Marco de Trabajo Aplicado a la Epidemiología en Santander

Autor:

Yuly Andrea Ramírez Sierra

Trabajo de Investigación para Optar el Título de

Magister en Ingeniería Industrial

Director:

Henry Lamos Díaz

Ph.D en Matemática - Física

Universidad Industrial de Santander

Facultad de Ingenierías Físico-Mecánicas

Escuela de Estudios Industriales y Empresariales

Maestría en Ingeniería Industrial

Bucaramanga

2019

Dedicatoria

A aquellos que trabajan desde las diferentes ramas del conocimiento en la propuesta y desarrollo de proyectos que buscan tener un impacto positivo en la sociedad y sobre todo contribuir al bienestar de ésta, integrando diferentes disciplinas y construyendo relaciones que no solo involucren académicos y expertos sino a todos los interesados, con el fin de proponer estrategias integrales de mejora

Agradecimientos

El desarrollo de un trabajo de investigación es un proceso gratificante y retador que lleva a explorar capacidades y a desarrollar la visión que se tiene sobre lo que sucede a nuestro alrededor y sobre todo a observar y analizar el impacto que tiene el aplicar este conocimiento en contribuciones al bienestar de la sociedad, de esta manera agradezco la oportunidad de cursar un programa de investigación que me permitió crecer en muchos aspectos y ha dejado lecciones vitales y útiles.

Agradezco al profesor Henry Lamos, por apoyarme en esta travesía y ser un guía en el desarrollo de este proyecto, contribuyendo a mi desarrollo personal y profesional.

Agradezco los espacios ofrecidos en los semilleros de investigación del Grupo de Optimización de Sistemas Productivos, Administrativos y Logísticos -OPALO - de la Escuela de Estudios Industriales y Empresariales, ya que me permitió compartir y motivar a estudiantes de pregrado a involucrarse en las diferentes líneas de investigación y así fortalecer cada vez la investigación tanto del grupo como de la Escuela.

Agradezco el apoyo de personitas especiales como Tatiana Rodríguez, Johana Rueda, Diana Osma y David Puentes, que con sus enseñanzas y consejos contribuyeron a que en esta etapa de mi vida no solo me fortaleciera a nivel profesional sino también personal.

Agradezco a Pedro Ramírez y a Marly Ramírez, por su apoyo incondicional y por la linda hermandad que nos une.

Agradezco la oportunidad de codirigir algunos proyectos de investigación de pregrado, en los cuáles se exploraron otras aplicaciones de la analítica de datos y por tanto se desarrollaron otras habilidades en esta línea de interés en investigación.

Agradezco al Ministerio de Salud y Protección Social y al IDEAM por compartirme bases de datos importantes para integrar dentro del repositorio de estudio de mi proyecto.

Tabla de contenido

Introducción 16
1. Objetivos 20
1.1 Objetivo general 20
1.2 Objetivos Específicos 20
2. Marco teórico 22
2.1 Marco de trabajo 22
2.2 Proceso de extracción de conocimiento 22
2.3 Big data 22
2.4 Herramientas Big data 26
2.4.1 Apache Hadoop. 26
2.4.2 Apache Spark. 26
2.5 Minería de datos 29
2.6 Aprendizaje automático 29
3. Revisión de literatura 31
3.1 Big data y vigilancia epidemiológica 31
3.2 Mapas de riesgos 37
3.3 Factores asociados a eventos de salud pública 39
3.3.1 Enfermedades infecciosas 40
3.3.2 Enfermedades crónicas o enfermedades no transmisibles (ENT).. 44
3.4 Métodos de monitoreo de enfermedades 47
3.5 Modelos matemáticos para la vigilancia epidemiológica 49
4. Marco de trabajo 60
4.1 Caracterización del área de estudio 61
4.2 Proceso de extracción de conocimiento 63
4.2.1 Integración y recopilación de la información. 63
4.3 Procesamiento, evaluación e interpretación de la información. 65
4.4 Difusión y uso 66

5. Estudio de caso 67
5.1 Caracterización del área de estudio 67
5.2 Proceso de extracción de conocimiento 72
5.2.1 Integración y recopilación de la información. 72
5.2.2 Procesamiento, evaluación e interpretación de la información. 83
5.3 Curva de aprendizaje 92
5.4 Importancia de las variables para los modelos 94
5.4.1 Visualización de la información. 96
Conclusiones 96
Recomendaciones 100
Referencias bibliográficas 101

Lista de figuras

Figura 1. Componentes de Spark. 28

Figura 2. Cadena de factores para las causas de enfermedades crónicas..... 45

Figura 3. Marco de trabajo para la vigilancia epidemiológica en Santander, Colombia 62

Figura 4. Comportamiento del dengue, dengue grave y mortalidad por dengue, periodo 2007–2016 70

Figura 5. Tiempos de procesamiento agrupamiento nítido..... 74

Figura 6. Tiempos de procesamiento agrupamiento difuso 75

Figura 7. Tiempos de procesamiento agrupamiento k-medias difuso 78

Figura 8. Representación de los niveles de incidencia de dengue en Santander periodo 2007 – 2016 81

Figura 9. Árbol de decisión con partición gini 85

Figura 10. Árbol de decisión con partición información 85

Figura 11. Bosque aleatorio comparando Exactitud vs N° de predictores 85

Figura 12. Error vs N° de observaciones en nodos terminales 86

Figura 13. Evolución del error out-of-bag vs número de árboles 86

Figura 14. Modelos de regresión logística multinomial - Exactitud vs Decay 87

Figura 15. Sensibilidad por cada clase vs 89

Figura 16. Recuperación por cada clase vs los modelos ajustados..... 89

Figura 17. Especificidad por cada clase vs los modelos ajustados 90

Figura 18. Precisión por cada clase vs los modelos ajustados 91

Figura 19. Curva de aprendizaje árbol de decisión CART 92

Figura 20. Curva de aprendizaje bosques aleatorios 93

Figura 21. Curva de aprendizaje regresión logística multinomial 93

Figura 22. Importancia de variables en árbol de decisión 95

Figura 23. Importancia de variables bosque aleatorio 95

Figura 24. Importancia de variables en regresión logística multinomial 95

Lista de tablas

Tabla 1. Factores de riesgo para enfermedades infecciosas y/o cronicas 47

Tabla2. Método de análisis y actores asociados a eventos 58

Tabla 3. Núcleos de Desarrollo Provincial en Santander 68

Tabla 4. Casos de dengue y dengue grave en el departamento de Santander 69

Tabla 5. Variables del caso de estudio 72

Tabla 6. Tiempos de procesamiento (segundos)- índices para agrupamiento nítido y difuso 75

Tabla 7. Resultados de los mejores algoritmos de agrupamiento 77

Tabla 8. Tiempo procesamiento k-medias difuso 78

Tabla 9. Agrupaciones de municipios por año del periodo 2007-2016 79

Tabla 10. Niveles de incidencia 80

Tabla 11. Perfil de los niveles de incidencia del dengue en Santander, periodo 2007- 2016..... 82

Tabla 13. Resultados de los clasificadores 88

Tabla 14. Proporciones de presencia de clases en los datos 88

Lista de apéndices

(Ver apéndices adjuntos en el CD y pueden visualizarlos en la Base de Datos de la Biblioteca UIS)

Apéndice A. Marco de trabajo para la vigilancia epidemiológica en Santander

Apéndice B. Arquitectura Big Data

Apéndice C. Eventos de interés en salud pública

Apéndice D. Descripción de variables

Apéndice E. Variables seleccionadas

Apéndice F. Índices para validación del agrupamiento

Apéndice G. Niveles de incidencia del dengue en Santander, periodo 2007 – 2016

Apéndice H. Perfil de los niveles de incidencia del dengue en Santander, periodo 2007-2016

Glosario

- **Brote:** es una epidemia limitada a incrementos localizados en la incidencia de una enfermedad, ejemplo: en una villa, pueblo, o institución; aumento significativo es utilizado algunas veces como un eufemismo para brote.
- **Enfermedad endémica:** es la presencia constante de una enfermedad o agente infeccioso dentro de un área geográfica o grupo de población (Porta, 2008)
- **Epidemia:** es la ocurrencia en una comunidad o región de casos de una enfermedad, comportamiento específico relacionado a salud, u otros eventos relacionados con la salud, claramente superiores a la expectativa normal (Porta, 2008).
- **Epidemiología:** es el estudio de la ocurrencia y distribución de estados o eventos relacionados con la salud en poblaciones específicas, incluyendo el estudio de los determinantes que influyen en dichos estados, y la aplicación de este conocimiento al control de los problemas de salud.
- **Evento:** manifestación de una enfermedad o un suceso principalmente patógeno (Reglamento Sanitario Internacional (2005), 2016). Estados y eventos relacionados con la salud incluye enfermedades, causas de muerte, comportamientos, reacciones a programas preventivos, y provisión y uso de los servicios de salud (Porta, 2008).
- **Pandemia:** es una epidemia que ocurre a nivel mundial o sobre un área amplia, cruzando fronteras internacionales y por lo general afectando a un gran número de personas (Porta,

2008).

- **Patógeno:** es un organismo que tiene la capacidad de causar una enfermedad (Porta, 2008).
- **Salud pública:** es uno de los esfuerzos organizados por la sociedad para proteger, promocionar y restaurar la salud de las personas. Es la combinación de ciencias, habilidades y convicciones que están dirigidas al mantenimiento y a mejorar la salud de todas las personas a través de acciones colectivas o sociales (Porta, 2008).

RESUMEN

TÍTULO: MARCO DE TRABAJO APLICADO A LA EPIDEMIOLOGÍA EN SANTANDER*

AUTOR: YULY ANDREA RAMÍREZ SIERRA**

PALABRAS CLAVE: SAL UD PÚBLICA, VIGILANCIA EPIDEMIOLÓGICA, MINERÍA DE DATOS, APRENDIZAJE AUTOMÁTICO, BIG DATA

DESCRIPCIÓN:

La vigilancia es una característica esencial de la práctica epidemiológica, y es el resultado de la recolección, validación, análisis e interpretación de datos de salud y enfermedades. El valor de la vigilancia implica la entrega efectiva y eficiente de información útil. Así, los sistemas de vigilancia deben ser flexibles tanto al incremento en las necesidades de información, como a la utilización de tecnologías apropiadas que permitan difundir datos oportunamente. Entonces, para afrontar la brecha en el conocimiento de la salud, la vigilancia de la salud pública puede beneficiarse de los avances en ciencias de la información, tecnologías y el aumento de bases de datos y fuentes de datos.

Por consiguiente, en esta investigación se construye un marco de trabajo como una guía para la vigilancia epidemiológica, en las etapas de análisis e interpretación de datos, y difusión de la información, con apoyo del paradigma big data, adoptando técnicas de minería de datos para descubrir patrones y tendencias, y de aprendizaje automático para construir modelos predictivos, con el fin de apoyar la toma de decisiones de las autoridades de salud pública.

El marco de trabajo se define teniendo en cuenta las etapas claves del descubrimiento de conocimiento en bases de datos (KDD), considerando cinco fases tales como: caracterización del área de estudio; integración y recopilación de la información; selección, limpieza y transformación; procesamiento, evaluación e interpretación de la información, y difusión y uso. Finalmente, se realiza un estudio de caso para un evento de salud pública crítico de Santander, Colombia como lo es el virus del dengue.

---

* Trabajo de grado

** Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director Henry Lamos Díaz, Doctor en Matemática - Física

ABSTRACT

TITTLE: FRAMEWORK APPLIED TO EPIDEMIOLOGY IN SANTANDER[*]

AUTHOR: YULY ANDREA RAMÍREZ SIERRA[**]

KEYWORDS: PUBLIC HEALTH, EPIDEMIOLOGICAL SURVEILLANCE, DATA MINING, MACHINE LEARNING, BIG DATA

DESCRIPTION:

Surveillance is an essential feature of epidemiological practice, and it is the result of the collection, validation, analysis and interpretation of health and disease data. The value of surveillance implies the effective and efficient delivery of useful information. Besides, surveillance systems should must be flexible to the increase in information needs, as well as using the appropriate technologies to disseminate data opportunely. Therefore, to address the gap in health knowledge, public health surveillance can benefit from advances in information sciences, technologies and the increase of databases and data sources.

In this research, a framework is constructed as a guide for epidemiological surveillance, in the stages of analysis and interpretation of data, and dissemination of information, with support of the big data paradigm, adopting data mining techniques to discover patterns and trends, and machine learning to build predictive models, in order to support the decision making of public health authorities.

The framework is based on the key stages of knowledge discovery in databases (KDD). It has five phases sush as: characterization of the study area; integration and collection of information, selection, cleaning and transformation; processing, evaluation and interpretation of information and dissemination and use. Finally, we made a case study is made for a critical public health event in Santander (Colombia) such as the dengue virus.

---

[*] Master Thesis

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director Henry Lamos Díaz, Doctor en Matemática - Física

Introducción

En el sector salud el campo de la epidemiología es considerado el centro de la salud pública (Ogino et al., 2015), y la medicina preventiva (Galea, 2013), tiene como desafío en la era de los grandes datos, no solo comprender las causas de la salud de la población sino optimizar las intervenciones para mejorarla (Mooney, Westreich, & El-Sayed, 2015). Se destaca, que a través de la historia la epidemiología se ha preocupado por tres aspectos: la extensión espacial de los brotes, el progreso de la enfermedad y los posibles controles, la identificación del origen de la enfermedad y la comparación con brotes ocurridos en el pasado (Govindaraju, Raghavan, & C.R., 2013).

Sin embargo, se requiere que la epidemiología amplíe su alcance más allá de la perspectiva histórica de la etiología para abarcar la detección temprana de la enfermedad, el tratamiento, el pronóstico y convertirse en una traducción más eficaz de los descubrimientos científicos, generando un impacto en la salud individual y poblacional, retos que le permitirán enfocarse en la precisión del análisis de los factores de riesgo para abordar cuestiones de gran importancia social actual, como la economía de los servicios de salud, el envejecimiento de la población, la creciente carga de enfermedades crónicas comunes, la persistencia de las desigualdades sanitarias y la salud mundial (Khoury, 2015; Khoury et al., 2013).

En este sentido, la epidemiología computacional del Big Data es un área interdisciplinar

emergente, que utiliza modelos computacionales y big data para entender y controlar la difusión espacio-temporal de una enfermedad a través de la población (Govindaraju et al., 2013). De manera que se está optando por desarrollar tecnologías en bases de datos, así como aplicar técnicas de minería de datos y aprendizaje automático, que permiten manejar cantidades masivas de información, construir modelos y apoyar la toma de decisiones. Así, se busca utilizar herramientas que tengan la capacidad de afrontar el reto de escalabilidad que se presenta con los grandes datos (Manuel & Sesmero, 2015). Se resalta que dentro del Big Data, la analítica aporta técnicas robustas, y las herramientas y técnicas de visualización tienen cada vez más importancia, dado que permiten apoyar la toma de decisiones (McAfee & Brynjolfsson, 2012); en el caso de la analítica, el análisis predictivo comprende una variedad de técnicas que predicen resultados futuros basados en datos históricos y actuales (Gandomi & Haider, 2015).

Así, los grandes datos permiten identificar patrones de enfermedades para realizar las intervenciones en el estado previo con el fin de actuar oportunamente (Docherty & Lone, 2015). En este sentido, la Organización Mundial de la Salud (OMS) manifiesta que todos los países deben detectar y responder de forma adecuada a las amenazas de enfermedades emergentes y con tendencia a producir epidemias, con el fin de reducir su impacto en la salud y la economía de la población mundial.

En el contexto nacional, en la ley 1122 del 2007 mediante la cual se realizan ajustes al Sistema General de Seguridad Social en Salud de Colombia, se establece que la salud pública está guiada por políticas que buscan garantizar la salud de la población mediante acciones dirigidas de manera individual y colectiva, conformando indicadores de condiciones de vida, bienestar y desarrollo del

país; de manera que se ha establecido el Plan Decenal de Salud Pública 2012-2021, como una guía que define la interacción de diferentes actores y sectores públicos, privados y comunitarios para crear condiciones que garanticen el bienestar integral y la calidad de vida en Colombia (Ministerio de Salud y de Proteccion Social, 2012). Adicionalmente para el 2021 en el PDSP, se proyecta un Sistema de Vigilancia en Salud Pública en todo el territorio nacional, en donde se integren los sistemas de vigilancia y control sanitario, e inspección, en coordinación con las entidades territoriales.

Adicionalmente, en Colombia una de las funciones de la Dirección de Epidemiología y Demografía es promover, orientar y dirigir la elaboración del estudio de la situación de salud, por tanto se ha fortalecido el proceso de construcción de informes sobre el Análisis de la Situación de la Salud (ASIS), los cuales se realizan con el propósito de definir necesidades, prioridades, políticas en salud y evaluar su pertinencia y cumplimiento, para formular estrategias de promoción, prevención, control de daños a la salud y construir escenarios prospectivos de salud (Ministerio de Salud y de Proteccion Social, 2014).

De lo anteriormente mencionado, se observa la importancia de adoptar el paradigma big data en epidemiología, así como la oportunidad de integración de bases de datos para descubrir patrones novedosos que aporten conocimiento de apoyo a la toma de decisiones estratégicas en salud pública.

Por consiguiente en esta investigación se propone un marco de trabajo como guía a la vigilancia epidemiológica en Santander en las etapas de análisis e interpretación de datos, y difusión de la

información, teniendo en cuenta la aplicación del paradigma big data en eventos de salud pública y los diferentes factores que influyen en estos eventos, así mismo se adoptan técnicas de análisis, como lo son la minería de datos y el aprendizaje automático para aprovechar esta información con fines descriptivos y también predictivos, por un lado descubrir patrones y tendencias y por otro construir un modelo que caracterice el comportamiento de un evento de salud pública crítico de Santander y apoye a los tomadores de decisiones.

En el documento inicialmente se detallan los objetivos de investigación, seguido por un marco conceptual y un marco teórico que menciona aspectos importantes sobre Big Data, minería de datos y aprendizaje automático, la revisión de literatura mediante la cual se identifican aspectos relevantes del Big Data aplicado a salud y a epidemiología, y factores asociados a eventos de salud pública. Después se detallan los componentes del marco de trabajo, el estudio de caso y las conclusiones.

1. Objetivos

1.1 Objetivo general

Desarrollar un marco de trabajo, bajo el paradigma big data para la vigilancia de eventos epidemiológicos en el departamento de Santander mediante técnicas de aprendizaje automático.

1.2 Objetivos Específicos

Identificar los factores asociados al comportamiento de eventos epidemiológicos, mediante una revisión literaria.

Consolidar diferentes bases de datos que describan eventos epidemiológicos para conformar el repositorio objeto de estudio.

Desarrollar un plan de gestión de datos para la adquisición, almacenamiento, preprocesamiento y procesamiento de datos del estudio.

Diseñar modelos de aprendizaje automático y minería de datos para el comportamiento de eventos epidemiológicos.

Desarrollar una herramienta que permita probar los modelos diseñados para los datos objeto de estudio, como apoyo a toma de decisiones.

2. Marco teórico

2.1 Marco de trabajo

Un marco de trabajo está orientado a los usos del conocimiento y las herramientas existentes que proporcionan una base metodológica al estudio que se esté llevando a cabo (Robertson, 2017); por ejemplo, Ali et al. (2016), proponen un marco de trabajo que busca establecer un estándar y proporcionar los detalles necesarios para la futura implementación del sistema de vigilancia de enfermedades infecciosas en regiones con recursos limitados.

2.2 Proceso de extracción de conocimiento

El proceso de extracción de conocimiento (KDD por sus siglas en inglés), se define como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos" (Fayyad, Piatetsky-Shapiro, & Smyth, 1996)

2.3 Big data

En la era digital actual, debido al desarrollo del internet y de las tecnologías, el volumen de los datos ha aumentado cada día proveniente de diferentes fuentes (Fahad et al., 2014). Por lo tanto, el concepto de Big Data se adopta para captar el significado de la tendencia de explosión de los

datos, y ha generado un nuevo campo de investigación atractivo para diversos sectores como industrias, gobiernos y comunidad de investigadores (Hu, Wen, Chua, & Li, 2014).

La definición de Big Data es diversa y ha sido difícil llegar a un consenso (H. Hu et al., 2014), sin embargo, se destacan tres definiciones: por una lado la consultora tecnológica IDC (International Data Corporation) expresa: "Las tecnologías de Big Data describen una nueva generación de tecnologías y arquitecturas, diseñadas para extraer económicamente valor desde volúmenes muy grandes de una amplia variedad de datos, para permitir la captura, el descubrimiento y/o el análisis a alta velocidad" (Gantz & Reinsel, 2011). Por otro lado, Gartner en 2011 define el término como: "Big Data excede el alcance de los entornos de hardware de uso común para capturar, gestionar y procesar los datos dentro de un tiempo transcurrido tolerable para su población de usuarios" (Aguilar L. J., 2013, pág. 3).

Los Big Data exceden la capacidad de las herramientas tradicionales utilizadas para la gestión y el procesamiento de los datos (Docherty & Lone, 2015; Martínez, 2015; Tan, Gao, & Koch, 2015), similarmente en un informe de McKinsey (Manyika et al., 2011) se menciona que: "Big Data son conjuntos de datos cuyo tamaño está más allá de la capacidad de las típicas herramientas de software de base de datos para capturar, almacenar, administrar y analizar", la cual se considera que es una definición subjetiva porque no define a Big data en términos de una métrica, pero engloba un aspecto dinámico ya sea en el tiempo o en el sector, respecto a la magnitud del conjunto de datos para ser considerado grande (H. Hu et al., 2014).

El Instituto Nacional de Estándares y Tecnología (NIST) (como se citó en Hu et al., 2014, pág

654), manifiesta que “Big Data es en donde el volumen de datos, velocidad de adquisición, o representación de datos limita la habilidad para desempeñar análisis efectivo usando aproximaciones relacionales tradicionales o requiere el uso de escalamiento horizontal significativo para un procesamiento eficiente”.

Los Big Data se caracterizan por los siguientes aspectos:

a. **Variedad**: relacionada con la heterogeneidad estructural de los datos, en donde los avances tecnológicos permiten usar varios tipos de datos como: estructurados, semiestructurados y no estructurados.
a. **Volumen**: relacionado con el tamaño de los datos (Tan et al., 2015), el cual ha ido aumentado a través de los años, por ejemplo entre 1970s y 1980s se tratan desde megabyte a gigabyte, y en el 2011 se empieza a generar información con una magnitud de exabytes (H. Hu et al., 2014).
b. **Velocidad**: relacionada con la velocidad de generación y análisis los datos (Assunção, Calheiros, Bianchi, Netto, & Buyya, 2015), presentándose el gran reto del análisis en tiempo real, como consecuencia del aumento acelerado de teléfonos inteligentes y sensores que han contribuido a la generación de datos (Gandomi & Haider, 2015).
c. **Veracidad**: es un reto dado el aumento de las diferentes fuentes de datos, porque se debe garantizar la calidad de la información capturada para generar conocimiento válido y confiable; según Gandomi & Haider (2015) la fiabilidad se convierte en un reto para Big Data, y se trata mediante herramientas y análisis para la gestión y extracción de datos inciertos.
d. **Valor:** consiste en obtener conocimiento de grandes bases de datos de manera eficiente y

rentable (Aguilar L. J., 2013), es decir, implica transformar la información en conocimiento que aporte valor económico a las compañías y a la sociedad (Gabrielsson, Politis, Dahlstrand, & Patents, 2016).

Es importante resaltar, que en el desarrollo y gestión de Big Data se presentan retos en cuanto a seguridad, calidad y gestión de los datos, privacidad, la inversión justificada que implica tener definición clara del problema para evaluar las necesidades de herramientas y tecnologías de gestión y análisis de datos, y por último la falta de científicos de datos calificados (Lee, 2017). Adicionalmente, se identifican tres tendencias (Grossglauser & Saner, 2014): la incertidumbre de los datos, relacionada con el riesgo de una simplificación excesiva; la escalabilidad que se da por el crecimiento exponencial del volumen de información, tendencia que ha sido tratada por el servicio en la nube, caracterizado por la flexibilidad de aportar recursos cuando es necesario y así se reducen las inversiones de capital; y, la energía es el recurso más crítico en el universo de los Big Data, dado que las necesidades energéticas de la infraestructura de la computación superarán otras industrias, por tanto, se están realizando esfuerzos para reducir esta huella energética.

En conclusión, la gestión de grandes bases de datos presenta dificultades en la captura, almacenamiento, búsqueda, análisis, compartición y visualización (Pérez M, María, 2015), así la virtualización y la computación en la nube han avanzado significativamente y han facilitado el desarrollo de plataformas más eficientes para apoyar la gestión de estos datos (Aguilar, 2013). Las tecnologías de gestión de datos se clasifican en sistemas de archivos, tecnologías de bases de datos, y modelos de procesamiento (H. Hu et al., 2014).

2.4 Herramientas Big data

A continuación, se presenta una breve descripción de algunos marcos de software que apoyan el paradigma big data:

2.4.1 Apache Hadoop. Hadoop, ha atraído la atención considerable tanto de la industria como de los académicos. Además, ha sido durante mucho tiempo el pilar del movimiento de Big Data y es un marco de software de código abierto que admite el almacenamiento y procesamiento masivo de datos (H. Hu et al., 2014). Una de sus principales características es la capacidad para explotar la informática local de datos, es decir, ofrece la posibilidad de acercar las aplicaciones a los datos, lo cual reduce las congestiones de red y aumenta el rendimiento global del sistema (Cattaneo, Giancarlo, Petrillo, & Roscigno, 2018). Hadoop es especialmente adecuado para la gestión y el análisis de big data, al caracterizarse por la escalabilidad, eficiencia de costos, flexibilidad y tolerancia a fallas.

El marco de trabajo Apache Hadoop está soportado por el paradigma MapReduce, el cual es un modelo de programación que permite el paralelismo automático y la distribución de aplicaciones de cómputo a gran escala. El modelo computacional consta de dos funciones definidas por el usuario, llamadas Mapear y Reducir (Apache™ Hadoop®, n.d.)

2.4.2 Apache Spark. Apache Spark es un marco de trabajo de código abierto y permite la computación distribuida para grandes y complejos conjuntos de datos. Spark y MapReduce son linealmente escalables y tolerantes a las fallas, pero Spark puede ser hasta 100 veces más rápido para ciertas aplicaciones y proporciona interfaces de máquina intuitivas (H. Hu et al., 2014).

Apache Spark™ es una plataforma de computación en cluster de acceso abierto. Se encuentra catalogada como una herramienta big data según The Apache Software Foundation (por sus siglas en inglés, ASF) (The Apache Software Foundation, 2017).

