Reglas de asociación aplicadas al análisis de contenido de los tweets sobre enfermedades transmitidas por vectores en Santander, Colombia.

Cristian Eduardo Rodríguez Angarita

Juan Camilo Rojas Mariño

Trabajo de grado para optar el título de Ingeniero Industrial

Director:

Henry Lamos Díaz

PhD en Física-Matemática

Codirector:

Yuly Andrea Ramírez Sierra

Ingeniera Industrial

Universidad Industrial de Santander

Facultad de Ingeniería Fisicomecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2019

AGRADECIMIENTOS

Gracias a Dios por la sabiduría y entendimiento para poder culminar esta etapa satisfactoriamente.

Gracias al profesor Henry Lamos Díaz por confiar en nosotros, por darnos la oportunidad de crecer en conjunto, por el constante acompañamiento y orientación durante la realización de este proyecto.

Gracias a nuestra codirectora Yuly Andrea Ramírez, por su paciencia, entrega, dedicación y amplio conocimiento compartido para con nosotros.

Gracias al grupo de investigación OPALO, por permitirnos hacer parte de este valioso equipo.

Gracias a la Universidad Industrial de Santander, por formarnos como personas íntegras y profesionales a lo largo de estos años.

DEDICATORIA

Primeramente, dedico este trabajo a Dios por guiarme y permitirme culminar esta etapa de la mejor manera, a mis padres, Aníbal y Zonia, por brindarme todos estos años su constante educación, dedicación y esfuerzo día a día, sin ellos esto no habría sido posible, a Ximena por brindarme su amor y apoyo incondicional durante toda esta etapa, y a mi hija Lucía por ser mi mayor felicidad y motivación para salir adelante

Juan Camilo Rojas

A Dios por orientar mi camino, fortalecerme espiritualmente y ser mi guía incondicional

A mis padres por su apoyo, sus valores, su perseverancia y constancia para formarme como persona y profesional con valores éticos, morales y espirituales.

A mis hermanos para que siempre luchen por sus metas y logren cumplir todo lo que se proponen.

Cristian Eduardo Rodríguez

Tabla de contenido

Lista de figuras

Lista de tablas

Lista de apéndices

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

Resumen

TÍTULO: REGLAS DE ASOCIACIÓN APLICADAS AL ANÁLISIS DE CONTENIDO DE LOS TWEETS SOBRE ENFERMEDADES TRANSMITIDAS POR VECTORES EN SANTANDER, COLOMBIA*

AUTORES: Cristian Eduardo Rodríguez Angarita

Juan Camilo Rojas Mariño**

PALABRAS CLAVE: Aprendizaje automático, Reglas de asociación, Máquinas de soporte vectorial, Twitter, Enfermedades transmitidas por vectores.

DESCRIPCIÓN:

Las redes sociales permiten generar gran cantidad de datos que pueden ser procesados por medio de técnicas de minería de datos y de aprendizaje automático para obtener conocimiento de valor y apoyar la toma de decisiones. Twitter a través de la interfaz de programación de aplicaciones, permite extraer efectivamente estos datos y mediante la aplicación de técnicas de representación del texto se proceda a descubrir patrones útiles, novedosos y válidos, por ejemplo, estos datos permiten caracterizar poblaciones con algún brote epidémico durante un tiempo determinado.

En esta investigación se aplican algoritmos de clasificación supervisada y modelos de reglas de asociación para el análisis de contenido de los tweets relacionados con información sobre Enfermedades Transmitidas por Vectores (ETV) tanto en Colombia como en el Departamento de Santander. De esta forma, se clasifican los tweets que son publicados por los diferentes usuarios de la red social, con etiquetas de "válido" o "inválido" según corresponda, siendo el modelo de máquinas de soporte vectorial con función de kernel lineal el que mejor representa los datos, con una exactitud del 90,7%. Posteriormente, se generan las reglas de asociación que cumplan con unos mínimos establecidos de soporte y confianza, para así, extraer y visualizar las relaciones entre los términos relacionados con cada una de las ETV, identificando las palabras que tienden a presentarse principalmente cuando se habla de una enfermedad en específico, con el fin de identificar posibles relaciones que sean de interés para abordar la salud pública de la población.

---

* Trabajo de grado

** Facultad de Ingenierías Fisicomecánicas. Escuela de Estudios Industriales y Empresariales. Director: PhD. Henry Lamos Díaz, Codirector: Ing. Yuly Andrea Ramírez Sierra

Abstract

TITLE: ASSOCIATION RULES APPLIED TO THE CONTENT ANALYSIS OF TWEETS ON VECTOR-BORNE DISEASES IN SANTANDER, COLOMBIA*

AUTHORS: Cristian Eduardo Rodríguez Angarita

Juan Camilo Rojas Mariño**

KEYWORDS: Machine Learning, Association Rules, Support Vector Machines, Twitter, Vector-borne diseases.

DESCRIPTION:

Social networks allow the generation of a large amount of data that can be processed by means of data mining and machine learning techniques to obtain valuable knowledge and support decision-making. Twitter through the application programming interface, allows to extract these data effectively and through the application of techniques of representation of the text it is come to discover useful, novel and valid patterns, for example, these data allow to characterize populations with some epidemic outbreak during a certain time.

In this research, supervised classification algorithms and models of association rules are applied to content analysis of tweets related to information on Vector-borne Diseases in both Colombia and the Department of Santander. In this way, the tweets that are published by the different users of the social network are classified, with "valid" or "invalid" labels as appropriate, being the model of Support Vector Machines with linear kernel function the one that best represents the data, with an accuracy of 90.7%. Subsequently, the association rules that comply with established minimum support and confidence are generated, in order to extract and visualize the relationships between the terms related to each of the Vector-borne Diseases, identifying the words that tend to occur mainly when talking about a specific disease, in order to identify possible relationships that are of interest to address the public health of the population.

*Bachelor Thesis

**Faculty of Physicomechanical Engineering. Industrial and Business School. Director: PhD. Henry Lamos Diaz, Co-director: Eng. Yuly Andrea Ramírez Sierra

Introducción

Actualmente hay una fuerte tendencia al uso de redes sociales y conectividad, donde los usuarios de la web se entusiasman cada vez más al interactuar, compartir y trabajar en comunidad a través de medios colaborativos en línea; esta inteligencia colectiva se ha extendido a diversas áreas con un impacto creciente en la vida cotidiana, como la educación, la salud, el comercio y el turismo, lo cual permite un crecimiento en el uso de las redes sociales (Hussain & Cambria, 2018). Las redes sociales generan una enorme cantidad de datos que deben ser procesados por medio de técnicas de análisis con el fin de obtener conocimiento de valor y apoyar la toma de decisiones. En el área de la salud, específicamente en salud pública se usan las redes sociales para identificar enfermedades transmitidas por vectores (ETV) y así realizar vigilancia epidemiológica, con el fin de establecer estrategias de control y favorecer las intervenciones de la salud pública en determinada área.

La vigilancia tradicional de enfermedades ha sido un aspecto clave en las entidades de salud pública durante mucho tiempo y es reconocida como una herramienta útil para evaluar, predecir y mitigar el impacto de los brotes de enfermedades infecciosas; dicha vigilancia se basa en datos recopilados por instituciones de salud, que consisten en datos de morbilidad y mortalidad, datos de laboratorio, investigaciones de campo, encuestas y datos demográficos. La revolución actual de la internet y los teléfonos móviles ha permitido que los datos epidemiológicos estén disponibles de forma más rápida y amplia, y que el público genere nuevos datos a partir de publicaciones o conexiones con otros usuarios (Salathé, 2016); la gente recurre cada vez más a la internet para

obtener información de salud, por ejemplo una encuesta de “Pew research center” muestra que el 61% de los adultos estadounidenses buscan información de salud en línea y el 41% de este grupo ha leído sobre las experiencias médicas de otras personas publicadas en las redes sociales; llegando a la conclusión que más de la mitad piensa que la información en línea afectaba una decisión sobre cómo tratar una enfermedad o condición (Fox & Jones, 2009).

Con el valor de la información en línea, cada vez más los investigadores aprovechan el uso de las redes sociales como fuente complementaria y en algunos casos reemplaza las prácticas existentes en salud pública. La red social Twitter conecta a una amplia variedad de usuarios para compartir información en torno a muchos temas e intercambiar ideas en línea, su característica clave es la naturaleza pública de sus tweets, permitiendo la recuperación de estos datos por medio de una interfaz de programación de aplicaciones (por sus siglas en inglés, API) (Burton, Tanner, Giraud-Carrier, West, & Barnes, 2012). Por tanto, esta fuente de información, Twitter, exige un análisis de contenido que puede llevarse a cabo mediante el uso de técnicas de minería de datos y de aprendizaje automático para aprovechar la cantidad de información que se genera.

En el presente trabajo se propone construir modelos de reglas de asociación y algoritmos de clasificación supervisada de aprendizaje automático para el análisis de contenido de los tweets relacionados con información sobre ETV tanto en Colombia como en el Departamento de Santander; así, se busca identificar relaciones de los principales eventos de salud pública mediante el uso de la información obtenida a partir de la API de Twitter con apoyo de la herramienta RStudio. Inicialmente se clasifica la información que es publicada en Twitter por las diferentes autoridades de salud pública en Colombia y en el Departamento de Santander, utilizando datos

etiquetados manualmente como “Válido” o “Inválido” según el contenido de cada tweet, con el fin de distinguir mejor los tweets que parecen describir casos reales de ETV de aquellos que no lo hacen, de igual manera con la clasificación de “Válido” e “Inválido” se busca identificar focos de las enfermedades dentro de Colombia y el Departamento de Santander para reconocer cuales son las zonas más propensas a una infección de este tipo, así, se busca construir modelos de clasificación que son validados mediante la precisión y exactitud de sus resultados.

El documento se estructura en tres secciones: en la primera sección, se presenta la revisión de literatura con el fin de dar soporte al tema de investigación tratado, seguido se detalla el modelo a desarrollar de máquinas de soporte vectorial (SVM, por sus siglas en inglés) y el algoritmo de reglas de asociación con sus respectivos resultados y validación, finalmente en la última sección se presentan las conclusiones y recomendaciones para futuros trabajos de investigación. A continuación, se presenta la Tabla 1. que evidencia el cumplimiento a cada objetivo planteado en la investigación.

Tabla 1.
Cumplimiento de objetivos

Objetivos específicos	***Cumplimiento***
Realizar una revisión de literatura sobre las reglas de asociación y algoritmos de clasificación supervisada de aprendizaje automático utilizados para el análisis de texto de redes sociales, teniendo en cuenta información sobre eventos de salud pública.	Capítulo 3
Identificar relaciones de los principales eventos de salud pública en Santander, a partir de los datos obtenidos de la red social Twitter, para desarrollar modelos de reglas de asociación.	Capítulo 5 Capítulo 6
Validar los modelos construidos mediante las técnicas aplicadas, para evaluar la precisión y exactitud de los resultados de clasificación.	Capítulo 7
Elaborar un artículo de carácter publicable resumiendo los hallazgos de la investigación.	Apéndice B

1. Planteamiento del problema

El desarrollo acelerado de la tecnología web 2.0 y el aumento en la disponibilidad de acceso a la internet, ha llevado a aumentar el uso de las redes sociales debido al constante intercambio de información en línea. Twitter, es una plataforma de red social líder y famosa a nivel mundial que se estableció en 2006, actualmente cuenta con 313 millones de usuarios y 500 millones de tweets por día, posee una capacidad de 140 caracteres para que sus usuarios publiquen mensajes cortos (tweets), cuenta con conversaciones, chats, informes de noticias, entre otros. Diariamente se están generando grandes cantidades de datos, las personas discuten temas relacionados sobre la salud, permitiendo el uso de estos datos a entidades interesadas para el procesamiento y análisis de los mismos con el fin de dar respuesta a problemas de salud pública, convirtiéndose Twitter en una herramienta de información potencial (Kagashe, Yan, & Suheryani, 2017).

Asimismo, las redes sociales pueden explotarse para caracterizar una tendencia de salud, como brotes epidémicos en una población durante un tiempo determinado. Twitter permite extraer efectivamente datos que luego se procesan y analizan utilizando estrategias estándar de aprendizaje automático (Yin, Fabbri, Rosenbloom, & Malin, 2015), por ejemplo los contenidos no estructurados de Twitter se extraen a través de una API, y así como se puede obtener información de un usuario en particular, también es posible consultar una temática de interés, con el propósito de realizar análisis descriptivo, de contenido o de las redes de comunidades.

En este sentido, las entidades de salud pública y otras partes interesadas pueden hacer uso completo de esta información para diseñar estrategias eficientes y efectivas para la vigilancia de epidemias, entendiendo la vigilancia como la recopilación sistemática, gestión, análisis e interpretación de los datos relacionados con las epidemias, con el fin de difundir la información y agilizar las acciones de los programas de prevención y promoción en salud pública, y así por ejemplo responder a brotes de enfermedades que requieren atención, intervención y prevención durante periodos críticos en donde se afecta la salud de la población (Kagashe et al., 2017).

Entre otras aplicaciones que han complementado la vigilancia tradicional de los eventos de salud pública con el fin de construir políticas o estrategias para el control y prevención de enfermedades se han puesto una serie de enfoques "infodemiológicos" para obtener una vigilancia casi en tiempo real a través de motores de búsqueda en línea como lo es Google y el uso de redes sociales como Twitter (Nagar et al., 2014), este enfoque de vigilancia ha sido acogido principalmente para el estudio de enfermedades infecciosas, teniendo en cuenta que por ejemplo, la gripe estacional causa aproximadamente entre tres y cinco millones de casos graves y entre 250.000 a 500.000 muertes a nivel mundial.

Así, en el Departamento de Santander, el área estratégica de inteligencia epidemiológica es fortalecida por entidades como el Observatorio de Salud Pública de Santander y la Secretaría de Salud Departamental, estas entidades contribuyen en la realización de análisis de la información, actividades de integración, vigilancia tradicional con el fin de establecer estrategias de control y favorecer las intervenciones de la salud pública en el Departamento. Dentro de las ETV, son de especial interés en salud pública la enfermedad de Chagas, Dengue, Leishmaniasis, y Malaria,

debido a las condiciones de vida de la población colombiana, como condiciones geográficas donde residen las personas y el nivel socioeconómico al cual pertenecen (Ochoa, 2014).

Según el boletín epidemiológico semanal de la Secretaría de Salud de Santander, en Colombia se evidencia un incremento para el año 2017 en la notificación de los eventos de Leishmaniasis Cutánea y un decremento en la de Chagas, Chikungunya, Dengue, Dengue Grave, mortalidad por Dengue, Leishmaniasis Mucosa, Malaria y Zika con relación al año 2016.

Entonces, dado el auge de las redes sociales y de los datos no estructurados que estas redes generan, se presenta un gran reto para los analistas relacionado con el uso de las herramientas y técnicas que permitan descubrir conocimiento válido y novedoso, adoptando técnicas de minería de datos y de aprendizaje automático para construir modelos descriptivos y predictivos, respectivamente. Así, se destacan técnicas como clustering, reglas de asociación, análisis de grafos, redes bayesianas, SVM, entre otras.

En resumen, con la presente investigación se busca aplicar minería de texto y un clasificador supervisado de aprendizaje automático para la predicción de los tweets válidos/inválidos para analizar la información sobre eventos de salud pública en Colombia y en el Departamento de Santander, que apoyan a las autoridades gubernamentales a diseñar estrategias eficientes y efectivas para la vigilancia de epidemias, clasificando los tweets que las diferentes autoridades de salud pública y usuarios de la red social publican en Twitter; esta clasificación corresponde a tweets válidos o inválidos, considerando que un tweet válido es aquel directamente relacionado

con la presencia de ETV, y un tweet es considerado inválido cuando su contenido no indica la presencia de casos de ETV.

2. Objetivos

2.1 Objetivo general

Construir modelos de reglas de asociación y de aprendizaje automático, para el análisis de contenido de los tweets relacionados con información sobre enfermedades transmitidas por vectores en Santander, Colombia.

2.2 Objetivos específicos

- Realizar una revisión de literatura sobre las reglas de asociación y algoritmos de clasificación supervisada de aprendizaje automático utilizados para el análisis de texto de redes sociales, teniendo en cuenta información sobre eventos de salud pública.
- Identificar relaciones de los principales eventos de salud pública en Santander, a partir de los datos obtenidos de la red social Twitter, para desarrollar modelos de reglas de asociación.
- Validar los modelos construidos mediante las técnicas aplicadas, para evaluar la precisión y exactitud de los resultados de clasificación.
- Elaborar un artículo de carácter publicable resumiendo los hallazgos de la investigación.

3. Revisión de la literatura

La revolución digital de las plataformas de redes sociales ha convertido a Twitter en una fuente común relacionada con el cuidado de la salud, constituyendo un área nueva para que los investigadores utilicen el análisis de contenido por medio de modelos estructurados que puedan responder a la naturaleza de los datos digitales recuperados. Estos enfoques también se conocen como infodemiología, vigilancia o investigación digital de detección de enfermedades (Hamad, Savundranayagam, Holmes, Kinsella, & Johnson, 2016) que son complementarios a los métodos existentes para recopilar información y mejorar las acciones de salud pública, asignando los recursos de manera más eficiente (Dunn, Leask, Zhou, Mandl, & Coiera, 2015).

Wang, Kraut, & Levine (2015), mencionan las tres razones principales para realizar análisis de contenido asistido por computadora: en primer lugar, los modelos de aprendizaje automático permiten a los investigadores implementar métodos comparables para desafiar, verificar o ampliar los resultados de otros; en segundo lugar, los datos a gran escala que pueden analizarse mediante métodos de aprendizaje automático permite a los investigadores realizar análisis más detallados, examinar patrones de investigación y responder a preguntas más sutiles. En tercer lugar, abre oportunidades para intervenciones en tiempo real.

Dentro de la vigilancia basada en las redes sociales manteniendo el contexto de las enfermedades infecciosas, Bernardo et al., (2013) identifican cuatro áreas temáticas: funciones, capacidad, ventajas y desafíos (Figura 1). A continuación, se explica cada una de estas áreas:

A. Funciones

Las funciones se relacionan con los aspectos metodológicos de los programas. Una variedad de algoritmos de búsqueda utilizados para sintetizar y filtrar la información no estructurada de una variedad de fuentes basadas en la web como lo son noticias, medios sociales, blogs y consultas de motores de búsqueda. El fin de estos programas es la detección temprana y control efectivo de enfermedades infecciosas.

