mismo dataset, usando las mismas métricas de desempeño:

- Porcentaje de instancias o registros correctamente clasificados, que se interpreta como Eficacia de la técnica para clasificación.
- Matriz de confusión, que indica la cantidad de instancias correctamente clasificadas y la cantidad de instancias "confundidas" o mal clasificadas, para cada valor del atributo clasificador.

También se utilizaron los mismos parámetros de ejecución:

- Cantidad de folds (Ejecuciones del algoritmo consecutivas para entrenamiento del modelo, cada vez con un conjunto de registros escogidos aleatoriamente dentro del total del dataset): 5.
- Conjunto de entrenamiento: 66 % del total de registros del dataset.

Para el caso de *clustering*, se ejecutó la técnica con la opción de evaluación de clasificación. En el caso de árboles de decisión, la evaluación se ejecuta por defecto.

1 Weka 3: Data Mining Software in Java. www.cs.waikato.ac.nz

Inicialmente se realizaron algunas pruebas para determinar los parámetros de discretización, la correlación entre atributos y la reducción de valores en atributos no numéricos. Como resultado de ello, se decidió lo siguiente:

- Discretizar en dos bins, es decir, dos grupos con respecto a la media aritmética, uno con los valores menores a ella y el otro, con los valores mayores.
- Para el atributo clasificador, Caracterización, se dejaron solo dos valores: Hurto a personas (por ser el de mayor ocurrencia) y agrupar como Otros a los restantes tipos.
- Para el atributo Modalidad, se realizó una reducción de valores similar al anterior, teniendo al final los valores Atraco, halado, factor de oportunidad y otros.
- Ignorar los atributos de Año y DíaMes, porque están correlacionados y su aporte al modelo no es significativo.

Los resultados de eficacia obtenidos en las pruebas más significativas se muestran en la Tabla a continuación.

PRUEBA	CLUSTERING	ÁRBOL J48
Con los atributos: Caracterización (con todos sus valores originales), Modalidad, Mes, DíaSemana, Longitud y Latitud.	Eficacia: 39,6 %	Eficacia: 57,3 %
Con los atributos: Caracterización (con los valores Hurto a personas y Otros), Modalidad (con los valores Atraco, Halado, Factor de oportunidad y Otros), DíaSemana, Longitud y Latitud.	Eficacia: 53 %	Eficacia: 59,7 %
Con los atributos: Caracterización, Modalidad (Con los mismos valores de la prueba anterior), Longitud y Latitud, DíaSemana.	Eficacia: 51 %	Eficacia: 59,7 %

Tabla 1. Eficacia obtenida en las pruebas de las técnicas de minería de datos

En las primeras pruebas, en los modelos obtenidos los datos espaciales no fueron especialmente significativos; en la última prueba sí se logró que la longitud y latitud fueran parte importante en ellos.

Los modelos obtenidos se resumen en las imágenes a continuación.

```
=== Confusion Matrix ===

   a    b   <-- classified as
6201 3313 |    a = HURTO A PERSONAS
5299 6535 |    b = OTROS
```

Figura 3.1. Matriz de confusión para el árbol J48

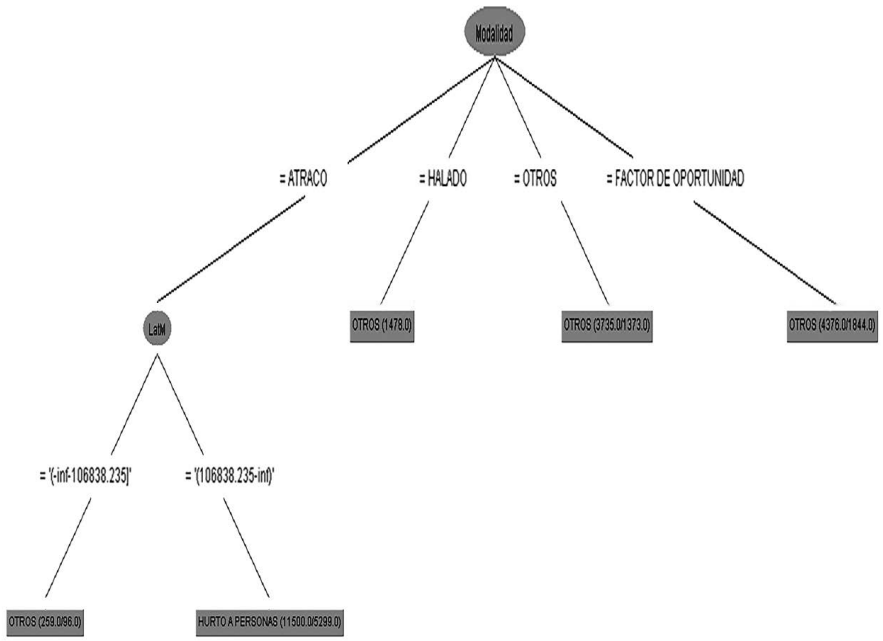

Figura 3.2. Árbol de clasificación J48

En la Figura 3.1 se puede apreciar que una gran cantidad de registros fueron clasificados erróneamente por el árbol de decisión como Hurto a personas, siendo que realmente eran de otro tipo. Los registros originalmente etique-

tados como Otros y que fueron mal clasificados por el modelo representan una menor proporción.