Spark admite programas con computación en memoria y un modelo de flujo de datos acíclico, permite la combinación de datos streaming y de procesamiento por lotes, a diferencia de Hadoop que solo se puede utilizar para aplicaciones por lotes. Además, no se limita al marco de trabajo basado en MapReduce, pero conserva las funciones de los marcos de trabajo basados en este paradigma (Cattaneo et al., 2018). Por otra parte, Spark proporciona APIs de alto nivel en Java, Scala, Python y R (Spark™, n.d.):

2.4.2.1 Arquitectura de Spark.

Spark tiene una arquitectura *maestro/esclavo* en donde hay tres procesos principales: worker, executor y driver. Un worker ejecuta un nodo esclavo; un executor es un servicio lanzado para una aplicación en un nodo worker que ejecuta tareas y almacena los datos en memoria o en disco; el Driver se ejecuta en el nodo maestro y gestiona los ejecutores usando un gestor de recursos de cluster (Cattaneo et al., 2018). Como gestores de clúster se encuentran:

- Clúster independiente: administrador de clúster incluido con Spark que facilita la instalación de un clúster.
- Apache Mesos: administrador de clúster general que también puede ejecutar Hadoop MapReduce y aplicaciones de servicio.
- Hadoop YARN: administrador de recursos en Hadoop 2.

Así, las aplicaciones Spark se ejecutan como conjuntos independientes de procesos en un cluster

coordinado por el *SparkContext* en el programa principal que recibe el nombre de *driver program.* Para ejecutar procesos en un clúster, SparkContext se conecta a los administradores de clúster que asignan recursos a través de las aplicaciones, y una vez establecida esta conexión, Spark adquiere ejecutores en los nodos del clúster y les envía el código de aplicación. Finalmente, SparkContext envía tareas a los ejecutores (ver Figura 1) (Spark™, 2018a).

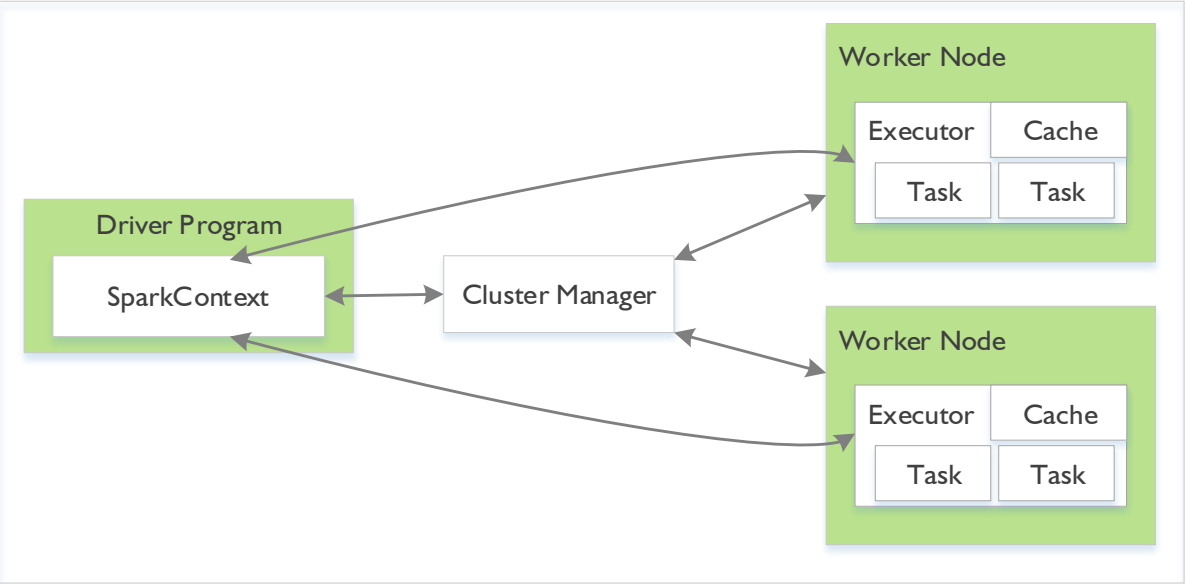

Figura 1. Componentes de Spark. Adaptado de Spark™, A. (2018). Cluster Mode Overview - Spark 2.0.2 Documentation. Retrieved from http://spark.apache.org/docs/latest/cluster-overview.html

Adicionalmente, la principal característica de Apache Spark es un conjunto de datos distribuido resistente (Resilient Distributed Dataset, RDD), que es una colección de elementos particionados a través de nodos del cluster que puede ser operado en paralelo.

Todas las trasformaciones en Spark son lentas (lazy), porque no calculan el resultado inmediatamente, sino que recuerdan la transformación aplicada a algún conjunto de bases de datos. Es decir, las transformaciones se calculan cuando una acción requiere que un resultado sea

retornado en el driver program (Spark™, 2018b).

2.5 Minería de datos

La minería de datos consiste en el análisis de enormes conjuntos de datos con el objetivo de identificar correlaciones y patrones válidos, nuevos, potencialmente útiles y comprensibles, y estas relaciones o resúmenes constituyen el modelo de los datos que se han analizado (Mdaghri, Yadari, Benyoussef, & Kenz, 2016). Existen diferentes formas de representar los modelos y cada una de ellas determina el tipo de técnica que puede usarse para inferirlos, es decir, los modelos pueden ser de tipo predictivos o descriptivos (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004).

Los modelos predictivos estiman valores futuros o desconocidos de variables objetivos o dependientes, usando otras variables denominadas variables dependientes o predictivas, para lo cual se utilizan técnicas de clasificación que es una de las tareas predictivas más utilizada, así como la regresión. En contraste, los modelos descriptivos, permiten identificar patrones que explican o resumen los datos; es decir, sirven para explorar las propiedades de los datos, mediante el agrupamiento (clustering), las correlaciones y las reglas de asociación.

2.6 Aprendizaje automático

Aprendizaje automático es un subcampo de la inteligencia artificial (Singh et al., 2018) y comprende un conjunto de métodos que pueden detectar patrones automáticamente en datos y entonces utiliza los patrones descubiertos para predecir datos fututos. Según Tom Mitchell “una computadora aprende de la experiencia (E) con respecto a alguna clase de tareas (T) y medida de

desempeño (P), si su desempeño en las tareas T, medida mediante P, mejora con la experiencia E" (citado por Lugo-Reyes, Maldonado-Colín, & Murata, 2014).

Además, se caracteriza por la representación y la generalización, es decir, que los datos analizados estén representados por el modelo, y que este modelo funcione bien frente a nuevos datos. En algunos casos el aprendizaje automático utiliza minería de datos como un paso previo para mejorar la exactitud de la predicción, o como parte del aprendizaje no supervisado. Es decir, que mediante la minería de datos se podrían extraer metadatos de interés (Lugo-Reyes et al., 2014).

Los modelos de aprendizaje automático, se categorizan como (Singh et al., 2018):

- **Aprendizaje supervisado:** el modelo tiene un conjunto de datos con etiquetas que corresponden a un vector de características, y además contienen las etiquetas de salida esperadas, así que, su meta es generar una función que mapea el vector de características a las etiquetas de salida. Dentro de los modelos basados en este aprendizaje se encuentras: máquinas de vectores de soporte, regresión lineal, bosques aleatorios, árboles de decisión y redes neuronales convolucionales.

- **Aprendizaje no supervisado:** en este tipo de aprendizaje, los conjuntos de datos no contienen información sobre las etiquetas de salida, así que, el objetivo de estos modelos es descubrir la relación entre las observaciones y/o revelar las variables latentes. Como modelos basados en este aprendizaje, se encuentran: k-medias, mapas auto-organizados, y las Redes Generativas Antagónicas (GANs, por sus siglas en inglés).

3. Revisión de literatura

3.1 Big data y vigilancia epidemiológica

La vigilancia de la salud pública es el resultado de la recolección, validación, análisis e interpretación de datos de salud y enfermedades (Amato-Gauci & Ammon, 2008; Tsui, Wong, Jiang, & Lin, 2011), de gran importancia para que los interesados basados en la evidencia implementen estrategias encaminadas a la prevención y control de enfermedades o brotes de enfermedades, siendo esencial tanto la difusión de información a los interesados como la garantía de la calidad, validez y comparabilidad de los datos, con el fin de apoyar la planeación, implementación y evaluación de las prácticas de salud pública (Amato-Gauci & Ammon, 2008; Porta, 2008; *Reglamento Sanitario Internacional (2005)*, 2016). Según la OMS, la vigilancia de la salud pública puede apoyar en los siguientes aspectos:

- Servir como un sistema de alertas tempranas para obstaculizar las emergencias en salud pública.
- Documentar el impacto de una intervención, o cortar el progreso a través de metas específicas.
- Monitorear y clarificar la epidemiología de los problemas de salud, permitir que se establezcan prioridades e informar las políticas y estrategias de salud pública.

Según el diccionario de epidemiología de Porta (2008), la vigilancia es la característica esencial

de la práctica epidemiológica y de salud pública, en donde los datos son recolectados sistemáticamente, generalmente usando métodos reconocidos por su practicidad, uniformidad y rapidez. Con esta información, se observan tendencias en el tiempo, lugar y personas, estos cambios pueden ser observados o anticipados con el fin de tomar acciones apropiadas incluyendo medidas de investigación y control. Las fuentes de datos deben relacionar directamente a enfermedades o a factores que influyen en las enfermedades (Porta, 2008).

Partiendo de que el valor de la vigilancia implica la entrega efectiva y eficiente de información útil, los sistemas de vigilancia se deben caracterizar por ser flexibles al incremento en las necesidades de información, así como utilizar las tecnologías apropiadas que permitan difundir datos oportunamente. Por tanto, la brecha en el conocimiento de la salud puede ocurrir cuando los sistemas de vigilancia carecen de oportunidad, completitud, fácil adaptación, o eficiencia, y para afrontar esto, la vigilancia de la salud pública puede beneficiarse de los avances en ciencias de información, tecnologías y el aumento de bases de datos y fuentes de datos (Hall, Correa, Yoon, & Braden, 2012).

Muchos brotes de enfermedades y errores médicos anormales pueden ser evitados con la estandarización efectiva del cuidado de la salud, el mejoramiento y los sistemas de vigilancia (Tsui et al., 2011), y se resalta que la propagación de enfermedades infecciosas tiene impacto significativo a nivel individual y social, de ahí que los sistemas de datos son importantes para aproximarse a la prevención, detección, respuesta y gestión de brotes de enfermedades infecciosas de plantas, animales y humanos (O'Shea, 2017) .

Los sistemas de vigilancia tradicional están basados en reportar obligatoriamente alguna enfermedad a la autoridad de salud central, y algunas ocasiones se monitorean síndromes no específicos como marcadores de enfermedades específicas (Green & Kaufman, 2002; O'Shea, 2017). Pero estos sistemas no tienen la capacidad de identificar nuevas amenazas a la salud pública, aunque aseguran adecuadas respuestas ante los riesgos identificados (Paquet, Coulombier, Kaiser, & Ciotti, 2006). En estos sistemas tradicionales se realiza la vigilancia pasiva, la cual implica que las autoridades de salud reciben reportes de enfermedad o condiciones por parte de médicos, laboratorios, y otros proveedores de salud según lo requerido por la legislación de salud pública (O'Connell, Zhang, Leguen, Llau, & Rico, 2010).

La meta del sistema tradicional de enfermedades infecciosas implica identificar los cambios en la incidencia, ya sea en forma de brote agudo o la identificación de un cambio en las tendencias a largo plazo (Green & Kaufman, 2002). La evolución del brote identificado se monitorea por tiempo, región geográfica y por las características demográficas de los casos (Green & Kaufman, 2002). Las señales de cambios en la incidencia de morbilidad pueden ser la ocupación de camas de hospitales, uso de la unidad de cuidados intensivos, muestras enviadas a laboratorios, ausencias a la escuela y al trabajo, ventas de medicamentos, factores meteorológicos, y datos de animales (Green & Kaufman, 2002).

Esta forma tradicional aunque es importante para la vigilancia de enfermedades trasmisibles, no garantiza la rápida identificación de emergencias, así que se buscan otras formas de complementar esta información como la vigilancia sindrómica o actividades de monitoreo, recolectando datos estructurados que son compilados como indicadores (Amato-Gauci & Ammon,

2008). La vigilancia sindrómica, está relacionada con la adquisición automática de datos y la generación de alertas, monitorear los indicadores en tiempo real o cercano al tiempo real para detectar brotes de enfermedades (O'Connell et al., 2010).

Las posibles fuentes de información para los sistemas de vigilancia sindrómica son: visitas de los médicos de atención primaria (centinela), admisiones del departamento de emergencias, especialistas de enfermedades infecciosas, sospecha de muerte por enfermedades infecciosas, e informes de brotes a las oficinas de salud (Green & Kaufman, 2002). Es decir, se recolectan datos de pre-diagnósticos relacionados con el objetivo de detectar alertas tempranas de brote de enfermedades y epidemias que sería posible usando datos de vigilancia tradicional como diagnósticos confirmados e informes de laboratorio (Jefferson et al., 2008).

En el siglo XX aumentó significativamente la movilidad mundial y esta tendencia continua, lo que implica que personas, especies vivas y productos agrícolas crucen las fronteras influyendo en el incremento de la probabilidad de que los patógenos circulantes en un área se trasladen a otras regiones, por lo que es importante actuar hacia la prevención y control de enfermedades teniendo en cuenta que además del impacto de las infecciones sobre los individuos, también las enfermedades tienen un impacto social que puede desestabilizar economías, gobiernos, poblaciones e instituciones sociales (Hartley et al., 2010). Según Tsui et al. (2011), es necesario llevar a cabo investigaciones orientadas al diseño y modelamiento de la transmisión efectiva de enfermedades y la vigilancia de la salud pública, considerando que en el siglo XXI, la vigilancia de la salud pública debe tener en cuenta aspectos como: léxico común; necesidades de vigilancia global; la informática; trabajadores calificados; acceso y uso de datos; y gestión, almacenamiento

y análisis de datos (Hall et al., 2012).

Según Hall et al., (2012), nuevas fuentes de datos o de información, tales como datos clínicos no estructurados, fuentes de internet, medios de comunicación, y datos ambientales y de cambio climático proporcionan oportunidades para mejorar el conocimiento de la salud dado que otorga más detalles de los eventos de interés con respecto a tiempo, lugar y persona. Los medios de comunicación, y otras fuentes informales de información se han adoptado como fuentes valiosas de alertas de salud pública (Paquet et al., 2006). Así, se habla de vigilancia digital que puede aportar conocimiento a la salud pública partiendo del análisis de la información de salud almacenada digitalmente, y considerando la distribución y los patrones que rigen el acceso a estos datos (Milinovich, Williams, Clements, & Hu, 2014).

El control de los brotes epidémicos se ha facilitado con los avances de las ciencias modernas como nuevos medicamentos, nuevos tratamientos y la infraestructura cooperativa para el control y vigilancia de enfermedades, y la expansión de las tecnologías a nivel mundial tales como computadores y smartphone, lo que ha llevado a una oportunidad para recolectar y comunicar información relacionada con salud, por tanto, los datos de vigilancia en tiempo real son claves para la identificación rápida de las emergencias relacionadas con la salud pública, para optimizar la disposición de recursos como respuesta a estos eventos, y, para planear medidas de mitigación y detención (Paolotti et al., 2014).

En la vigilancia de salud pública se destaca el término, alerta temprana que es definida como la identificación de una enfermedad específica en un estado temprano en la historia natural de la

enfermedad, usualmente aparece para mejorar la supervivencia, y es justificada solamente si mejora significativamente los resultados a nivel individual o de la comunidad (Porta, 2008), así mismo debe tener como objetivos: intervenciones rápidas y ejecutar actividades de control (Hartley et al., 2010). Una forma de mejorar la detección temprana es monitorear el comportamiento de las consultas relacionadas con salud que realizan las personas en los buscadores en línea (Ginsberg et al., 2009). El Reglamento Sanitario Internacional-2005 propone un marco de trabajo legal internacional para la detección de alertas, informe y respuesta a brotes de enfermedades infecciosas. Además, las naciones miembros de la OMS están obligadas a desarrollar y mantener la vigilancia, informe, notificación, verificación y capacidades de respuesta. Cualquier nación debe reportar a la OMS dentro de 24 horas, la información relacionada con brotes de preocupación internacional (Hartley et al., 2010).

Adicionalmente, en la salud pública se destaca la importancia de evaluar el impacto de la enfermedad en la población lo cual está relacionado con la pérdida de años de vida saludables, que es expresado mediante indicadores como: años de vida saludables perdidos (HEALYs, por sus siglas en inglés), años de vida ajustados por discapacidad (DALYs, por sus siglas en inglés) o años de vida ajustados por calidad (QALYs, por sus siglas en inglés). Estos indicadores expresan el impacto que las muertes y la morbilidad causan en la sociedad, y permiten comparar el impacto de los factores de riesgo o enfermedades (Porta, 2008). Los años de vida ajustados por discapacidad son el resultado de la suma de los años de vida con discapacidad y los años de vida perdidos (Abajobir et al., 2017).

3.2 Mapas de riesgos

El objetivo de los mapas de riesgos es ubicar los lugares en los cuales se ha observado la enfermedad, es decir, se busca identificar el nicho fundamental del organismo objeto de estudio (Hay, George, Moyes, & Brownstein, 2013), dado que por los factores evolutivos, biogeográficos y ecológicos un organismo no explota todo el espacio geográfico disponible, por ejemplo, en un estudio realizado por Hay, et al (2013), utilizan el conocimiento de un experto para delimitar a nivel mundial el alcance del nicho del evento del dengue y así realizar el mapeo de este.

Los mapas permiten obtener información de ubicación y resaltar el vector con mayor presencia en determinada área, pero esto necesita complementarse con información sobre el comportamiento de las especies para que los interesados tengan información suficiente que apoye la toma de decisiones (Sinka et al., 2012). Cabe resaltar que los mapas de riesgos rápidamente se vuelven obsoletos, así que el aprendizaje automático y las fuentes abiertas facilitan la posibilidad de actualizar continuamente el atlas de las enfermedades infecciosas, lo que lleva a tener mapas de riegos dinámicos que son de gran valor para los profesionales de la salud, de manera que con el paradigma Big data se presenta un cambio conceptual de lo estático a mejores y evolucionados mapas de riesgos, transformando la inteligencia espacial, la vigilancia y la preparación (Hay et al., 2013).

Entonces, las búsquedas en la web relacionadas con salud, datos de microblogs, datos recolectados de múltiples fuentes y las cuentas personales son oportunidades para monitorear en tiempo real la salud de la población, y el volumen, la velocidad y variedad de estas fuentes de

información incrementará rápidamente y transformará la habilidad de crear puntos de referencia geográficos para un rango de enfermedades (Hay et al., 2013). Por ejemplo, en Colombia a partir del mapeo con la predicción del vector aedes aegypti en el área rural de dos municipios de Cundinamarca, se destaca la importancia del componente geográfico para la prevención y control de enfermedades, dado que complementa el análisis de la dinámica de la enfermedad en tiempo y espacio (Angulo, Esteban, Urbano, Hincapié, & Núñez, 2017).

Por otro lado, la epidemiología del paisaje o eco-epidemiología, se preocupa por analizar la interacción entre las características del paisaje con la enfermedad y el riesgo de enfermedad (Robertson, 2017). La información eco-epidemiológica es de gran importancia y debería ser incluida en los mapas de riesgos para dar recomendaciones sobre la priorización de intervenciones, así como también para identificar áreas de riesgo por visitantes o personas que retornan de otros lugares (Rodriguez-Morales et al., 2017).

Robertson (2017), propone un marco de trabajo para epidemiología panorámica geocomputacional, conformado por la caracterización del ensamblaje que corresponde a patrones espaciales como: patrones de composición (ejm: porcentaje de bosques, ríos y montañas, diversidad de especies, entre otros) y patrones de configuración (ejm: longitud de la frontera forestal, tasa de área-perímetro, panorama mixto), seguidamente se destacan las funciones de caracterización de los flujos (patrones de actividad movimiento de vectores, y de animales), los modelos matemáticos, herramientas de análisis, después se realiza el mapeo para evaluar los resultados, analizando los riesgos, posibles brotes y la sensibilidad del análisis.

3.3 Factores asociados a eventos de salud pública

El análisis de la mala salud puede dar conocimiento de oportunidades y prioridades para la prevención, investigación, políticas y desarrollo (Forouzanfar et al., 2016), teniendo en cuenta que la prevención puede mejorar significativamente la salud humana, y tiene como objetivo central la reducción o modificación de los factores de riesgo que influyen en la salud de la población; así, tanto la cuantificación de estos riesgos como el monitoreo actualizado de la evidencia basada en factores de riesgo son claves para la salud pública y fundamentales para modificar el riesgo individual a través de la atención primaria y la autogestión (Abajobir et al., 2017). Cabe resaltar que un factor de riesgo es un atributo, una característica, o exposición de un individuo que lleva a incrementar la probabilidad de desarrollar una enfermedad (World Health Organization, 2017).

El estudio sobre el Impacto Global de las Enfermedades del año 2016 (GBD por sus siglas en inglés), menciona que el factor de riesgo comportamental explica el 32,7% de los DALYs, seguido por los factores metabólicos que explican el 16,8%, y los riesgos ambientales y ocupacionales el 13,1% (Abajobir et al., 2017).

Dado que las tendencias actuales son involucrar la geografía en estudios ambientales y de salud humana, existe la necesidad de desarrollar teorías, herramientas y bases de datos para apoyar la investigación que relaciona el ambiente y la salud (Robertson, 2017). Se resalta que el calentamiento global y el clima han presentado anormalidades significativas en el corto plazo (Wu, Lee, Fu, & Hung, 2008). Los cambios a nivel ambiental se dan por el deterioro del ecosistema, la pérdida de la biodiversidad, agotamiento del ozono estratosférico y el cambio climático (Semenza

& Menne, 2009).

3.3.1 Enfermedades infecciosas. Según Robertson (2017), el impacto global de enfermedades se ha ido desplazando desde las enfermedades transmisibles a no transmisibles y lesiones, aunque las enfermedades infecciosas continúan produciendo persistente y episódicas amenazas a la salud pública. Así, la vigilancia y el control de enfermedades transmisibles es fundamental para el estatus de salud de una comunidad (O'Connell et al., 2010).

Las epidemias de enfermedades infecciosas entre humanos y otros animales se debe a la transmisión de un patógeno ya sea directamente entre huéspedes (hosts) o indirectamente a través del ambiente o huéspedes intermediarios, de ahí que la eficiencia de la transmisión depende del nivel de infección del huésped infectado y la susceptibilidad de los individuos no infectados y que están expuestos a la infección.

Según Grassly & Fraser (2008), la infecciosidad comprende tres componentes, el biológico que está relacionado con la carga viral o bacteriana del patógeno infeccioso; comportamental que se relaciona con los patrones de contacto de un individuo infectado y los patrones de contacto del huésped o vector intermediario, este comportamiento depende de la enfermedad y de la ruta de transmisión; y el componente ambiental depende de la ubicación y ambiente del individuo infectado, dado que el ambiente es importante para que el patógeno sobreviva fuera del huésped así como para que los huéspedes o vectores intermediarios sobrevivan. De manera que las variaciones climáticas en temperatura y precipitaciones conducen a patrones estacionales de la incidencia de enfermedades para muchas infecciones.

El estudio de la estacionalidad de las enfermedades infecciosas trae beneficios como: mejor entendimiento de la biología y ecología del patógeno y el huésped, sistemas de vigilancia más precisos, mejoramiento en la capacidad de predicción de epidemias y pandemias, y la importancia de entender el efecto del cambio climático sobre las enfermedades infecciosas. Los mecanismos que afectan la estacionalidad son el comportamiento de la población, las interacciones patógeno – patógeno, efectos ambientales sobre los patógenos y efectos ambientales sobre el huésped. Un evento climático que influye en la ocurrencia de enfermedades infecciosas transmisibles es la Oscilación Sureña de El Niño (ENSO, por sus siglas en inglés) (Fisman, 2007).

Además, la susceptibilidad del individuo también tiene componentes biológicos, comportamentales y ambientales, por ejemplo la susceptibilidad del individuo podría depender de la memoria inmune, patrones de contacto y ubicación, así que el contacto entre un patógeno infeccioso y la susceptibilidad del individuo llevan a una probabilidad de transmisión que es función de la infecciosidad y la susceptibilidad (Grassly & Fraser, 2008)

En el caso de la vigilancia basada en mosquitos que consiste en la recolección sistemática de muestras de mosquitos y la prueba de reservas de mosquitos para detectar arbovirus, se busca evaluar el estado de la transmisión y apoyar la toma de decisiones. La transmisión de enfermedades arbovirales depende de las interacciones entre virus, mosquitos y huéspedes (hosts), y estas interacciones son afectadas por factores abióticos como temperatura, precipitaciones, y humedad, y factores bióticos como abundancia de huéspedes vertebrados y mosquitos vectoriales (Gu, Unnasch, Katholi, Lampman, & Novak, 2008). Por ejemplo, el ZIKA y otras enfermedades

arbovirales podrían estar asociadas con múltiples factores incluyendo cambios en las condiciones ambientales y sociales que posiblemente influyan en su caracterización epidemiológica (Rodriguez-Morales et al., 2017).

El dengue, se considera que es una amenaza endémica para las regiones más tropicales, (Wu, Lee, Fu, & Hung, 2008), con el más alto riesgo en las Américas y Asia (Bhatt et al., 2013). Este virus es transmitido por el mosquito Aedes y estudios han mostrado que el tamaño de la población del mosquito depende de la presencia de lugares adecuados y de las condiciones climáticas (Wu et al., 2008). Por ejemplo, en México se evalúa el impacto de las variables meteorológicas y los indicadores climáticos sobre la incidencia del dengue en dos municipalidades de dos estados de Veracruz desde 1995 a 2003 (Hurtado-Díaz, Riojas-Rodríguez, Rothenberg, Gomez-Dantés, & Cifuentes, 2007)

En un estudio en el cual se analizan 335 enfermedades infecciosas entre 1940 y 2004, se concluye que las enfermedades infecciosas emergentes (EIDs, por sus siglas en inglés), se correlacionan significativamente con los factores socio-económicos, ambientales y ecológicos, además se destaca que la mayoría de patógenos involucrados en los eventos infecciosos son bacterias (54,3%), la mayoría de eventos EID (60,3%) son causados por patógenos zoonoticos originados en especies de vida salvaje y de no vida salvaje, por patógenos resistente a los medicamentos y patógenos que nacen de vectores (Jones et al., 2008).