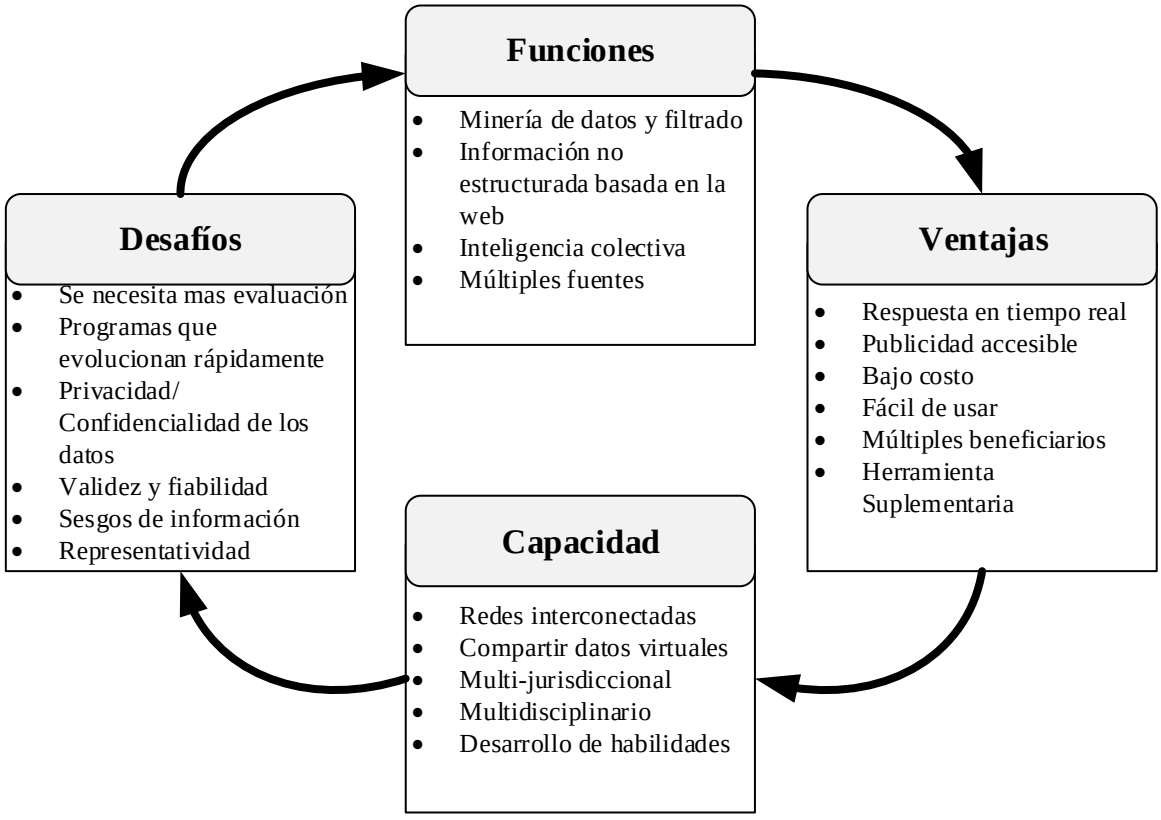

Figura 1. Características principales para la vigilancia de enfermedades infecciosas. Adaptado de (Bernardo et al., 2013, p.8). Scoping review on search queries and social media for disease surveillance: A chronology of innovation.

Por otra parte, la mayor parte de investigaciones en el área de clasificación de sentimientos de Twitter, ha utilizado desde el 2009 enfoques basados en aprendizaje automático como clasificadores bayesianos y SVM (Ji, Chun, Wei, & Geller, 2015). Así mismo, Prieto, Matos,

Álvarez, Cacheda, & Oliveira (2014) aplican varias técnicas de aprendizaje automático en los conjuntos de datos obtenidos para filtrar los datos, entre las que se identifican las SVM, clasificadores bayesianos, árboles de decisión y vecinos más cercanos, utilizando el software WEKA, una herramienta libre para la minería de datos y el aprendizaje automático. Greaves, Ramirez-Cano, Millett, Darzi, & Donaldson (2013), utilizan cuatro métodos diferentes más rápidos y precisos para evaluar los resultados: clasificadores bayesianos, arboles de decisión, agregación de bootstrap y SVM. Papalexakis, Faloutsos, & Sidiropoulos (2016) afirman que en la minería de datos, existe una gran variedad de herramientas denominadas descomposiciones o factorizaciones de tensor, herramientas muy potentes y versátiles que pueden modelar una gran variedad de datos heterogéneos y de múltiples aspectos, que tienen la capacidad de extraer una estructura significativa y latente en los datos, las cuales son técnicas muy potentes, populares y exitosas para obtener un rendimiento óptimo.

Por otro lado, Chen, Lin, & Chang, (2015) integran la minería de reglas de asociación para organizar las palabras clave extraídas en metadatos. Nikfarjam, Sarker, O'Connor, Ginn, & Gonzalez, (2015), aplican la minería de reglas de asociación, una técnica de minería de datos que busca aprender los patrones de lenguaje que usan los pacientes para informar reacciones adversas a medicamentos. AL-Rubaiee, Qiu, & Li (2017) mencionan que los métodos de minería de textos implican técnicas como la categorización de texto, el resumen, la recuperación de información, la agrupación de documentos, la detección de temas y la extracción de conceptos. Para la clasificación y extracción de las reglas de asociación, los autores presentan dos marcos novedosos con el objetivo de definir tanto la estructura general como el sentimiento dentro de un corpus acumulado; la metodología desarrollada en el trabajo de investigación tiene como objetivo la

extracción de las reglas de asociación y visualización de sentimientos positivos y negativos en el texto financiero árabe planteando inicialmente el modelo de análisis de sentimiento árabe, seguido por el pre-procesamiento del texto árabe, a continuación se clasifica el texto en sentimientos positivos o negativos y se evalúa el modelo para dar cumplimiento al primer marco metodológico, después se crea el modelo de reglas de asociación de texto árabe, seguido del pre-procesamiento del texto, la búsqueda del conjunto frecuente de elementos para crear y visualizar reglas de asociación árabe para cada clase, dando así cumplimiento del segundo marco metodológico.

Los investigadores Jain & Kumar (2018), han propuesto una técnica inteligente basada en la teoría de conjuntos aproximados para identificar al sospechoso del virus de la gripe porcina (H1N1) al considerar los síntomas significativos indicados en los tweets. Las dos sub tareas principales del trabajo propuesto son: Identificación de atributos condicionales (síntomas) importantes de los tweets y el descubrimiento de reglas de decisión que caracterizan la dependencia entre los valores de los atributos condicionales y los atributos de decisión. El método propuesto por los autores se basa en la identificación de tweets relevantes que contienen síntomas relacionados con el H1N1 y se analiza utilizando la teoría de conjuntos aproximados para la identificación de casos sospechosos. A partir de una gran cantidad de atributos del conjunto de datos, se deben extraer atributos significativos que apoyen la toma de decisiones. El sistema se divide en cinco fases: fase de recolección de datos, preprocesamiento de tweets, identificación de síntomas, sistema de información y fase de identificación de sospechosos. En el procesamiento de los datos se aplican técnicas de clasificación como SVM, Naïve Bayes, bosques aleatorios y árbol de decisión. Los resultados se explican en términos de exactitud, sensibilidad, precisión y medida F1 (proporciona el rendimiento general de un clasificador).

B. Capacidad

La capacidad se relaciona con el desarrollo de los programas en la práctica y permite la recopilación, el intercambio, la evaluación adecuada y la comunicación de los datos en amplias zonas geográficas. Los programas de redes sociales permiten la interacción entre las redes internacionales de salud pública, seguridad alimentaria y otros profesionales a través de redes virtuales, para facilitar intervenciones y favorecer la infraestructura de respuesta de salud pública (Greene, 2010; Lopes et al., 2010).

Dentro de algunos ejemplos que capturan datos de Twitter para analizar la correlación de dos o más fuentes de información, se destacan: Aslam et al., (2014), quienes, durante la temporada 2013-2014 recopilan los tweets que contienen la palabra clave "gripe" generados por once ciudades de Estados Unidos para mejorar la correlación de los tweets con las tasas de enfermedades tipo influenza (ILI) proporcionadas por los centinelas de cada ciudad. Al final del periodo de recolección, usan 159.802 tweets para los análisis de correlación con las tasas de ILI proporcionadas por el Departamento de salud de la ciudad o condado correspondiente. Los dos métodos usados para observar las correlaciones son: filtrar los tweets por tipo (no retweets, retweets, tweets con un URL, tweets sin URL) y el uso de un clasificador de aprendizaje automático que determina si el Tweet es "válido" o de un usuario que probablemente está enfermo de gripe. El clasificador de aprendizaje automático arroja las mayores correlaciones para muchas de las ciudades para casos de influenza confirmados y tweets que el clasificador determina como válidos. Finalmente, el estudio demuestra una mayor precisión en el uso de Twitter como herramienta de vigilancia complementaria para la influenza con el uso de SVM como método de filtrado y clasificación, el cual obtiene mayores correlaciones para la temporada de influenza.

Allen, Tsou, Aslam, Nagel, & Gawron (2016) en su investigación, presentan un marco mejorado para monitorear brotes de influenza usando la plataforma de red social Twitter, por medio de métodos de filtrado espacial especificando latitud, longitud y un radio proporcionado por la interfaz de programación de aplicaciones de búsqueda de Twitter; así recopilan, filtran y analizan mensajes de Twitter de las treinta ciudades más pobladas de Estados Unidos durante la temporada de gripe 2013-2014. El ruido de los datos es tratado al excluir retweets y URL, para desarrollar un modelo de clasificación de aprendizaje automático utilizando SVM; para entrenar el modelo utilizan tweets al azar que contienen la palabra "gripe", realizando previamente transformaciones estadísticas (TF-IDF) para cada palabra o para cada par de palabras, este procedimiento permite al clasificador reconocer automáticamente palabras clave y así identificar mejor los tweets que indican casos reales de influenza y aquellos mensajes que son irrelevantes para la enfermedad. Los resultados del procedimiento se comparan con los informes de brotes de gripe nacionales, regionales y locales, que revelan una correlación estadísticamente significativa entre las dos fuentes de datos.

C. Ventajas

Las ventajas de los programas de vigilancia para las enfermedades infecciosas son: identificación en tiempo real, acceso público, gratuitos, interfaz familiar y facilidad de uso; de igual manera se debe tener en cuenta que monitorear preocupaciones de salud pública con los sistemas de vigilancia tradicionales es costoso, poseen cobertura limitada, ruido, ambigüedad y demoras significativas. Entonces, para abordar estas desventajas, existen mensajes de Twitter sin costo, se publican en tiempo real y se generan en todo el mundo y este contenido se analiza por medio de los enfoques de aprendizaje automático buscando descubrir información relevante y generar conocimiento (Ji

et al., 2015; Lampos, Yom-Tov, Pebody, & Cox, 2015). Así, la capacidad de procesar grandes volúmenes de datos automáticamente utilizando el procesamiento de lenguaje natural y algoritmos de aprendizaje automático, han abierto nuevas oportunidades para abordar algunas limitaciones relacionadas con la vigilancia tradicional (Sarker et al., 2015).

Sparks, Robinson, Power, Cameron, & Woolford (2017), mencionan que la red social Twitter ofrece datos e información sobre las etapas iniciales de una enfermedad en una determinada población por medio de los tweets publicados en las cuentas personales, siendo posible identificar los primeros síntomas de malestar inclusive antes de ingresar a un centro de salud, además de que Twitter puede proporcionar información adicional que no está disponible en los informes de dichos Departamentos.

También cabe destacar que Jain & Kumar (2018), en su investigación enfocada a identificar casos sospechosos de H1N1, resaltan que los datos de las redes sociales ofrecen desafíos únicos y oportunidades para el monitoreo y la vigilancia de la salud pública, y teniendo en cuenta que el retraso en la identificación del comienzo de una epidemia infecciosa da como resultado un gran daño a la sociedad, entonces, para manejar los casos de epidemia de manera efectiva, se necesita un sistema de diagnóstico de bajo costo, preciso y oportuno. La Internet es uno de los recursos más importantes en el campo de los sistemas de vigilancia para el seguimiento de brotes de enfermedades, ya que brinda una oportunidad para el cálculo rápido y de bajo costo de los datos en comparación con los sistemas de vigilancia tradicional existentes. Nair, Shetty, & Shetty (2018) desarrollan un sistema de predicción del estado de la salud en tiempo real basado en el big data, implementando un modelo de aprendizaje automático. En este sistema, el usuario por medio de

Twitter expresa su estado de salud y la aplicación recibe la información en tiempo real estableciendo una conexión con la API de Twitter; a partir de estos datos se extraen los atributos y se aplica el modelo de árbol de decisión, para clasificar a un usuario con presencia o ausencia de enfermedad, en función de sus atributos de salud. A partir de estos resultados, con un solo tweet, un usuario recibe información sobre su estado de salud de forma instantánea sin costo adicional.

D. Desafíos

Finalmente, los desafíos que conlleva el uso de las redes sociales como fuente de información para la vigilancia, se relacionan con la confiabilidad de los datos y su validez, por tanto Brownstein, Freifeld, & Madoff, (2010); Carneiro & Mylonakis, (2009); Greene, (2010) recomiendan la necesidad de filtrar el ruido de fondo para garantizar la confiabilidad de los datos y evitar sesgos; además hay segmentos de poblaciones sin acceso a la internet, especialmente en los países en desarrollo y otro desafío se relaciona con la propiedad de los datos, la privacidad y confidencialidad de la información del usuario.

Los datos a gran escala permiten realizar análisis más detallados de información permitiendo implementar técnicas de aprendizaje automático destacando el uso de algoritmos de clasificación supervisada tales como: clasificadores bayesianos, SVM, arboles de decisión y vecinos más cercanos, para sintetizar y filtrar la información no estructurada de las fuentes basadas en la web, siendo importante, el uso de técnicas de limpieza y captura de datos para filtrar el ruido y garantizar la confiabilidad de la información obtenida. Así mismo, se resalta el uso de la minería de reglas de asociación para encontrar relaciones entre la información, favoreciendo la toma de decisiones y las intervenciones en tiempo real.

Por consiguiente, se propone emplear en esta investigación las SVM como modelo de clasificación de la información capturada a partir de Twitter, sin olvidar aplicar técnicas de limpieza de datos adecuadas para filtrar información irrelevante y garantizar la confiabilidad de la misma. Además, aplicar una técnica de minería de datos, tal como las reglas de asociación, con el fin de identificar las relaciones entre la información capturada a través de la red social y extraer conocimiento válido, que permita responder a la naturaleza de los datos digitales y ser una fuente complementaria de vigilancia tradicional en salud pública.

4. Marco teórico

4.1 Minería de datos

La minería de datos estudia métodos y algoritmos, y nace como una disciplina con capacidad de extracción automática de información para el adecuado análisis de grandes cantidades de datos e información.

Además, combina técnicas semiautomáticas de inteligencia artificial, análisis estadístico, bases de datos y visualización gráfica, para la obtención de información que no esté representada explícitamente en los datos (Beltrán, 2014, p.18). La minería de datos con el propósito de soportar los procesos de toma de decisiones con un mayor conocimiento, descubre relaciones, tendencias, desviaciones comportamientos atípicos y patrones y así se puede ubicar a esta disciplina en el nivel más alto de los procesos tecnológicos de análisis de datos (Beltrán, 2014). El término Minería de Datos Inteligente refiere específicamente a la aplicación de métodos de aprendizaje automático basados en la estadística, para descubrir y enumerar patrones presentes en los datos. Cabe resaltar

que en la medida en que se incrementa la cantidad de información almacenada en las bases de datos, los métodos estadísticos empiezan a enfrentar problemas de eficiencia y escalabilidad y es aquí donde aparece el concepto de minería de datos (Felgaer et al., 2003).

Esta disciplina es un conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y comprensible, a partir de grandes conjuntos de datos, con objeto de predecir de forma automatizada tendencias y comportamientos, y describir de forma automatizada modelos previamente desconocidos (Perichinsky et al., 2003, p.2). Una de las diferencias entre al análisis de datos tradicional y la minería de datos es que el primero supone que las hipótesis ya están construidas y validadas contra los datos, mientras que el segundo supone que los patrones e hipótesis son automáticamente extraídos de los datos. Las tareas de la minería de datos se pueden clasificar en dos categorías: minería de datos descriptiva y minería de datos predicativa (Felgaer et al., 2003).

4.2 Minería de texto

La minería de texto es el área de investigación más reciente del procesamiento automático de textos, y es definida como el proceso automático de descubrimiento de patrones interesantes en una colección de textos. Estos patrones no deben existir explícitamente en ningún texto de la colección y deben surgir de relacionar el contenido de varios de ellos (Gómez, 2005, p.64).

El proceso de minería de texto consiste de dos etapas principales: una etapa de pre-procesamiento y una etapa de descubrimiento. La primera etapa, tiene como objetivo transformar los documentos de entrada a información estructurada o semiestructurada para que sean más

consistentes y facilitar la representación para su posterior análisis (Xia & Huan, 2013, p.388), mientras que en la segunda etapa las representaciones intermedias se analizan con el objetivo de descubrir en ellas algunos patrones interesantes o nuevo conocimiento. Entonces, el tipo de métodos aplicados en la etapa de pre-procesamiento, define el tipo de representaciones intermedias construidas, y en función de dicha representación se determinan los métodos usados en la etapa de descubrimiento para la identificación de patrones (Gómez, 2005, p.64).

4.2.1 Preprocesamiento de texto. Tiene como objetivo hacer que los documentos de entrada sean más consistentes para facilitar la representación de texto, lo cual es necesario para la mayoría de las tareas de análisis de texto. La recopilación de datos debe ir acompañada de una limpieza e integración para que estén en condiciones óptimas de análisis; la limpieza de datos tiene como fin, eliminar el mayor número posible de datos erróneos o inconsistentes e irrelevantes para la aplicación de técnicas de minería de datos. Esta limpieza, en muchos casos puede detectar y solucionar problemas de datos no resueltos durante la fase de integración, como valores anómalos y faltantes.

Los métodos tradicionales de preprocesamiento de texto incluyen la eliminación de palabras de parada (en inglés, stop word removal) y la derivación (en inglés, Stemming). Stop words removal se encarga de eliminar palabras que tienen alta frecuencia pero no representan el tema del cual trata un determinado documento; Stemming reduce las palabras inflexionadas (o algunas veces derivadas) a su raíz, base o forma de raíz (Xia & Huan, 2013, p.388).

La forma más común de modelar documentos es transformarlos en vectores numéricos dispersos y luego tratarlos con operaciones algebraicas lineales. Esta representación se llama bolsa de palabras (por sus siglas en inglés BOW) o Modelo de espacio vectorial "Vector Space Model" (por sus siglas en inglés, VSM). En estos modelos básicos de representación de texto, la estructura lingüística dentro del texto es ignorada y conduce a una "maldición estructural"; en el modelo BOW, una palabra se representa como una variable separada que tiene un peso numérico de importancia variable. El esquema de ponderación más popular es Frecuencia de términos / Frecuencia de documentos inversos (por sus siglas en inglés, TF-IDF) (Xia & Huan, 2013, p.389).