La Figura 3.2 muestra gráficamente el árbol J48 obtenido. En él se puede apreciar que la mayor cantidad de Hurto a personas en Modalidad de Atraco que tienen más probabilidad de ocurrir en la zona norte de la ciudad.

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 42926.0

Initial starting points (random):

Cluster 0: ATRACO,'\'(-74.987087-inf)\'','\'(10.683823-inf)\'','marzo     ','sábado   '
Cluster 1: ATRACO,'\'(-74.987087-inf)\'','\'(10.683823-inf)\'','marzo     ','martes   '

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                Cluster#
Attribute                          Full Data           0                1
                                   (21348.0)     (17145.0)         (4203.0)
=============================================================================
Modalidad                             ATRACO        ATRACO           ATRACO
Longitud                  '(-74.987087-inf)' '(-74.987087-inf)' '(-74.987087-inf)'
Latitud                    '(10.683823-inf)'  '(10.683823-inf)'  '(10.683823-inf)'
Mes                             diciembre          diciembre          agosto
DiaSemana                         sábado             sábado           martes
```

Figura 3.3. Clústeres obtenidos

La técnica de *clustering* se aplicó construyendo dos grupos, utilizando distancia euclidiana para determinar la pertenencia de cada registro a algún grupo o clúster. La Figura 3.3 muestra el centroide de cada uno de ellos. Se puede apreciar que la modalidad Atraco es dominante, en un solo cuadrante de la ciudad (el cual puede interpretarse como noroccidental, con respecto al punto medio); la diferencia entre ambos clústeres reside en el mes y día de la semana.

CONCLUSIONES

La aplicación de técnicas de minería de datos de *clustering* y clasificación mediante árboles de decisión a datos georreferenciados, para el caso de

estudio en particular de Hurtos en la ciudad de Barranquilla, arrojó resultados medianamente satisfactorios, pues el mejor modelo obtenido, un árbol de decisión J48, solo clasifica correctamente el 59 % de las instancias. Se puede inferir de ello que si se utiliza el modelo para predecir la ocurrencia de un hurto en una zona, esta tiene menos del 60 % de confiabilidad.

Las razones de este resultado son múltiples; entre ellas:

- Si se manipulan los datos de posición geográfica (longitud y latitud) como valores numéricos se simplifica su interpretación, pero también queda reducido al manejo de zonas, no a sitios específicos, debido a la discretización necesaria para este tipo de datos.
- Los atributos no numéricos, como Caracterización, Modalidad, DíaSemana tienen originalmente muchos valores distintos; las técnicas usadas no son eficientes manejando estos datos, lo cual hace necesario su reducción, usando solamente los valores de mayor ocurrencia y agrupando los restantes en un solo valor. Esta forma de reducción de valores puede llevar implícita un sesgo, al ser de forma perceptiva.

A pesar de no haber obtenido modelos altamente eficaces en predicción, sí se observa que es posible aplicar técnicas distintas al *clustering*, recurrentemente usado en varios estudios previos, para construirlos. Se deben identificar otras formas de tratamiento a los datos espaciales, que permitan el manejo de zonas más específicas. Así mismo, el resultado de *clustering* se puede usar como buen descriptor de los datos usados.

REFERENCIAS BIBLIOGRÁFICAS

Abril Frade, D. & Pérez Castillo, J. N. (2007). Estado actual de las tecnologías de bodega de datos y OLAP aplicadas a base de datos espaciales. *Ingeniería e Investigación, 27*(1), 58-67.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D. & Levine, A. (1999). Broad patterns of gene expression revealed by *clustering* analysis

of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*.

Austin, S. B., Melly, S. J., Patel, S. B., Buka, S. & Gortmaker, S. (2005). *Clustering* of fast-food restaurants around schools: a novel application of spatial statistics to the study of food environments. *American Journal of Public Health, 95*(9), 1575-1581.

Cangrejo Aljure, D. & Agudelo, J. G. (2011). Minería de datos espaciales. *Revista Avances en Sistemas e Informática*, *8*(3), 71-77.