Según, Jones et al., (2008), la aparición de enfermedades infecciosas es producto en gran medida de cambios antropogénicos y demográficos, y es un "costo" oculto del desarrollo

económico humano, así que, los esfuerzos para conservar áreas ricas en diversidad de vida silvestre mediante la reducción de la actividad antropogénica pueden tener un valor agregado en la reducción de la probabilidad de aparición futura de enfermedades zoonóticas. Así mismo, destaca la reubicación de recursos para "vigilancia inteligente" de puntos calientes o hotspots de enfermedades emergentes en latitudes bajas.

Lopez et al., (2014), destaca que la integración de datos de salud, clima y ambiente apoyados por los sistemas de información geográfica e imágenes satelitales, junto con herramientas computacionales facilitan el diseño y desarrollo de sistemas de alerta temprana para influenza epidémica. Teniendo en cuenta que la influenza es un problema de salud pública a nivel mundial, ya que se presenta de forma estacional, zoonotica o pandémica, e influye en el aumento de frecuencia de consultas médicas, de admisiones al hospital, de muertes, y de ausencia en las escuelas y lugares de trabajo (Koppeschaar et al., 2017).

Por ejemplo, en el sistema de vigilancia Influenzanet se analiza la información registrada voluntariamente por los participantes en relación con datos sociodemográficos, médicos, preguntas de comportamiento, código postal de residencia y lugar de trabajo. Como factores de riesgo para adquirir una infección de influenza se destaca principalmente el tener contacto directo o indirecto con una persona infectada (Koppeschaar et al., 2017); además, tener una enfermedad crónica como asma, diabetes, enfermedad del corazón y/o una condición inmune comprometida, vivir con niños, ser mujer, pertenecer a un grupo de edad joven, tener mascotas en casa (gatos y/ perros) y ser fumador (van Noort et al., 2015).

En el caos de la Tuberculosis, por ejemplo, en China se realiza un estudio en el cual se concluye que la población migrante contribuye a la prevalencia de Tuberculosis en los 18 distritos de Beijing y recomiendan incluir en próximos estudios el analizar la correlación entre la Tuberculosis y la economía (Jia et al., 2008).

3.3.2 Enfermedades crónicas o enfermedades no transmisibles (ENT). El término "enfermedades no transmisibles" se utiliza para hacer la distinción entre estas afecciones y las enfermedades infecciosas o "transmisibles". Se caracterizan porque las epidemias tardan décadas en establecerse por completo, dado que tienen su origen a edades tempranas; requieren un enfoque sistemático a largo plazo para el tratamiento; y, dada su larga duración, existen múltiples oportunidades para la prevención (WHO, 2017). Las principales enfermedades no transmisibles que contribuyen a la carga de muerte y morbilidad por enfermedades no transmisibles son: enfermedades cardiovasculares, cáncer, enfermedades respiratorias crónicas y diabetes (WHO, 2017).

Chen et al (2016), proponen la cadena de factores para las causas de las enfermedades crónicas (Figura 2), y manifiestan que las causas principales de las enfermedades crónicas incluyen tres tipos de factores: factores no cambiantes, factores cambiantes y factores duros para ser cambiados. La edad y la herencia son factores no cambiantes y explican el 20% de las causas de las enfermedades crónicas, así como otros factores que impactan en las causas de las enfermedades crónicas son los ambientales, tendencias económicas y sociales a nivel mundial, que incluyen: el envejecimiento de la población, la razón directa del aumento del número de pacientes con estas enfermedades, la urbanización que empeora la polución ambiental, y la globalización ha llevado a que nuevas tecnologías, móviles y redes sociales favorezcan la vida urbana, pero han motivado los

hábitos no saludables.

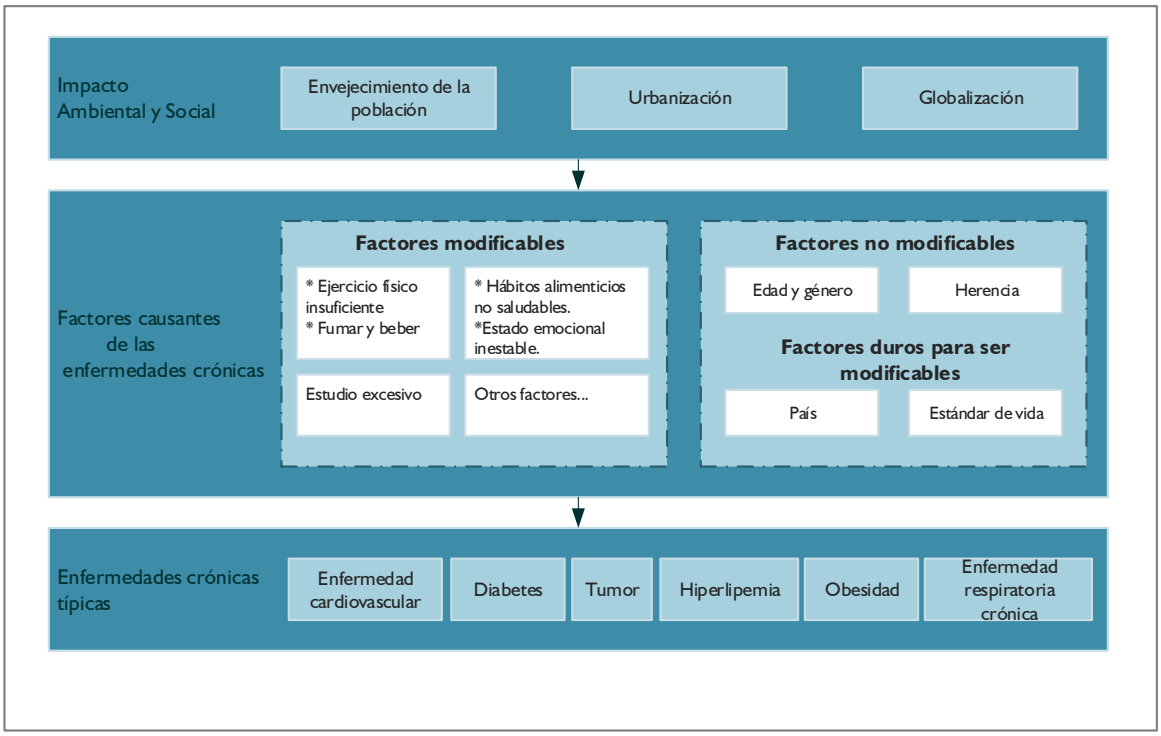

Figura 2. Cadena de factores para las causas de enfermedades crónicas.Adaptado de Chen, M., Ma, Y., Song, J., Lai, C.-F., & Hu, B. (2016). Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring. *Mobile Networks and Applications*, *21*(5), 825–845. https://doi.org/10.1007/s11036-016-0745-1

Muchas de las causas de muertes por enfermedades crónicas están relacionadas con los estilos de vida, ya que el 50 % de los problemas médicos depende del estilo de vida, tales como hábitos de vida, estilos de alimentación y ejercicio; mientras los cirujanos y las instalaciones médicas, pueden resolver solo el 10% de los problemas médicos, y el 20% restante se encuentra en aspectos biológicos como la herencia (Chen et al., 2016).

Las enfermedades no transmisibles tienen presencia a nivel mundial pero el mayor impacto se presenta en las poblaciones pobres y vulnerables; según un informe de la Organización Mundial

de la Salud del año 2014, estas enfermedades son la principal causa de muerte a nivel mundial, por ejemplo, en el año 2012 se presentaron 38 millones (68%) de los 56 millones de defunciones, más del 40% fueron muertes prematuras ocurridas antes de los 70 años de edad, y casi las tres cuartas partes de defunciones y la mayoría de los fallecimientos prematuros (82%) se produjeron en países de ingresos bajos y medios (Organizacion Mundial de la Salud, 2014).

La organización Mundial de la Salud, (OMS) propone nueves metas para el 2025, relacionadas con las enfermedades no transmisibles con el fin de lograr los objetivos de desarrollo sostenible, enfocándose a reducir el impacto de factores de riesgo que influyen en las enfermedades crónicas, dentro de los factores de riesgo resalta el consumo de alimentos grasosos, el bajo consumo de frutas y verduras, el sobrepeso, la obesidad, el uso nocivo del alcohol, la inactividad física, el estrés psicológico y los determinantes socioeconómicos.

Además, la OMS plantea una guía para los países que opten por aplicar una encuesta de factores de riesgo de enfermedades no transmisibles (ENT) siguiendo el enfoque STEPwise de la OMS para la vigilancia, lo cual les permite: planificar y preparar el alcance de la encuesta, la muestra y el entorno; capacitar al personal; realizar el trabajo de campo; capturar y analizar los datos recopilados, e informar y difundir los resultados. Así, según el Observatorio de Salud Pública de Santander, en un estudio realizado en el año 2015 basándose en este método STEPwise, descubren que los factores de riesgo biológicos como el sobrepeso y la obesidad, la hiperglicemia, entre otros factores; y comportamentales como el bajo consumo de fruta y verduras, la inactividad física y el consumo de tabaco y alcohol, influyen significativamente en el desarrollo de enfermedades crónicas de la población santandereana.

En la Tabla 1 se muestra el consolidado de los factores de riesgo para cada uno de los grupos de enfermedades infeccionas y crónicas.

Tabla 1.
Factores de riesgo para enfermedades infecciosas y/o crónicas

Enfermedad	Factores	
Enfermedades infecciosas	• Comportamiento del vector • Ambientales • Temperaturas • Precipitaciones • Humedad • Huéspedes vertebrados • Cantidad de vectores • Componente biológico del patógeno • Ambiente del individuo infectado • Comportamiento de la población • Patrones similares entre enfermedades infecciosas • Oscilación Sureña de El Niño (ENSO)	• Susceptibilidad del individuo • Factores socioeconómicos: densidad de la población, uso de medicamentos antibióticos, prácticas agrícolas, nivel de pobreza • Habitad del vector • Cambios antropogénicos • Latitud, longitud y altitud del área de estudio • Temperatura de la superficie de la tierra • La elevación del nivel del mar • El aumento de la temperatura de la superficie del mar
Enfermedades crónicas	• La edad • La herencia • Falta de actividad física • Fumar • Consumir alcohol • Hábitos alimenticios: bajo consumo de frutas y verduras, el consumo excesivo de sal • Estado emocional • Estudio excesivo	• Ambientales y sociales: envejecimiento de la población, urbanización y globalización • Lugar de residencia • Estándar de vida • Ocupación • Nivel de pobreza • Obesidad • Sobrepeso • Presión arterial elevada • Alta glucosa en la sangre • Lípidos sanguíneos anormales

3.4 Métodos de monitoreo de enfermedades

Los métodos de monitoreo de la propagación de enfermedades se pueden clasificar en técnicas de vigilancia espacial, temporal y espaciotemporal. La vigilancia temporal implica monitorear la ocurrencia de eventos en una región a lo largo del tiempo, con el fin de detectar cambios en el tiempo respecto a la frecuencia de ocurrencia de los eventos. Por otro lado, la vigilancia espaciotemporal está relacionada con el monitorio de la ocurrencia de eventos en varias regiones,

para detectar la frecuencia de ocurrencia con los cambios de tiempo y ubicación (Tsui et al., 2011). Así, los retos en la vigilancia espaciotemporal son estimar la cobertura y la magnitud del brote (Tsui et al., 2011).

En la planeación internacional y en la vigilancia de la salud pública se le ha dado poca atención a la epidemiología espacial y se presenta una brecha entre la información de eventos y el mapeo de la distribución global de enfermedades infecciosas. La transmisión de enfermedades infecciosas está relacionado con conceptos de proximidad espacial y espacio-temporal, y en el caso de enfermedades no transmisibles es importante la proximidad a los factores de riesgo (Pfeiffer et al., 2008).

Varios estudios de epidemiología espacial se interesan por identificar los factores que contribuyen a la distribución del riesgo espacial, así se destacan el uso de modelos tipo regresión o modelamiento jerárquico Bayesiano (Robertson, 2017). Generalmente los métodos utilizados han sido los métodos de regresión basados en árboles, aunque en recientes años se han ido reemplazando por métodos de aprendizaje automático como bosque aleatorio y árboles de regresión aumentados, dado que estos métodos se ven menos afectados por los valores perdidos, la no linealidad, la autocorrelación, la falta de independencia y los supuestos de distribución de los métodos paramétricos (Pfeiffer & Stevens, 2015)

Adicionalmente, se resalta que el desarrollo de los sistemas de información geográfica, ha proveído a las autoridades de salud pública un nuevo conjunto de herramientas para monitorear y responder a los retos de la salud, debido a que estos sistemas permiten localizar casos y

exposiciones a tendencias espaciales, identificación de clusters de enfermedades, correlación entre diferentes conjuntos espaciales y pruebas de hipótesis, así que con estos sistemas se realiza la visualización y mapeo de los datos (Carroll et al., 2014).

3.5 Modelos matemáticos para la vigilancia epidemiológica

La integración de modelos matemáticos con métodos estadísticos rigurosos para estimar parámetros importantes y probar hipótesis utilizando los datos disponibles ha llevado a darle importancia a la epidemiologia matemática. Estos modelos deberían ser tan simples como sea posible pero no tan simples que lleven a alterar las conclusiones al considerar la complejidad innecesaria o la sobre simplificación (Grassly & Fraser, 2008).

La regresión y las series de tiempo son métodos reconocidos por el modelamiento de patrones de referencia, la meta consiste en obtener el modelo matemático que permita describir la relación entre las variables respuesta y los factores, adicionalmente, los métodos de pronóstico y la regresión no paramétrica han proporcionado herramientas efectivas para el modelamiento de patrones de referencia y capturar patrones estacionales complejos (Tsui et al., 2011).

Los sistemas de vigilancia de salud pública son diseñados para facilitar la detección de comportamiento anormal de enfermedades infecciosas y otros eventos adversos a la salud, los modelos de series de tiempo buscan predecir el comportamiento epidemiológico modelando datos históricos de vigilancia. Zhang et al (2014), comparan cuatro series temporales para nueve tipos de enfermedades infecciosas: brucelosis, gonorrea, síndrome renal de fiebre hemorrágica, hepatitis A, hepatitis B, fiebre escarlata, esquistosomiasis, sífilis y fiebre tifoidea. Además, se tienen en

cuenta factores como precipitaciones, temperatura, tiempo de vacaciones, etc. Las cuatro series comparadas son: dos métodos de descomposición como: suavizamiento exponencial y regresión; modelo ARIMA (Promedio móvil autorregresivo integrado), y máquinas de soporte vectorial.

Cabe resaltar que las máquinas de soporte vectorial podrían conducir a un mejor rendimiento en aplicaciones prácticas dado que emplean el principio de minimización del riesgo estructural, que cuenta con mayor capacidad de generalización y es superior al principio empírico de minimización del riesgo que es utilizado por las redes neuronales tradicionales. Las medidas de evaluación utilizada son el error medio absoluto (MAE), el error promedio medio absoluto (MAPE) y el error medio cuadrático (RMSE).

El pronóstico de las series de tiempo es una herramienta con muchas aplicaciones a la vigilancia. Los modelos lineales generalizados (GLMs) son populares herramientas para el monitoreo de enfermedades. Los métodos de regresión adaptativa tienen la capacidad de tomar el efecto de las vacaciones para un pronóstico más preciso. Es importante tener presente que, para los cambios dinámicos, es esencial estudiar los algoritmos de vigilancia que tengan dos "propiedades conflictivas": ser robustos ante varios patrones en el ajuste de los procesos de referencia y sensibilidad en la detección de cambios en el proceso (como brotes). Como consecuencia del avance de las tecnologías de información en salud, sistemas de información médica y sistemas de recolección de datos sindrómicos y salud pública, hay grandes retos y oportunidades para robustecer los algoritmos de vigilancia en términos de monitorear y tener la trazabilidad de los cambios de patrones (Tsui et al., 2011).

Por ejemplo, para los juegos olímpicos del año 2000 en Sidney, el objetivo de los sistemas de

vigilancia de salud pública fue detectar brotes de enfermedades emergentes o patrones inusuales o heridas que podrían requerir intervenciones antes, durante y después de los juegos. Concluyendo que los sistemas de vigilancia del futuro necesitan ser flexibles para capturar la información inesperada (Jorm, Thackway, Churches, & Hills, 2003)

En relación con los modelos de nicho ecológico, por ejemplo, González et al., (2010), construyen un modelo utilizando el algoritmo de máxima entropía para la distribución de las especies de vectores, con el fin de predecir la distribución geográfica de las especies y la estimación de la población humana proyectada que potencialmente podría estar expuesta a la leishmaniasis en 2020, 2050 y 2080 en Norte América. Se utilizan datos de proyecciones de población y se toma como variables exploratorias los datos topográficos (elevación, pendiente y aspecto) y patrones climáticos como: temperatura y precipitaciones. Por otro lado, Lopez et al., (2014) desarrollan un modelo basado en una regresión ponderada geográficamente para predecir el impacto de la influenza epidemiológica en Vellore, India durante el periodo 2013 – 2014, utilizando la epidemiología espacial y datos ecológicos del evento epidémico H1N1 del 2009-2010. La regresión ponderada geográficamente, es un modelo no-estacionario utilizado para estimar los patrones locales, contrario a los modelos de regresión tradicional que se enfocan en los patrones globales.

Similarmente, Hay et al., (2013), construyen un mapa del alcance definitivo de la enfermedad (presencia, ausencia, e intermedio), este mapa se combina con el mapa de presencias conocidas para generar los puntos de pseudoausencia, Entonces, los datos de presencia y pseudoausencias son utilizados para el análisis, junto con las covarianzas ambientales para predecir el riesgo de la

enfermedad en una escala de bajo a alto riesgo. Las observaciones de ausencia mejoran el modelado de las relaciones entre la ocurrencia de las especies y factores ambientales (Stokland, Halvorsen, & Støa, 2011) (Hay et al., 2013). Generalmente los datos de ausencias no se encuentran disponibles, entonces se utiliza el método de "pseudo-ausencias", que es un intermedio entre los modelos de solo-presencia y de presencia-ausencia (Chefaoui & Lobo, 2008).

Sinka et al. (2012), obtienen un mapa que representa las distribuciones de los vectores dominantes de la malaria, indicando la probabilidad de presencia de los vectores a partir de variables climáticas y ambientales definidas como predictores, que indica dónde el ambiente es adecuado para que exista la especie. Bhatt et al., (2013), analizan la probabilidad de ocurrencia de infección, pero del dengue (riesgo del dengue), para lo cual se construye un modelo de distribución de la enfermedad usando arboles de regresión aumentado, y se define una relación entre la probabilidad de ocurrencia del dengue y la incidencia evidente y no evidente del dengue utilizando un modelo jerárquico Bayesiano.

Los árboles de clasificación y de regresión (CART) son métodos de análisis de datos exploratorios basados en árboles que han sido muy útil para identificar y estimar relaciones jerárquicas complejas en los contextos ecológicos y médicos (W. Hu, O'Leary, Mengersen, & Choy, 2011).

Por ejemplo, un modelo CART Bayesiano es utilizado para modelar el problema de la infección por criptosporidiosis en Queensland, Australia. Los datos tenidos en cuenta son los casos notificados, la fecha de inicio y el lugar de inicio de los casos notificados de infección por criptosporidiosis, la edad y el sexo de los pacientes y la fecha del examen de laboratorio, datos del

clima (temperatura diaria y precipitación diaria) y del índice socioeconómico para las áreas. Con este modelo se busca estimar la distribución espacial del riesgo de la enfermedad y se destaca que facilita la adaptación de interacciones no lineales complejas, como la combinación de variables ambientales y sociológicas para ayudar a explicar los patrones espaciales de una enfermedad. Se compara el modelo Bayesiano CART con el modelo Autoregresivo Condicional espacial Bayesiano (CAR), concluyendo que el modelo CART bayesiano utilizado para la identificación y estimación de la distribución espacial del riesgo de la enfermedad puede ser útil en la vigilancia y evaluación de enfermedades infecciosas y en la toma de decisiones sobre prevención y control (W. Hu et al., 2011).

Por otro lado, (Wu et al., 2008) para la vigilancia del dengue, aplican la transformación de ondículas para el pre-procesamiento de datos, estas salidas de la transformación de ondículas son entradas para el algoritmo genético con el fin de validar y confirmar las características seleccionadas del modelo y generar un conjunto de factores que mejor contribuyen. Seguidamente, de la información del algoritmo genético, se construye un clasificador utilizando máquinas de vectores de soporte, resolviéndolo como un problema de programación cuadrática. Finalmente, se aplican algoritmos de regresión (Regresión lineal, Función de base radial (RBF, por sus siglas en inglés) y máquinas de vectores de regresión (SVR, por sus siglas en inglés) para realizar una predicción del modelo. Concluyendo que el SVR es más estable en términos de pronósticos comparado con la regresión lineal simple y que los factores climáticos son importantes en afectar el brote del dengue en Singapur.

En Noumea (Nueva Caledonia, Archipiélago de Nueva Zelanda) buscan estudiar la dinámica

del dengue a través de la interacción entre el huésped humano, el vector mosquito y los virus que son influenciados por factores climáticos y ambientales, con el fin de proponer un sistema de alerta temprana. Así, utilizan modelos no-lineales multivariados para construir modelos explicativos mediante los cuales se identifican las condiciones sostenibles para la ocurrencia de la epidemia, y modelos predictivos para apoyar a las autoridades de salud pública a anticipar el riesgo de brote del dengue, para lo cual utilizan Máquinas de Soporte Vectorial, tomando como entrada datos meteorológicos para predecir a cuál de las dos clases posibles pertenece (año epidémico o año no epidémico). La selección del modelo más relevante se logra utilizando el Criterio de Información de Akaike, el cual permite no solo recompensar la bondad de ajuste, sino que también incluye una penalización que evita el sobreajuste. Además, se analiza la correlación entre años epidémicos y no epidémicos utilizando el test de correlación de Sperman. (Descloux et al., 2012)

En Colombia, se construyen modelos de regresión Poisson no-lineal simple y múltiple para determinar la asociación entre la media de los índices entomológicos (Índice Aedic, Índice Breteau, Índice de recipientes) del 2013 – 2015 y las tasas de incidencia de la infección del virus ZIKA en las áreas urbana y rural de Pereira, Risaralda. En este modelo se tiene en cuenta información entomológica, ambiental y densidad humana, los patrones de viaje y los datos de casos de infecciones transmitidas por vectores, como el ZIKA, concluyendo que este modelo conduce a una herramienta valiosa que se puede utilizar para detectar puntos críticos también para infecciones como dengue, chikungunya y malaria, entre otros (Rodriguez-Morales et al., 2017). En otro estudio se evalúa mediante una regresión Poisson no-lineal la asociación entre las variaciones climáticas y los casos de dengue en el departamento de Risaralda, Colombia, teniendo en cuenta datos climáticos basados en un índice macroclimático global como el Índice del Niño Oceánico (ONI,

por sus siglas en inglés), y variables microclimáticas como precipitaciones o pluviometría (MM) (Quintero-Herrera et al., 2015).

Angulo et al. (2017), destacan que el análisis espacio-temporal de las enfermedades transmitidas por vectores son un complemento a los métodos utilizados para la predicción de áreas de riesgo de presencia de vectores, y se complementa aún más al adicionar variables climáticas, geográficas, sociales y epidemiológicas.

Otros modelos utilizados para la predicción de brotes de dengue son las reglas de asociación difusas teniendo en cuenta variables como datos de incidencia, datos climáticos/meteorológicos (precipitaciones, temperatura del día, temperatura de la noche, Índice de vegetación de diferencia normalizada (NDVI, por sus siglas en inglés), anomalías de la temperatura de la superficie del mar (SSTA, por sus siglas en inglés), Índice de oscilación del sur (SOI, por sus siglas en inglés)), datos socioeconómicos (estabilidad política, sanidad, agua, y electricidad), datos de altitud y demográficos. El SOI está basado en la diferencia Tahití (Polinesia Francesa) la cual influye en la fuerza de los vientos predominantes del este, estos datos proporcionan una medida del efecto del clima de la Oscilación del Sur de El Niño (Buczak, Koshute, Babin, Feighner, & Lewis, 2012).

La minería de reglas de asociación difusas se han aplicado también con el fin de crear modelos predictores para la malaria, extrayendo relaciones entre datos epidemiológicos, meteorológicos, climáticos y socioeconómicos de Corea del Sur, utilizando métricas de desempeño como: valores predictivos positivos (proporción de predicciones positivas que son brotes - PPV), valor predictivo negativo (proporción de predicciones negativas que no son brotes - NPV), sensibilidad o

probabilidad de detección (proporción de brotes predichos correctamente), la especificidad (proporción de predicciones correctas que no son brotes), y la prueba F que considera PPV y sensibilidad. Adicionalmente, se compara el desempeño del método de reglas de asociación con clasificadores como árboles de decisión, bosque aleatorio y máquinas de soporte vectorial, y también aplica el suavizamiento exponencial de Holt-Winters. Las técnicas de machine learning utilizan múltiples variables, el método exponencial suavizado solamente utiliza el número de casos de la enfermedad. Por lo tanto se destaca el desempeño del algoritmo de reglas de asociación para predecir incidencia de la enfermedad alta/media/baja debida a los mosquitos no solo para la malaria sino para otras enfermedades que nacen de mosquitos (Buczak et al., 2015).

Otra forma de aplicar los modelos matemáticos a la vigilancia de eventos relacionados con salud pública, se encuentran los creados a partir de la información registrada en las consultadas realizadas en internet, por ejemplo Ginsberg et al (2009), realizan un estudio sobre la influenza utilizando como variable explicativa la probabilidad de que una búsqueda en internet realizada desde la misma región en la que se hace la visita al médico esté relacionada con la Influenza.

En Pakistán proponen un sistema de apoyo de decisiones analíticas visuales para la vigilancia de enfermedades infecciosas, el cual provee un automatizado mecanismo para la adquisición de datos de enfermedades infecciosas, detección de brotes y gestión de respuestas ante epidemias. Así, para categorización de los síndromes de las enfermedades utilizan redes neuronales y lógica difusa, y la predicción de las enfermedades se realiza utilizando regresión de soporte vectorial (SVR, por sus siglas en inglés).