4.3 Reglas de asociación

Son técnicas populares que se centran en expresar patrones de datos de una base de datos; estos patrones son de utilidad para conocer el comportamiento general del problema que genera la base de datos permitiendo tener más información que pueda apoyar la toma de decisiones. Estas reglas expresan patrones de comportamiento entre los datos en función de la aparición conjunta de valores de dos o más atributos y su característica principal es tratar con atributos nominales (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, p.237). Las reglas de asociación tienen importantes aplicaciones prácticas en diversas áreas tales como redes de telecomunicaciones, gestión de riesgos y mercados, y control de inventarios (Kotsiantis & Kanellopoulos, 2006); estas aplicaciones por lo general tienen asociadas gran volumen de datos, por lo que la eficiencia es un factor clave en el aprendizaje de reglas de asociación. Los algoritmos de aprendizaje trabajan en la búsqueda de reglas que cumplan unos requisitos mínimos en estas medidas. El aprendizaje de reglas de asociación se divide normalmente en dos fases: la extracción de los conjuntos de ítems que cumplen con la cobertura requerida desde los datos, y la generación de las reglas con base en estos

conjuntos; la siguiente fase consiste en la creación de reglas a partir de conjuntos de ítems frecuentes (Orallo et al., 2004).

Las reglas de asociación son declaraciones si / cuando, que ayudan a exponer las relaciones entre datos aparentemente no relacionados, es decir, una regla de asociación tiene dos componentes, un antecedente (si) y un consecuente (luego). Un antecedente es un conjunto de elementos que se encuentra en el conjunto de datos, y un consecuente es un conjunto de elementos encontrado en la incorporación con el antecedente (AL-Rubaiee et al., 2017). Dentro de las métricas de evaluación de las reglas, se tienen:

4.3.1 Cobertura o soporte (support). Se define como el número de instancias que la regla predice correctamente (también se utiliza el porcentaje) (Orallo et al., 2004), es decir, indica el número de veces que la regla se podría aplicar correctamente.

A continuación, se presenta un ejemplo de esta medida (Tabla 2), en donde cada fila hace referencia a todos los productos comprados bajo una misma cesta de compra. Las letras A, B, C y D hacen referencia a 4 productos (ítems) distintos. Puede observarse que, los artículos A y B, aparecen en 2 de las 5 cestas de compra, los artículos B y C en 3 de las 5 cestas, por lo tanto, el soporte de los ítemset {A, B} es del 40% y del ítemset {B, C} del 60%.

$$soporte(A \rightarrow B) = P(A \cup B)$$

Tabla 2
Ejemplo de medida Soporte

TID	Ítems	Soporte
1	ABC	
2	ABD	Soporte {AB} = 2/5 = 40%
3	BC	Soporte {BC} = 3/5 = 60%
4	AC	Soporte {ABC} = 1/5 = 20%
5	BCD	

4.3.2 Confianza (confidence). También llamada precisión, mide el porcentaje de veces que la regla se cumple cuando se puede aplicar (Orallo et al., 2004), es decir, cuándo se expresa por la siguiente ecuación:

$$confianza(A \rightarrow B) = P(B|A) = \frac{soporte(A \cup B)}{soporte(A)}$$

A continuación**,** se presenta un ejemplo de esta medida (Tabla 3), en donde, de las 3 cestas de compra que incluyen A, 2 incluyen B, por lo tanto, la regla "clientes que compran el artículo A también compran B", se cumple, acorde a los datos, un 66% de las veces. Esto significa que la confianza de la regla {A → B} es del 66%. Así mismo, de las 2 cestas de compra que incluyen A y B, 1 incluye C, por lo tanto, la regla "clientes que comprar los artículos A y B también compran C", se cumple, acorde a los datos, un 50% de las veces. Esto significa que la confianza de la regla {AB → C} es del 50%.

Tabla 3
Ejemplo de medida Confianza

TID	Ítems	Confianza
1	ABC	
2	ABD	Confianza {A → B} = 2/3 = 66%
3	BC	Confianza {B → C} = 3/4 =75%
4	AC	Confianza {AB → C} = 1/2 = 50%
5	BCD	

4.3.3 Mejora de la confianza (Lift). Mide la proporción entre la confianza de una regla y la confianza esperada para el producto consecuente (Hernández Orallo & Ferri Ramírez, 2006), y se representa por la siguiente expresión:

$$lift(A \rightarrow B) = \frac{soporte(A \cup B)}{soporte(A) * soporte(B)}$$

4.4 Algoritmo Apriori

Su nombre se debe a que usa conocimiento apriori para la generación de conjuntos de ítems frecuentes (De Moya Amaris & Rodríguez Rodríguez, 2003, p.98). El funcionamiento de este algoritmo se basa en la búsqueda de los conjuntos de ítems con determinada cobertura. Para ello, en primer lugar, se construyen simplemente los conjuntos formados por solo un ítem que supera la cobertura mínima. Este conjunto se utiliza para construir conjuntos de dos ítems, y así sucesivamente hasta que se llegue a un tamaño en el cual no existan conjuntos de ítems con la cobertura requerida (Orallo et al., 2004, p.240).

Normalmente el aprendizaje de reglas de asociación se divide en dos fases:

- La extracción de los conjuntos de ítems que cumplan con la mínima cobertura requerida de los datos.
- La generación de las reglas a partir de estos conjuntos frecuentes.

4.5 Algoritmo FP-Growth

Se basa en el crecimiento o extensión de los conjuntos frecuentes de ítems. En un primer recorrido de la colección de datos, se obtienen los conjuntos frecuentes de ítems de tamaño uno (1). En un segundo recorrido de la colección de datos, se inserta cada conjunto de ítems, con los ítems

ordenados descendentemente de acuerdo a su soporte, en una estructura de datos compacta denominada FP-tree, también conocida como árbol de prefijos. De esta forma, prefijos iguales, de conjuntos de ítems diferentes, comparten la misma rama del árbol. Luego, a partir de esta estructura se generan los conjuntos de ítems frecuentes recorriendo recursivamente las ramas del árbol (Rodríguez González, Martínez Trinidad, Carrasco Ochoa, & Ruiz Shulcloper, 2009, p.6).

4.6 Algoritmo ECLAT

ECLAT (Equivalence Class Transformation) introducido en 1997; realiza una búsqueda profunda y adopta una posición vertical para representar el conjunto de elementos de la base de datos, en el que cada componente está representado por un ID transaccional denominado TIDset y sus transacciones contienen los elementos (Zaki, Parthasarathy, Ogihara, & Li, 1997). Se basa en realizar una agrupación en clúster entre los elementos para acercarse a los conjuntos de elementos frecuentes, lo que reduce el proceso de generación de conjuntos de elementos candidatos (Andrea, Hernández, Milena, Herrera, & Rodríguez Rodríguez, 2016, p.213).

4.7 Aprendizaje automático (en inglés, Machine Learning).

Es una rama de la inteligencia artificial que permite construir modelos con la capacidad de generalizar comportamientos a partir del suministro de información con el objetivo de detectar patrones. Se ocupa de desarrollar técnicas capaces de aprender, es decir, extraer de forma automática conocimiento subyacente en la información. Constituye junto con la estadística el corazón del análisis inteligente de los datos. Los principios seguidos en el aprendizaje automático y en la minería de datos son los mismos: la máquina genera un modelo a partir de ejemplos y lo usa para resolver el problema (Ruiz Sánchez, 2006, p.12).

Según Mitchell (1997) “un programa de computadora aprende de la experiencia E con respeto a alguna clase de tareas T y medida de rendimiento P, si su desempeño en tareas en T, medida por P, mejora con la experiencia E” (p. 2).

4.7.1 Aprendizaje no supervisado. Es un método de aprendizaje automático en donde un modelo es ajustado a las observaciones, se caracteriza porque no hay un conocimiento apriori y trata los objetos de entrada como un conjunto de variables aleatorias, siendo construido un modelo de densidad para el conjunto de datos. En este caso, los datos de entrada no vienen acompañados de una salida deseada, siendo objetivo encontrar patrones que ocurran con más generalidad (Palma, Tomás, & Morales, 2008), es decir se desea descubrir si hay agrupaciones y diferencias entre estos grupos.

En el aprendizaje no supervisado, se han desarrollado algoritmos de agrupamiento, entre los que se pueden señalar K-medias, agrupamiento espacial basado en densidad de aplicaciones con ruido (DBSCAN), pedido de puntos para identificar la estructura de agrupación (OPTICS), Máxima Esperanza (EM), agrupación aplicada a grandes datos (CLARA).

Por ejemplo, las redes neuronales con aprendizaje no supervisado (también conocido como auto supervisado) no requieren influencia externa para ajustar los pesos de las conexiones entre sus neuronas. La red no recibe ninguna información por parte del entorno que le indique si la salida generada en respuesta a una determinada entrada es o no correcta. Estas redes deben encontrar las características, regularidades, correlaciones o categorías que se puedan establecer entre los datos que se presenten en su entrada. Existen varias posibilidades en cuanto a la interpretación de la

salida de estas redes, que dependen de su estructura y del algoritmo de aprendizaje empleado (Matich, 2001, p.21).

4.7.2 Aprendizaje supervisado. Es una formalización de la idea de aprender a partir de ejemplos; el modelo recibe dos conjuntos de datos, un conjunto de entrenamiento y un conjunto de prueba. El objetivo del modelo es desarrollar una regla, un programa o un procedimiento para clasificar ejemplos sin etiqueta (en el conjunto de prueba) con la mayor precisión posible a partir de un conjunto de ejemplos etiquetados (en el conjunto de entrenamiento) (Erik G., 2014). Esas relaciones sirven para realizar la predicción de datos cuya etiqueta es desconocida (Rodríguez & Díaz, 2009, p.76).

El algoritmo de aprendizaje supervisado parte de una serie de observaciones para optimizar los parámetros de entrada y producir una salida muy similar a la del modelo de entrenamiento (Palma et al., 2008). Dado un conjunto de ejemplos de entrada X y su correspondiente valor deseado Y para cada uno de ellos, se trata de establecer una función (modelo/hipótesis) f(), que permita predecir el valor Y* =f(X*) para nuevos ejemplos X* . Si f() es una función discreta, se habla de clasificación supervisada; en general, clasificación multiclase, y si f() sólo puede tomar dos valores (ejemplos positivos y negativos) se habla de clasificación biclase o aprendizaje de conceptos (concept learning) ; si f() es una función continua, se habla de modelos de regresión ,si f() es una probabilidad, se habla de estimación de distribuciones de probabilidad (F. D. Gómez, 2014).

Dentro de los algoritmos de aprendizaje supervisado más utilizados se encuentran: SVM, Regresión Lineal, Regresión Logística, Clasificador Naive Bayes, Análisis discriminante lineal, Árbol de decisión, K-vecinos más cercanos y Redes neuronales.

4.8 Máquinas de soporte vectorial.

Las SVM son un conjunto de algoritmos pertenecientes a la rama de aprendizaje supervisado que están relacionados con problemas de clasificación y regresión en el cual dado un conjunto de muestras se construye un modelo que prediga nuevos conjuntos de muestras.

Una máquina de soporte vectorial primero mapea los puntos de entrada a un espacio de características de una dimensión mayor (i.e.: si los puntos de entrada están en $\Re^2$ entonces son mapeados por la SVM a $\Re^3$) y encuentra un hyperplano que los separe y maximice el margen m entre las clases en este espacio como se observa en la Figura 2.

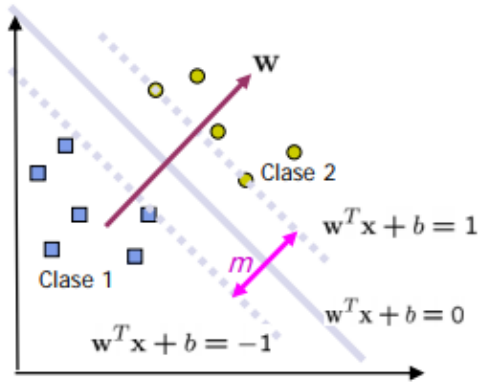

Figura 2. Máquinas de Soporte Vectorial. Adaptado de (Betancour, 2005, p.67). http://revistas.utp.edu.co/index.php/revistaciencia/article/view/6895/4139

Maximizar el margen *m* es un problema de programación cuadrática (QP) y puede ser resuelto por su problema dual introduciendo multiplicadores de Lagrange. Sin ningún conocimiento del

mapeo, la SVM encuentra el hyperplano óptimo utilizando el producto punto con funciones en el espacio de características que son llamadas *kernels*. La solución del hyperplano óptimo puede ser escrita como la combinación de unos pocos puntos de entrada que son llamados vectores de soporte (Betancour, 2005, p.67). Así, el problema de encontrar el hiperplano equidistante a dos clases implica realizar la siguiente optimización con restricciones:

$$Maximizar\ \frac{1}{\|\omega\|}$$

$$Sujeto\ a\colon y_i(<\omega, x_i> + b) \geq 1$$

$$1 \leq i \leq N$$

Dónde: ω es el vector ortogonal al hiperplano $h(x) =< \omega, x > +b$ en un espacio D-dimensional $\Re^D$, b expresa el producto escalar habitual en $\Re^D$ y $\|\omega\|$ es la norma en $\Re^D$ asociada al producto escalar (es decir, $\|\omega\|^2 =< \omega, \omega >$).

Una Máquina de Soporte Vectorial (por sus siglas en inglés, SVM) aprende la superficie decisión de dos o más clases distintas de los puntos de entrada. Como un clasificador de una sola clase, la descripción dada por los datos de los vectores de soporte es capaz de formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con muy poco o ningún conocimiento de los datos fuera de esta frontera. Los datos son mapeados por medio de un kernel Gaussiano u otro tipo de kernel a un espacio de características en un espacio dimensional más alto, donde se busca la máxima separación entre clases. Esta función de frontera, cuando es traída de regreso al espacio de entrada, puede separar los datos en todas las clases distintas, cada una formando un agrupamiento (Betancour, 2005, p.67).

4.8.1 Función kernel. La construcción de las SVM se basa en la idea de transformar o proyectar un conjunto de datos pertenecientes a una dimensión *n* dada, hacia un espacio de dimensión superior aplicando una función kernel - kernel Trick (Alpaydin, 2010). El cual consiste en que, a partir del nuevo espacio creado, se operarán los datos como si se tratase de un problema de tipo lineal, resolviendo el problema sin considerar la dimensionalidad de los datos.

Como se observa en la Figura 3 en el espacio de entrada los datos no son linealmente separables, pero al aplicar una función kernel existe la posibilidad de transformar los datos a un espacio de mayor dimensión (espacio de características) en el que los puntos si puedan ser separados por un hiperplano.

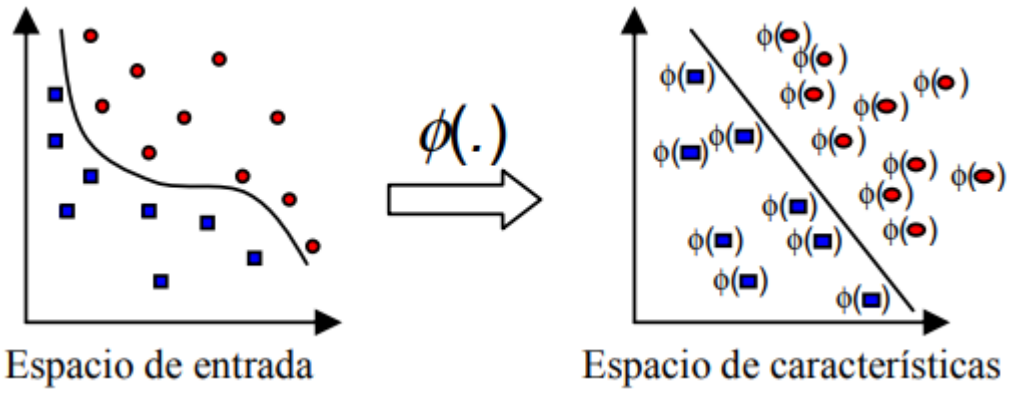

Figura 3. Idea del uso de un kernel para transformación del espacio de los datos. Adaptado de Gustavo A. Betancourt (2005). http://revistas.utp.edu.co/index.php/revistaciencia/article/view/6895/4139

Los algoritmos de SVM utilizan diferentes tipos de funciones del kernel. Estas funciones pueden ser de diferentes tipos. Por ejemplo, lineal, polinomial, función de base radial (RBF) y sigmoide. Un kernel polinomial, por ejemplo, permite modelar conjunciones de características según el orden del polinomio. Las funciones de base radial permiten seleccionar círculos (o hiperesferas), en contraste con el kernel lineal, que solo permite seleccionar líneas (o hiperplanos).

4.9 Validación cruzada

Es una técnica estadística de validación de modelos que permite evaluar cómo los resultados de un análisis estadístico se generalizarán a un conjunto de datos de prueba (Pérez-Planells, Delegido, Rivera-Caicedo, & Verrelst, 2015). Usualmente es aplicada en modelos predictivos y su principal objetivo es probar la capacidad del modelo para predecir nuevos datos que no se usan en la etapa de entrenamiento, con el fin de evitar problemas de sobreajuste o sesgo de selección. Esta técnica es recomendada como una de las mejores formas de probar un modelo (Rueda Restrepo & Cotes Torres, 2009), pues permite identificar los parámetros que mejor se ajustan al modelo bajo las condiciones particulares de cada conjunto de datos.

4.9.1 Validación cruzada K-fold. Una iteración de la validación cruzada K-fold se realiza de la siguiente manera: primero, la muestra original se divide en K subconjuntos de aproximadamente el mismo tamaño. De los K subconjuntos, un solo subconjunto se retiene como datos de validación para probar el modelo (este subconjunto se llama "testset"), y los subconjuntos K - 1 restantes se usan juntos como datos de entrenamiento ("trainset"). La capacitación y evaluación del modelo se repite K veces, y cada uno de los K subconjuntos se usa exactamente una vez como conjunto de prueba (FH Joanneum, 2005). El caso de una validación cruzada de 10-fold se ilustra en la Figura 4.

La forma básica de validación cruzada es la validación cruzada k-fold. Otras formas de validación cruzada son los casos especiales de validación cruzada aleatoria, validación cruzada dejando uno fuera (en inglés, LOOCV) o repetidas rondas de validación cruzada k-fold (Payam, Lei, & Huan, 2008).

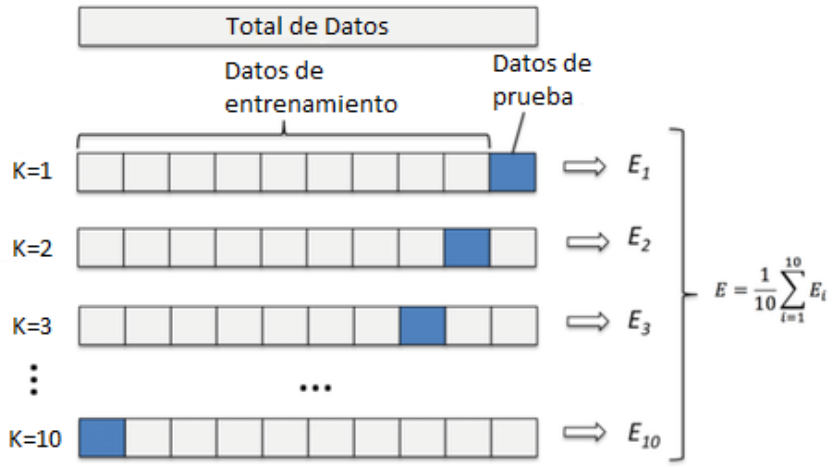

Figura 4. Validación Cruzada K-fold. Modificado de https://www.researchgate.net/figure/An-example-of-a-10-fold-cross-validation-cro17_fig1_322509110

4.9.2 Validación cruzada aleatoria. Este método consiste en dividir aleatoriamente el conjunto de datos de entrenamiento y el conjunto de datos de prueba. Para cada división la función de aproximación se ajusta a partir de los datos de entrenamiento y calcula los valores de salida para el conjunto de datos de prueba. El resultado final se corresponde a la media aritmética de los valores obtenidos para las diferentes divisiones. La ventaja de este método es que la división de datos entrenamiento-prueba no depende del número de iteraciones. Pero, en cambio, con este método hay algunas muestras que quedan sin evaluar y otras que se evalúan más de una vez, es decir, los subconjuntos de prueba y entrenamiento se pueden solapar (Moore, 2001).