Chau, M., Cheng, R., Kao, B. & Ng, J. (2006). Uncertain data mining: An example in *clustering* loc*ation data. In Paci*fic-Asia Conference on Knowledge Discovery and Data Mining. Berlin, Germany.

Chen, M.-S., Han, J. & Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering*, *8(6), 866-883*.

De la Hoz Franco E., Ortiz García A., Ortega Lopera J., De la Hoz Correa E., & Prieto Espinosa A. (2013) Network Anomaly Detection with Bayesian Self-Organizing Maps. In: Rojas I., Joya G., Gabestany J. (eds) *Advances in Computational Intelligence. IWANN 2013. Lecture Notes in Computer Science*, vol 7902. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-38679-4_53

Dueñas-Reyes, M. X. (2009). Minería de datos espaciales en búsqueda de la verdadera información. *Ingeniería y Universidad, 13*(1), 137-156.

Ester, M., Kriegel, H. & Xu, X. (1995). Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. *International Symposium on Spatial Databases*. Berlin: Germany.

Han, J., Pei, J. & Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier.

Hartigan, J. (1979). Algorithm AS 136: A k-means *clustering* algorithm. *Journal of the Royal Statistical Society, 28*(1), 100-108.

Jain, A. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651-666.

Larose, D. T. (2014). *Discovering knowledge in data: An introduction to data mining*. John Wiley & Sons.

Mantel, N. (1967). The detection of disease *clustering* and a generalized regression approach. *Cancer research, 27*(2), 209-220.

Ng, H., Ong, S., Foong, K., Goh, P. & Nowinski, W. L. (2006). Medical image segmentation using k-means clustering and improved watershed algorithm. *7th IEEE Southwest Symposium on Image Analysis and Interpretation*. Denver, USA.

Peng, W., Chen, J. & Zhou, H. (2009). An Implementation of ID3 --- Decision Tree Learning Algorithm.

Perversi, I. & Fernández, M. I. (2007). Aplicación de minería de datos para la exploración y detección de patrones delictivos en Argentina. En XIII Congreso Argentino de Ciencias de la Computación.

Rahm, E. & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering, 23*(4), 3-13.

Rajesh, D. (2011). Application of spatial data mining for agricultur. *International Journal of Computer Applications, 15*(2), 7-9.

Ruiz Gómez, J. (2016). *Análisis espacial y supervisado de datos criminalísticos utilizando un sistema de información geográfica*.

Suárez, J. (2016). *Técnicas de agrupamiento de minería de datos espaciales para la caracterización y análisis de los hurtos que afectan Bogotá*.

Swian, S. & Sarangi, S. (2013). Study of Various Classification Algorithms using Data Mining. *Int. J. Adv. Res. Sci. Technol, 2*(2), 110-114.

Tanasa, D. & Trousse, B. (2004). Advanced data preprocessing for intersites web usage mining. *IEEE Intelligent Systems, 19*(2), 59-65.

Wang, S. & Li., D. (2004*).* Concepts, principles and applications of spatial data mining and knowledge discovery.

Yuxun, L. & Niuniu, X. (2010). Improved ID3 algorithm. In *3rd IEEE International Conference in Computer Science and Information Technology (ICCSIT)*. CIUDAD, PAÍS.

Cómo citar este capítulo:
Sanmartín Mendoza, P., Castillo Vélez, J., Heredia Vizcaíno, D. & Quintero Méndez, V. (2018). Aplicación de técnicas de minería de datos sobre datos georreferenciados para obtener un modelo predictivo: Caso de estudio: Hurtos en la ciudad de Barranquilla. En V. Quintero, D. Heredia Vizcaíno, P. Sanmartín Mendoza, R. Romero, & J. Castillo Vélez, *Datos, Información, Tendencias, tres miradas sobre un contexto cambiante* (pp.77-93). Barranquilla: Ediciones Universidad Simón Bolívar.

Acerca de los autores

Vladimir QUINTERO MÉNDEZ
Profesor Investigador, Universidad Simón Bolívar (Barranquilla)
Vquintero2@unisimonbolivar.edu.co

Diana HEREDIA VIZCAÍNO
Profesor Investigador, Universidad Simón Bolívar (Barranquilla)
dianahv@unisimonbolivar.edu.co

Paul SANMARTÍN MENDOZA
Profesor Investigador, Universidad Simón Bolívar (Barranquilla)
psanmartin@unisimonbolivar.edu.co

Jhonny CASTILLO VÉLEZ
Estudiante de Maestría en Ingeniería de Sistemas y Computación, Universidad Simón Bolívar (Barranquilla)
jonny.castillo@unisimonbolivar.edu.co

Ronald ROMERO ZÚÑIGA
Estudiante de Maestría en Ingeniería de Sistemas y Computación, Universidad Simón Bolívar (Barranquilla)
ronald.romeroz@gmail.com