Otra de las técnicas aplicadas son las redes bayesianas, las cuales son sistemas probabilísticos que han tenido acogida en aplicaciones biomédicas para diagnosticar enfermedades y estudiar redes celulares complejas, entre muchas otras aplicaciones (Berchialla et al., 2012). Por ejemplo, en un estudio en particular realizado en Europa, se construye una red bayesiana automática con el fin de analizar los datos relativos a las lesiones debidas a la inhalación, ingestión y aspiración de cuerpos extraños en los niños, teniendo en cuenta variables tanto del objeto extraño (volumen, tipo de objeto extraño, forma y consistencia) como del niño (edad y sexo), que son comparadas con el riesgo de sufrir una hospitalización. Como fuente de datos se toma los datos recolectados de un estudio de lesiones debidas a objetos extraños en niños de 0 a 14 años de edad de 19 hospitales de Europa; este modelo se le realiza un análisis de sensibilidad para caracterizar la incertidumbre asociada. Se destaca que la ventaja de utilizar estas técnicas es debido a la capacidad de integrar el modelo con la interpretación gráfica (Berchialla et al., 2012).

Finkelstein & Jeong, (2017), utilizan 7001 informes de pacientes con asma recolectados utilizando un sistema de telegestión domiciliaria, en el cual incluía información diaria sobre los síntomas respiratorios, disturbios del sueño debido al asma, limitaciones de actividad física, presencia de resfriado, y uso de medicamentos; estos datos se utilizan con el objetivo de explorar los datos y construir algoritmos de aprendizaje automático que predigan las exacerbaciones del asma. Así, se utiliza una red Bayesiana adaptativa, un clasificador Bayesiano Naive y máquinas de soporte vectorial.

El clasificador bayesiano Naive analiza los datos históricos y calcula la probabilidad condicional para los valores de clase observando la frecuencia de los atributos y combinándolos con otros valores atributo, además se asume que los atributos son independientes. La Red

Bayesiana Adaptativa se basa en redes bayesianas, en donde se utiliza un grafo acíclico dirigido conformado por nodos que representan los atributos y son instancia con probabilidad condicional. En el caso de las máquinas de soporte vectorial, estas son útiles en instancias con límites de clase no lineales y evitan el sobreajuste. Dentro de las medidas utilizadas para evaluar los modelos se encuentran, la exactitud, la sensibilidad y la especificidad (Finkelstein & Jeong, 2017).

En la Tabla 2, se encuentran un resumen de los modelos anteriormente mencionados.

Tabla2.
Método de análisis y actores asociados a eventos

Autor	Evento	Método	Datos	Objetivo	Validación
(Hay et al., 2013)	Dengue	Árbol de regresión aumentado (Boosted Regresion Tree)	• Casos de ocurrencia de las enfermedades obtenidas de la literatura. • GenBank • Bases de datos de población en donde la enfermedad ha sido reportada. • Covarianzas ambientales (Datos de temperatura. Y de precipitaciones)	Predecir la probabilidad de que el dengue se produzca en cada lugar y, generar un mapa de riesgos con una medida de la incertidumbre.	
(Wu et al., 2008)	Dengue	Análisis de ondículas Máquinas de vectores de soporte, Algoritmo genético, Regresión lineal SVR	• Casos de dengue • Máxima, media y mínima temperatura • Máxima, media y mínima humedad • Intensidad de las precipitaciones • Datos de nubosidad	Estudiar la correlación entre los factores climáticos y las tendencias de casos de dengue	Validación cruzada en la aplicación del algoritmo genético. Error medio cuadrado (MSE, por sus siglas en inglés) Coeficiente de determinación (R2)
(Rodriguez-Morales et al., 2017)	ZIKA	Regresión Poisson Múltiple y simple	• Datos entomológicos (Indice Aedico, Índice de depósito e índice Breteau) • Datos ambientales (precipitaciones) • Densidad de población humana • Patrones de viaje • Datos de caso de infectados • Datos económicos de la región	Obtener estimaciones de tasas de incidencia del ZIKA en el municipio de Pereira, la capital del departamento de Risaralda y la ciudad principal de la región del Triángulo del Café de Colombia; y desarrollo de mapas basados en SIG	Significancia estadística.

Tabla 2 (Continuación)

Autor	Evento	Método	• Datos	Objetivo	Validación
(Ginsberg et al., 2009)	Influenza	Modelo lineal utilizando log-odds	• Consultas en internet • Datos de la red de vigilancia centinela de influenza de Estados Unidos.	Estimar la probabilidad de que una visita al médico en una región en particular está relacionada con la Influenza entre 2003 y 2007.	Se prueba el modelo en datos desde el 2007 al 2008, y se calcula la correlación con el porcentaje de datos observados por la red de vigilancia.
(Sinka et al., 2012)	Malaria	Árbol de regresión aumentado (Boosted Regresion Tree)	• Especies de vectores dominantes de malaria identificadas de la literatura. • Bases de datos de vectores. • Variables ambientales • Variables climáticas • Opinión de expertos en mapas	Predecir los mapas de distribución.	
(Bhatt et al., 2013)	Dengue	• Árbol de regresión aumentado (Boosted Regresion Tree) • Modelo jerárquico Bayesiano	• Ocurrencias geolocalizas del periodo 1960 – 2012, de la literatura y fuentes en línea como Healthmap y Promed. • Variables de precipitación • Índice de idoneidad de la temperatura para la transmisión del dengue • Índice de vegetación/humedad • Demarcaciones de las áreas urbanas y periurbanas. • Métrica de accesibilidad urbana • Indicador de pobreza.	Definir la probabilidad de ocurrencia de infección del dengue (riesgo de dengue) dentro de cada uno de los pixeles globales 5 km x 5 km	
(Jones et al., 2008)	Enfermedades infecciosas Emergentes	Regresión logística multivariada	• Patógenos, nombre de patógenos asociados con el evento EID. • Año de primera aparición de la enfermedad. • Tipo de patógeno • Tipo de trasmisión: no-zoonótica o zoonótica • Tipo de Zoonótico • Resistencia a medicamentos • Modo de transmisión • Factores causales: desarrollo económico y uso de tierra, tecnología e industria, • Análisis espacial: densidad de la población, crecimiento de la población, latitud, lluvia (promedio de lluvia por año, mm)	Construir mapa del riesgo de la distribución relativa de un evento emergente de enfermedades infecciosas	
(Lopez et al., 2014)	Influenza	Regresión ponderada geográficamente (GWR, por sus siglas en inglés)	• Precipitaciones • Temperatura • Velocidad del viento • Humedad • Población	Desarrollar un modelo espacial de big data para predecir el impacto epidemiológico de la influenza epidémica en Vellore, India.	

Tabla 2 (Continuación)

Autor	Evento	Método	Datos	Objetivo	Validación
(Buczak et al., 2012)	Dengue	• Reglas de asociación difusas • Regresión logística	• Temperatura • Precipitaciones • Altitud • Demográficos • Estabilidad política • Índice de vegetación de diferencia normalizada (NDVI, por sus siglas en inglés) • Anomalías de la temperatura de la superficie del mar (SSTA, por sus siglas en inglés) • Índice de oscilación del sur (SOI por sus siglas en inglés)	Construir un clasificador para la predicción del dengue	Valores predictivos positivos Valores predictivos negativos Sensibilidad Especificidad
(Quintero-Herrera et al., 2015)	Dengue	• Regresión Poisson no-lineal	• Datos de incidencia del dengue • Variables macro climáticas • Variables micro climáticas • Imágenes satelitales	Evaluar la asociación entre las variables climáticas y los casos de dengue	

4. Marco de trabajo

Este marco de trabajo se construye con base en el proceso de extracción de conocimiento (KDD por sus siglas en inglés), dado que este proceso permite realizar la preparación de los datos para el análisis y consecuentemente extraer patrones y modelos, que mediante un proceso de evaluación e interpretación se convierten en conocimiento.

De acuerdo con las características de Big Data, las organizaciones necesitan procesos eficientes para convertir los datos en conocimiento significativo. Así, el proceso de extraer información de los datos comprende cinco etapas que forman dos subprocesos: gestión y análisis de datos; por una lado, la gestión de datos se relaciona con los procesos y tecnologías que apoyan la adquisición, almacenamiento, preparación y recuperación de los datos para el respectivo análisis; por otro lado,

la analítica se refiere a las técnicas utilizadas para analizar y adquirir el conocimiento de grandes datos (Gandomi & Haider, 2015),

El marco de trabajo se construye con el fin de proponer una guía para las etapas de análisis, interpretación de datos y difusión de la información relacionada con la vigilancia epidemiológica; se consideran las principales características del paradigma big data, se utilizan técnicas de análisis, como lo son la minería de datos y el aprendizaje automático, para aprovechar esta información no solo con fines descriptivos sino también predictivos, es decir, por un lado descubrir patrones y tendencias y por otro construir un modelo predictivo que apoye la toma de decisiones. A continuación se mencionan las fases que conforman el marco de trabajo (*Figura 3*), resaltando que en el Apéndice A, se encuentra la explicación detallada de cada fase, y en el Apéndice B se explica la arquitectura big data que apoya este marco.

4.1 Caracterización del área de estudio

La caracterización del área de estudio está relacionada con información que presente una descripción sobre el área de estudio en cuanto a división política, extensión y características en general, con el fin de entender los aspectos generales del área de estudio y los objetivos de análisis específicos para esta área. Además, se selecciona el evento crítico de salud pública en Santander en el cual se enfocará el proceso de descubrimiento de conocimiento; seguidamente se detalla la epidemiología del evento y su impacto en la población

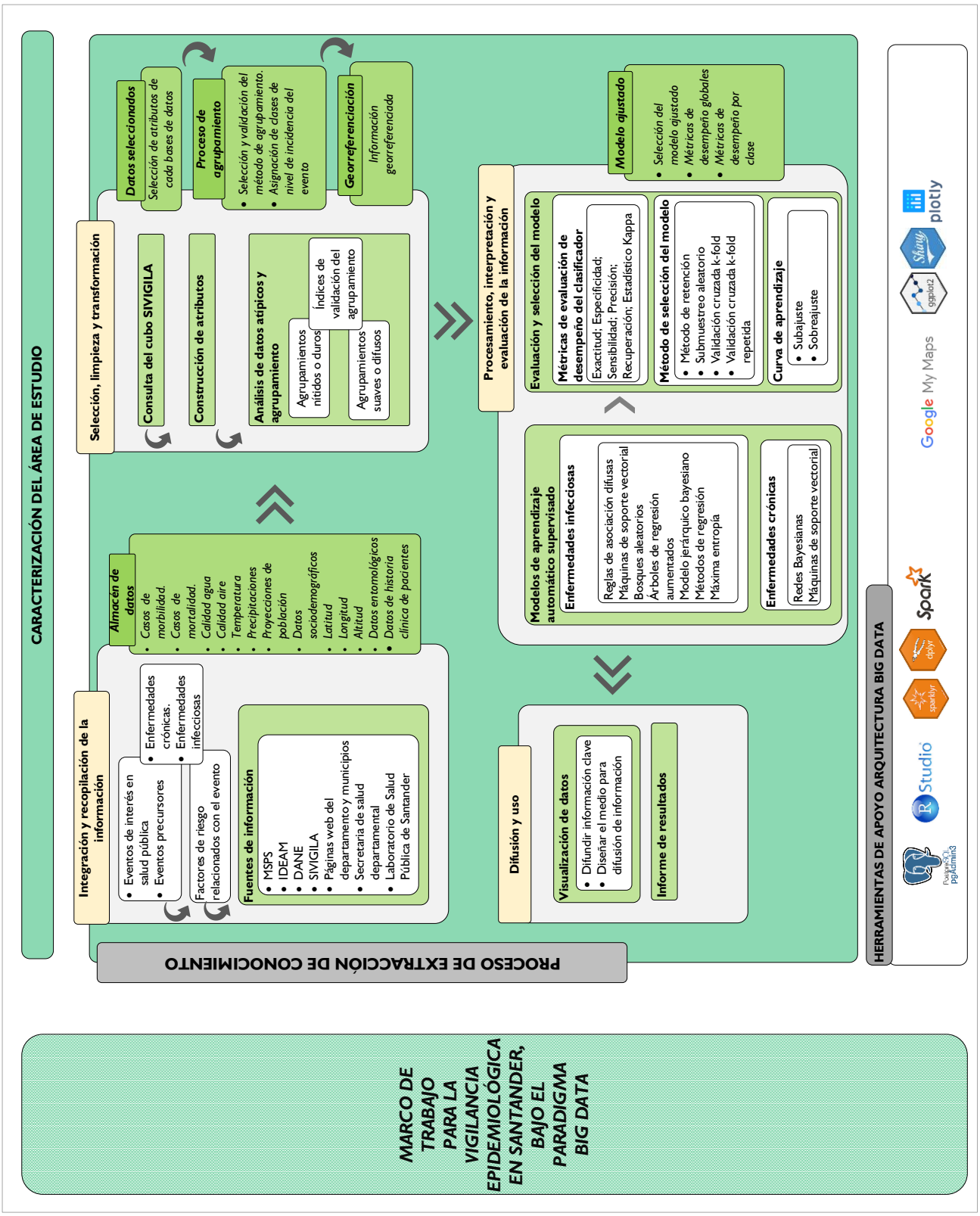

Figura 3. Marco de trabajo para la vigilancia epidemiológica en Santander, Colombia

4.2 Proceso de extracción de conocimiento

Esta fase permite orientar los aspectos claves e importantes a seguir para realizar el proceso de descubrimiento de conocimiento en bases de datos, detallando las siguientes actividades:

4.2.1 Integración y recopilación de la información. En esta fase se definen las variables a considerar teniendo en cuenta los factores mencionados en la Tabla 1, con el fin de identificar posibles fuentes de datos en donde se encuentre la información requerida. Dentro de las entidades que poseen información que se podría vincular al análisis, se destacan: el Ministerio de Salud Y Protección Social, la cuenta de alto costo (CAC), la Secretaria de Salud de Santander, el Centro de Investigaciones en Enfermedades Tropicales de la Universidad Industrial de Santander, el Instituto de Hidrología, Meteorología y estudios ambientales (IDEAM), el Departamento Administrativo Nacional de Estadística (DANE), y las páginas web de municipios y el departamento.

4.2.1.1 Selección, limpieza y transformación. La calidad del conocimiento descubierto no depende solamente de la técnica de análisis utilizada sino también de la calidad de los datos analizados, después de la recopilación de las diferentes bases de datos; en esta fase se realizan las siguientes actividades:

4.2.1.2 Consulta de fuentes. Consiste en la exploración de las fuentes de datos, con el fin de seleccionar las variables a incluir en el análisis, así que, inicialmente se revisa la fuente que proporciona información sobre el número de casos del evento de interés a analizar, después, este número de casos se transforma a una tasa de ocurrencia del evento por cada uno de los municipios del departamento; la interpretación a este cálculo es dada como un indicador de incidencia del

evento, que corresponde al número de casos del evento que se desarrolla en una población durante un período de tiempo determinado.

4.2.1.3 Construcción de atributos. Construir atributos a partir de los atributos originales, es decir realizar la modificación de algunos datos, para facilitar el procesamiento de la información o para involucrar otros atributos de interés en salud pública.

4.2.1.4 Análisis de datos atípicos y agrupamiento. El análisis de clúster o de agrupamiento es una herramienta útil que permite descubrir grupos no conocidos previamente en los datos (Han, Kamber, & Pei, 2012); se supone que en los datos existe un "verdadero" o agrupamiento natural, de manera que al recopilar objetos similares en grupos se intenta reconstruir el grupo desconocido con la esperanza de que cada agrupación encontrada represente un tipo o categoría real de objetos, en donde los puntos de datos dentro un grupo tengan alta similaridad entre sí en comparación con los puntos de datos que pertenecen a otros grupos. Sin embargo, la asignación de objetos a las clases y la descripción de estas clases son desconocidas (Doring, Lesot, & Kruse, 2006; Grekousis & Thomas, 2012).

Así que se aplican los algoritmos de k-medias y k-medias difuso analizando las variables relacionadas con: área de ocurrencia, edad, grupo ocupacional, sexo, tipo de régimen de afiliación a salud, con el fin de descubrir agrupaciones de municipios del departamento de Santander durante el periodo 2007 – 2016, es decir, se busca identificar municipios que tengan características similares para ser etiquetados en un grupo en común y que presenten diferencias con los municipios que conforman los otros grupos. En este proceso de agrupamiento, cada uno de los algoritmos aplicados se evalúa mediante índices de validación que definen la compactación (los

miembros de cada uno de los grupos deberían estar tan cerca a cada uno de los otros miembros como sea posible), y la separación (los grupos deberían estar ampliamente separados).

Después del proceso de agrupamiento, se selecciona el algoritmo con el mejor desempeño expresado mediante los índices, resaltando que al seleccionar el mejor algoritmo también se selecciona la cantidad de grupos a conformar. Seguidamente, de obtener el algoritmo ajustado más adecuado a cada una de las bases de datos consideradas, se calcula el promedio de tasa de incidencia del dengue por cada uno de los grupos, y se asigna una etiqueta de nivel de incidencia del evento objeto de estudio. Las etiquetas consideradas para el nivel de incidencia son: bajo, muy bajo, medio, alto y muy alto.

Finalmente, para visualizar los municipios que pertenecen a cada nivel de incidencia por cada uno de los años, se realiza la georreferenciación de las agrupaciones descubiertas previamente. Este proceso se lleva a cabo con el apoyo de R y Google Maps, de manera que en R se almacena la información de cada grupo en un archivo con extensión klm, para ser importado desde Google Maps y visualizar los resultados en mapas del departamento de Santander.

4.3 Procesamiento, evaluación e interpretación de la información.

El objetivo de esta fase es descubrir nuevo conocimiento para apoyar la toma de decisiones. Se busca obtener un modelo con base en los datos recopilados, que, describa los patrones y tendencias de los datos, permita tener un mejor entendimiento de estos o explicar situaciones pasadas.

De acuerdo con el tipo de datos y con apoyo del software R, y el paquete sparklyr para interactuar con el marco de trabajo de Spark, se obtiene un modelo utilizando técnicas de aprendizaje automático para tareas de clasificación, con el fin de evaluar el riesgo del evento, teniendo en cuenta que este provee el riesgo de que ocurra un evento en condiciones específicas (Corley et al., 2014). Para esto se relacionan los niveles de incidencia del evento y el comportamiento de variables que permitan descubrir patrones para caracterizar a estas etiquetas (niveles de incidencia); así mismo, se identifican aquellas variables importantes que tienen mayor influencia en la generación de la clasificación.

Los algoritmos recomendados para el proceso de clasificación son los mencionados en la Tabla2, y se seleccionan para aplicación por lo menos dos algoritmos que permitan involucrarse en el proceso de análisis; dentro de estos algoritmos se destaca el desempeño de los árboles de decisión, los bosques aleatorios y los métodos de regresión con fines de clasificación. Después del ajuste de cada uno de los modelos utilizando métodos como validación cruzada para identificar sus mejores parámetros y se compara su desempeño utilizando métricas de evaluación como exactitud, precisión, recuperación, sensibilidad y especificidad, con el fin de seleccionar el modelo que se destaque en el desempeño de estas métricas. Además, se construye la curva de aprendizaje de estos modelos para observar el sobreajuste o subajuste.

4.4 Difusión y uso

Con el apoyo de Shiny, el software R, y la plataforma shyniapp.io, se construir una aplicación que muestra los resultados más importantes del estudio. Así, en esta última fase, se documentan los

resultados con el fin de apoyar a las autoridades de salud pública a nivel municipal y departamental.

5. Estudio de caso

De acuerdo con el marco de trabajo propuesto en el capítulo 5 y la arquitectura big data (ver Apéndice B), a continuación, se realiza la aplicación de estas fases a un evento de interés en salud pública del departamento de Santander, Colombia.

5.1 Caracterización del área de estudio

El área de estudio es Santander, uno de los 32 departamentos que conforman la República de Colombia. Según los Lineamientos y Directrices de Ordenamiento Territorial para el Departamento de Santander al 2030 – LDOTSA, adoptados por el Decreto 264 del 20 de agosto de 2014 emitido por la Gobernación de Santander, el departamento tiene una extensión de 30.537 km^2, participando con el 2,7% de la extensión territorial nacional y el 40% de la región nororiental, cuenta con diversos pisos térmicos con alturas entre 100 y 4200 msnm, y las temperaturas promedio oscilan entre los 4 °C y 28 °C.

Santander, está conformado por 87 municipios y se agrupan en seis (6) Provincias Administrativas y de Planificación (PAP), reorganizadas en ocho núcleos de desarrollo provincial (NDP): Área

Metropolitana, Comunera, García Rovira, Guanentá, Mares, Carare Opón y Soto Norte (Tabla 3) (Gobernación de Santander, 2014; Santander, 2015).

Tabla 3.
Núcleos de Desarrollo Provincial en Santander

Provincia	Capital	Municipios
Comunera	El Socorro	Chima, Confines, Contratación, El Guacamayo, Galán, Gámbita, Guadalupe, Guapotá, Hato, Oiba, Palmar, Palmas del Socorro, Simacota, Socorro y Suaita
García Rovira	Málaga	Capitanejo, Carcasí, Cerrito, Concepción, Enciso, Guaca, Macaravita, Málaga, Molagavita, San Andrés, San José de Miranda y San Miguel
Guanentá	San Gil	Aratoca, Barichara, Cabrera, Cepitá, Coromoro, Curití, Charalá, Encino, Jordán, Mogotes, Ocamonte, Onzaga, Páramo, Pinchote, San Gil, San Joaquín, Valle de San José y Villanueva
Mares	Barrancabermeja	Barrancabermeja, Betulia, El Carmen de Chucurí, Puerto Wilches, Sabana de Torres, San Vicente de Chucurí y Zapatoca
Metropolitana	Bucaramanga	Floridablanca, Girón, Piedecuesta, Bucaramanga, Lebrija, Los Santos, Santa Bárbara y Rionegro
Soto Norte	Matanza	Tona, California, Charta, El Playón, Matanza, Suratá, y Vetas
Vélez	Vélez	Aguada, Albania, Barbosa, Bolívar, Chipatá, El Peñón, Florián, Guavatá, Güepsa, Jesús María, La Belleza, La Paz, Puente Nacional, Vélez, San Benito y Sucre
Carare – Opón	Cimitarra	Cimitarra, Landázuri, Santa Helena del Opón y Puerto Parra

Nota: Provincias históricas del departamento. Adaptado de Gobernación de Santander. (2014). Lineamientos y Directrices de Ordenamiento Territorial del Departamento de Santander. Bucaramanga, Santander. Retrieved from http://historico.santander.gov.co/index.php/gobernacion/documentacion/finish/359-lineamientos-y-directrices-del-ordenamiento-territorial/5969-lineamientos-y-directrices-de-ordenamiento-territorial-del-departamento-de-santander

5.1.1.1 Selección del evento de salud pública. A continuación, se detalla el análisis de los eventos de interés en salud pública que se presentan en el departamento de Santander, relacionando el comportamiento a nivel municipal y haciendo comparaciones con otros departamentos del país.

En el departamento de Santander el mayor evento de salud pública ocurrido en el periodo 2007-2016 es el dengue con el 33,26% (88.421) del total de casos reportados en el departamento, seguido por la varicela con el 15,66% (41.627) de los casos reportados; el dengue grave se encuentra en la quinta posición con el 3,93% (10.458) del total de casos de eventos de salud pública de Santander.

A nivel nacional en este mismo periodo (2007 – 2016), en cuanto a casos de dengue, Santander se ubica como el segundo departamento en el cual ocurren la mayoría de los casos (11,55%), después del departamento del Valle del Cauca en el que ocurre el 15,02% (114.957 casos).

En los municipios de Santander el evento del dengue varía entre 2 y 31.255 casos, presentándose la mayor cantidad de casos en la ciudad de Bucaramanga (35,35%) y el mínimo número de casos en los municipios Macaravita y Albania (0,0023%). Con respecto a los casos de dengue grave, el municipio en el cual ocurren la mayoría de los casos (39,83%) es Bucaramanga y el menor número de casos en los municipios: Macaravita, San Andrés, El Peñón, Molagavita, Bolívar, San Miguel, Palmar y Carcasí, cada uno con el 0,01% (1) de los casos. Adicionalmente, en el periodo 2007-2016 el número total de casos de mortalidad por dengue ha sido de 103, presentándose el 33,01% (34) de estos en el municipio de Bucaramanga. En la siguiente tabla se muestran los 10 municipios con mayor presencia de casos de dengue y de dengue grave, respecto al total de casos presentados por cada uno de estos eventos en el departamento de Santander.

Tabla 4.
Casos de dengue y dengue grave en el departamento de Santander

Dengue			Dengue grave		
Municipio	**Número de casos**	**Frecuencia relativa**	**Municipio**	**Número de Casos**	**Frecuencia relativa**
68001 - Bucaramanga	31255	35,35%	68001 - Bucaramanga	4165	39,83%
68276 - Floridablanca	17387	19,66%	68276 - Floridablanca	2022	19,33%
68307 - Girón	8120	9,18%	68307 - Girón	1003	9,59%
68547 - Piedecuesta	6673	7,55%	68547 - Piedecuesta	899	8,60%
68081 - Barrancabermeja	6406	7,24%	68679 - San Gil	694	6,64%
68679 - San Gil	2406	2,72%	68081 - Barrancabermeja	239	2,29%
68755 - Socorro	1971	2,23%	68406 - Lebrija	193	1,85%
68406 - Lebrija	1680	1,90%	68755 - Socorro	165	1,58%
68077 - Barbosa	949	1,07%	68077 - Barbosa	113	1,08%
68689 - San Vicente De Chucurí	829	0,94%	68615 - Rionegro	75	0,72%

Nota. Comportamiento del dengue, dengue grave y mortalidad por dengue, periodo 2007 – 2016. Fuente: Cubo SIVIGILA de la base de datos SISPRO.

En la siguiente gráfica se evidencia que el comportamiento del dengue es variable, presentándose el menor número de casos (2053) en el año 2011, y dos picos de mayor presencia correspondientes a los años 2010 (18417 casos) y 2013 (15892 casos); se observa una tendencia creciente desde el año 2008 hasta el año 2010 y del año 2011 al año 2013, y una tendencia decreciente en la presencia de casos desde el año 2013 al año 2016. En relación con el dengue grave, se observa que disminuye significativamente del año 2010 (23,9%) al año 2011 (1,06%), sin embargo, del año 2011 al año 2014 hay una tendencia creciente, pero para los años 2015 y 2016 se presenta la menor cantidad de casos, 44 y 69 casos, respectivamente.