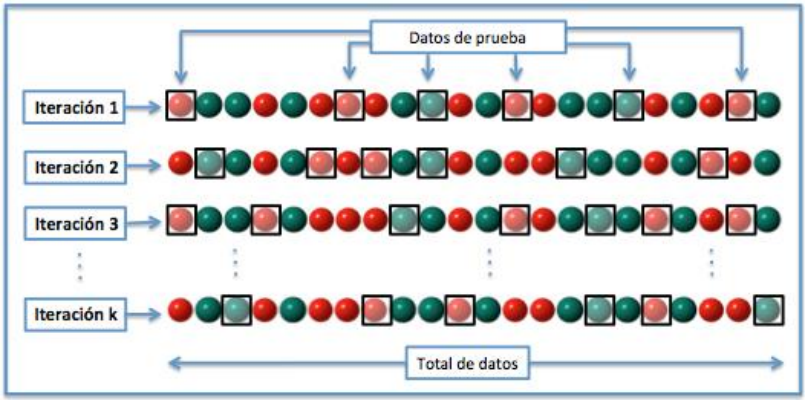

Figura 5. Validación cruzada aleatoria. Adaptado de Joanneum, 2005-2006

4.9.3 Validación cruzada dejando uno fuera. La validación cruzada dejando uno fuera o Leave-one-out cross-validation (LOOCV) implica separar los datos de forma que para cada iteración tengamos una sola muestra para los datos de prueba y todo el resto conformando los datos de entrenamiento. La evaluación viene dada por el error, y en este tipo de validación cruzada el error es muy bajo, pero en cambio, a nivel computacional es muy costoso, puesto que se tienen que realizar un elevado número de iteraciones, tantas como N muestras tengamos y para cada una analizar los datos tanto de entrenamiento como de prueba (Elkan, 2011).

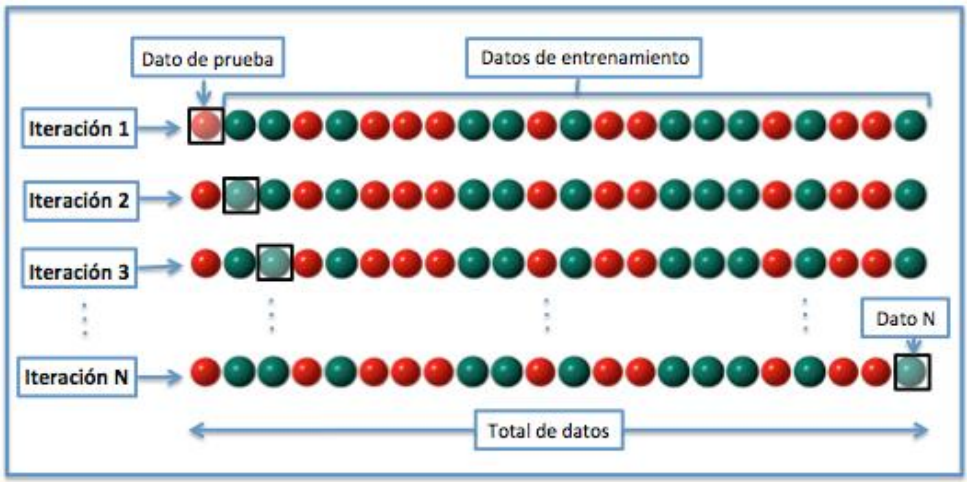

Figura 6. Validación cruzada dejando uno fuera (LOOCV). Adaptado de Joanneum, 2005-2006

4.10 Matriz de confusión

Una matriz de confusión (Kohavi y Provost, 1998) contiene información sobre las clasificaciones reales y las pronosticadas por un sistema de clasificación. El rendimiento de tales sistemas se evalúa comúnmente utilizando los datos de la matriz. La siguiente tabla muestra la matriz de confusión para un clasificador de dos clases.

Tabla 4
Matriz de Confusión

		Predicho	
		Negativo	**Positivo**
Real	**Negativo**	a	b
	Positivo	c	d

- a es el número de predicciones correctas de que una instancia es negativa
- b es el número de predicciones incorrectas de que una instancia es positiva
- c es el número de predicciones incorrectas que una instancia es negativa
- d es el número de predicciones correctas de que una instancia es positiva

Las métricas asociadas a la matriz de confusión para un clasificador de dos clases son las siguientes:

- Exactitud (Accuracy, AC): es la proporción del número total de predicciones que son correctas.

$$AC = \frac{a + d}{a + b + c + d}$$

- Sensibilidad (Recall or True positive rate): es la proporción de casos positivos que se identifican correctamente.

$$Recall = \frac{d}{c + d}$$

- Tasa de falsos positivos (False positive rate, FP): es la proporción de casos negativos que se clasifican incorrectamente como positivos.

$$FP = \frac{b}{a + b}$$

- Especificidad (Especificity or True negative rate, TN): se define como la proporción de casos negativos que se clasifican correctamente.

$$TN = \frac{a}{a + b}$$

- Tasa de falsos negativos (False negative rate, FN): es la proporción de casos positivos que se clasifican incorrectamente como negativos.

$$FN = \frac{c}{c + d}$$

- Precisión (P): es la proporción de casos positivos pronosticados que son correctos.

$$P = \frac{d}{b + d}$$

- F-Measure (F): es el promedio ponderado de la precisión y el recall.

$$F = \frac{2(Recall * P)}{Recall + P}$$

4.11 Salud pública

De acuerdo con la Ley 1122 de 2007 del Ministerio de Salud y Protección Social, la salud pública está constituida por un conjunto de políticas que busca garantizar de manera integrada, la salud de la población por medio de acciones dirigidas tanto de manera individual como colectiva ya que sus resultados se constituyen en indicadores de las condiciones de vida, bienestar y desarrollo. Dichas acciones se realizarán bajo la rectoría del Estado y deberán promover la participación responsable de todos los sectores de la comunidad (Ministerio de Salud, n.d.).

4.12 Enfermedades transmitidas por vectores.

Las ETV son enfermedades infecciosas transmitidas por mosquitos (vector) que están previamente infectados por el virus y se transmiten a una persona sana a través de la picadura de dicho mosquito.

Los vectores son animales que transmiten patógenos, entre ellos parásitos, de una persona (o animal) infectada a otra y ocasionan enfermedades graves en el ser humano. Las enfermedades vectoriales representan un 17% de la carga mundial estimada de enfermedades infecciosas. La más mortífera de todas ellas (el paludismo) causó 627 000 muertes en 2012. No obstante, la enfermedad de este tipo con mayor crecimiento en el mundo es el Dengue, cuya incidencia se ha multiplicado por 30 en los últimos 50 años (Organización Mundial de la Salud., n.d.).

Las ETV son problemas importantes de salud pública, particularmente en regiones tropicales y subtropicales y lugares donde el acceso a sistemas seguros de agua potable y saneamiento es un desafío. Se consideran enfermedades de los pobres, son endémicas en grupos de bajos ingresos o en áreas donde existe un círculo vicioso de enfermedades y pobreza. Algunos están asociados con un alto estigma, discriminación y tabú. Si bien todas estas enfermedades son prevenibles y curables, la carga de la enfermedad y el impacto económico son muy altos e inaceptables en esta era moderna. Algunas de estas enfermedades son fatales si no se tratan, mientras que otras dejan a los pacientes discapacitados (World Health Organization & Regional Office for South-East Asia, 2014).

4.12.1 Dengue. El Dengue es una infección vírica transmitida por la picadura de las hembras infectadas de mosquitos del género Aedes. Hay cuatro serotipos de Virus del Dengue (DEN 1, DEN 2, DEN 3 y DEN 4). El Dengue se presenta en los climas tropicales y subtropicales de todo el planeta, sobre todo en las zonas urbanas y semiurbanas. Los síntomas aparecen 3 y 14 días (promedio de 4 a 7 días) después de la picadura infectiva. El Dengue es una enfermedad que afecta a lactantes, niños pequeños y adultos (Organización Mundial de la Salud, n.d.-a).

Los síntomas son una fiebre elevada (40Cº) acompañada de dos de los siguientes: dolor de cabeza muy intenso, dolor detrás de los globos oculares, dolores musculares y articulares, náuseas, vómitos, agrandamiento de ganglios linfáticos o sarpullido (Organización Mundial de la Salud, n.d.-a).

4.12.2 Chagas. La enfermedad de Chagas, también llamada tripanosomiasis americana, es una enfermedad potencialmente mortal causada por el parásito protozoo *Trypanosoma cruzi*. Se encuentra sobre todo en zonas endémicas de 21 países de América Latina, donde se transmite a los seres humanos principalmente por las heces de insectos triatomíneos conocidos como *Triatoma infestans*, según la zona geográfica. La infección también se puede adquirir mediante transfusión de sangre, transmisión congénita (de la madre infectada a su hijo) y órganos donados, aunque estos modos de transmisión son menos frecuentes (Organización Mundial de la Salud, n.d.-b).

Los síntomas pueden ser fiebre, dolor de cabeza, agrandamiento de ganglios linfáticos, palidez, dolores musculares, dificultad para respirar, hinchazón y dolor abdominal o torácico en la fase aguda y trastornos cardiacos y alteraciones digestivas (típicamente, agrandamiento del esófago o del colon), neurológicas o mixtas en la fase crónica. Con el paso de los años, la infección puede causar muerte súbita o insuficiencia cardiaca por la destrucción progresiva del músculo cardiaco (Organización Mundial de la Salud, n.d.-b).

4.12.3 Zika. El Virus de Zika es un flavivirus transmitido por mosquitos que se identificó por vez primera en macacos, a través de una red de monitoreo de la fiebre amarilla. Posteriormente, en 1952, se identificó en el ser humano en Uganda y la República Unida de Tanzania. Se han

registrado brotes de enfermedad por este virus en África, las Américas, Asia y el Pacífico (Organización Mundial de la Salud., 2016).

El Virus de Zika se transmite a las personas principalmente a través de la picadura de mosquitos infectados del género Aedes, y sobre todo de *Aedes aegypti* en las regiones tropicales. Los mosquitos *Aedes* suelen picar durante el día, sobre todo al amanecer y al anochecer, y son los mismos que transmiten el Dengue, la fiebre Chikungunya y la fiebre amarilla. Asimismo, es posible la transmisión sexual, y se están investigando otros modos de transmisión, como las transfusiones de sangre (Organización Mundial de la Salud., 2016).

4.12.4 Chikungunya. Es una enfermedad vírica transmitida al ser humano por mosquitos. Se describió por primera vez durante un brote ocurrido en el sur de Tanzania en 1952. Se trata de un virus ARN del género alfavirus, familia Togaviridae. "Chikungunya" es una voz del idioma Kimakonde que significa "doblarse", en alusión al aspecto encorvado de los pacientes debido a los dolores articulares (Organización Mundial de la Salud, 2017).

El virus se transmite de una persona a otras por la picadura de mosquitos hembra infectados. Generalmente los mosquitos implicados son *Aedes aegypti* y *Aedes albopictus*, dos especies que también pueden transmitir otros virus, entre ellos el del Dengue. Estos mosquitos suelen picar durante todo el periodo diurno, aunque su actividad puede ser máxima al principio de la mañana y al final de la tarde. Ambas especies pican al aire libre, pero *Aedes. aegypti* también puede hacerlo en ambientes interiores. La enfermedad suele aparecer entre 4 y 8 días después de la picadura de un mosquito infectado, aunque el intervalo puede oscilar entre 2 y 12 días (Organización Mundial de la Salud, 2017).

4.12.5 Leishmaniasis. La Leishmaniasis es causada por un protozoo parásito del género Leishmania, que cuenta con más de 20 especies diferentes. Se conocen más de 90 especies de flebotominos transmisores de Leishmania. La enfermedad se presenta en tres formas principales: Leishmaniasis visceral, Leishmaniasis cutánea y Leishmaniasis mucocutánea (Organización Mundial de la Salud., 2018a).

La Leishmaniasis se transmite por la picadura de flebótomos hembra infectados. Su epidemiología depende de las características de la especie del parásito, las características ecológicas locales de los lugares donde se transmite, la exposición previa y actual de la población humana al parásito y las pautas de comportamiento humano. Hay unas 70 especies animales, entre ellas el hombre, que son reservorios naturales de *Leishmania* (Organización Mundial de la Salud., 2018a).

4.12.6 Malaria. Es causada por parásitos del género *Plasmodium* que se transmiten al ser humano por la picadura de mosquitos hembra infectados del género *Anopheles*, los llamados vectores del paludismo. Hay cinco especies de parásitos causantes del paludismo en el ser humano, si bien dos de ellas - *Plasmodium falciparum* y *Plasmodium vivax* - son las más peligrosas (Organización Mundial de la Salud., 2018b).

La Malaria es una enfermedad de alto poder epidémico que es endémica en una gran parte del territorio nacional, en áreas localizadas por debajo de los 1.500 m.s.n.m. Constituye un evento cuya vigilancia, prevención y control revisten especial interés en salud pública (Ministerio de Salud y Proteccion Social & Federacion Medica Colombiana, 2013).

5. Proceso de descubrimiento de conocimiento

A continuación, se detallan las actividades a realizar para la captura, pre-procesamiento y procesamiento de la información objeto de estudio.

5.1 Extracción de datos

La recopilación de datos se realiza a través de la API oficial de Twitter con apoyo del software R, herramienta que permiten capturar tweets de Colombia y el Departamento de Santander por medio de las palabras clave relacionadas con ETV, tales como: "Dengue", “Zika”, “Chagas”, “Malaria”, "Chikungunya" y “Leishmaniasis”, seleccionadas a partir del informe de indicadores básicos del Observatorio de Salud Pública de Santander (2017), el cual está diseñado para ser utilizado como material de consulta por parte de los entes territoriales, universidades, investigadores, entidades públicas y privadas relacionadas con la salud en el Departamento, en aras de contribuir al monitoreo, evaluación y planeación de su gestión en salud; la formulación de proyectos de investigación, planes y programas de salud, y el diseño de políticas públicas en salud.

Por tanto, la captura de la información se realiza teniendo en cuenta radios de búsqueda variables y coordenadas geográficas (Latitud y Longitud) de los centroides de los Departamentos a escala nacional y los centroides de los municipios de Santander. En la Tabla 5 y Tabla 6, se presentan los radios y coordenadas geográficas definidas para Colombia y Santander, respectivamente:

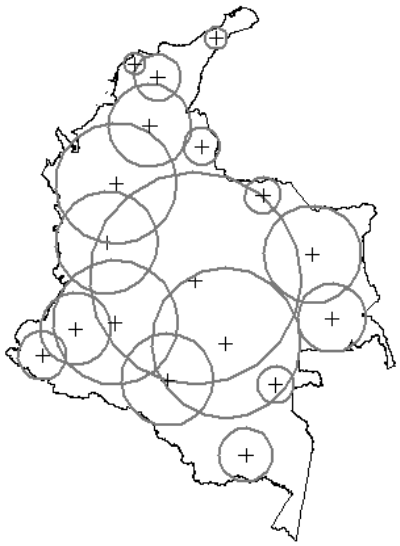

Figura 7. Radios definidos para Colombia

Tabla 5.
Coordenadas geográficas para radios definidos en Colombia

Colombia			
N°	**Latitud**	**Longitud**	**Radio (Km)**
1	3,9027	-73,08277	550
2	-1,546228	-71,50213	150
3	0,7985562	-73,95947	250
4	0,6462456	-70,56141	100
5	1,924532	-72,1286	400
6	2,727843	-68,81661	180
7	4,713557	-69,414	250
8	1,571093	-77,8702	130
9	2,396835	-76,82423	190
10	2,570143	-75,58435	330
11	6,569577	-70,96732	90
12	5,083342	-75,84215	260
13	6,922796	-75,56499	300
14	8,745178	-74,50852	200
15	8,095141	-72,88189	90
16	10,24738	-74,26176	110
17	10,67701	-74,96522	50
18	11,47687	-72,42951	50

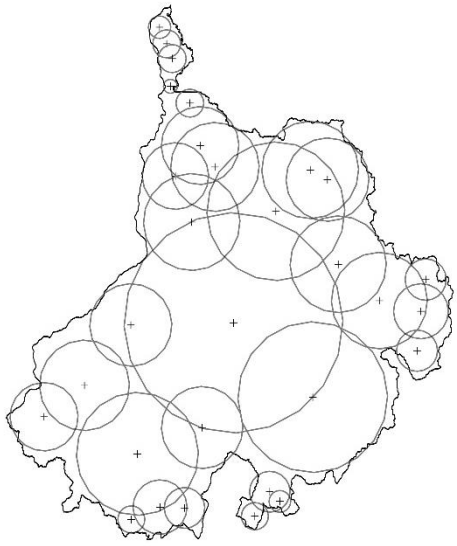

Figura 8. Radios definidos para Santander, Colombia

Tabla 6.
Coordenadas geográficas para radios definidos en Santander, Colombia

Santander			
N°	**Latitud**	**Longitud**	**Radio (Km)**
1	6,693633	-73,48601	80
2	7,406542	-73,57132	32
3	6,096727	-73,92718	45
4	6,40973	-74,1703	33
5	7,348778	-73,05439	30
6	6,568122	-72,64322	15
7	6,796483	-72,81613	35
8	5,924384	-73,3194	15
9	6,962456	-73,00517	35
10	7,20406	-73,29138	50
11	6,355369	-73,12212	55
12	6,217374	-73,62977	30
13	6,685279	-73,95761	30
14	5,854288	-73,82342	20

De acuerdo con los radios y las coordenadas identificadas tanto a nivel nacional como departamental, se recopilan en total 5.884 y 194 tweets, respectivamente, durante un periodo de 3 meses, distribuidos por cada enfermedad como se muestra en las siguientes tablas:

Tabla 7.
Número de Tweets recopilados en Colombia

Colombia	
Enfermedades Transmitidas por Vectores	**Tweets**
Dengue	2374
Malaria	1938
Leishmaniasis	159
Zika	971
Chagas	186
Chikungunya	256
Total	**5.884**

Tabla 8.
Número de Tweets recopilados en Santander

Santander	
Enfermedades Transmitidas por Vectores	**Tweets**
Dengue	147
Malaria	0
Leishmaniasis	0
Zika	39
Chagas	0
Chikungunya	8
Total	**194**

Así mismo, se recopilan tweets de entidades oficiales que desempeñan labores relacionadas con la salud pública en Colombia, tales como:

Tabla 9.
Entidades oficiales relacionadas con salud pública

Cuenta	**Usuario**	**Descripción**
Ministerio de Salud	MinSaludCol	Es un ente regulador que determina normas y directrices en materia de temas de salud pública, asistencia social, población en riesgo y pobreza.
Instituto Nacional de Salud	INSColombia	Organismo Público Ejecutor del Ministerio de Salud de Colombia cuya principal labor es la investigación de los problemas prioritarios de salud que afectan a la comunidad colombiana.
Asociación Colombiana de Salud Publica	saludpublicacol	Propicia el debate público, el intercambio de experiencias, moviliza e incide de forma efectiva en los grandes temas de la salud pública nacional y su relación con los diálogos regionales y mundiales.
Representación en Colombia de la Organización Panamericana de la Salud	OPSOMS_Col	Brinda cooperación técnica en protección social y salud, fortaleciendo la articulación intersectorial, para el mejoramiento de las condiciones de salud y bienestar de la población.

Continuación Tabla 9.
Entidades oficiales relacionadas con salud pública

Cuenta	Usuario	Descripción
Federación Médica Colombiana	FedemedicaC	Promueve el concepto de la salud como derecho fundamental, defiende los principios de Seguridad Social y Salud Pública como obligación del Estado.
World Mosquito Program Colombia	WMPColombia	World Mosquito Program es una iniciativa sin ánimo de lucro. El programa ha desarrollado un método de control biológico innovador y autosostenible para reducir la transmisión de virus como el Dengue, el Zika y el Chikungunya, por el mosquito Aedes aegypti.

A partir de las cuentas seleccionadas se obtiene un máximo de 3.200 tweets para cada uno de los usuarios especificados en la Tabla 9, por tanto, es necesario filtrar los resultados obtenidos por palabras clave previamente definidas para cada enfermedad y así obtener un total de 793 tweets distribuidos por enfermedad transmitida por vectores como se muestra en la Tabla 10:

Tabla 10.
Número de tweets recopilados de las cuentas oficiales

Enfermedades Transmitidas por Vectores	N° Tweets
Dengue	277
Malaria	88
Leishmaniasis	13
Zika	231
Chagas	80
Chikungunya	104
Total	**793**

Por lo tanto, como resultado de la extracción de datos, se obtiene una base de datos de 8 columnas y 6.677 tweets. Cada columna representa una variable y cada fila corresponde al contenido de cada tweet. En la Tabla 11 se describen las variables de la base de datos.

Tabla 11.
Variables de la base de datos

Variable	Descripción
Fecha	Fecha y hora de creación del tweet
Usuario	Nombre de usuario que publica el tweet
Texto	Texto del tweet publicado
Idioma	Idioma del tweet publicado
Ubicación	Ubicación geográfica del tweet
Radio	Radio de la zona a la que pertenece el usuario que publico el tweet
ETV	Etiqueta asignada por tipo de enfermedad transmitida por vector a la que pertenece el tweet
Clasificación	Etiqueta de válido o inválido para la clasificación del tweet.

En la base de datos se presenta la variable clasificación que corresponde a etiquetas asignadas a cada tweet con el fin de identificar qué cantidad de información de la publicada en esta red social aporta información importante a la vigilancia de ETV, y ser una fuente adicional para la infodemiología, o investigación digital de detección de enfermedades con el fin de apoyar la toma de decisiones en cuanto a planes estratégicos de salud pública y es una oportunidad para apoyar la vigilancia tradicional.

Allen et al., (2016) han demostrado en su investigación el uso de clasificación de aprendizaje automático utilizando SVM, con el objetivo de identificar aquellos tweets que indiquen casos reales de influenza. Siendo así, para la variable "Clasificación" se considera un tweet "válido" aquel que contenga términos o palabras que indiquen posibles casos de ETV y adicionalmente, su contenido posee términos relacionados con salud pública de ETV. Para el caso contrario, un tweet es considerado "inválido" cuando su contenido no indica la presencia de casos de ETV y posee términos jocosos, ofensivos e irrelevantes para el análisis. En la Tabla 12 se presentan ejemplos para aquellos tweets etiquetados como válidos e inválidos.

Tabla 12.
Ejemplo de clasificación de Tweets

Tweet	Clasificación
"JAJAJAJJAJAJA con solo ver esta foto me dio Dengue, Zika y Chikungunya"	Inválido
"Casi llenan el Palma Zika, un clásico."	Inválido
"Chéryshev en muletas y con Dengue hemorrágico hace más que Messi y Neymar en sus selecciones."	Inválido
"Comuna Norte, la más afectada con casos de Dengue en #Bucaramanga"	Válido
"#TopBLU Con masiva fumigación, Bucaramanga le declara la guerra al Dengue, Zika y Chikungunya"	Válido
"25 de 90 casos ya son positivos para Dengue vía @vanguardiacom"	Válido

5.2 Preprocesamiento de datos

La calidad del conocimiento descubierto no solo depende del algoritmo de minería utilizado, sino también de la calidad de los datos minados, estos deben ser íntegros, completos y consistentes, para obtener una información apta para analizar. La preparación de datos tiene como objetivo la eliminación del mayor número posible de datos innecesarios (truncados, duplicados, inconsistentes, irrelevantes), con el fin de presentar los datos de la manera más apropiada para la minería de datos y para la aplicación de modelos de aprendizaje automático (Orallo et al., 2004). Asimismo, se destaca la representación del texto como una matriz de término frecuencia (TF-IDF), con el fin de cuantificar la temática del texto, así como la importancia de cada término que lo conforma, así, se reduce la importancia de aquellos términos que aparecen en varios tweets y que, por lo tanto, no aportan información selectiva.

Para un total de 6.677 datos (Tweets) obtenidos como se detalla en la fase previa, con el apoyo del software R, se realiza la respectiva limpieza teniendo en cuenta los siguientes aspectos:

- Eliminación de tweets duplicados, teniendo en cuenta que de la información considerada algunos radios de búsqueda se superponen lo que trae como consecuencia repetición en la información obtenida
- Filtración de tweets por lenguaje (español), debido a que existen usuarios de otros países que realizan publicaciones en otro idioma en las coordenadas especificadas de búsqueda y afecta el preprocesamiento de los datos, es información relacionada con hechos específicos de la zona de la que proceden.
- Eliminación de caracteres extraños, signos de puntuación, dígitos, emoticones y espacios en blanco múltiples.
- Eliminación de páginas web (palabras que empiezan con "htttp.")
- Eliminación de palabras sin significado como artículos, pronombres, preposiciones, etc., lo cual recibe el nombre de palabras de parada Stopwords. porque son palabras más comunes en un idioma que no agregan información vital al estudio. Ejemplo: *de, la, que, el, los, su, una, le, para*, etc.
- Convertir el texto a minúscula.
- Reducir las palabras a su raíz (en inglés, Stemming), con el fin de agrupar aquellas palabras que pueden ser representadas por una raíz y sigue conservando el significado, por ejemplo: *acab* para las palabras *acaban, acaba, acabo* y *acabar*.

A partir de la limpieza realizada se obtiene un total de 2.209 tweets, distribuidos de la siguiente manera:

Tabla 13.
Número de Tweets finales por ETV

Enfermedades Transmitidas por Vectores	N° Tweets
Dengue	882
Malaria	533
Leishmaniasis	51
Zika	436
Chagas	113
Chikungunya	194
Total	**2209**

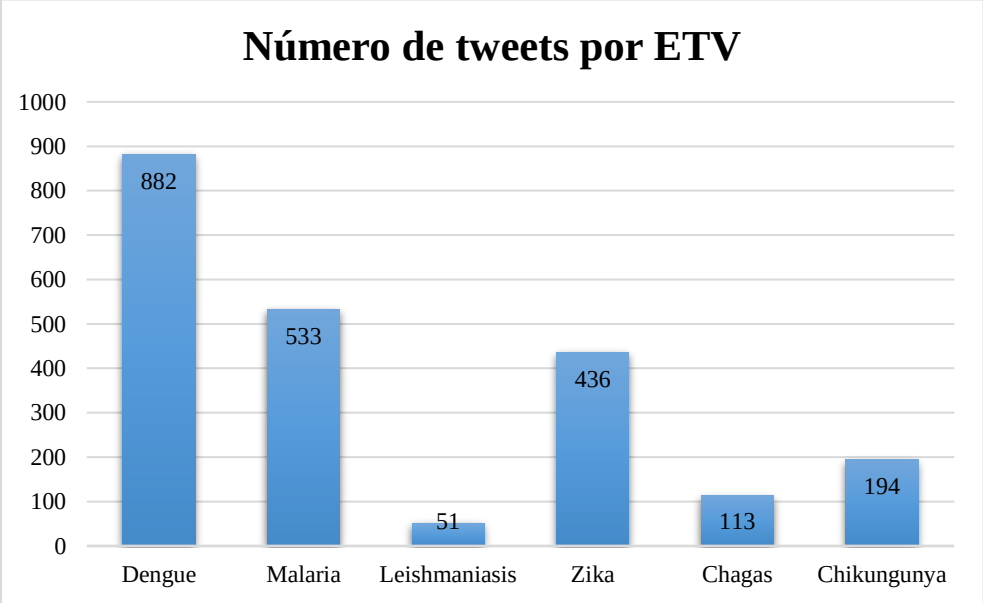

Figura 9. Distribución del Número de Tweets por ETV

5.3 Matriz termino-frecuencia (TF-IDF)

La frecuencia de termino – frecuencia inversa de documento (por sus siglas en inglés Tf-Idf, Term frequency – Inverse document frequency) es una medida numérica que expresa cuan relevante es una palabra para un documento en una colección. El valor de Tf-idf aumenta proporcionalmente a la frecuencia de una palabra (de ahí TF) pero es compensada por la frecuencia de esa palabra en la colección de documentos, lo que permite manejar el hecho de que algunos palabras son más

comunes en la colección que otras (Blanco & Sanz, 2016). Como Tf-idf es el producto de dos medidas, a continuación, se presenta su formulación.

La frecuencia de término (TF) calcula el número de repeticiones de una palabra *i* (término) en un documento *j*, algunos algoritmos de clasificación de texto normalizan la frecuencia de término al dividir con la frecuencia del término más común en el documento (Trstenjak, Mikac, & Donko, 2009). Otra de las consideraciones para calcular la frecuencia del término corresponde a:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Donde: $n_{i,j} = es\ el\ número\ de\ ocurrencias\ de\ un\ término\ t_i\ en\ un\ documento\ d_j$

La frecuencia inversa de documento (IDF) de un término t_i está dada de la siguiente manera (D.Manning, Raghavan, & Schütze, 2009):

$$idf_i = log_2 \frac{|D|}{|\{d \mid t_i \in d\}|}$$

Donde:

$|D| = Número\ total\ de\ documentos\ en\ una\ colección.$

$|\{d \mid t_i \in d\}| = Número\ de\ documentos\ donde\ aparece\ el\ termino\ t_i.$

Por tanto, el esquema de ponderación TF-IDF asigna al término *i* un peso en el documento *j* dado por:

$$tf_{ij}idf_i = tf_{ij} \times idf_i$$

En otras palabras, TF-IDF asigna al término t_i un peso en el documento d_j que es más alto cuando t_i aparece muchas veces dentro de un pequeño número de documentos y más bajo cuando el término aparece en prácticamente todos los documentos.

En esta investigación se emplea la función Term-Document Matrix perteneciente a la librería tm del software R y se obtienen los siguientes resultados.

Tabla 14.
Características de la Matriz Termino Documento

DocumentTermMatrix	**documentos: 2209, términos: 5040**
Entradas no dispersas	25687/11107673
Dispersión	100%
Longitud máxima del término	39

Asimismo, en la Tabla 15 se muestra un ejemplo de la matriz TF-IDF que hace parte de todos los documentos (tweets) analizados. Las filas contienen los documentos o en este caso los tweets y en las columnas los términos con su respectivo TF-IDF, por tanto, el término *Dengue* posee un valor de 0,05544678, menor con respecto al término *mosquito* que cuenta con un valor de 0,1621863, es decir que en esta muestra de tweets la palabra *Dengue* es más común en la colección de documentos que la palabra mosquito.

Tabla 15.
Ejemplo de la Matriz TF-IDF

Documentos	**Términos**							
	Caso	Chikungunya	Dengue	Enfermedad	Malaria	Mosquito	Salud	Zika
1032	0	0	0	0	0.09440349	0	0	0
1235	0	0	0	0	0.10434070	0	0	0
1315	0	0	0	0	0	0	0	0
1317	0	0	0	0	0	0	0	0
1582	0	0	0	0	0	0	0	0
2182	0	0	0	0	0	0	0	0

Continuación tabla 15
Ejemplo de la Matriz TF-IDF

Documentos	**Términos**							
	Caso	Chikungunya	Dengue	Enfermedad	Malaria	Mosquito	Salud	Zika
571	0	0	0.05544678	0	0	0.1621863	0	0
599	0	0	0	0	0	0	0	0
870	0	0	0	0	0	0	0	0
937	0	0	0	0	0	0	0	0

5.4 Procesamiento de la información

En esta etapa con los 2.209 tweets obtenidos después de la etapa de pre-procesamiento, se busca predecir si un tweet es reconocido como válido o inválido según corresponda, con el fin de identificar la presencia de posibles casos de ETV y ser una fuente complementaria de la vigilancia tradicional; para lo cual, se parte de la etiquetación manual para después aplicar un método de aprendizaje supervisado que identifique la relación entre los atributos considerados y las etiquetas de clasificación (válido o inválido) y así evaluar el modelo más apropiado para esta clasificación, así que dentro de los métodos mencionados en la sección 3 (revisión de literatura) el más utilizado es la máquina de soporte vectorial, resaltando que es una técnica que mediante sus diferentes kernel permite evaluar cuál es el modelo que representa la información objeto de estudio y tiene la capacidad de clasificar nueva información.

Las características de los recursos de cómputo empleados en la ejecución del modelo son:

- Procesador: Inter® Core™ i5-4200M CPU @ 2.50GHz
- Memoria instalada (RAM): 8,00 GB
- Tipo de sistema: Sistema operativo de 64 bits, procesador x64

Con el apoyo de la librería "e1071" del software R se ejecutan las SVM teniendo en cuenta las siguientes variaciones: SVM con kernel lineal, SVM con kernel polinomial, SVM con kernel radial y SVM con kernel sigmoide.

A continuación, se presentan los diferentes parámetros para cada tipo de kernel:

$$Kernel\ Lineal\ \rightarrow U' *\ V$$

$$Kernel\ Polinomial\ \rightarrow (Gamma * U' * V + Coef0)^{degree}$$

$$Kernel\ Radial\ \rightarrow Exp(-Gamma * |U - V|^2)$$

$$Kernel\ Sigmoide\ \rightarrow \tanh(Gamma * U' * V + Coef0)$$

El método empleado para determinar el valor óptimo de los parámetros que corresponden a cada tipo de kernel es la validación cruzada. En la siguiente tabla se presenta la red de parámetros considerados:

Tabla 16.
Rango de valores definidos para cada parámetro

Parámetro	**Rango**									
Costo	0,5	1	1,5	2	2,5	3	3,5	4	4,5	5
	5,5	6	6,5	7	7,5	8	8,5	9	9,5	10
	10,5	11	11,5	12	12,5	13	13,5	14	14,5	15
	15,5	16	16,5	17	17,5	18	18,5	19	19,5	20
	20,5	21	21,5	22	22,5	23	23,5	24	24,5	25
	25,5	26	26,5	27	27,5	28	28,5	29	29,5	30
Gamma	0,5	1	2	4						
Coef0	0,5	1	1,5	2						
Degree	0,5	1	1,5	2	2,5	3				

El parámetro Costo determina la penalización permitida sobre el margen de separación, para valores grandes de C, la optimización elegirá un hiperplano de menor margen si ese hiperplano hace un mejor trabajo para que todos los puntos de entrenamiento se clasifiquen correctamente. En caso contrario, un valor muy pequeño de C hará que el optimizador busque un hiperplano de separación de mayor margen, incluso si ese hiperplano clasifica erróneamente más puntos. Gamma define hasta dónde llega la influencia de un conjunto de datos de entrenamiento, los valores bajos indican una mayor distancia entre las observaciones y el modelo resultante se comportará de manera similar a un modelo lineal; el parámetro gamma se puede ver como el inverso del radio de influencia de las muestras seleccionadas por el modelo como vectores de soporte. Coef0 es el coeficiente que acompaña la variable en cada función de kernel y el Degree es el grado del polinomio necesario para la función del kernel de tipo polinomial y controla la flexibilidad del clasificador resultante.

Se debe tener en cuenta que valores muy grandes del parámetro Costo provoca un sobreajuste en el modelo, es decir, el algoritmo de aprendizaje se ajusta perfectamente a los datos de entrenamiento dados y será incapaz de reconocer nuevos datos de entrada; valores muy cercanos a 0 ocasionan una menor penalización a las observaciones permitiendo así un mayor porcentaje de datos mal clasificados. El parámetro gamma controla el comportamiento del kernel (Radial, Sigmoide y Polinomial), a mayor valor la función se hace más flexible. Valores muy grandes del parámetro Degree lleva el modelo a problemas de sobreajuste y el parámetro Coef0 por lo general toma valores entre 0 y 2, valores más grandes ocasionan problemas de sobreajuste en el modelo.

5.5 Validación cruzada

Se utiliza la técnica de validación cruzada para evaluar los resultados de los 2.209 tweets utilizados para la investigación y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Esta técnica consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones, para de esa manera estimar la precisión del modelo que se llevará a la práctica.

Con la técnica K iteraciones (K-fold cross-validation) de la librería "e1071" del software R se selecciona la mejor combinación de parámetros para los 4 tipos de kernel disponibles utilizando un K=10 como se muestra en la siguiente tabla.

Tabla 17.
Mejores parámetros para cada tipo de kernel

Kernel	Costo	Gamma	Degree	Coef0	Tiempo de ejecución
Radial	15,5	0,5	~	~	94 min
Lineal	19	~	~	~	21 min
Polinomial	7	0,5	2	2	1924 min
Sigmoide	2	0,5	~	0,5	220 min

Una vez realizada la validación cruzada, se evidencia en la Figura 10 el comportamiento del parámetro Costo para el tipo de kernel lineal, en donde su mejor rendimiento se da cuando este asume un valor de 19 con un error mínimo de 0,08487. A diferencia de otros tipos de kernel como el Polinomial, donde intervienen más parámetros como lo son el gamma, Degree y coef0 (Figura 11), en donde se evidencia el valor óptimo del Costo en 7 con un mínimo error de 0,0724319. Cabe resaltar que los valores óptimos de cada parámetro obtenidos por el procedimiento de validación cruzada son utilizados con el fin de entrenar y validar el modelo para cada tipo de kernel de las SVM.