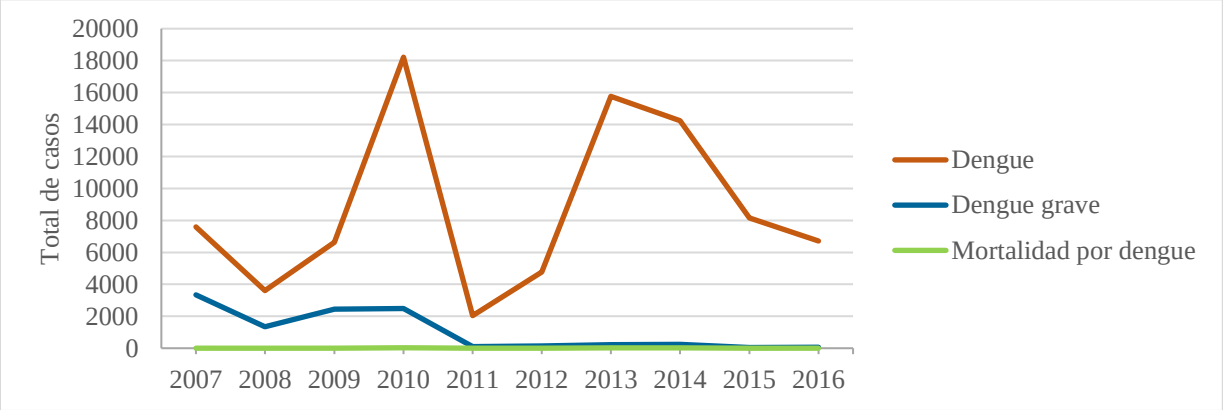

Figura 4. Comportamiento del dengue, dengue grave y mortalidad por dengue, periodo 2007 – 2016. Fuente: Cubo SIVIGILA de la base de datos SISPRO.

Por lo tanto, con la información detallada anteriormente se aplica el marco de trabajo propuesto al análisis del dengue, con el fin de mejorar el conocimiento para los tomadores de decisiones acerca del evento crítico tanto a nivel departamental como nacional.

5.1.1.2 Epidemiología del dengue. El dengue es una infección vírica transmitida por mosquitos hembra principalmente de la especie Aedes Aegypti y, en menor grado, de Aedes Albopictus; estos mosquitos también transmiten la fiebre chikungunya, la fiebre amarilla y la

infección por el virus de Zika. El dengue se presenta en los climas tropicales y subtropicales y sobre todo en zonas urbanas y semiurbanas (OMS, 2017) .

Se conocen cuatro serotipos del arbovirus causante del dengue (DEN-1, DEN-2, DEN-3 y DEN-4) y es la virosis humana más importante transmitida por artrópodos (Organización Panamericana de la Salud, 2015). Cuando una persona se recupera de la infección adquiere inmunidad de por vida contra el serotipo en particular. Sin embargo, la inmunidad cruzada a los otros serotipos es parcial y temporal. Las infecciones posteriores causadas por otros serotipos aumentan el riesgo de padecer el dengue grave (INS, 2017; WHO, 2018).

Tras un periodo de incubación del virus que dura entre 4 y 10 días, un mosquito infectado puede transmitir el agente patógeno durante toda la vida. Las personas infectadas sintomáticas y asintomáticas son los portadores y multiplicadores principales del virus, y los mosquitos se infectan al picarlas. Tras la aparición de los primeros síntomas, las personas infectadas con el virus pueden transmitir la infección (durante 4 o 5 días; 12 días como máximo) a los mosquitos Aedes.

El mosquito Aedes aegypti vive en hábitats urbanos y se reproduce principalmente en recipientes artificiales. A diferencia de otros mosquitos, este se alimenta durante el día; los periodos en que se intensifican las picaduras son el principio de la mañana y el atardecer, antes de que oscurezca. En cada periodo de alimentación, el mosquito hembra pica a muchas personas.

El dengue grave (conocido anteriormente como dengue hemorrágico) fue identificado por vez primera en los años cincuenta del siglo pasado durante una epidemia de la enfermedad en Filipinas y Tailandia. Hoy en día, afecta a la mayor parte de los países de Asia y América Latina y se ha

convertido en una de las causas principales de hospitalización y muerte en los niños y adultos de dichas regiones (OMS, 2018).

5.2 Proceso de extracción de conocimiento

En esta fase se detalla la integración y recopilación de información, la selección, limpieza y transformación de la información relacionada con el dengue en el departamento de Santander, así como el procesamiento y evaluación de esta información y la difusión y el uso de este conocimiento.

5.2.1 Integración y recopilación de la información. A continuación, se detallan las actividades importantes para conformar el repositorio objeto de estudio.

5.2.1.1 Definir variables. De acuerdo con las recomendaciones del marco de trabajo, las variables de estudio se muestran en la siguiente tabla:

Tabla 5.
Variables del caso de estudio

Variable	Fuente	Unidades
Casos de dengue	Cubo SIVIGILA	Número de casos de incidencia
Características asociadas a los casos	Cubo SIVIGILA	Número de casos de incidencia
Información geográfica	Páginas web de los municipios y del departamento	Latitud y longitud de cada municipio
Precipitaciones	IDEAM	Precipitaciones en milímetros
Datos poblacionales	DANE	Habitantes por año a nivel: total, cabecera, área rural dispersa y centro poblado.
Mapa para georreferenciación	DANE	Shapefile con coordenadas geográficas
Altitud	Páginas web de cada municipio	metros sobre el nivel del mar (msnm)

5.2.1.2 Selección, limpieza y transformación. Después de la recopilación de las diferentes bases de datos, se procede a realizar las siguientes actividades:

5.2.1.2.1 Transformación de variables. Para efecto de los análisis, los números de casos de ocurrencia del dengue en el departamento de Santander durante el periodo 2007 - 2016 se transforman a tasas de ocurrencia del evento por 100.000 habitantes, este cálculo se realiza para cada uno de los municipios teniendo en cuenta la estimación y proyección de población nacional, departamental y municipal total por área 1985-2020, elaborada por el DANE.

5.2.1.2.2 Cálculo de índices de validación interna del agrupamiento. Se utilizan dos algoritmos de agrupamiento difuso: algoritmo k-medias difuso y algoritmo Gustafson-Kesel, para la aplicación de agrupamiento difuso, y los algoritmos Hartigan, MacQueen, Lloyd y Floyd para aplicar el agrupamiento K-medias. El procesamiento de los algoritmos se realiza con apoyo del software R, de Spark, así como también se utiliza Google My Maps para visualizar las agrupaciones de municipios. El análisis se efectúa para cada año del periodo 2007-2016.

Para la elección del número adecuado de grupos se evalúan cuatro índices, calculados para cada uno de los algoritmos, y de acuerdo con los criterios de mejor adecuación de estos índices se escoge el número de grupos a conformar por cada año. Estos grupos están comprendidos por los municipios que tienen mayor similaridad en características como: área de ocurrencia, edad en quinquenios, sexo y tipo de régimen de afiliación a salud, variables que se detallan en la Tabla 5.

Al realizar el cálculo de los índices, se comparan los tiempos de procesamiento realizando ejecuciones en R y en Spark; en la Tabla 6 , se observan los tiempos de procesamiento en segundos

obtenidos para cada una de las cuatro veces que se ejecuta el cálculo de los índices en cada una de las herramientas, además se presenta el promedio por cada uno de los índices en cada año, y en la Figura 5 y Figura 6 se observan las comparaciones de estos tiempos. En el cálculo de los índices correspondientes al algoritmo K-medias los tiempos en Spark son mayores que los de R, y similarmente ocurre en el cálculo del algoritmo k-medias difuso, pero en el algoritmo Gustafson Kessel los tiempos en Spark son menores que los tiempos de ejecución en R; cabe resaltar que las iteraciones para este último eran mayores a las de los otros algoritmos, de alrededor de 10^6 iteraciones por defecto que establece el programa R.

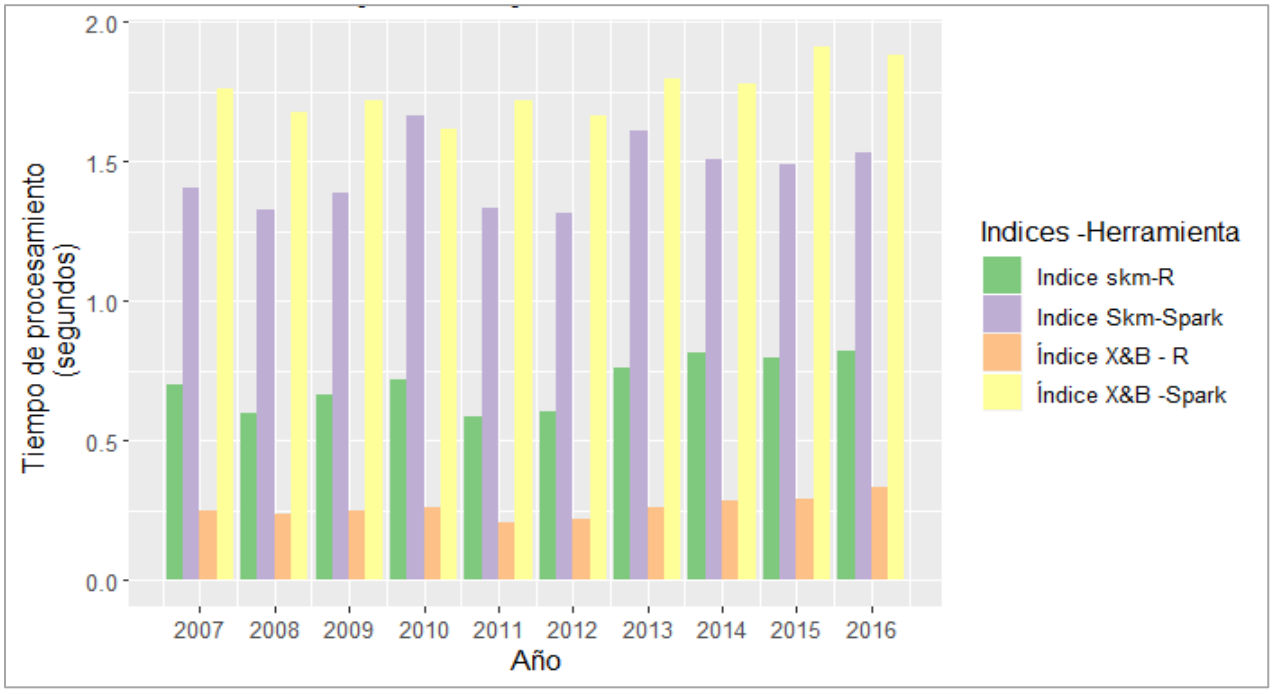

Figura 5. Tiempos de procesamiento agrupamiento nítido

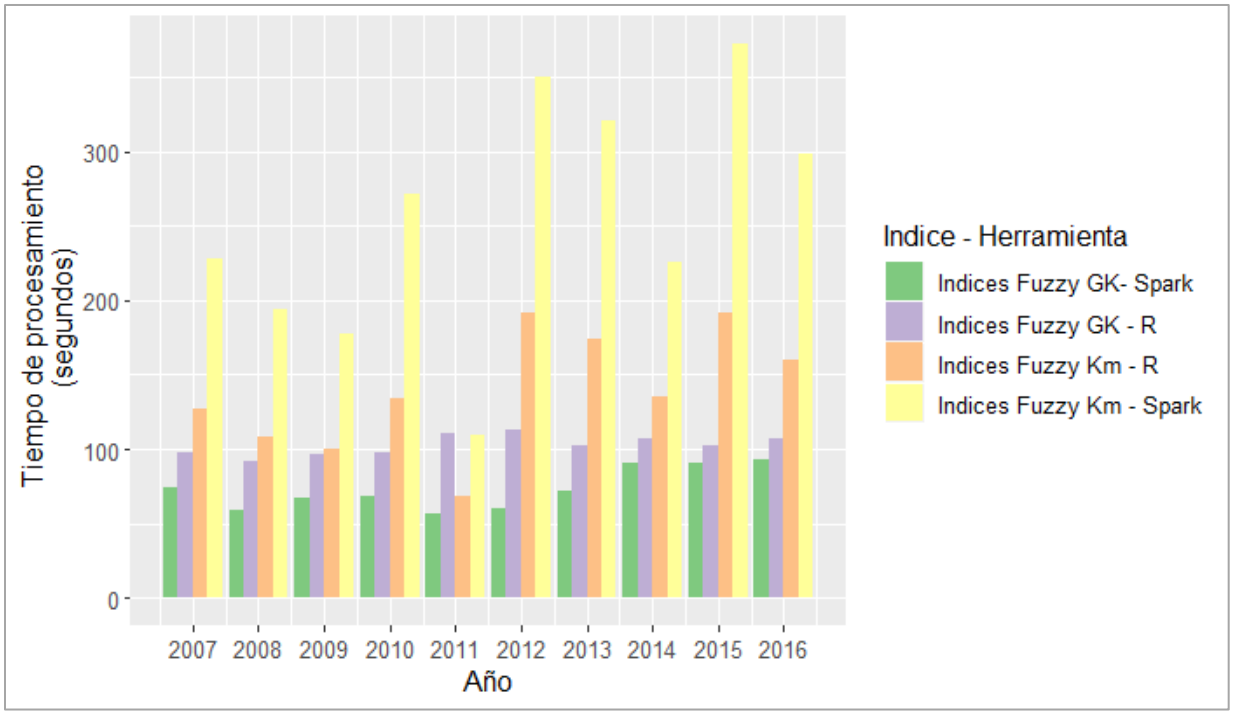

Figura 6. Tiempos de procesamiento agrupamiento difuso

Tabla 6.

Tiempos de procesamiento (segundos)- índices para agrupamiento nítido y difuso

Año	Indice skm		Índice X&B		Fuzzy Km		Fuzzy GK	
	R	Spark	R	Spark	R	Spark	R	Spark
2007	0,707	1,419	0,283	1,771	128,452	226,017	98,400	73,285
	0,704	1,380	0,241	1,740	125,724	224,939	96,169	73,028
	0,714	1,404	0,239	1,766	127,450	233,380	97,914	72,939
	0,667	1,419	0,237	1,766	126,339	230,307	96,911	75,255
Promedio	0,237	3,900	0,004	9,613	126,991	228,661	97,349	73,627
2008	0,625	1,425	0,256	1,719	109,642	197,866	91,515	59,813
	0,596	1,304	0,219	1,684	108,285	194,011	91,602	59,066
	0,614	1,292	0,252	1,662	109,003	193,415	92,053	58,667
	0,544	1,290	0,222	1,649	107,500	191,663	91,931	59,085
Promedio	0,124	3,095	0,003	7,931	108,608	194,239	91,775	59,158
2009	0,654	1,431	0,229	1,751	100,051	180,086	97,306	66,752
	0,652	1,391	0,237	1,700	99,674	175,888	97,608	67,536
	0,668	1,391	0,274	1,709	100,387	176,295	95,864	66,691
	0,670	1,325	0,248	1,722	101,884	176,798	96,216	65,987
Promedio	0,191	3,670	0,004	8,757	100,499	177,267	96,749	66,741
2010	0,865	2,253	0,253	0,948	148,069	276,751	97,397	69,141
	0,679	1,552	0,270	1,794	153,853	269,466	97,505	68,448
	0,659	1,404	0,269	1,829	146,919	269,466	97,416	68,354
	0,673	1,460	0,246	1,882	89,121	270,052	96,971	67,748
Promedio	0,261	7,168	0,005	5,851	134,491	271,434	97,322	68,423

Tabla 6 (Continuación)

Año	Indice skm	Índice X&B	Fuzzy Km	Fuzzy GK	Año	Indice skm	Índice X&B	Fuzzy Km
	R	Spark	R	Spark		R	Spark	R
2011	0,612	1,406	0,204	1,742	69,952	109,480	110,064	57,277
	0,573	1,271	0,216	1,739	68,083	110,478	112,280	56,983
	0,579	1,330	0,201	1,664	68,664	109,795	110,125	55,482
	0,572	1,314	0,206	1,720	68,448	107,743	110,150	55,364
Promedio	0,116	3,122	0,002	8,668	68,787	109,374	110,655	56,276
2012	0,605	1,303	0,206	1,686	187,829	352,496	113,011	59,483
	0,619	1,297	0,224	1,644	192,172	352,455	112,834	58,966
	0,611	1,357	0,227	1,671	191,756	349,912	114,591	61,389
	0,590	1,293	0,212	1,647	192,751	348,671	109,186	58,514
Promedio	0,135	2,965	0,002	7,627	191,127	350,884	112,406	59,588
2013	0,715	2,183	0,288	1,817	173,880	320,751	101,191	71,814
	0,724	1,464	0,251	1,804	171,626	320,181	101,657	71,577
	0,738	1,415	0,253	1,780	174,122	319,474	102,851	71,231
	0,870	1,384	0,255	1,786	174,551	321,351	102,149	71,330
Promedio	0,333	6,259	0,005	10,418	173,545	320,439	101,962	71,488
2014	0,789	1,545	0,297	1,807	133,263	222,075	108,050	91,062
	0,793	1,522	0,274	1,767	133,672	223,041	106,991	90,437
	0,833	1,477	0,267	1,759	136,934	231,057	107,441	90,122
	0,838	1,490	0,288	1,793	137,762	226,012	107,692	90,819
Promedio	0,437	5,175	0,006	10,071	135,408	225,546	107,543	90,610
2015	0,805	1,560	0,288	1,875	193,625	369,586	102,799	91,324
	0,779	1,470	0,288	1,911	189,152	374,853	103,361	89,791
	0,770	1,471	0,310	1,891	192,013	372,147	102,743	88,471
	0,825	1,458	0,280	1,969	194,209	374,445	101,978	91,481
Promedio	0,398	4,921	0,007	13,344	192,250	372,758	102,720	90,267
2016	0,840	1,602	0,340	1,910	157,530	297,192	107,592	93,420
	0,812	1,491	0,436	1,860	160,387	300,307	108,655	93,245
	0,805	1,511	0,282	1,841	159,870	300,637	105,492	91,441
	0,825	1,513	0,283	1,923	161,824	298,532	106,166	93,021
Promedio	0,453	5,463	0,012	12,577	159,902	299,167	106,976	92,782

En resumen, en el Apéndice F, se presentan los resultados de los índices de evaluación, eligiendo para cada uno de los agrupamientos de k-medias y difuso el mejor algoritmo y el número de grupos asociados a valores que tiendan a cero en los índices de entropía (PE) y de Xie-Bie (XB), y valores que tiendan a 1 en los índices de partición (PC) y en el índice silueta (S) . En la Tabla 7, se observan los mejores índices para cada uno de los algoritmos de agrupamiento por cada año, relacionando el número de agrupaciones que se recomienda conformar. Para los años del periodo 2007 – 2016,

excepto para los años 2010 y 2013, el número de agrupaciones a conformar es de 2, en comparación con los 4 y 3 grupos que se deben construir con los datos de los años 2013 y 2010, respectivamente. El mejor algoritmo para cada uno de los años es el k-medias difuso, teniendo en cuenta que presenta mejores valores en los índices, y se evidencia que el algoritmo Gustafson-Kessel tiene el peor desempeño con respecto a los otros algoritmos.

Tabla 7.

Resultados de los mejores algoritmos de agrupamiento

Año	Método	Grupos	Indice_PC	Indice_PE	Indice_XB	Indice_S
2007	K-medias (Hartigan, Forgy, Lloyd, MacQueen)	2	1,0000	0	0,6679	0,6143
	K-medias difuso	**2**	**0,7879**	**0,3337**	**0,3145**	**0,8103**
	Gustafson_kessel	2	0,6036	0,5766	Inf	-0,0135
2008	K-medias (Hartigan, Forgy, Lloyd, MacQueen)	2	1,0000	0	0,6079	0,5819
	K-medias difuso	**2**	**0,8192**	**0,2923**	**0,1938**	**0,8329**
	Gustafson_kessel	2	0,6149	0,5624	Inf	0,0384
2009	K-medias (Forgy; Lloyd; MacQueen)	5	1,0000	0	0,2760	0,6558
	K-medias difuso	**2**	**0,8518**	**0,2464**	**0,1353**	**0,8748**
	Gustafson_kessel	2	0,6149	0,5624	Inf	-0,0013
2010	K-medias (Hartigan, Forgy, Lloyd, MacQueen)	2	1,0000	0	0,4718	1,1693
	K-medias difuso	**3**	**0,6343**	**0,6431**	**0,7560**	**0,7983**
	Gustafson_kessel	2	0,6036	0,5766	Inf	0,0003
2011	K-medias (Hartigan)	10	1,0000	0	0,3680	1,3884
	K-medias difuso	**2**	**0,6927**	**0,4596**	**0,8042**	**0,7115**
	Gustafson_kessel	2	0,6149	0,5624	Inf	-0,0073
2012	K-medias (Forgy; Lloyd; MacQueen)	5	1,0000	0	0,4187	1,9887
	K-medias difuso	**2**	**0,7510**	**0,3871**	**0,3527**	**0,7128**
	Gustafson_kessel	2	0,6149	0,5624	Inf	0,0037
2013	K-medias (Hartigan, Forgy, Lloyd, MacQueen)	2	1,0000	0	1,2724	0,5113
	K-medias difuso	**4**	**0,5696**	**0,8093**	**1,4414**	**0,7835**
	Gustafson_kessel	2	0,6149	0,5624	Inf	-0,0028
2014	K-medias (Forgy; Lloyd; MacQueen)	3	1,0000	0	0,8598	0,4576
	K-medias difuso	**2**	**0,7501**	**0,3914**	**0,3519**	**0,7353**
	Gustafson_kessel	2	0,6149	0,5624	Inf	-0,0355

Tabla 7 (Continuación)

Año	Año	Año	Año	Año	Año	Año
	K-medias (Forgy; Lloyd; MacQueen)	2	1,0000	0	0,6754	0,5970
2015	**K-medias difuso**	**2**	**0,7442**	**0,3951**	**0,4363**	**0,7770**
	Gustafson_kessel	2	0,6149	0,5624	Inf	0,0044
	K-medias (Hartigan-Wong)	2	1,0000	0	1,4725	0,4828
2016	**K-medias difuso**	**2**	**0,7683**	**0,3669**	**0,3228**	**0,7910**
	Gustafson_kessel	2	0,6149	0,5624	Inf	0,0166

Posteriormente, al ejecutar el algoritmo k.medias difuso para cada uno de los años teniendo en cuenta el número de grupos que se recomienda previamente, se analiza el tiempo de procesamiento de este método tanto en R como en Spark, obteniendo los tiempos expuestos en la Tabla 8 y Figura 7. Se observa que el tiempo de ejecución en Spark es menor que el de ejecución en R, sobretodo en el año 2013 se presenta gran diferencia entre estas dos herramientas, así, se resalta la eficiencia de Spark para ejecutar las tareas que conduzcan a la obtención de resultados en el menor tiempo posible en comparación con el procesamiento en R.

Tabla 8.
Tiempo procesamiento k-medias difuso

Año	R	Spark
2007	3,7628	1,5746
2008	3,6099	1,4237
2009	3,7340	3,5841
2010	5,3962	3,4705
2011	3,5171	1,5634
2012	3,4557	1,4206
2013	77,6386	4,9553
2014	3,6039	1,9009
2015	3,8219	1,7851
2016	3,7175	1,5344

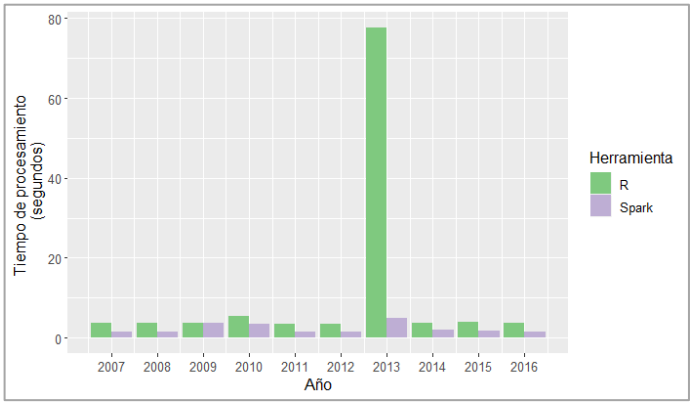

Figura 7. Tiempos de procesamiento agrupamiento k-medias difuso

Después de aplicar el agrupamiento k-medias difuso, en la

Tabla *9* se muestran las agrupaciones obtenidas por cada año y el número de municipios que las conforman. Adicionalmente, en la Tabla 10 se presentan los niveles de incidencia del dengue como: Muy alta, alta, media, baja y muy baja, etiquetas que son establecidas a partir del cálculo de los quintiles de las incidencias promedio anual de dengue por 100.000 habitantes que se presentan en las agrupaciones

En la

Tabla *9,* se muestra que la incidencia promedio de 1391,89 casos y 1226,31 casos de dengue por 100.000 habitantes, son las mayores tasas de incidencia promedio y se dan en los grupos 3 y 1 de los años 2010 y 2013, respectivamente. Además, en estos grupos se encuentran aproximadamente el 15% de los municipios del departamento. En contraste, el menor promedio de incidencia del dengue (29,74 casos por 100.000 hab.) se presenta en el grupo de municipios número 1 del año 2008, que está conformado aproximadamente por el 81% de los municipios del departamento. Se evidencia que, en los niveles de incidencia alta y muy alta, se encuentran grupos conformados por pocos municipios en comparación a los grupos de municipios con niveles de incidencia muy baja.