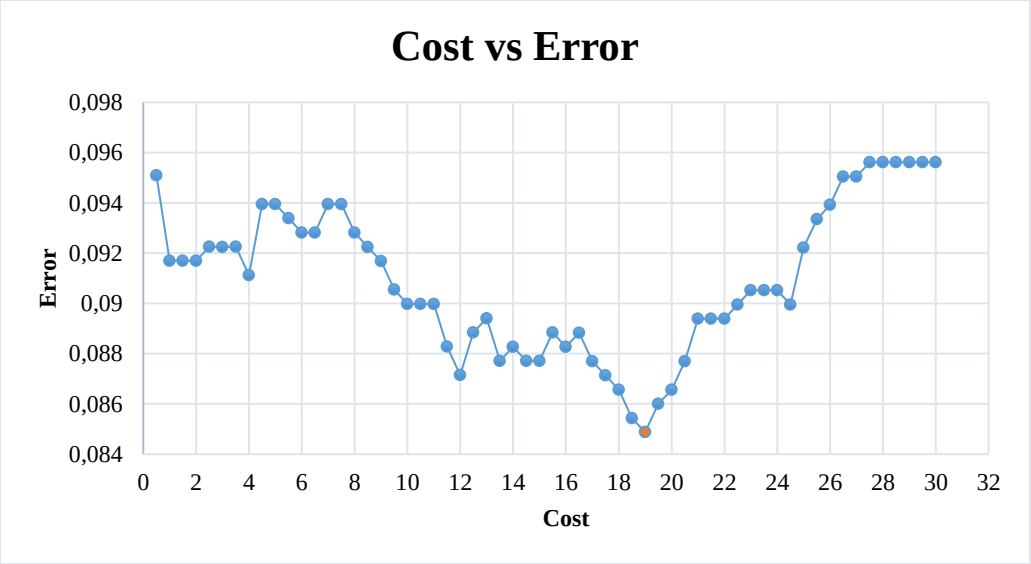

Figura 10. Cost vs Error para el kernel Lineal

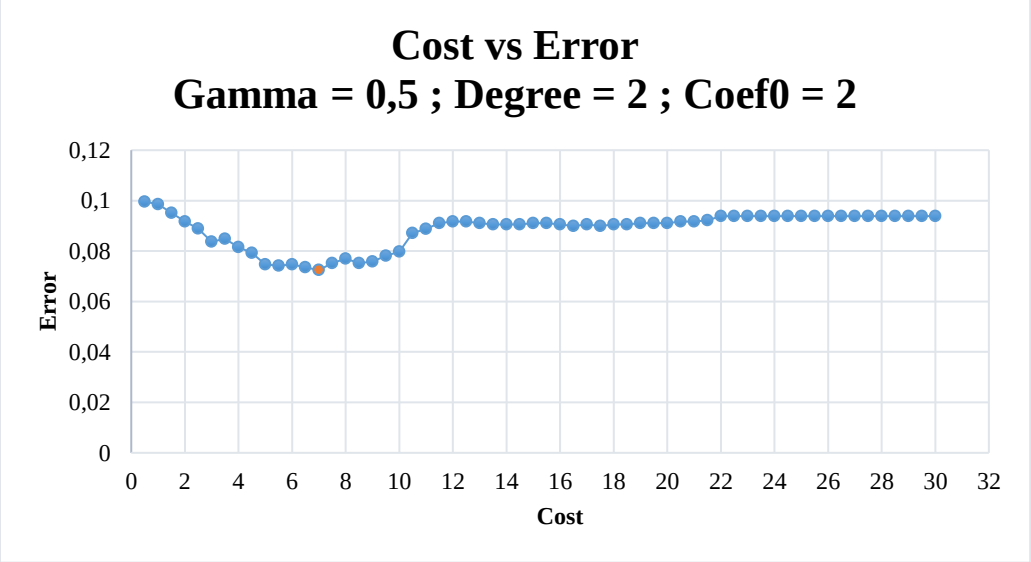

Figura 11. Cost vs Error para el kernel Polinomial

Para el kernel de tipo base radial como se ilustra en la Figura 12, se evidencia una disminución del error a medida que aumenta su parámetro Costo, teniendo en cuenta su parámetro gamma con un valor de 0.5, puesto que son los dos parámetros a considerar en este tipo de kernel. Así, en un valor de costo de 15.5 se alcanza el mínimo error de 0.129593.

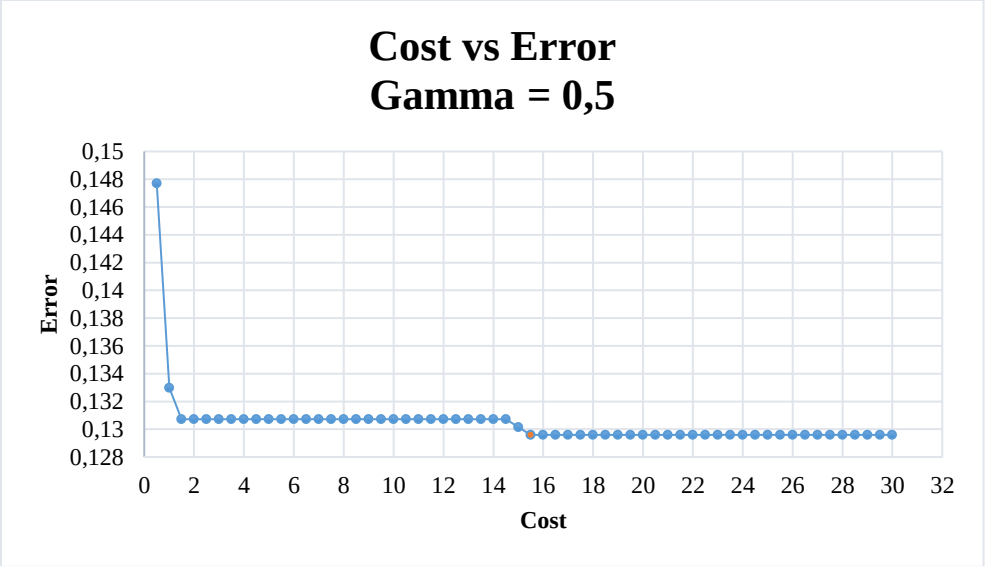

Figura 12. Cost vs Error para el kernel Base Radial

En la Figura 13, se muestra la variación del parámetro Costo para el kernel de tipo sigmoide, incluyendo los parámetros gamma y coef0 puesto que también afectan a este tipo de kernel. La variación del error presenta una tendencia creciente, a medida que aumenta el Costo, el error aumenta; con un valor de 2 se obtiene un error mínimo de 0.09678.

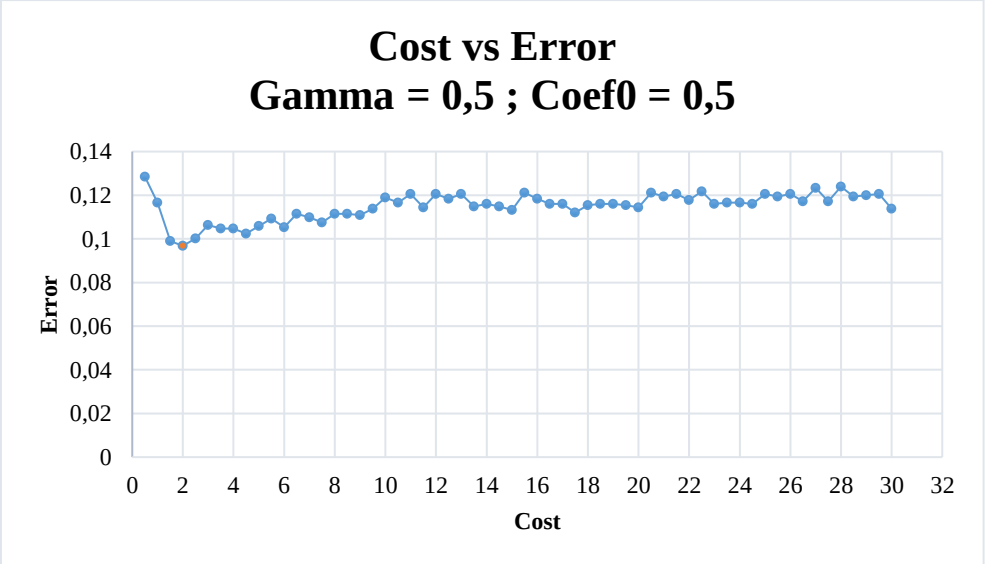

Figura 13. Cost vs Error para el kernel Sigmoide

5.6 Resultados de clasificación

A partir de la selección de la mejor combinación de parámetros para cada uno de los tipos de kernel, se procede a ejecutar el modelo de SVM a partir de una muestra aleatoria de datos para entrenamiento (80%) y el restante para datos de prueba (20%), esto con el fin de obtener sus respectivas métricas de validación y así realizar comparaciones entre los resultados de clasificación de los kernel, seleccionando aquel que permita obtener una clasificación con mayor precisión y exactitud.

La matriz de confusión contiene información sobre las clasificaciones reales y pronosticadas por las SVM para la variable "clasificación". A continuación, se presentan los resultados de predicción obtenidos para cada tipo de kernel:

- Para un kernel lineal con un valor de costo igual a 19 se obtiene la siguiente matriz de confusión.

Tabla 18.
Matriz de confusión kernel Lineal

	Predicho	
Observado	**Inválido**	**Válido**
Inválido	38	21
Válido	20	363

- Para un kernel Polinomial con un valor de costo igual 7, gamma igual a 0,5 degree igual a 2 y coef0 igual a 2 se obtiene la siguiente matriz de confusión.

Tabla 19.
Matriz de confusión kernel Polinomial

	Predicho	
Observado	**Inválido**	**Válido**
Inválido	39	20
Válido	13	370

- Para un kernel Radial con un valor de costo igual a 15.5 y un gamma igual a 0,5 se obtiene la siguiente matriz de confusión.

Tabla 20.
Matriz de confusión kernel Base Radial

	Predicho	
Observado	**Inválido**	**Válido**
Inválido	9	50
Válido	4	379

- Para un kernel Sigmoide con un valor de costo igual a 2, gamma igual a 0.5 y coef0 igual a 0.5 obtiene la siguiente matriz de confusión.

Tabla 21.
Matriz de confusión kernel Sigmoide

	Predicho	
Observado	**Inválido**	**Válido**
Inválido	31	28
Válido	12	371

En la Tabla 22 se exponen los resultados de las 4 métricas obtenidas para cada tipo de kernel, Exactitud (Accuracy), Precisión, sensibilidad (Recall) y medida-F (F-Measure), con el fin de comparar y evaluar el modelo que represente una mejor clasificación de los datos.

Tabla 22.
Métricas de evaluación del modelo

	Kernel			
Métricas	**Lineal**	**Polinomial**	**Radial**	**Sigmoide**
Exactitud	0,9072398	0,9253394	0,8778281	0,90950226
Precisión	0,9453125	0,9487179	0,8834499	0,92982456
Sensibilidad	0,9477807	0,9660574	0,9895561	0,96866841
Medida-F	0,946545	0,9573092	0,9334975	0,9488491

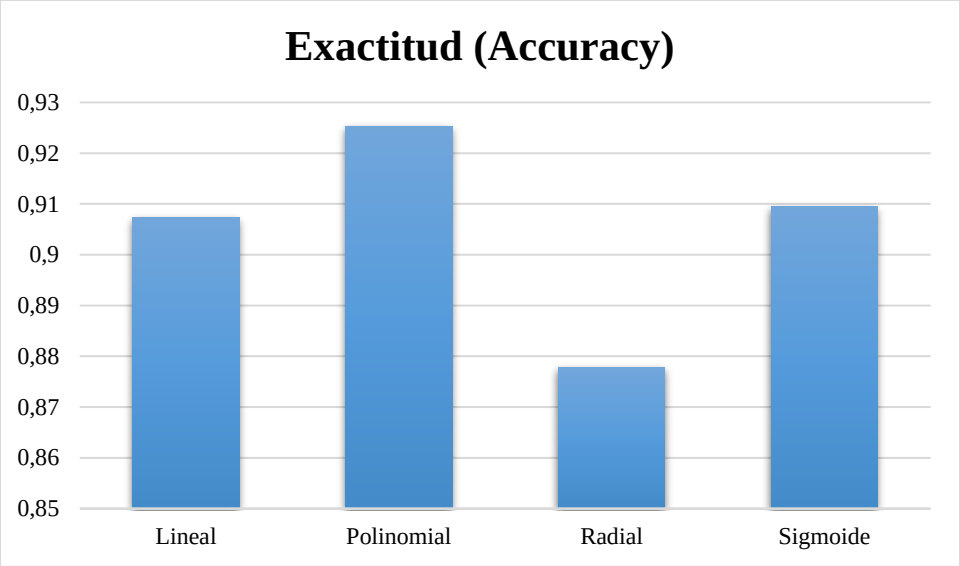

Figura 14. Accuracy

Como se observa en la Figura 14, para la métrica Exactitud (Accuracy), la cual define la proporción del número total de predicciones que son correctas en cuanto a la clasificación para la etiqueta válido como inválido, el kernel que mejores resultados brinda es el Polinomial. Se resalta que con el kernel radial se obtiene un valor menor al 90% en esta métrica, siendo este el menor valor respecto a los demás kernel.

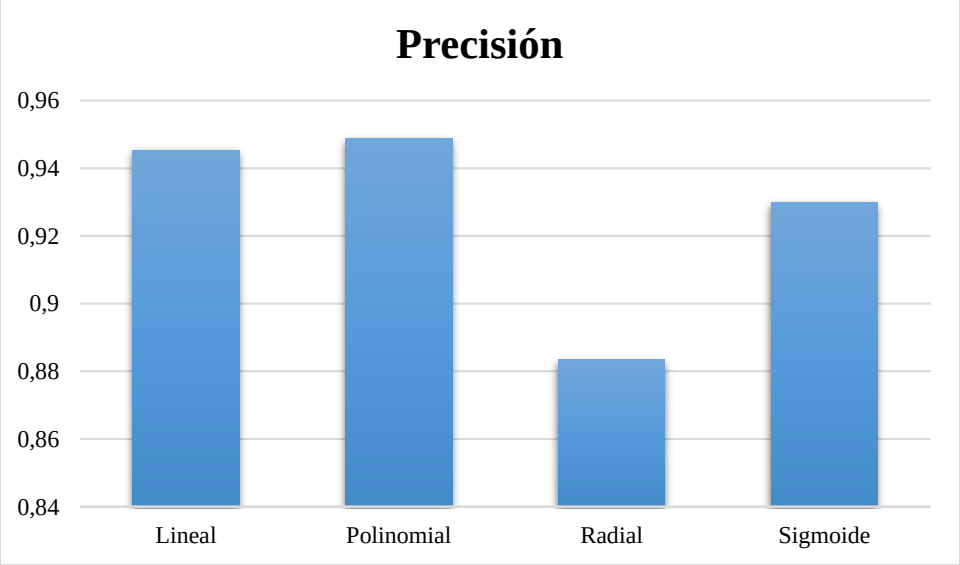

Figura 15. Precisión

Como se observa en la Figura 15, para la métrica precisión, la cual define la proporción de casos positivos pronosticados que son correctos, el kernel que mejores resultados brinda es el Polinomial con un valor de 94.8%, seguido del kernel lineal con un valor de 94.5%. Similar a la métrica de exactitud, el kernel radial presenta el menor valor.

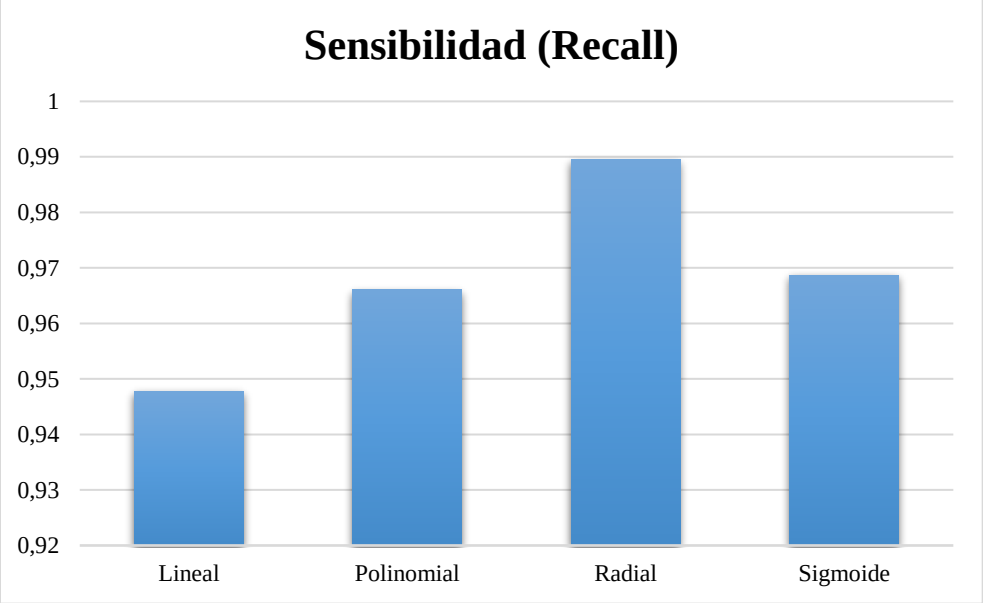

Figura 16. Recall

Como se observa en la Figura 16, para la métrica sensibilidad (Recall), la cual define la proporción de casos positivos que se identificaron correctamente, los dos kernel que mejores resultados brindan son el Radial y el Sigmoide con un valor de 0,98 y 0,96 respectivamente. Se resalta que todos los kernel obtienen un valor superior al 95% en esta métrica

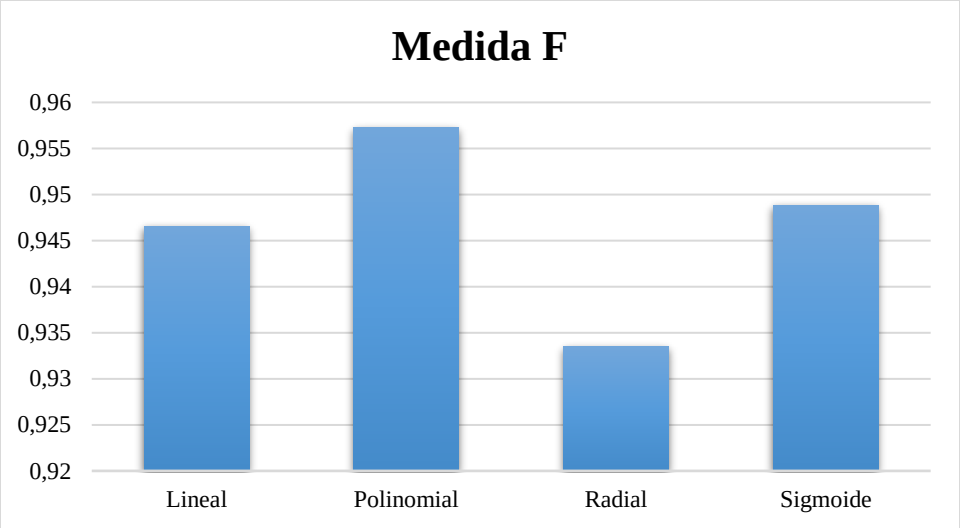

Figura 17. F-Measure

Como se observa en la Figura 17, para la métrica medida F (F-Measure), la cual mide el rendimiento general de un clasificador y controla la importancia relativa entre la precisión y la sensibilidad (Recall), el kernel que mejor resultados brinda es el Polinomial con un valor de 0,95 seguido del lineal con un valor de 0,94.