Tabla 9.
Agrupaciones de municipios por año del periodo 2007-2016

Año	Casos de dengue por 100.000 hab.	Total de casos	Grupo	Número de municipios	% municipios por grupo	Promedio de incidencia anual de dengue por 100.000 hab.	Nivel de incidencia según el periodo 2007-2016
2007	254,94	113	1	17	19,54%	400,32	Baja
			2	70	80,46%	43,58	Muy Baja
2008	236,85	3.664	1	72	82,76%	29,74	Muy Baja

			2	15	17,24%	297,39	Media
2009	285,53	6.616	1	11	12,64%	449,75	Alta
			2	76	87,36%	40,21	Muy baja

Tabla 9 (Continuzación)

Año	Casos de dengue por 100.000 hab.	Total de casos	Grupo	Número de municipios	% municipios por grupo	Promedio de incidencia anual de dengue por 100.000 hab.	Nivel de incidencia según el periodo 2007-2016
2010	434,58	18.417	1	56	64,37%	128,38	Media
			2	17	19,54%	643,97	Alta
			3	14	16,09%	1226,31	Muy alta
2011	114,02	2.053	1	55	63,22%	16,64	Muy baja
			2	32	36,78%	135,97	Media
2012	60,54	4.797	1	64	73,56%	39,2	Muy baja
			2	23	26,44%	322,24	Media
2013	405,8	15.892	1	13	14,94%	1391,89	Muy alta
			2	19	21,84%	413,05	Alta
			3	46	52,87%	70,39	Baja
			4	9	10,34%	958,69	Muy alta
2014	91,99	14.313	1	24	27,59%	753,61	Muy alta
			2	63	72,41%	107,21	Baja
2015	75,89	8.192	1	22	25,29%	672,19	Muy alta
			2	65	74,71%	89,5	Baja
2016	113,29	6.865	1	21	24,14%	637,06	Alta
			2	66	75,86%	101,54	Baja

Tabla 10.
Niveles de incidencia

Quintil	Incidencia promedio de dengue por 100.000 habitantes	Nivel de incidencia
1	0 – 54,304	Muy baja
2	54,304 – 124,146	Baja
3	124,146 – 400,23	Media
4	400,23 – 661,16	Alta
5	661,16 – 1391,89	Muy alta

En la Figura 8, se muestra que el año 2007 presenta agrupaciones de municipios con niveles de incidencia de dengue muy bajos y bajos, en contraste con los años 2008, 2011 y 2012, que además de tener niveles de incidencia muy bajos o bajos también tienen nivel de incidencia en categoría media. Para los años 2009, y del 2014 al 2016 hay niveles de incidencia tanto bajos como altos o

muy altos. En contraste, en los años 2010 y 2013, se presenta el mayor promedio de incidencia de dengue por 100.000 habitantes, y se caracterizan por niveles de incidencia media, alta y muy alta.

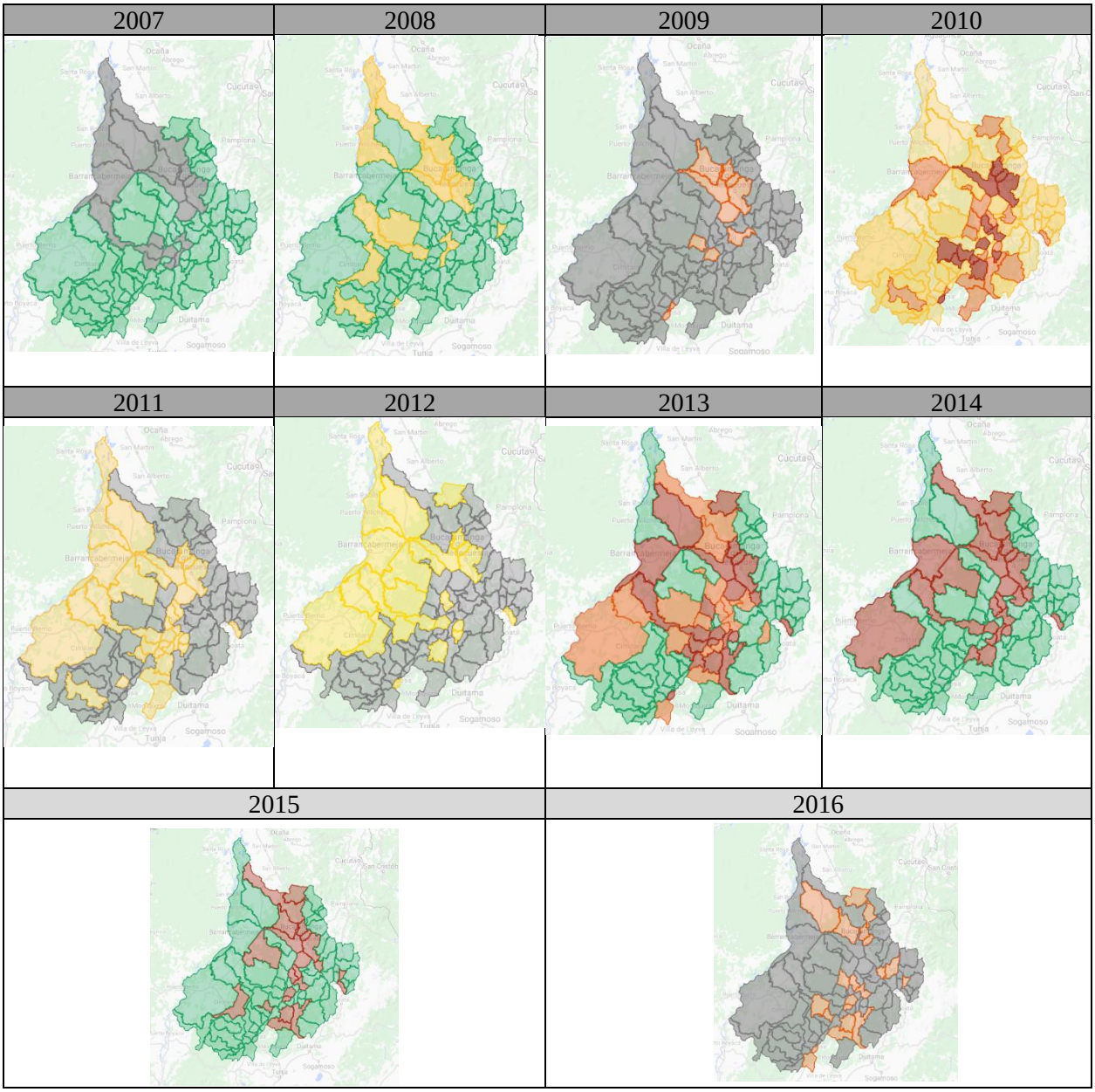

Figura 8. Representación de los niveles de incidencia de dengue en Santander periodo 2007 – 2016[1]

[1] Mapas construidos con apoyo de Google My Maps (2007 -2009,2010-2012, 2013-2015 y 2016)

5.2.1.3 Descripción de los niveles de incidencia del dengue. De acuerdo con los niveles de incidencia y la información del Apéndice G, a continuación, se muestra el perfil que caracteriza a cada uno de los niveles de incidencia.

Tabla 11.
Perfil de los niveles de incidencia del dengue en Santander, periodo 2007- 2016

Nivel de incidencia	Descripción
Muy baja	Nivel de incidencia promedio entre 0 y 54,304 casos por 100.000 habitantes. Presencia de casos de dengue en personas que residen en el área cabecera. La distribución de la población en estos municipios es mayor en el área rural y centro poblado que en el área cabecera. Mayor incidencia en personas del sexo masculino, en personas afiliadas al régimen de salud subsidiado, contributivo, especial y las no afiliadas. Las entidades prestadoras del servicio son en su mayoría públicas. Personas en el ciclo vital de la juventud. En el rango de edad de 5 a 29 años y de 35 a 39 años se presenta la mayor proporción de casos. Agrupa municipios con una altitud promedio de 1655 m.s.n.m.
Baja	El nivel de incidencia promedio de dengue se encuentra entre 54,304 y 124,146 casos por 100.000 habitantes. Las personas afectadas son residentes del área cabecera, aunque también se destaca la presencia de casos en centros poblados. Hay mayor incidencia en los rangos de edad de 0 a 4 años y de 10 a 29 años, así como en el sexo masculino. Personas en las etapas del ciclo vital juventud o adultez. Se destaca la presencia de dengue en mujeres gestantes. Las personas están afiliadas al régimen subsidiado, contributivo o excepción. Las entidades prestadoras del servicio son de carácter público. Algunos casos son confirmados por laboratorio. Se agrupan municipios que se caracterizan por una altitud promedio de 1522,36 m.s.n.m.
Media	Nivel de incidencia promedio entre 124,146 y 400,23 casos por 100.000 habitantes. Personas son residentes del área cabecera. Mayor presencia en el sexo masculino. La distribución de la población es mayor en el área cabecera que en las otras áreas. Se resalta la incidencia de dengue en madres comunitarias. El rango de edad de 5 a 34 años hay mayor presencia de casos. Personas en las etapas del ciclo vital juventud o adultez. Los regímenes de afiliación a salud predominantes son el subsidiado, contributivo y no afiliado. Se destacan las entidades públicas prestadoras del servicio. Se agrupan municipios con una altitud promedio de 1178,54 m.s.n.m.

Tabla 11 (Continuación)

Nivel de incidencia	Descripción
Alta	Nivel de incidencia promedio de dengue se encuentra entre 400,23 – 661,16 casos por 100.000 habitantes. Se destaca la presencia de casos en el área cabecera, en personas del sexo masculino y femenino, y en mujeres gestantes. El régimen de afiliación de las personas es subsidiado o contributivo, y las entidades prestadoras del servicio son públicas. Personas en las etapas del ciclo vital juventud o adultez. Mayor presencia de casos en los rangos de edad de 5 a 24 años y de 30 a 34 años. En cuanto a la distribución de la población, la mayoría se encuentra en el área cabecera, y los municipios que se agrupan se caracterizan con un promedio de altitud de 1276,9 m.s.n.m.
Muy alta	Nivel de incidencia promedio de dengue entre 661,16 y 1391,89 casos por 100.000 habitantes. Presencia de casos en personas que residen en el área cabecera, personas del sexo masculino y femenino, y en mujeres gestantes. En el rango de edad de 0 a 24 años se da la mayoría de los casos. Personas en las etapas del ciclo vital juventud o adultez. Los regímenes se afiliación a salud son el subsidiado o contributivo. Las entidades prestadoras del servicio se caracterizan por ser públicas, privadas o mixtas. Se resalta la confirmación de casos por nexo epidemiológico. En cuanto a la distribución de la población es mayor en el área cabecera que en las otras áreas. La altitud de los municipios se caracteriza por el promedio de 1177,70 m.s.n.m.

5.2.2 Procesamiento, evaluación e interpretación de la información. Para el análisis de aprendizaje automático se consideran 47 variables predictoras relacionadas con el número de casos del evento por periodo epidemiológico, atributos sociodemográficos, características geográficas, precipitaciones, y la variable dependiente corresponde a las etiquetas de niveles de incidencia construidas a partir del proceso previo de agrupamiento.

Los modelos de aprendizaje automático supervisados para clasificación se obtienen con apoyo del software R y las funciones de sparklyr, con el fin de elegir a partir de árboles de decisión, bosques aleatorios y regresión logística, cuál es el más adecuado o que tiene la capacidad de

desempeñarse mejor en las diferentes métricas como exactitud, precisión, recuperación, sensibilidad y especificidad.

Inicialmente para cada uno de los métodos de clasificación se compara la exactitud obtenida entre el modelo en el que se seleccionan todas las variables consideradas y otro en el que no se consideran la latitud ni la longitud de los municipios. Para la optimización de los parámetros se utiliza validación cruzada repetida 10-fold, y para evaluar la capacidad predictiva del modelo se considera un 80% de los datos para el conjunto de entrenamiento y el restante 20% para el conjunto de prueba.

5.2.2.1 Árboles de decisión. De acuerdo con los resultados obtenidos mediante el método CART, en la Figura 9 y Figura 10 se muestran los resultados al considerar la partición gini y la partición información, respectivamente, junto con la variación del parámetro cp. Por tanto, se selecciona el modelo que considera todas las variables con el criterio de partición información teniendo en cuenta que presenta una mayor exactitud con un valor cercano al 55% y un valor de cp de 0,009859155.

5.2.2.2 Bosque aleatorios (random forest). Similar a lo ejecutado en los árboles de decisión, se aplica el bosque aleatorio utilizando la función *randomForest*; similarmente a lo obtenido en los árboles de decisión en la Figura 11 se muestra que el modelo 2, el cual considera todas las variables tiene un mejor desempeño que el modelo 1 en el que no se considera la latitud ni la longitud de los municipios; en este caso se compara la exactitud con el parámetro del número de predictores empleados obteniendo para una exactitud de aproximadamente el 63% un parámetro

óptimo de predictores a evaluar en cada división (mtry) correspondiente a 37.

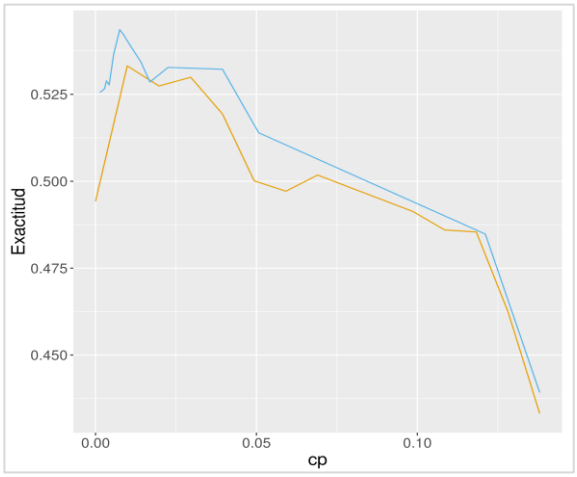

Figura 9. Árbol de decisión con partición gini

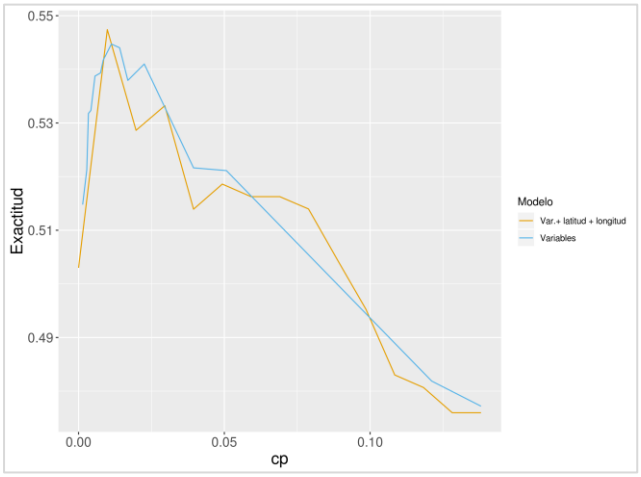

Figura 10. Árbol de decisión con partición información

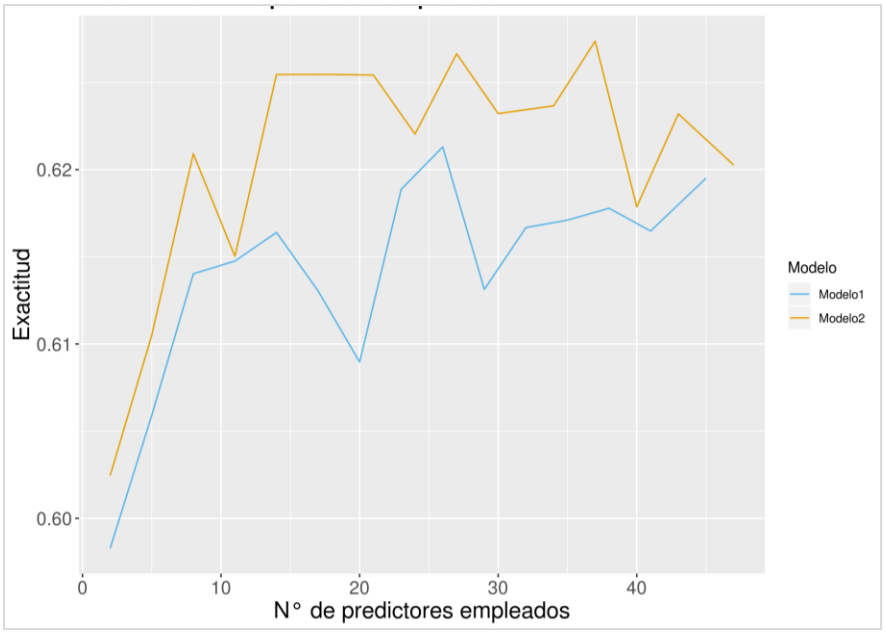

Figura 11. Bosque aleatorio comparando Exactitud vs N° de predictores

Después de seleccionar todas las variables inicialmente consideras, y de considerar un parámetro mtry de 37, se procede a optimizar los otros parámetros, el número de árboles y el número de observaciones en nodos terminales, obteniendo que el número óptimo de observaciones mínimas que deben contener los nodos terminales dado que minimizan el error corresponde a 6 o 21 **(**Figura 12**)**.

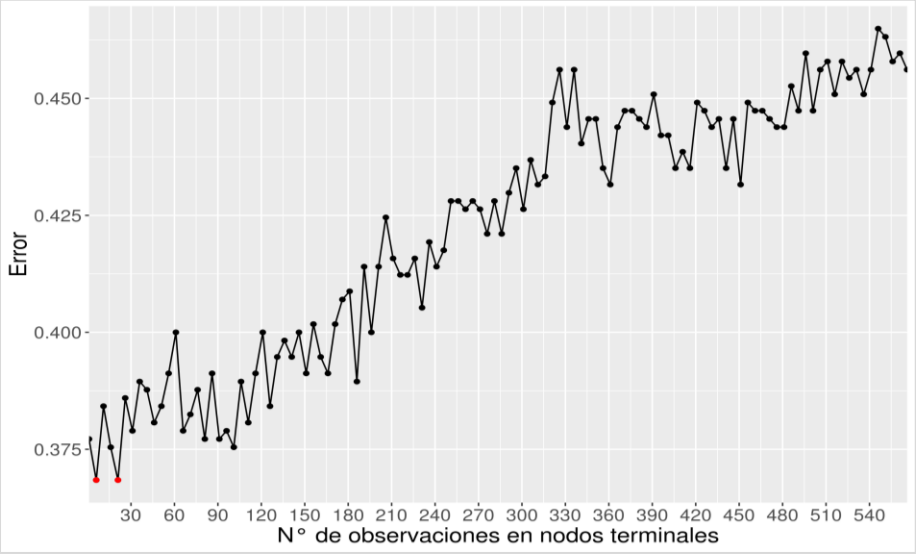

Figura 12. Error vs N° de observaciones en nodos terminales

El ajuste final del bosque aleatorio se realiza calculando el test de error mediante el error de clasificación error-out-of-bag (OOB por sus siglas en inglés), obteniendo así un número óptimo de árboles correspondiente a 82 árboles (Figura 13).

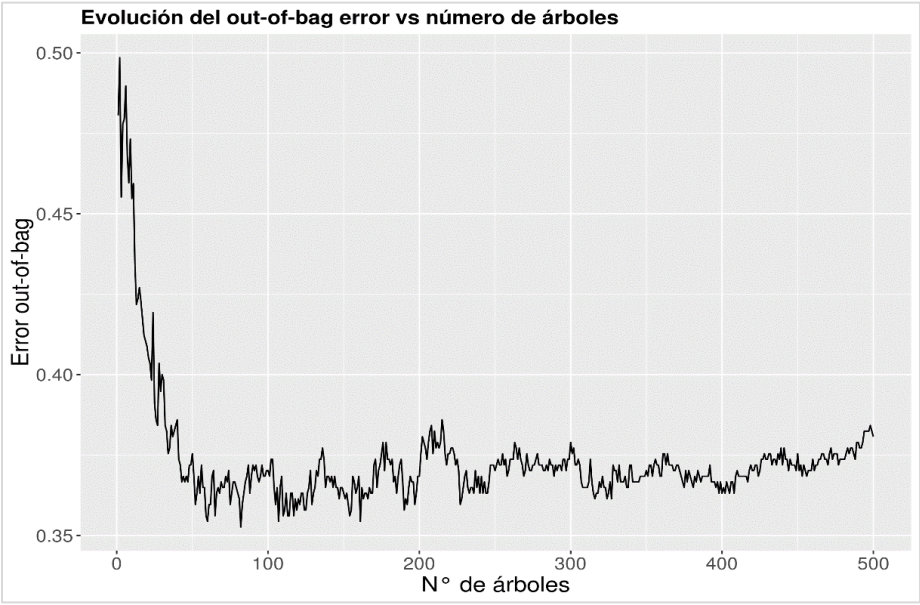

Figura 13. Evolución del error out-of-bag vs número de árboles

5.2.2.3 Regresión logística multinomial. En los resultados obtenidos para la regresión logística multinomial el conjunto de variables sin considerar latitud ni longitud permiten obtener mayor exactitud - como se muestra en el modelo 1 de la Figura 14, en comparación con el modelo 2 el cuál considera todo el conjunto de variables, sin embargo, al comparar los mayores valores de exactitud obtenidos por cada modelo, se encuentra que no hay diferencia significativa entre estos valores, por lo que se selecciona el modelo 1 con el fin de considerar las mismas variables que han sido consideradas en el árbol de regresión y en el bosque aleatorio. Por tanto, el valor óptimo del parámetro decay corresponde a 0.1.

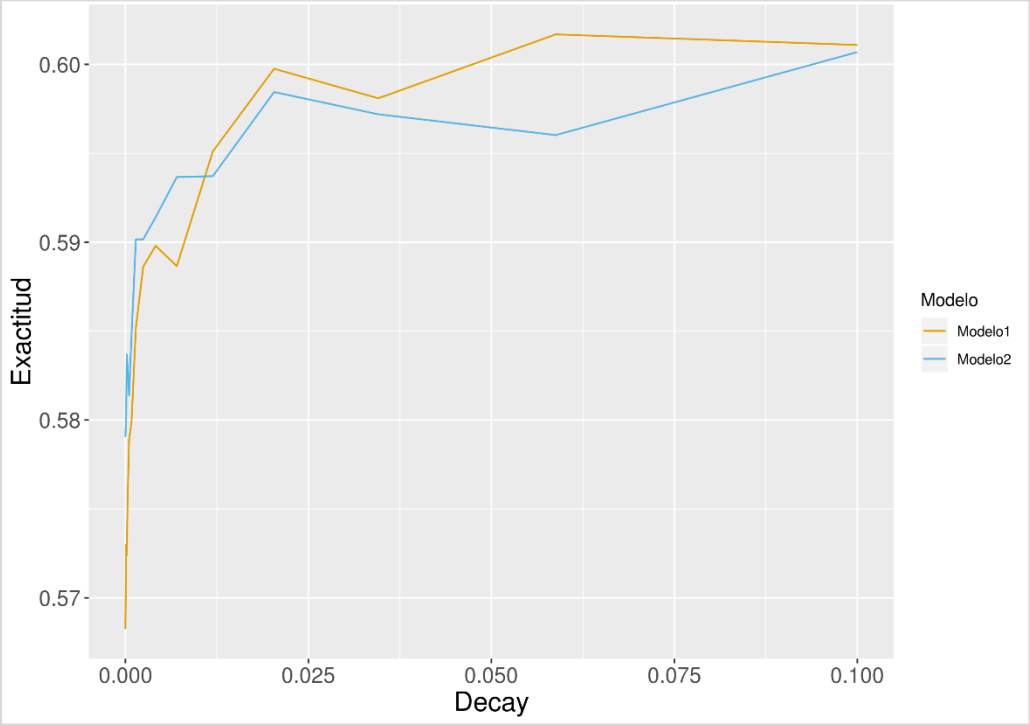

Figura 14. Modelos de regresión logística multinomial - Exactitud vs Decay

5.2.2.4 Comparación de los modelos. De acuerdo con los resultados obtenidos para seleccionar los parámetros adecuados para cada uno de los modelos, en la Tabla 12 se observa que en cuanto a las métricas globales de los modelos a un nivel de confianza del 95% la mayor

exactitud se obtiene en el clasificador de regresión logística multinomial, lo que lleva a deducir que este sería el clasificador más adecuado, aunque sea una valor bajo de exactitud.

Tabla 12.
Resultados de los clasificadores

Algoritmo	Parámetros	Modelo seleccionado	Exactitud	Intervalo de confianza Exactitud (95%)	P-valor
Árboles de decisión	Gini; Entropía Factor de costo de complejidad (cp); minsplit	Minsplit = 5; Cp = 0.009859155 ; Información	57,43	(47,44648, 66,45201)	$4{,}470037 \times 10^{-5}$
Bosque aleatorios	Mtry ntree nodezise	Mtry = 37; ntree = 82; nodesize = 6	58,93	(49,24, 68,14)	$8{,}654 \times 10^{-6}$
Regresión logística múltiple	Decay	Decay = 0,1	59,82	(0.5014, 0.6897)	$3{,}612 \times 10^{-6}$

Sin embargo, se procede a revisar las métricas obtenidas para cada una de las clases de niveles de incidencia analizando métricas como sensibilidad y especificidad, precisión y recuperación, teniendo en cuenta que las clases no están balanceadas como se muestra en la siguiente tabla, ya que al seleccionar un clasificador que clasifique bien las clases de incidencia muy alta o alta, como baja o muy baja, es un hecho que implica altos costos para los planes de salud pública, porque se puede llegar a invertir pocos recursos en planes preventivos en zonas que posiblemente necesiten otras estrategias de mitigación de presencia del evento.

Tabla 13.
Proporciones de presencia de clases en los datos

Clase	Porcentaje (%)
Muy alta	9,42
Alta	7,81
Media	14,48
Baja	29,54
Muy baja	38,73

Por tanto, al analizar la sensibilidad o recuperación de las clases en cada uno de los clasificadores, se muestra que la proporción de observaciones en la clase media son bien clasificadas por la regresión logística multinomial (41,2%), así, de las observaciones de incidencia muy baja el 90,7% son clasificados como pertenecientes a esta clase. Sin embargo, para la clase de incidencia muy alta el mejor clasificador es el árbol de decisión con una sensibilidad o recuperación de aproximadamente 91%, aunque este algoritmo no tiene mejor desempeño en la clasificación de las demás etiquetas en comparación con la regresión logística multinomial. En relación con el bosque aleatorio se tiene un pobre desempeño en clasificar las observaciones con nivel de incidencia media en comparación con los otros algoritmos, pero en cuanto a las clases de incidencia muy baja, muy alta, baja y alta, el modelo ajustado tiene la capacidad de clasificar el 74,4%, 63,6%, 75% y el 57,6%, respectivamente, como pertenecientes a esta clase.