Finalmente, el modelo SVM refleja una mejor clasificación de los datos para los kernel de tipo Lineal y Polinomial, ambos kernel presentan una exactitud (accuracy) mayor al 90% y en su precisión presentan una diferencia mínima del 0.34% entre ambos, sin embargo, bajo las mismas condiciones computacionales, el entrenamiento del kernel Polinomial requiere un tiempo de procesamiento significativamente mayor, siendo de 1924 minutos (32 horas), en comparación con el kernel lineal, 21 minutos, puesto que, la cantidad de parámetros involucrados en cada kernel es diferente. Por tanto, un kernel de tipo lineal resultará más adecuado para considerar en el apoyo de la toma de decisiones inmediatas y su rendimiento no difiere significativamente en relación con el rendimiento del kernel Polinomial.

6. Reglas de asociación en Enfermedades Transmitidas por Vectores

Las reglas de asociación son declaraciones si / entonces que ayudan a exponer relaciones entre datos aparentemente no relacionados. Una regla de asociación tiene dos componentes, un antecedente (si) y un consecuente (entonces). Un antecedente es un conjunto de elementos que se encuentra en el conjunto de datos, y un consecuente es un conjunto de elementos encontrado en la incorporación con el antecedente (M.-S. Chen, Han, & Yu, 1996).

Las reglas de asociación se generan mediante el análisis de datos para patrones frecuentes de si/entonces, el soporte y la confianza son los criterios para identificar las relaciones más importantes. El soporte es un desenlace de la frecuencia con la que aparecen los elementos en la base de datos y la confianza indica la cantidad de veces que las representaciones si /entonces son verdaderas (AL-Rubaiee et al., 2017).

Los investigadores Liu, Ma, & Wong (2000), expertos en minería de datos argumentan que algunos términos o ítems aparecen con mayor frecuencia en el conjunto de datos, mientras que otros rara vez aparecen. En este caso, los valores del soporte mínimo controlarán el descubrimiento de la regla. Si el soporte mínimo se establece en un valor alto, no se encontrarán las reglas que ocurren con poca frecuencia. De lo contrario, si el soporte mínimo se establece en un valor bajo, se generarán las reglas que ocurren con frecuencia.

Este estudio tiene como objetivo identificar las reglas de asociación para cada Enfermedad Transmitida por Vector definidas previamente, seleccionando los datos a partir de su variable clasificación, para así generar reglas de asociación por cada etiqueta "válido" o "inválido" según corresponda. A continuación, se muestra la variación de los parámetros, soporte y confianza, y la cantidad de reglas de asociación generadas para cada ETV, con el fin de identificar los parámetros que permitan generar los ítems más frecuentes y las reglas que cumplan con los mínimos establecidos. Se utiliza el algoritmo *apriori* de la librería "*arules"* con apoyo del software R, el cual permite encontrar de forma eficiente conjuntos de ítems frecuentes que sirven de base para generar reglas de asociación.

Así, teniendo en cuenta los parámetros soporte y confianza para cada ETV, se considera variar el soporte desde 2 hasta 15, tanto para una confianza del 70% como para una del 65%, con el fin de identificar los mejores parámetros que se ajusten al modelo y que permitan generar la mayor cantidad de reglas de asociación para descubrir relaciones entre los términos. En el apéndice A se presentan las tablas de los resultados obtenidos de la variación de los parámetros para cada enfermedad, y a continuación, se presentan las gráficas que representa el comportamiento de estos parámetros para cada enfermedad.

- **Variación del soporte con Confianza 70%**

Al variar el intervalo de soporte de 2 a 15 con una confianza del 70%, para la etiqueta válido, a medida que aumenta el soporte, las reglas de asociación generadas van disminuyendo, esto debido a que para valores de soporte altos, las reglas de asociación con menor frecuencia no son

consideradas. En el caso de la etiqueta inválido, a medida que aumenta el soporte, el número de reglas de asociación es constante hasta alcanzar un valor de 0 reglas con el máximo soporte considerado. Por ejemplo, para la enfermedad Dengue (Figura 18) cuya etiqueta es inválido, el número de reglas de asociación es constante con un valor de 2 reglas generadas hasta alcanzar un valor de 0. En la Figura 19, se muestran las reglas de asociación generadas para la enfermedad Malaria; la etiqueta válido con un soporte de 2 genera 1.364.762 reglas de asociación que van disminuyendo a medida que el soporte aumenta, para la etiqueta de inválido con un soporte de 2 se generan 1.798.842 reglas de asociación y a medida que aumenta el soporte el número de reglas permanece constante, hasta alcanzar un valor de 0 reglas generadas.

Así mismo, en la enfermedad Leishmaniasis (Figura 20) se observa un comportamiento inversamente proporcional del soporte para la etiqueta válido, respecto al número de reglas generadas, puesto que, el soporte define la frecuencia de ocurrencia de los ítems en el conjunto de datos en análisis, es decir, a mayor soporte, menor el número de términos o ítems seleccionados por el algoritmo para la generación de las reglas; respecto a la etiqueta inválido el algoritmo no reconoce reglas de asociación tanto para una confianza del 70% como del 65% dado que, no existe relación entre los términos de dicha etiqueta o la información es muy limitada. En la Figura 21 se observa el mismo comportamiento para la etiqueta válido en la enfermedad Zika, a diferencia de la etiqueta inválido donde se puede apreciar que cuando el soporte toma valores de 3 a 6 la gráfica se va normalizando hasta alcanzar un valor de 0 reglas de asociación generadas, que es cuando toma valores de soporte 7 o superiores.

Finalmente se evidencia en la Figura 22 el comportamiento de la enfermedad de Chagas, para la etiqueta válido existe una disminución significativa al aumentar el soporte de 3 a 4, con un valor de 156.423 a 960 reglas generadas respectivamente, por otra parte, la etiqueta inválido presenta una disminución mayor cuando su soporte es de 3; sin embargo, cabe resaltar que tanto para la etiqueta válido e inválido, el valor mínimo de la confianza es de 75%. Por otra parte, en la Figura 23 se observa el número de reglas de asociación generadas para la enfermedad Chikungunya con su respectiva variación de soporte para una confianza del 70%; para la etiqueta válido, su mayor disminución es dada por un soporte de 6 con un valor de 333 reglas generadas; para la etiqueta de inválido a medida que aumenta el soporte, las reglas de asociación se van normalizando con un valor de 0 a partir del soporte equivalente a 4.

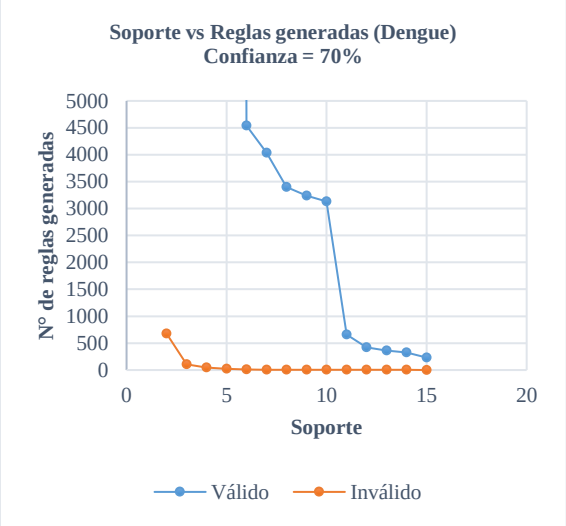

Figura 18. Soporte Vs Número de reglas generadas y Confianza 70%, Dengue

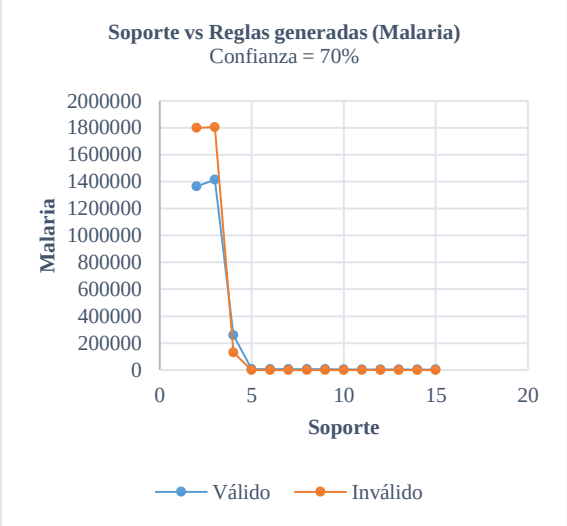

Figura 19. Soporte Vs Número de reglas generadas y Confianza 70%, Malaria

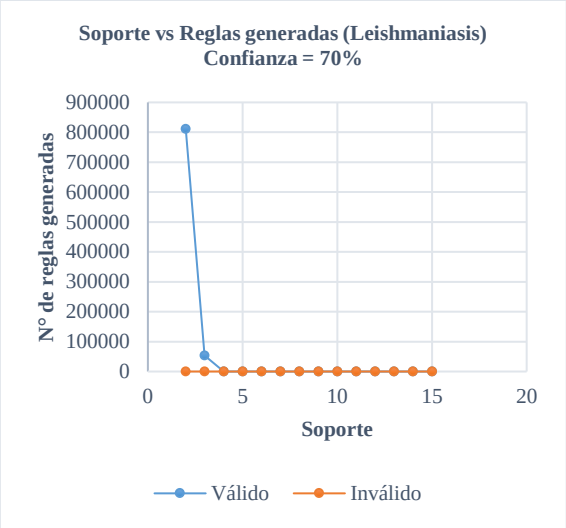

Figura 20. Soporte Vs Número de reglas generadas y Confianza 70%, Leishmaniasis

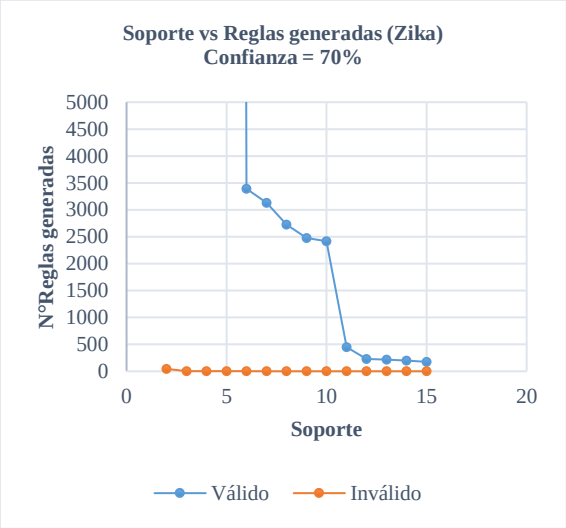

Figura 21. Soporte Vs Número de reglas generadas y Confianza 70%, Zika

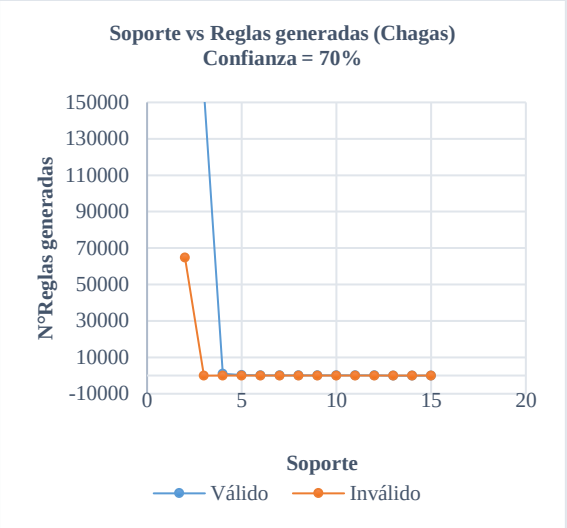

Figura 22. Soporte Vs Número de reglas generadas y Confianza 70%, Chagas

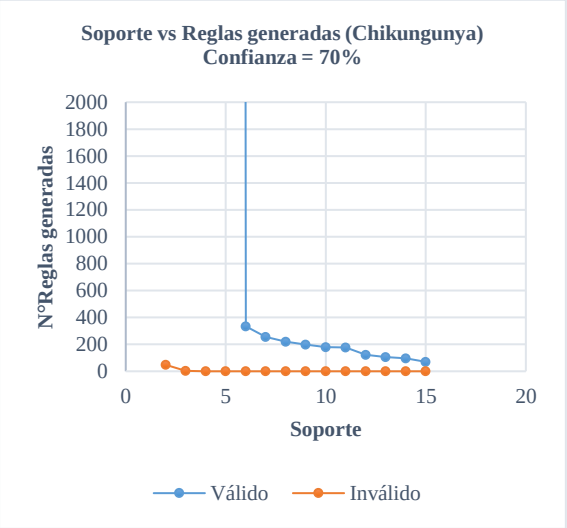

Figura 23. Soporte Vs Número de reglas generadas y Confianza 70%, Chikungunya

- **Variación soporte con confianza del 65%**

Con el fin de analizar el comportamiento del mismo intervalo de soporte desde 2 hasta 15 analizado con la confianza de 70%, a continuación, se presentan las gráficas para cada ETV, con una confianza del 65%, con el fin de identificar el comportamiento del número de reglas generadas

a partir de dicho parámetro, resaltando que con una confianza del 70%, la etiqueta inválido presenta un menor número de reglas generadas.

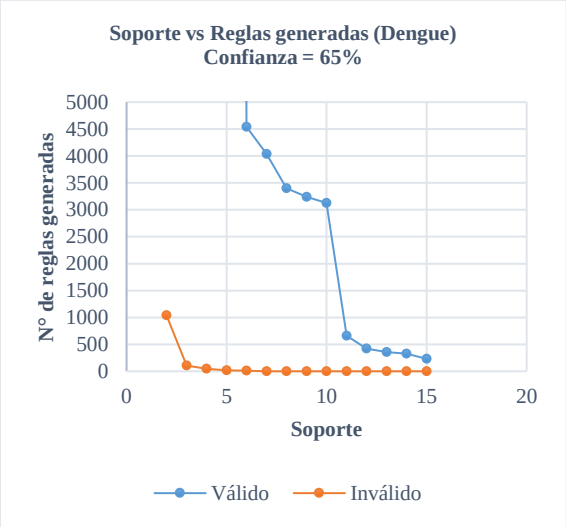

Figura 24. Soporte Vs Número de reglas generadas y Confianza 65%, Dengue

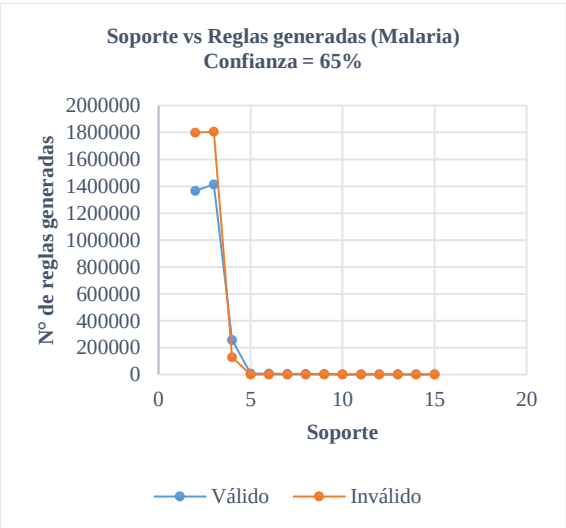

Figura 25. Soporte Vs Número de reglas generadas y Confianza 65%, Malaria

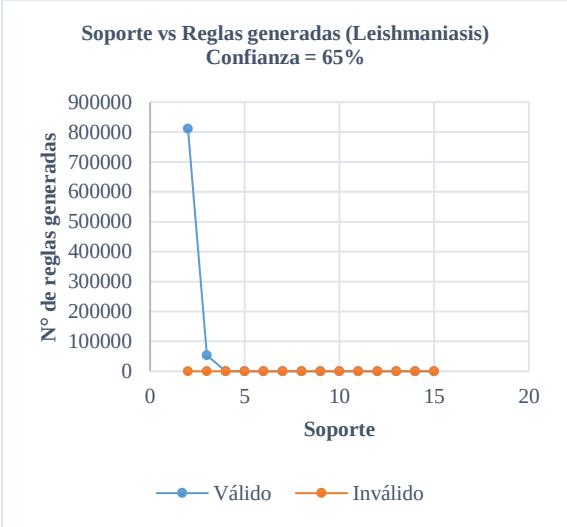

Figura 26. Soporte Vs Número de reglas generadas y Confianza 65%, Leishmaniasis

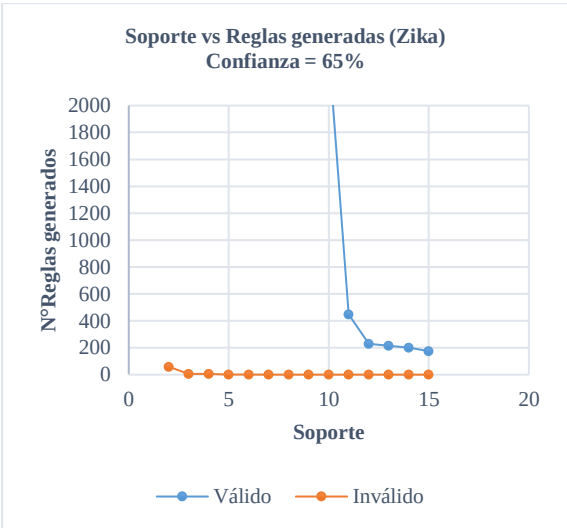

Figura 27. Soporte Vs Número de reglas generadas y Confianza 65%, Zika

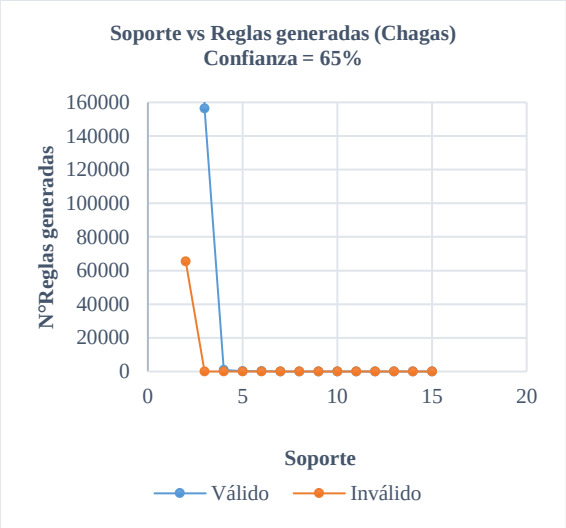

Figura 28. Soporte Vs Número de reglas generadas y Confianza 65%, Chagas

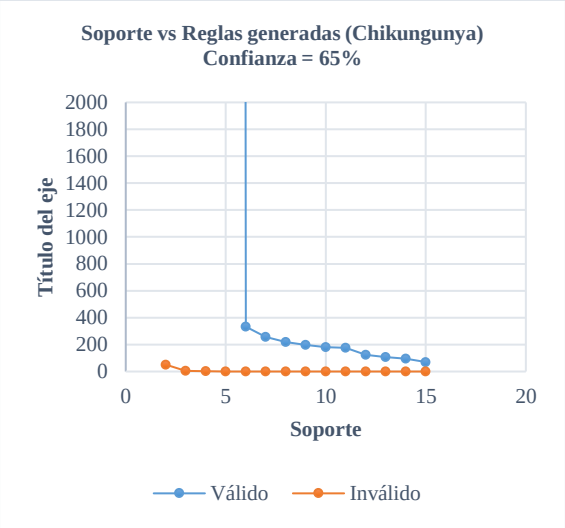

Figura 29. Soporte Vs Número de reglas generadas y Confianza 65%, Chikungunya

De acuerdo a los resultados anteriores, al disminuir la confianza de 70% a 65%, el número de reglas generadas por el algoritmo no se afectó significativamente, sin embargo, para la etiqueta válido, se presenta un comportamiento decreciente a medida que su soporte va aumentando, siendo similar al comportamiento obtenido para una confianza del 70%, por otra parte, la etiqueta inválida presenta una tendencia constante a medida que su soporte aumenta hasta alcanzar un valor de cero.