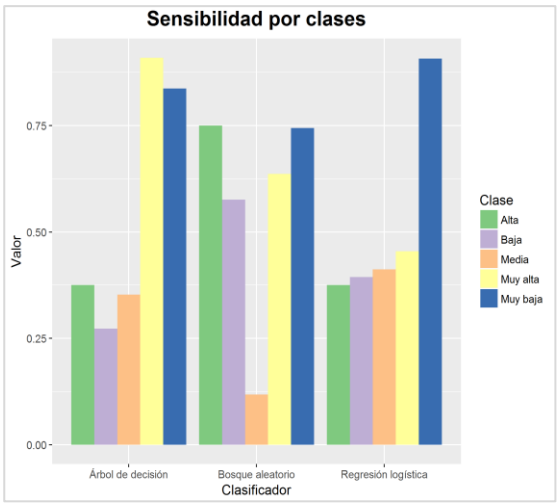

Figura 15. Sensibilidad por cada clase vs los modelos ajustados

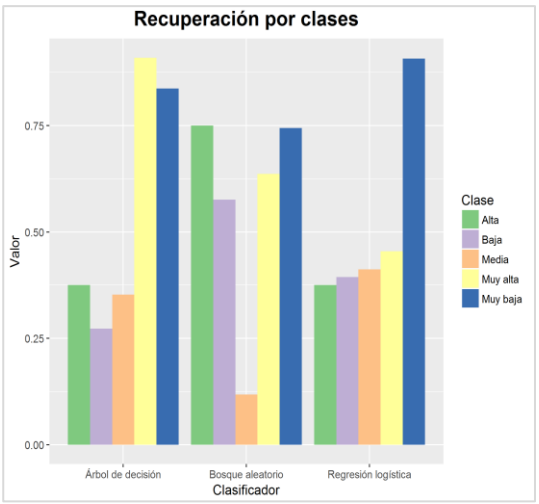

Figura 16. Recuperación por cada clase vs los modelos ajustados

En cuanto a la especificidad se muestra que, aunque todos los algoritmos ajustados tienen una exactitud entre 57%y 60%, los valores de especificidad son mayores al 60%, lo cual significa que

los algoritmos reconocen muy bien los datos que no pertenecen a la clase que ese está analizando. Por ejemplo, si se revisa la incidencia muy baja, el bosque aleatorio clasifica bien el 78,3% de las observaciones no pertenecientes a este nivel de incidencia tales como incidencia baja, media, alta y muy alta. Así, con el bosque aleatorio ajustado se obtienen valores más altos para esta métrica en la clasificación de los niveles de incidencia muy baja, muy alta y media, pero tiene el más bajo desempeño en la clase de nivel de incidencia baja (74,7%) y el árbol de decisión tiene el valor más alto al evaluar las clasificaciones de incidencia alta.

Por otro lado, en cuanto a la precisión de los modelos ajustados se muestra que el bosque aleatorio hay alta precisión en la clasificación de niveles de incidencia muy alta, alta y muy baja, en comparación con los otros algoritmos. Sin embargo, en cuanto a la clasificación de las clases baja y media tienen valores más bajos que los otros algoritmos.

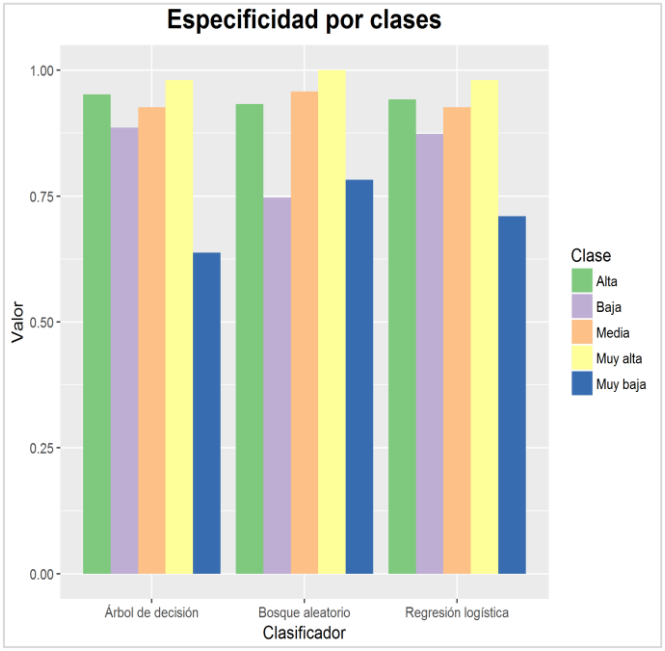

Figura 17. Especificidad por cada clase vs los modelos ajustados

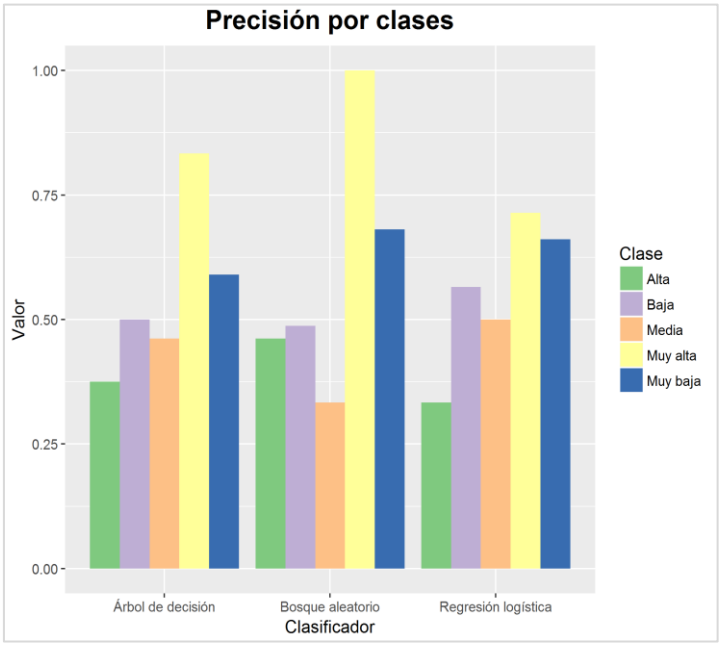

Figura 18. Precisión por cada clase vs los modelos ajustados

Por consiguiente, de acuerdo con los resultados de otras métricas de desempeño diferentes a la exactitud, el modelo ajustado que se podría adoptar es el de bosque aleatorio, dado que se destaca en su desempeño en cuanto a la clasificación de clases de incidencia alta, muy alta, y baja y muy baja, lo que permite considerar que las etiquetas asignadas por este clasificador permiten generar confianza en cuanto a los planes estratégicos a orientar en cuanto a salud pública, porque este clasificador permite identificar las clases de incidencia altas y muy altas que de no clasificarse bien traería como consecuencia el incurrir en más costos a largo plazo porque no se invertiría en planes preventivos hacia situaciones de alto riesgo sino que se implementarían planes preventivos como si fuese una zona de bajo riesgo. Además, en el caso de la clasificación de etiquetas de incidencia baja y muy baja, está apoyaría con alta confianza a las autoridades de salud pública.

Aunque el modelo de bosque aleatorio tiene un desempeño poco destacable en la clasificación

de niveles de incidencia media, al analizar este nivel de incidencia se presenta que tiene alto desempeño en clasificar las observaciones que no pertenezcan a esta clase.

5.3 Curva de aprendizaje

La evaluación del tamaño del conjunto considerado para entrenamiento y para prueba, en las curvas de aprendizaje se observa que a medida que aumenta la experiencia métrica se observa una exactitud entre el 50% y el 65%, y ninguno de los modelos presenta sobreajuste. En el comportamiento del árbol de decisión (Figura 19) se observa que el desempeño disminuye con la experiencia y que es inestable ante la variación del porcentaje de elementos asignados al conjunto de entrenamiento, así que presenta alta varianza en los dos conjuntos de entrenamiento y prueba.

El modelo de bosque aleatorio (Figura 20) ajustado presenta estabilidad en el comportamiento de entrenamiento y muestra un ajuste similar entre el conjunto de entrenamiento y de prueba. Por otro lado, el modelo ajustado de regresión logística multinomial (Figura 21) también muestra un ajuste similar entre el conjunto de entrenamiento y el de prueba.

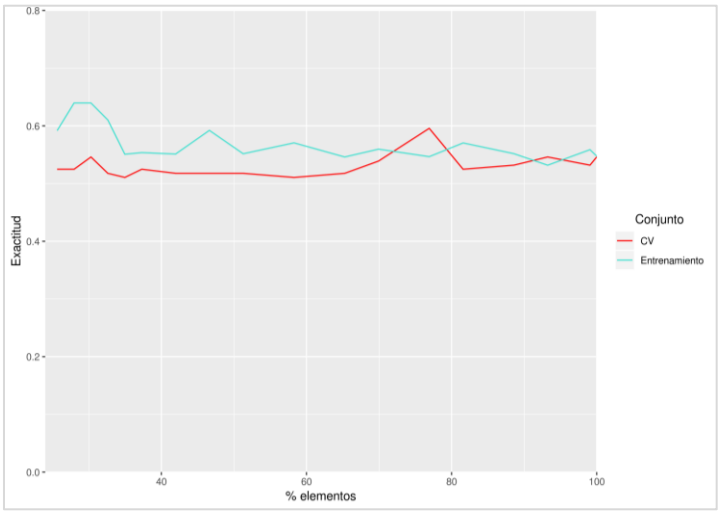

Figura 19. Curva de aprendizaje árbol de decisión CART

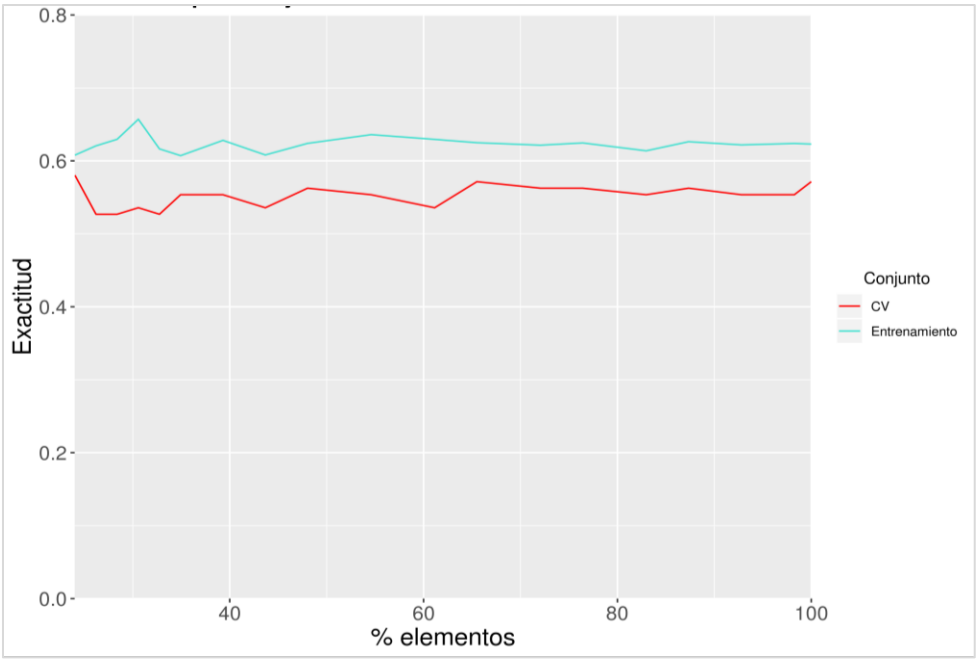

Figura 20. Curva de aprendizaje bosques aleatorios

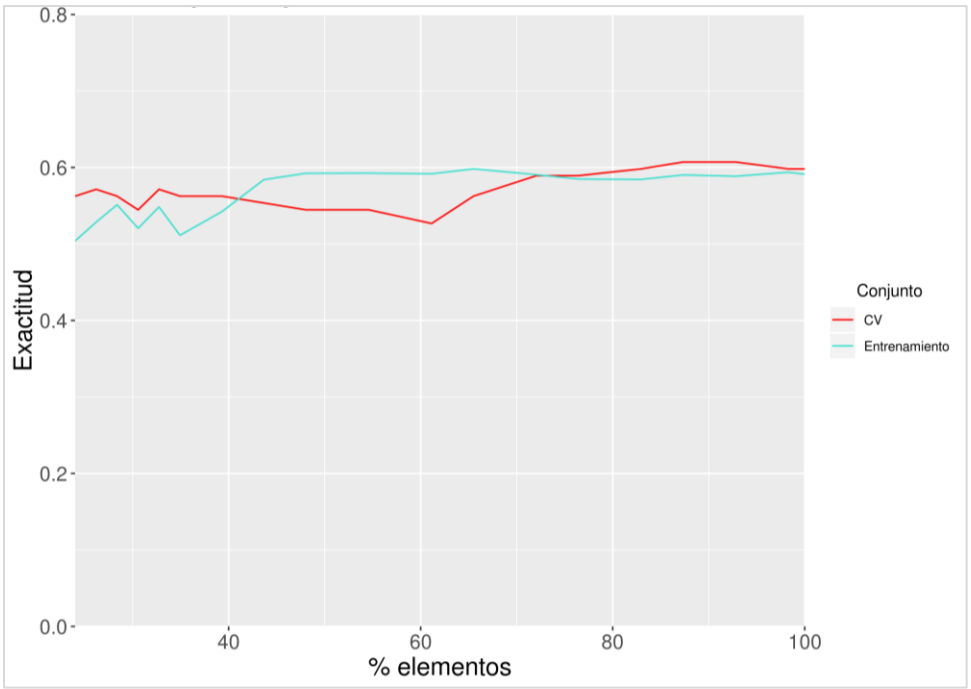

Figura 21. Curva de aprendizaje regresión logística multinomial

5.4 Importancia de las variables para los modelos

En el caso de los árboles de decisión, las variables con mayor influencia son sexo (femenino y masculino), régimen de afiliación subsidiado, los quinquenios de edad 10 a 14, 15 a 19, y de 35 a 39 años, además el área de ocurrencia cabecera presenta influencia destacable en el modelo ajustado de árbol de decisión (Figura 22). El modelo de bosque aleatorio, utilizando el índice de gini, que está relacionado con la disminución total en las impurezas de los nodos al dividirse en la variable en específico, promediando todos los árboles, se muestra que las variables predictores de mayor importancia son el sexo(femenino y masculino), similar al árbol de decisión, el régimen de afiliación contributivo, la precipitación total (en mm), la población del área de cabecera y centro poblado y del área cabecera, así mismo el quinquenio de edad más importante es de 10 a 14 años, el área de ocurrencia cabecera y la altitud (msnm) de los municipios.

El modelo ajustado de regresión logística multinomial muestra que las variables predictoras más importantes son los periodos epidemiológicos 1, 4, 2, 7,3, 13 y 5, así como la población del área cabecera, y los quinquenios de edad de 40 a 44, de 70 a 74 y de 75 a 79, así como el régimen de afiliación excepción (Figura 24).

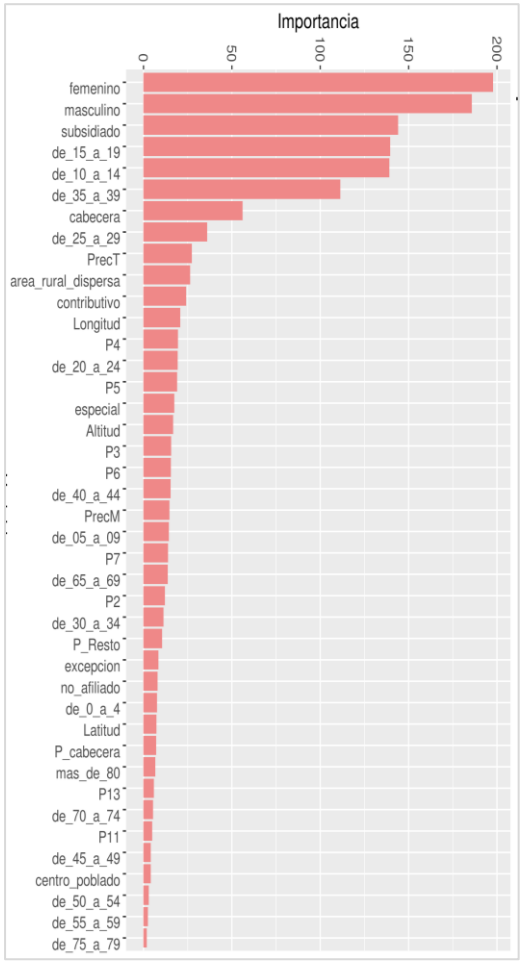

Figura 22. Importancia de variables en árbol de decisión

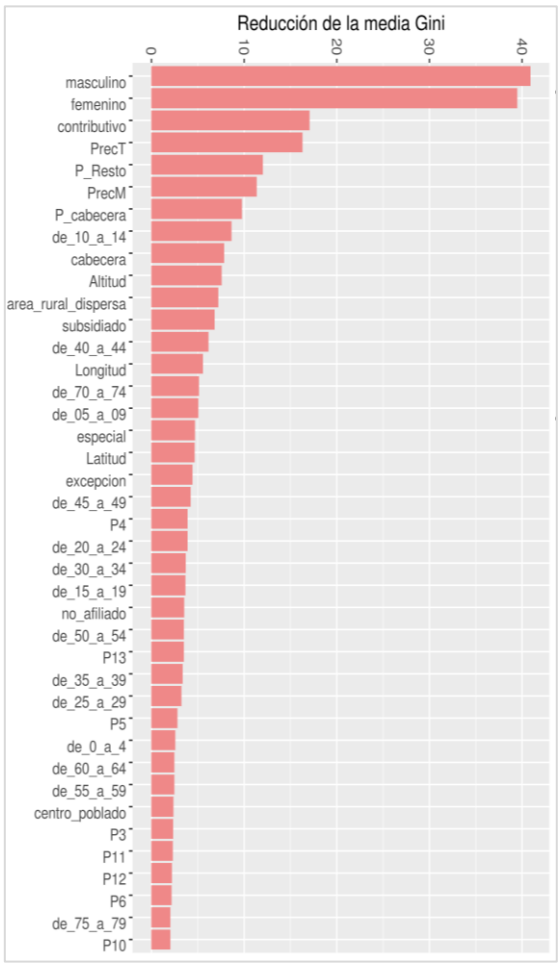

Figura 23. Importancia de variables bosque aleatorio

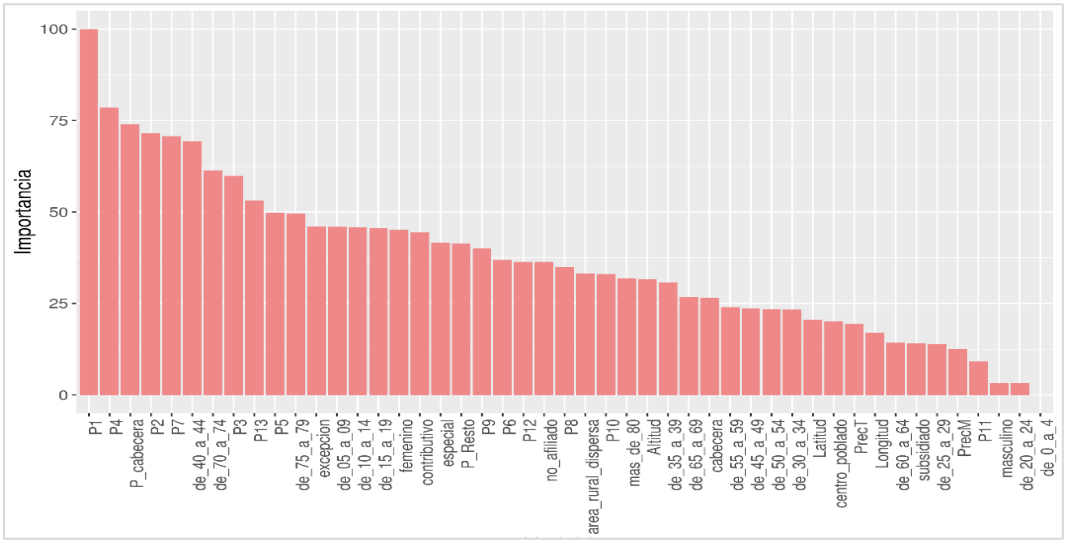

Figura 24. Importancia de variables en regresión logística multinomial

5.4.1 Visualización de la información. El análisis de información objeto de estudio, se utiliza para realizar la visualización de los resultados obtenidos, de manera que con apoyo del marco de trabajo Spark, el software R y shiny, se construye una aplicación compartida en la plataforma shinyapps.io, y se ubica en el enlace: *https://proyectodengue.shinyapps.io/Dengue_en_Santander/*. La aplicación está conformada por las siguientes secciones:

- **Sección "Información general":** en donde se encuentra gráficas y tablas que resumen las tasas de dengue en el periodo 2007 – 2016, en el departamento de Santander.
- **Sección "Nivel de incidencia":** se ubican los resultados del agrupamiento en un mapa en el cual se observan los niveles de incidencia por cada municipio por cada uno de los años del periodo 2007-2016.
- **Sección "Modelamiento":** se encuentra información sobre los resultados de los modelos de aprendizaje automático.

Conclusiones

La integración de herramientas big data de acceso abierto utilizadas en el trabajo, permiten evidenciar que estas son un soporte para la gestión de la información y para configurar un sistema integrado que interactúe y facilite el tratamiento de la información. Además, se evidencia que al comparar los tiempos de procesamiento en R vs los de Spark, estos últimos se destacan por presentar resultados en el menor tiempo.

El marco de trabajo diseñado teniendo en cuenta el paradigma big data y el proceso de descubrimiento de conocimiento, permite ser una guía clara para que los interesados en el tema de vigilancia epidemiológica lo consideren como posible referencia base que apoye las necesidades de analítica de la información.

La revisión de literatura que soporta el trabajo de investigación permite identificar que la mayoría de los estudios tenidos en cuenta han involucrado algoritmos con fines de regresión y no de clasificación, por tanto, el marco de trabajo propuesto permite integrar las diferentes etapas que conducen a aplicar algoritmos de clasificación, con el fin de analizar el nivel de incidencia del evento objeto de estudio.

El SIVIGILA en Colombia, es una base de datos en la cual reposa información con un rezago aproximadamente de dos años, lo cual dificulta que se realicen análisis con datos de lo ocurrido en el presente y así evaluar estrategias de salud pública con información actualizada. Sin embargo, se destaca que cuenta con un protocolo de validez de la información que da confiabilidad para considerarla como una fuente principal de datos, además, se facilita el acceso a todas las personas con iniciativas de investigación o proyectos definidos.

Se destaca el desempeño del algoritmo k-medias difuso en comparación con el algoritmo nítido k-medias, ya que el agrupamiento difuso da flexibilidad de que todas las instancias consideradas no pertenecen totalmente a un grupo, sino que también posee características de otras agrupaciones, así, el algoritmo k-medias difuso como método de minería de datos de agrupamiento, que se utiliza con fines descriptivos en esta investigación, permite ser un apoyo para asignar etiquetas de

clasificación de los niveles de incidencia del dengue en el periodo 2007-2016 de los diferentes municipios que conforman el departamento de Santander.

Las etiquetas de clasificación obtenidas para los niveles de incidencia del dengue a partir de los patrones identificados con variables sociodemográficas de los casos de dengue permiten comparar la evolución de la incidencia de esta enfermedad ETV a través del periodo evaluado (2007 -2016).

El agrupamiento permite realizar una caracterización de los niveles de incidencia del dengue e identificar las variables que más diferencian a cada uno de estos; así, esta clasificación es una fuente de información para la toma de decisiones relacionadas con las posibles acciones que se deben direccionar para mitigar la presencia de este evento, así como la evaluación de los planes de prevención actuales con el fin de contribuir a la salud de la población desde las diferentes iniciativas por parte de las entidades de salud pública.

Al darle prioridad a la clasificación de niveles de incidencia alta y muy alta, teniendo en cuenta que son situaciones críticas para la salud pública, el ajuste del bosque aleatorio presenta valores destacables en cuanto a la sensibilidad, y esto se soporta con los resultados de la especificidad en donde se evidencia que al evaluar la clase media del nivel de incidencia, este método identifica con mayor precisión las observaciones pertenecientes a las otras clases que no sean de incidencia, así, este modelo se considera el más adecuado para el problema objeto de estudio.

Las etiquetas asignadas por este clasificador bosque aleatorio generaría confianza en cuanto a los planes estratégicos a orientar en relación con salud pública, porque este clasificador permite

identificar las clases de incidencia altas y muy altas que de no clasificarse bien traería como consecuencia el incurrir en más costos a largo plazo porque no se invertiría en planes preventivos hacia situaciones de alto riesgo sino que se implementarían planes preventivos en zonas catalogadas erróneamente como de bajo riesgo. Además, en el caso de la clasificación de etiquetas de incidencia baja y muy baja, está apoyaría con alta confianza a las autoridades de salud pública.

El estudio de caso permite evidenciar que a partir del marco de trabajo propuesto y siendo este una guía de apoyo específico para el análisis de datos involucrando métodos de minería de datos y de aprendizaje automático bajo el paradigma big data, es importante integrar varios atributos característicos de los casos de eventos de interés en salud pública, para dar un mejor conocimiento o relacionar de manera integral del evento objeto de estudio considerando posibles factores influyentes sobre este y asociar estos datos con las diferentes zonas en donde se presentan, con el fin de ser apoyo a la toma de decisiones estratégicas de salud pública.

La aplicación mediante la cual se visualizan los resultados obtenidos en el caso de estudio, permite interactuar con diferentes funciones y principales objetivos de la analítica descriptiva y predictiva, dado que es posible explorar el nivel de incidencia del dengue por cada uno de los años y se encuentran asociadas otras variables que caracterizan a la zona, además, se muestra tanto, el perfil de incidencia al que corresponde cada municipio así como los municipios que pertenecen a cada nivel de incidencia; por otro lado, la sección de modelamiento muestra los resultados principales de los algoritmos al variar el conjunto de entrenamiento.

Recomendaciones

Aplicar este análisis por semana epidemiológica para analizar la evolución del nivel de incidencia del evento crítico de interés en salud pública.

Involucrar dentro del análisis las comunas o barrios, y la variable provincias del departamento.

Integrar datos entomológicos para que tanto profesionales de la salud pública como académicos incluyan estos datos en el procesamiento de la información.

Realizar comparación del desempeño entre los modelos ajustados, árboles de decisión, bosques aleatorios, regresión logística multivariada y reglas de asociación difusas.

Proponer un sistema de alerta temprana.

Realizar un análisis relacionado con la predicción del riesgo espacial de la enfermedad, en donde según, Hay et al., (2013), se construye un mapa del alcance definitivo de la enfermedad (presencia, ausencia, e intermedio). Se resalta que los datos de ausencias no se encuentran disponibles, entonces se utiliza el método de “pseudo-ausencias”, que es un intermedio entre los modelos de solo-presencia y de presencia-ausencia (Chefaoui & Lobo, 2008)

Referencias bibliográficas

Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abdulle, A. M., … Zodpey, S. (2017). Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, *390*(10100), 1345–1422. https://doi.org/10.1016/S0140-6736(17)32366-8

Ali, M. A., Ahsan, Z., Amin, M., Latif, S., Ayyaz, A., & Ayyaz, M. N. (2016). ID-Viewer: A visual analytics architecture for infectious diseases surveillance and response management in Pakistan. *Public Health*, *134*, 72–85. https://doi.org/10.1016/j.puhe.2016.01.006

Amato-Gauci, A., & Ammon, A. (2008). The surveillance of communicable diseases in the European Union – a long-term strategy (2008-2013). *Eurosurveillance*, *13*(26), 1–3. https://doi.org/https://doi.org/10.2807/ese.13.26.18912-en

Angulo, V. M., Esteban, L., Urbano, P., Hincapié, E., & Núñez, L. A. (2017). Distribución espacial del mosquito Aedes aegypti (Diptera: Culicidae) en el área rural de dos municipios de Cundinamarca, Colombia. *Biomédica*, *37*, 24. https://doi.org/http://dx.doi.org/10.7705/biomedica.v33i4.836

Apache™ Hadoop®. (n.d.). Apache Hadoop. Retrieved from http://hadoop.apache.org/

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A. S., & Buyya, R. (2015). Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, *79–80*, 3–15. https://doi.org/https://doi.org/10.1016/j.jpdc.2014.08.003

Azar, A. T., El-Said, S. A., & Hassanien, A. E. (2013). Fuzzy and hard clustering analysis for thyroid disease. *Computer Methods and Programs in Biomedicine*, *111*(1), 1–16.

https://doi.org/10.1016/j.cmpb.2013.01.002

Berchialla, P., Snidero, S., Stancu, A., Scarinzi, C., Corradetti, R., & Gregori, D. (2012). Understanding the epidemiology of foreign body injuries in children using a data-driven Bayesian network. *Journal of Applied Statistics*, *39*(4), 867–874. https://doi.org/10.1080/02664763.2011.623156

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2–3), 191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., … Hay, S. I. (2013). The global distribution and burden of dengue. *Nature*, *496*(7446), 504–507. https://doi.org/10.1038/nature12060

Bihani, P., & Patil, S. T. (2014). A comparative study of data analysis techniques. *International Journal of Emerging Trends & Technology in Computer Science*, *3*(2), 95–101.

Buczak, A. L., Baugher, B., Guven, E., Ramac-Thomas, L. C., Elbert, Y., Babin, S. M., & Lewis, S. H. (2015). Fuzzy association rule mining and classification for the prediction of malaria in South Korea. *BMC Medical Informatics and Decision Making*, *15*(1), 47. https://doi.org/10.1186/s12911-015-0170-6

Buczak, A. L., Koshute, P. T., Babin, S. M., Feighner, B. H., & Lewis, S. H. (2012). A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. *BMC Medical Informatics and Decision Making*, *12*(1), 124. https://doi.org/10.1186/1472-6947-12-124

Campello, R. J. G. B., & Hruschka, E. R. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, *157*(21), 2858–2875.

https://doi.org/10.1016/j.fss.2006.07.006

Carroll, L. N., Au, A. P., Detwiler, L. T., Fu, T. chieh, Painter, I. S., & Abernethy, N. F. (2014). Visualization and analytics tools for infectious disease epidemiology: A systematic review. *Journal of Biomedical Informatics*, *51*, 287–298. https://doi.org/10.1016/j.jbi.2014.04.006

Cattaneo, G., Giancarlo, R., Petrillo, U. F., & Roscigno, G. (2018). MapReduce in Computational Biology Via Hadoop and Spark. In *Reference Module in Life Sciences* (pp. 1–9). https://doi.org/10.1016/B978-0-12-809633-8.20371-3

Chefaoui, R. M., & Lobo, J. M. (2008). Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling*, *210*(4), 478–486. https://doi.org/10.1016/j.ecolmodel.2007.08.010

Chen, M., Ma, Y., Song, J., Lai, C.-F., & Hu, B. (2016). Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring. *Mobile Networks and Applications*, *21*(5), 825–845. https://doi.org/10.1007/s11036-016-0745-1

Corley, C. D., Pullum, L. L., Hartley, D. M., Benedum, C., Noonan, C., Rabinowitz, P. M., & Lancaster, M. J. (2014). Disease prediction models and operational readiness. *PLoS ONE*, *9*(3), 1–9. https://doi.org/10.1371/journal.pone.0091989

Crisci, C., Ghattas, B., & Perera, G. (2012). A review of supervised machine learning algorithms and their applications to ecological data. *Ecological Modelling*, *240*, 113–122. https://doi.org/10.1016/j.ecolmodel.2012.03.001

Descloux, E., Mangeas, M., Menkes, C. E., Lengaigne, M., Leroy, A., Tehei, T., … de Lamballerie, X. (2012). Climate-based models for understanding and forecasting dengue epidemics. *PLoS Neglected Tropical Diseases*, *6*(2). https://doi.org/10.1371/journal.pntd.0001470

Desgraupes, B. (2017). Clustering Indices. *CRAN Package*. Retrieved from cran.r-project.org/web/packages/clusterCrit

Desgraupes, B. (2018). Package ‘ clusterCrit .’ Retrieved from https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf

Docherty, A. B., & Lone, N. I. (2015). Exploiting big data for critical care research. *Current Opinion in Critical Care*, *21*(5), 467–72. https://doi.org/10.1097/MCC.0000000000000228

Doring, C., Lesot, M.-J., & Kruse, R. (2006). Data analysis with fuzzy clustering methods. *Computational Statistics & Data Analysis*, *51*, 192–214. https://doi.org/10.1016/j.csda.2006.04.030

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, *3*(3), 32–57. https://doi.org/10.1080/01969727308546046

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., … Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, *2*(3), 267–279. https://doi.org/10.1109/TETC.2014.2330519

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Finkelstein, J., & Jeong, I. cheol. (2017). Machine learning approaches to personalize early prediction of asthma exacerbations. *Annals of the New York Academy of Sciences*, *1387*(1), 153–165. https://doi.org/10.1111/nyas.13218

Fisman, D. N. (2007). Seasonality of Infectious Diseases. *Annual Review of Public Health*, *28*(1), 127–143. https://doi.org/10.1146/annurev.publhealth.28.021406.144128

Forouzanfar, M. H., Afshin, A., Alexander, L. T., Biryukov, S., Brauer, M., Cercy, K., … Zhu, J. (2016). Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, *388*(10053), 1659–1724. https://doi.org/10.1016/S0140-6736(16)31679-8

Gabrielsson, J., Politis, D., Dahlstrand, Å. L., & Patents, A. (2016). A formal definition of Big Data based on its essential features. *Library Review*, *65*(3), 122–135. https://doi.org/10.1108/LR-06-2015-0061

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gantz, J., & Reinsel, D. (2011). Extracting Value from Chaos State of the Universe: An Executive Summary. *IDC IView*, (June), 1–12. Retrieved from https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

Georgiev, G., Gueorguieva, N., Chiappa, M., & Krauza, A. (2016). Feature selection using Gustafson-Kessel fuzzy algorithm in high dimension data clustering. *Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015*, 1–6. https://doi.org/10.1109/ICMLA.2015.57

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. https://doi.org/10.1038/nature07634

Giordani, A. P., Ferraro, M. B., & Giordani, M. P. (2018). Fuzzy Clustering. Retrieved from https://cran.r-project.org/web/packages/fclust/fclust.pdf

Gobernación de Santander. (2014). Lineamientos y Directrices de Ordenamiento Territorial del Departamento de Santander. Bucaramanga, Santander. Retrieved from http://historico.santander.gov.co/index.php/gobernacion/documentacion/finish/359-lineamientos-y-directrices-del-ordenamiento-territorial/5969-lineamientos-y-directrices-de-ordenamiento-territorial-del-departamento-de-santander

González, C., Wang, O., Strutz, S. E., González-Salazar, C., Sánchez-Cordero, V., & Sarkar, S. (2010). Climate change and risk of leishmaniasis in North America: Predictions from ecological niche models of vector and reservoir species. *PLoS Neglected Tropical Diseases*, *4*(1). https://doi.org/10.1371/journal.pntd.0000585

Govindaraju, V., Raghavan, V. ., & C.R., R. (Eds.). (2013). *Handbook of Statistics: Big Data Analytics*. *Journal of Chemical Information and Modeling* (Vol. 53).

Grassly, N. C., & Fraser, C. (2008). Mathematical models of infectious disease transmission. *Nature Reviews Microbiology*, *6*, 477–487. https://doi.org/10.1038/nrmicro1845

Green, M. S., & Kaufman, Z. (2002). Surveillance for early detection and monitoring of infectious disease outbreaks associated with bioterrorism. *Israel Medical Association Journal*, *4*(7), 503–506.

Grekousis, G., & Thomas, H. (2012). Comparison of two fuzzy algorithms in geodemographic segmentation analysis: The fuzzy C-means and Gustafson-Kessel methods. *Applied Geography*, *34*, 125–136. https://doi.org/10.1016/j.apgeog.2011.11.004

Grossglauser, M., & Saner, H. (2014). Data-driven healthcare: from patterns to actions. *European Journal of Preventive Cardiology*, *21*(2 Suppl), 14–17. https://doi.org/10.1177/2047487314552755

Gu, W., Unnasch, T. R., Katholi, C. R., Lampman, R., & Novak, R. J. (2008). Fundamental issues

in mosquito surveillance for arboviral transmission. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *102*(8), 817–822. https://doi.org/10.1016/j.trstmh.2008.03.019

Gustafson, D., & Kessel, W. (1978). Fuzzy clustering with a fuzzy covariance matrix. In *1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes* (pp. 761–766). IEEE. https://doi.org/10.1109/CDC.1978.268028

Hall, H. I., Correa, A., Yoon, P. W., & Braden, C. R. (2012). Lexicon, Definitions, and Conceptual Framework for Public Health Surveillance. *Mmwr 2012*. Centers for Disease Control and Prevention (CDC). https://doi.org/su6103a5 [pii]

Hamerly, G., & Drake, J. (2015). Accelerating Lloyd's Algorithm for k-Means Clustering BT - Partitional Clustering Algorithms. In M. E. Celebi (Ed.) (pp. 41–78). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-09259-1_2

Han, J., Kamber, M., & Pei, J. (2012). Cluster Analysis. In *Data Mining* (pp. 443–495). https://doi.org/10.1016/B978-0-12-381479-1.00010-1

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

Hartley, D. M., Nelson, N. P., Walters, R., Arthur, R., Yangarber, R., Madoff, L., … Lightfoot, N. (2010). The landscape of international event-based biosurveillance. *Emerging Health Threats Journal*, *3*(1), 7. https://doi.org/10.3134/ehtj.10.003

Hay, S. I., George, D. B., Moyes, C. L., & Brownstein, J. S. (2013). Big Data Opportunities for Global Infectious Disease Surveillance. *PLoS Medicine*, *10*(4). https://doi.org/10.1371/journal.pmed.1001413

Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004). *Introducción a la*

minería de datos. (S. A. Pearson Educación, Ed.). Madrid.

Hu, H., Wen, Y., Chua, T.-S., & Li, X. (2014). Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *IEEE Access*, *2*, 652–687. https://doi.org/10.1109/ACCESS.2014.2332453

Hu, W., O'Leary, R. A., Mengersen, K., & Choy, S. (2011). Bayesian Classification and regression trees for predicting incidence of cryptosporidiosis. *PLoS ONE*, *6*(8), 1–8. https://doi.org/10.1371/journal.pone.0023903

Huang, S., Lin, Y., Yih, J., & Tseng, J. (2017). Fuzzy Clustering Algorithm Based on Mahalanobis Distances with Recursive Process, *4*. https://doi.org/10.6148/IJITAS.2017.1004.02

Hurtado-Díaz, M., Riojas-Rodríguez, H., Rothenberg, S. J., Gomez-Dantés, H., & Cifuentes, E. (2007). Short communication: Impact of climate variability on the incidence of dengue in Mexico. *Tropical Medicine and International Health*, *12*(11), 1327–1337. https://doi.org/10.1111/j.1365-3156.2007.01930.x

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jefferson, H., Dupuy, B., Chaudet, H., Texier, G., Green, A., Barnish, G., … Meynard, J. B. (2008). Evaluation of a syndromic surveillance for the early detection of outbreaks among military personnel in a tropical country. *Journal of Public Health*, *30*(4), 375–383. https://doi.org/10.1093/pubmed/fdn026

Jia, Z. W., Jia, X. W., Liu, Y. X., Dye, C., Chen, F., Chen, C. S., … Tang, J. W. (2008). Spatial analysis of tuberculosis cases in migrants and permanent residents, Beijing, 2000-2006. *Emerging Infectious Diseases*, *14*(9), 1413–1420. https://doi.org/10.3201/1409.071543

Jiawei, H., Micheline, K., & Pei, J. (2012). Clasification: Basic concepts. In *Data Mining:*

Concepts and Techniques (pp. 327–392). Elsevier.

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, *451*(7181), 990–993. https://doi.org/10.1038/nature06536

Jorm, L. R., Thackway, S. V, Churches, T. R., & Hills, M. W. (2003). Watching the Games: public health surveillance for the Sydney 2000 Olympic Games. *Journal of Epidemiology and Community Health*, *57*(2), 102–108. https://doi.org/10.1136/jech.57.2.102

Joseph, R. C., & Johnson, N. A. (2013). Big data and transformational government. *IT Professional*, *15*(6), 43–48.

Kapageridis, I. K. (2015). Variable lag variography using k-means clustering. *Computers and Geosciences*, *85*, 49–63. https://doi.org/10.1016/j.cageo.2015.04.004

Kautz, T., Eskofier, B. M., & Pasluosta, C. F. (2017). Generic performance measure for multiclass-classifiers. *Pattern Recognition*, *68*, 111–125. https://doi.org/10.1016/j.patcog.2017.03.008

Khoury, M. J. (2015). Planning for the Future of Epidemiology in the Era of Big Data and Precision Medicine. *American Journal of Epidemiology*, *182*(12), 1–5. https://doi.org/10.1093/aje/kwv228.Planning

Khoury, M. J., Lam, T. K., Ioannidis, J. P. A., Hartge, P., Spitz, M. R., Buring, J. E., … Schully, S. D. (2013). Transforming epidemiology for 21st century medicine and public health. *Cancer Epidemiology Biomarkers and Prevention*, cebp-0146. https://doi.org/10.1158/1055-9965.EPI-13-0146

Kubat, M. (2015). Performance Evaluation. In M. Kubat (Ed.), *An Introduction to Machine Learning* (pp. 213–233). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-20010-1_11

Kuhn, M., & Johnson, K. (2013a). Measuring Performance in Classification Models BT - Applied Predictive Modeling. In M. Kuhn & K. Johnson (Eds.) (pp. 247–273). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-6849-3_11

Kuhn, M., & Johnson, K. (2013b). Over-Fitting and Model Tuning BT - Applied Predictive Modeling. In M. Kuhn & K. Johnson (Eds.) (pp. 61–92). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-6849-3_4

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons*, *60*(3), 293–303. https://doi.org/10.1016/j.bushor.2017.01.004

Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

Lopez, D., Gunasekaran, M., Senthil Murugan, B., Kaur, H., & Abba. (2014). Spatial Big Data Analytics of Influenze Epidemic in Vellore, India. In *IEEE International Conference on Big Data, 2014* (pp. 19–24).

Lugo-Reyes, S. O., Maldonado-Colín, G., & Murata, C. (2014). Inteligencia artificial para asistir el diagnóstico clínico en medicina. *Revista Alergia Mexico*, *61*(2), 110–120.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*(233), 281–297. https://doi.org/citeulike-article-id:6083430

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., … Kozlowski, K. (2018). "Finding Groups in Data''': Cluster Analysis Extended Rousseeuw et al." *R Package Version*. Retrieved from http://cran.r-project.org/web/packages/cluster/index.html

Manuel, J., & Sesmero, M. (2015). "Big Data", aplicación y utilidad para el sistema sanitario. *Farmacia Hospitalaria: Organo Oficial de Expresion Cientifica de La Sociedad Espanola de*

Farmacia Hospitalaria, *39*(2), 69–70. https://doi.org/10.7399/fh.2015.39.2.8835

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R. ;, Roxburgh, C., & Hung, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. *McKinsey Global Institute*. https://doi.org/10.1080/01443610903114527

Martinez, R., Ordunez, P., Soliz, P. N., & Ballesteros, M. F. (2016). Data visualisation in surveillance for injury prevention and control: conceptual bases and case studies. *Injury Prevention*, *22*(Suppl 1), i27–i33. https://doi.org/10.1136/injuryprev-2015-041812

Martínez, S. J. M. (2015). " Big data"; application and use for the health system. *Farmacia Hospitalaria: Órgano Oficial de Expresión Científica de La Sociedad Española de Farmacia Hospitalaria*, *39*(2), 69.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Buiness Review*, *90*(10), 60+.

Mdaghri, Z. A., Yadari, M. El, Benyoussef, A., & Kenz, A. El. (2016). Study and analysis of Data Mining for Healthcare. *IEEE*, 77–82.

Milinovich, G. J., Williams, G. M., Clements, A. C. A., & Hu, W. (2014). Internet-based surveillance systems for monitoring emerging infectious diseases. *The Lancet Infectious Diseases*, *14*(2), 160–168. https://doi.org/10.1016/S1473-3099(13)70244-5

Ministerio de Salud y de Proteccion Social. (2012). *Plan decenal de salud pública (PDSP) 2012-2021*. Bogotá, Colombia. https://doi.org/10.1017/CBO9781107415324.004

Ministerio de Salud y de Proteccion Social. (2014). *Guía conceptual y metodológica para la construcción del ASIS de las Entidades Territoriales. Colombia 2014*. *Ministerio de Salud Y protección Social*. Bogotá, Colombia. Retrieved from https://www.minsalud.gov.co/salud/publica/epidemiologia/Paginas/analisis-de-situacion-de-

salud-.aspx

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, *9*(1), 15–24. https://doi.org/10.20982/tqmp.09.1.p015

Nidheesh, N., Abdul Nazeer, K. A., & Ameer, P. M. (2017). An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data. *Computers in Biology and Medicine*, *91*(June), 213–221. https://doi.org/10.1016/j.compbiomed.2017.10.014

O'Connell, E. K., Zhang, G., Leguen, F., Llau, A., & Rico, E. (2010). Innovative uses for syndromic surveillance. *Emerging Infectious Diseases*, *16*(4), 669–671. https://doi.org/10.3201/eid1604.090688

O'Shea, J. (2017). Digital disease detection: A systematic review of event-based internet biosurveillance systems. *International Journal of Medical Informatics*, *101*, 15–22. https://doi.org/10.1016/j.ijmedinf.2017.01.019

OMS. (2017). OMS | Dengue. *Who*. Retrieved from http://www.who.int/mediacentre/factsheets/fs270/es/%0Ahttp://www.who.int/topics/dengue/es/

Organizacion Mundial de la Salud. (2014). *Informe sobre la situación mundial de las enfermedades no transmisibles 2014*. *Who*. https://doi.org/ISBN: 978 92 4 156422 9

Organización Panamericana de la Salud. (2015). *Guías para la atención de enfermos en la región de las américas*. *Catalogación en la Fuente, Biblioteca Sede de la OPS* (Vol. dos). https://doi.org/NLM: WC680

Paolotti, D., Carnahan, A., Colizza, V., Eames, K., Edmunds, J., Gomes, G., … Vespignani, A.

(2014). Web-based participatory surveillance of infectious diseases: The Influenzanet participatory surveillance experience. *Clinical Microbiology and Infection*, *20*(1), 17–21. https://doi.org/10.1111/1469-0691.12477

Paquet, C., Coulombier, D., Kaiser, R., & Ciotti, M. (2006). Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. *Euro Surveillance : Bulletin Européen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, *11*(12), 212–214. https://doi.org/665 [pii]

Pfeiffer, D. U., Robinson, T. P., Stevenson, M., Stevens, kim B., Rogers, D. J. &, & Clements, A. C. . (2008). *Spatial Analysis in Epidemiology*. *Spatial Analysis in Epidemiology*. https://doi.org/10.1016/j.actatropica.2004.05.001

Pfeiffer, D. U., & Stevens, K. B. (2015). Spatial and temporal epidemiological analysis in the Big Data era. *Preventive Veterinary Medicine*, *122*(1–2), 213–220. https://doi.org/10.1016/j.prevetmed.2015.05.012

Ponce Cruz, P. (2010). *Inteligencia Artificial con aplicaciones a la Ingeniería* (Primera ed). México: Alfaomega.

Porta, M. (2008). *Dictionary of Epidemiology*. Cary, UNKNOWN: Oxford University Press, USA. Retrieved from http://ebookcentral.proquest.com/lib/bibliouis-ebooks/detail.action?docID=537605

PostgreSql. (2018). PostgreSQL: Acerca de.

Quintero-Herrera, L. L., Ramírez-Jaramillo, V., Bernal-Gutiérrez, S., Cárdenas-Giraldo, E. V., Guerrero-Matituy, E. A., Molina-Delgado, A. H., … Rodríguez-Morales, A. J. (2015). Potential impact of climatic variability on the epidemiology of dengue in Risaralda, Colombia, 2010-2011. *Journal of Infection and Public Health*, *8*(3), 291–297.

https://doi.org/10.1016/j.jiph.2014.11.005

Reglamento Sanitario Internacional (2005). (2016) (Tercera ed). Ginebra: Organización Mundial de la Salud. Retrieved from http://www.who.int/ihr/publications/9789241580496/es/

Robertson, C. (2017). Towards a geocomputational landscape epidemiology: surveillance, modelling, and interventions. *GeoJournal*, *82*(2), 397–414. https://doi.org/10.1007/s10708-015-9688-5

Rodriguez-Morales, A. J., Ruiz, P., Tabares, J., Ossa, C. A., Yepes-Echeverry, M. C., Ramirez-Jaramillo, V., … Grobusch, M. P. (2017). Mapping the ecoepidemiology of Zika virus infection in urban and rural areas of Pereira, Risaralda, Colombia, 2015–2016: Implications for public health and travel medicine. *Travel Medicine and Infectious Disease*, *18*(May), 57–66. https://doi.org/10.1016/j.tmaid.2017.05.004

RStudio. (n.d.-a). Funciones de IDE - RStudio. Retrieved from https://www.rstudio.com/products/rstudio/

RStudio. (n.d.-b). Shiny. Retrieved from https://shiny.rstudio.com/

RStudio. (2017). Take control of your R code. Retrieved from https://www.rstudio.com/products/rstudio/

Santander, S. de S. (2015). *Análisis de Situación de Salud con el Modelo de los Determinantes Sociales de Salud del departamento de Santander, 2015*. Santander. https://doi.org/10.1017/CBO9781107415324.004

Semenza, J. C., & Menne, B. (2009). Climate change and infectious diseases in Europe. *The Lancet Infectious Diseases*, *9*(6), 365–375. https://doi.org/10.1016/S1473-3099(09)70104-5

Singh, G., Al'Aref, S. J., Van Assen, M., Kim, T. S., van Rosendael, A., Kolli, K. K., … Min, J. K. (2018). Machine learning in cardiac CT: Basic concepts and contemporary data. *Journal*

of Cardiovascular Computed Tomography, *12*(3), 192–201. https://doi.org/10.1016/j.jcct.2018.04.010

Sinka, M. E., Bangs, M. J., Manguin, S., Rubio-Palis, Y., Chareonviriyaphap, T., Coetzee, M., … Hay, S. I. (2012). A global map of dominant malaria vectors. *Parasites & Vectors*, *5*(1), 69. https://doi.org/10.1186/1756-3305-5-69

Slonim, N., Aharoni, E., & Crammer, K. (2013). Hartigan's k-means versus Lloyd's k-means - Is it time for a change? In *Twenty-Third International Joint Conference on Artificial Intelligence* (pp. 1677–1684).

Smith, M., Szongott, C., Henne, B., & Von Voigt, G. (2012). Big data privacy issues in public social media. In *In Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference* (pp. 1–6). IEEE. https://doi.org/10.1109/DEST.2012.6227909

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Spark™, A. (n.d.). Lightning-fast unified analytics engine. Retrieved April 21, 2018, from https://spark.apache.org/

Spark™, A. (2018a). Cluster Mode Overview - Spark 2.0.2 Documentation. Retrieved from https://spark.apache.org/docs/2.0.2/

Spark™, A. (2018b). RDD Programming Guide. Retrieved from https://spark.apache.org/docs/latest/rdd-programming-guide.html

Tan, S. S.-L., Gao, G., & Koch, S. (2015). Big Data and Analytics in Healthcare. *Methods of Information in Medicine*, 546–547. Retrieved from https://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSear

ch&qid=8&SID=Q1J6h1m5tXLIfHqsJ4x&page=1&doc=8&cacheurlFromRightClick=no

The Apache Software Foundation. (2017). Apache Projects List. Retrieved from https://projects.apache.org/projects.html

The R Fundation. (2018). ¿Qué es R? Retrieved from https://cran.r-project.org/

Tsui, K. L., Wong, S. Y., Jiang, W., & Lin, C. J. (2011). Recent research and developments in temporal and spatiotemporal surveillance for public health. *IEEE Transactions on Reliability*, *60*(1), 49–58. https://doi.org/10.1109/TR.2010.2104192

ur Rehman, M. H., Chang, V., Batool, A., & Wah, T. Y. (2016). Big data reduction framework for value creation in sustainable enterprises. *International Journal of Information Management*, *36*(6), 917–928.

van Noort, S. P., Codeco, C. T., Koppeschaar, C. E., van Ranst, M., Paolotti, D., & Gomes, M. G. M. (2015). Ten-year performance of Influenzanet: ILI time series, risks, vaccine effects, and care-seeking behaviour. *Epidemics*, *13*, 28–36. https://doi.org/10.1016/j.epidem.2015.05.001

Wang, Y. (2005). A multinomial logistic regression modeling approach for anomaly intrusion detection. *Computers and Security*, *24*(8), 662–674. https://doi.org/10.1016/j.cose.2005.05.003

Wani, A. H., & Shashirekha, H. L. (2018). Clustering: A Novel Meta-Analysis Approach for Differentially Expressed Gene Detection BT - Proceedings of International Conference on Cognition and Recognition. In D. S. Guru, T. Vasudev, H. K. Chethan, & Y. H. S. Kumar (Eds.) (pp. 119–126). Singapore: Springer Singapore.

WHO. (2017). WHO STEPS Surveillance Manual: The WHO STEPwise approach to noncommunicable disease risk factor surveillance. WHO. Retrieved from http://www.who.int/chp/steps/STEPS_Manual.pdf?ua=1

Wu, Y., Lee, G., Fu, X. J., & Hung, T. (2008). Detect climatic factors contributing to dengue outbreak based on wavelet, support vector machines and genetic algorithm. In *World Congress on Engineering 2008 Vols IIi* (Vol. 1, pp. 303–307). https://doi.org/10.1.1.149.3221

Xie, X. L., & Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/34.85677

Zhang, X., Zhang, T., Young, A. A., & Li, X. (2014). Applications and comparisons of four time series models in epidemiological surveillance data. *PLoS ONE*, *9*(2), 1–16. https://doi.org/10.1371/journal.pone.0088075