Después de la variación de los parámetros soporte y confianza, se seleccionan los valores de soporte y confianza para cada ETV, con el fin de visualizar relaciones entre los términos y generar las reglas de asociación a partir de estos valores establecidos (Ver Tabla 23).

Los valores de confianza y soporte establecidos se deben a que a medida que el soporte toma valores altos se excluyen las reglas de asociación de menor frecuencia las cuales son las que permiten un estudio más efectivo, de igual forma sucede con los valores de confianza, si la confianza toma valores mayores al 70% las reglas de asociación que se van a obtener serán aquellas

con antecedentes formados por conjuntos de ítems de palabras muy frecuentes lo cual sesga el análisis de este estudio.

Tabla 23
Parámetros para reglas de asociación

Enfermedad Transmitida por Vector	Soporte	Confianza
Dengue	0,0022	70%
Malaria	0,0037	70%
Leishmaniasis	0,039	70%
Zika	0,0045	70%
Chagas	0,017	70%
Chikungunya	0,010	70%

Así, teniendo en cuenta los parámetros seleccionados, se procede a identificar los ítems más frecuentes que sirven de base para generar las reglas de asociación que cumplan con los mínimos establecidos para cada ETV, seleccionando los datos a partir de su variable clasificación para identificar las relaciones por cada etiqueta “válido” o “inválido” según corresponda, a partir de las reglas generadas.

6.1 Dengue

Al realizar un análisis de los 10 ítems más frecuentes (Tabla 24), después de la enfermedad Dengue, que obtiene un valor de soporte mayor respecto de los demás términos, se encuentra el ítem Zika con un soporte de 0,195, al igual que el conjunto de Dengue-Zika y Chikungunya-Dengue, es decir que los términos Dengue y Zika aparecen en aproximadamente el 20% de los tweets en comparación con el 17% con el que se presentan los términos Chikungunya-Dengue al mismo tiempo, lo cual evidencia que la enfermedad del Dengue cuando es tratada en este medio de difusión, se relaciona con otras enfermedades como Zika o Chikungunya en el periodo de estudio.

Tabla 24
Ítems más frecuentes para Dengue

Ítems	Soporte	Frecuencia absoluta
Dengue	0,93424036	824
Zika	0.19501134	172
Dengue, Zika	0.19501134	172
Chikungunya	0.16893424	149
Chikungunya, Dengue	0.16666667	147
casos	0.14965986	132
casos, Dengue	0.14965986	132
salud	0.10657596	94
Dengue, salud	0.10090703	89
mosquitos	0.08616780	76

La asociación entre los términos que se relacionan con la etiqueta válido se muestra en la Tabla 25, por ejemplo, los términos casos, salud, mosquito o Aedes se relacionan con la palabra Dengue; el mayor soporte se obtiene de la relación "{casos, Dengue} => {válido}" con un valor de 0,1462 y una confianza del 98%, es decir que esta regla se cumple en el 14,6% de los tweets con una probabilidad del 98% de que al estar presente los términos casos y Dengue, está presente la etiqueta válido, así mismo, las reglas generadas superan una confianza del 94% y existe una probabilidad del 100% de que se cumpla la regla que relaciona la presencia de términos como *mosquito, transmisor, Dengue* y *Chikungunya*, haciendo referencia a la amenaza que representa el mosquito Aedes aegypti, transmisor del Dengue, Zika y Chikungunya.

Tabla 25
Reglas de Asociación para etiqueta válido en Dengue

Reglas de asociación	Soporte	Confianza	Frecuencia absoluta
{casos, Dengue} => {válido}	0,146258503	98%	129
{Dengue, salud} => {válido}	0,095238095	94%	84

Continuación Tabla 25.

Reglas de Asociación para etiqueta válido en Dengue

Reglas de asociación	Soporte	Confianza	Frecuencia absoluta
{Dengue, mosquito} => {válido}	0,062358277	95%	55
{Aedes, Dengue} => {válido}	0,060090703	96%	53
{Chikungunya, Dengue, enfermedades} => {válido}	0,041950113	100%	37
{Chikungunya, Dengue, mosquito, transmisor} => {válido}	0,020408163	100%	18
{Dengue, red, simposio} => {válido}	0,012471655	100%	11
{atención, Dengue, grave, manejo, prevención} => {válido}	0,011337868	100%	10
{atención, Dengue, diagnóstico, manejo, prevención, red} => {válido}	0,011337868	100%	10
{Aedes, mortalidad, prevención, red} => {válido}	0,011337868	100%	10
{Dengue, municipal, salud, secretaria} => {válido}	0,009070295	100%	8
{casos, Dengue, noticias} => {válido}	0,006802721	100%	6
{aguas, Dengue, Zika} => {válido}	0,006802721	100%	6

La Tabla 26 muestra la asociación entre los términos que están relacionados con la etiqueta inválido, por ejemplo, *"jajaja", "sobreviví", "basura", "amiga"* o *"polideportivo"* hacen parte del antecedente de las reglas. La regla con mayor soporte para la etiqueta inválido es *"{Dengue, jajaja} => {inválido}*", es decir que esta regla se cumple en el 0,34% de los tweets con una probabilidad del 100% de que al estar presente los términos "Dengue" y "jajaja", está presente la etiqueta inválido. Así mismo, se evidencian reglas con menor soporte, pero con igual confianza, tal como, *"{eléctricas, raquetas} => {inválido}"* con una probabilidad del 100% de que al estar presentes los términos eléctricas y raquetas, está presente la etiqueta inválido; el significado de la relación hace referencia al torneo internacional de prevención del Dengue con raquetas eléctricas.

Tabla 26

Reglas de Asociación para etiqueta Inválido en Dengue

Reglas de asociación	Soporte	Confianza	Frecuencia absoluta
{Dengue, jajaja} => {inválido}	0,003401	100%	3
{Dengue, sobreviví} => {inválido}	0,002268	100%	2

Continuación tabla 26

Reglas de Asociación para etiqueta Inválido en Dengue

Reglas de Asociación	Soporte	Confianza	Frecuencia absoluta
{basura, Dengue} => {inválido}	0,002268	100%	2
{amiga, clínica, iba} => {inválido}	0,002268	100%	2
{lado, periférico, polideportivo} => {inválido}	0,002268	100%	2
{agua, allí, Dengue, lado, natación, piscina} => {inválido}	0,002268	100%	2
{eléctricas, raquetas} => {inválido}	0,002268	100%	2
{cama, Dengue} => {inválido}	0,002268	100%	2
{Dengue, porquería} => {inválido}	0,002268	100%	2
{lado, polideportivo, practicaban} => {inválido}	0,002268	100%	2

6.2 Malaria

En la Tabla 27 se pueden observar los ítems más frecuentes para la enfermedad Malaria, seleccionando aquellos que tienen un valor de soporte mayor respecto a los demás términos, después de la enfermedad Malaria, se puede evidenciar que el termino casos con soporte de 0,1275 y el conjunto casos-Malaria con un soporte del 0,1257 están presentes en aproximadamente el 12% de los tweets, en comparación con el 7% con el que se presentan los términos Malaria y salud. También se puede observar la presencia del término sarampión lo cual hace referencia a un plan de vacunación emitido por el gobierno en la zona norte colombiana para dicha enfermedad y otras como la Malaria, Dengue, Zika y Chikungunya; también hace referencia a una epidemia de sarampión y Malaria dentro de comunidades indígenas colombianas.

Tabla 27

Ítems más frecuentes para Malaria

Término (Ítem)	Soporte	Frecuencia absoluta
Malaria	0.93621013	499
casos	0.12757974	68
casos, Malaria	0.12570356	67
Malaria, sarampión	0.11632270	62
Malaria, salud	0.07317073	39

En la Tabla 28 se pueden observar algunos ejemplos de las reglas de asociación generadas con un nivel de confianza del 70% y un soporte de 2. Para el consecuente válido el nivel de confianza de la regla oscila aproximadamente entre 81% y 100%, presentando así una frecuencia absoluta que varía de 21 a 32. La regla de asociación *"{Malaria, plan, vacunación} => {válido}"* presenta un soporte de 0,060 y un nivel de confianza del 100%, lo que quiere decir que los antecedentes *Malaria, plan y vacunación* están presentes en el 6% de los tweets y con una probabilidad del 100% se cumple que aparece el consecuente *válido*, lo cual permite comprobar que de la información recolectada con el término Malaria hay presencia de términos relacionados con salud pública en cuanto a lo sucedido en el periodo de estudio ya que está relacionado con una incidencia destacable de esta enfermedad, haciendo referencia a un plan de vacunación del gobierno contra el sarampión, difteria y Malaria.

Tabla 28
Reglas de Asociación para etiqueta válido en Malaria

Regla de asociación	Soporte	Confianza	Frecuencia absoluta
{mosquito} => {válido}	0,04315197	100%	23
{enfermedad} => {válido}	0,03939962	88%	21
{Malaria, mosquito} => {válido}	0,04315197	100%	23
{gobierno, vacunación} => {válido}	0,04878049	100%	26
{enfermedades, Malaria} => {válido}	0,05065666	82%	27
{Malaria, salud} => {válido}	0,06003752	82%	32
{Malaria, plan, vacunación} => {válido}	0,06003752	100%	32
{gobierno, Malaria, plan} => {válido}	0,04878049	100%	26
{gobierno, Malaria, sarampión, vacunación} => {válido}	0,04878049	100%	26
{difteria, gobierno, Malaria, plan, vacunación} => {válido}	0,04878049	100%	26

Para la siguiente regla de asociación: *"{top, Wimbledon} => {inválido}"* con un soporte de 0,01125 y un nivel de confianza del 75%, lo que quiere decir que los antecedentes *top* y *Wimbledon* están presentes en aproximadamente el 1% de los tweets y existe una probabilidad del 75% que su consecuente sea inválido, haciendo referencia a un tweet con contenido jocoso que no hace

referencia a un caso válido de ETV. La regla de asociación *"{gabaldon, Malaria} => {inválido}"* también se considera como inválido debido a que hace referencia a Arnoldo Gabaldón Carrillo, un médico, investigador y político venezolano que convirtió a Venezuela en el primer país en erradicar la Malaria; aunque se hace énfasis en la enfermedad Malaria, Venezuela no hace parte del sector que se está analizando en este estudio y el tweet no evidencia un posible caso de la enfermedad sino es de carácter informativo.

Tabla 29
Reglas de Asociación para etiqueta Inválido en Malaria

Regla de asociación	Soporte	Confianza	Frecuencia absoluta
{top, Wimbledon} => {inválido}	0,01125704	75%	6
{hambre, Malaria} => {inválido}	0,01125704	86%	6
{gabaldon, Venezuela} => {inválido}	0,01500938	100%	8
{gabaldon, Malaria} => {inválido}	0,01876173	100%	10
{gabaldon, Malaria, Venezuela} => {inválido}	0,01500938	100%	8

6.3 Leishmaniasis

La Tabla 30 muestra los 5 ítems más frecuentes para la enfermedad Leishmaniasis, que obtienen un valor de soporte mayor respecto de los demás términos, después del termino Leishmaniasis, se puede observar que los ítems que mayor soporte presentan son enfermedades o salud, es decir, están presentes en aproximadamente el 20% de los tweets, siendo así los términos de mayor ocurrencia dentro del conjunto de datos. Sin embargo, el término corregimientos está presente en aproximadamente el 12% de los tweets, indicando la incidencia de dicha enfermedad en 8 corregimientos del Departamento.

Tabla 30
Ítems más frecuentes para Leishmaniasis

Ítems	Soporte	Frecuencia absoluta
Leishmaniasis	0.82352941	42
Enfermedades	0.19607843	10
Salud	0.19607843	10
Corregimientos	0.11764706	6
Tratamiento	0.09803922	5

En la Tabla 31 se observan las reglas generadas por el algoritmo, cuyo mayor soporte es de 0,1372 generándose la relación *"{enfermedades, Leishmaniasis} => {válido}"* con una confianza del 100%, es decir que los antecedentes *enfermedades y Leishmaniasis* están presentes en aproximadamente el 14% de los tweets y con una probabilidad del 100% se cumple que aparece el consecuente *válido*. Así mismo, se presenta una variación de la frecuencia absoluta de las relaciones entre valores de 3 a 7, sin embargo, para frecuencias bajas se generan reglas con confianza del 100%, cuyo antecedente está conformado por términos tales como *acciones, control, Dengue, endémicas, intensifican, Leishmaniasis, sanitario* y *veredas,* indicando así la presencia de acciones de control vectorial y sanitario en veredas endémicas de enfermedades como Dengue y Leishmaniasis.

Tabla 31
Reglas de asociación para etiqueta válido en Leishmaniasis

Reglas de asociación	Soporte	Confianza	Frecuencia absoluta
{descubren, nuevo} => {válido}	0,078431	100%	4
{Leishmaniasis, tratamiento} => {válido}	0,098039	100%	5
{endémicas, veredas} => {válido}	0,078431	100%	4
{casos, Leishmaniasis} => {válido}	0,098039	100%	5
{corregimientos, salud} => {válido}	0,078431	100%	4
{enfermedades, Leishmaniasis} => {válido}	0,137255	100%	7
{control, Dengue, vectorial} => {válido}	0,078431	100%	4
{Leishmaniasis, muere} => {válido}	0,098039	100%	5

Continuación tabla 31

Reglas de asociación para etiqueta válido en Leishmaniasis

Reglas de asociación	Soporte	Confianza	Frecuencia absoluta
{acciones, control, Dengue, endémicas, intensifican, Leishmaniasis, sanitario, veredas} => {válido}	0,058824	100%	3
{Leishmaniasis, ministerio, salud} => {válido}	0,058824	100%	3

6.4 Zika

En la Tabla 32 se pueden observar los ítems más frecuentes para la enfermedad Zika, seleccionando aquellos que tienen un valor de soporte mayor respecto de los demás términos, después del termino Zika se puede evidenciar el conjunto de ítems Dengue-Zika y virus-Zika con un soporte de 0,34403 y 0,19495 respectivamente, es decir, los términos Dengue y Zika aparecen en un 34% de los tweets, en comparación con el 19% con el que se presentan los términos virus y Zika, lo que permite concluir que la enfermedad del Zika cuando es tratada, se está relacionando al mismo tiempo con la enfermedad Dengue en el periodo de estudio

Tabla 32

Ítems más frecuentes para Zika

Término (Ítem)	Soporte	Frecuencia absoluta
Zika	0,8876147	387
Dengue, Zika	0,3440367	150
virus	0,1972477	86
virus, Zika	0,1949541	85
salud	0,1284404	56

En la Tabla 33 se pueden observar las reglas que tienen mayor relevancia dentro de dicho conjunto de datos, demostrando así la relación que tienen los antecedentes con su respectivo consecuente. Por ejemplo, para la enfermedad Zika se tiene la siguiente regla de asociación *"{Dengue, mosquitos, Zika} => {válido}"* con un soporte de 0,080 y un nivel de confianza del

100%, lo que quiere decir que los términos *Dengue*, *mosquitos y Zika* del antecedente están presentes en el 8% de los tweets y su consecuente sea válido con una probabilidad del 100%, para corroborar que determinado tweet hace referencia a una Enfermedad Transmitida por Vector.

Tabla 33
Reglas de asociación para etiqueta válido en Zika

Regla de asociación	Soporte	Confianza	Frecuencia absoluta
{enfermedad} => {válido}	0,03440367	94%	15
{transmisor} => {válido}	0,04816514	100%	21
{prevenir, Zika} => {válido}	0,04357798	100%	19
{transmisión, Zika} => {válido}	0,03440367	100%	15
{Aedes, aegypti} => {válido}	0,04357798	100%	19
{Aedes, aegypti, mosquito} => {válido}	0,03669725	100%	16
{control, Dengue, Zika} => {válido}	0,05504587	100%	24
{Dengue, mosquitos, Zika} => {válido}	0,08027523	100%	35
{Chikungunya, salud, Zika} => {válido}	0,03899083	100%	17
{Aedes, aegypti, mosquito, Zika} => {válido}	0,03669725	100%	16

La Tabla 34 muestra la asociación entre los términos que están relacionados con la etiqueta inválido, por ejemplo, *"ver", "dio", "cualquier",* o *"ahora"* hacen parte del antecedente de las reglas. Las reglas presentan igual soporte y confianza con un valor de 0,0045 y 100% respectivamente, por ejemplo, de la regla "*{ahora, dos, Zika} => {inválido}"* se cumple, acorde a los datos, un 100% de las veces, que los términos *ahora*, *dos* y *Zika* estén presentes en el 0,4% de los tweets y su consecuente sea inválido, a su vez, el contenido de los tweets no refleja un posible caso de dicha enfermedad y posee términos irrelevantes para esta investigación.

Tabla 34
Reglas de asociación para etiqueta inválido en Zika

Regla de asociación	Soporte	Confianza	Frecuencia absoluta
{ver, Zika} => {inválido}	0,00458716	100%	2
{dio, toda} => {inválido}	0,00458716	100%	2

Continuación tabla 34

Reglas de asociación para etiqueta inválido en Zika

Regla de asociación	Soporte	Confianza	Frecuencia absoluta
{cualquier, si} => {inválido}	0,00458716	100%	2
{cualquier, Zika} => {inválido}	0,00458716	100%	2
{ahora, dos, Zika} => {inválido}	0,00458716	100%	2
{cualquier, si, Zika} => {inválido}	0,00458716	100%	2

6.5 Chagas

La Tabla 35 muestra los 5 ítems más frecuentes para la enfermedad Chagas, que obtienen un valor de soporte mayor respecto de los demás términos, después de la enfermedad Chagas, los términos Chagas y enfermedad son los que presentan mayor soporte, con un valor de 0.30, es decir, los términos están presentes en aproximadamente el 30% de los tweets, en comparación con el 19% con el que aparecen los términos Chagas y salud al mismo tiempo, lo cual evidencia la relación de dicha enfermedad con términos de salud pública que afectan a la población.

Tabla 35

Ítems más frecuentes para Chagas

Ítems	Soporte	Frecuencia absoluta
Chagas	0,79646018	90
Enfermedad	0,31858407	36
Chagas, Enfermedad	0,30088496	34
Salud	0,20353982	23
Chagas, Salud	0,19469027	22

La Tabla 36 muestra la asociación entre los términos que se relacionan con la etiqueta válido, por ejemplo, los términos *atención, integral, municipios* o *vector* se relacionan con la palabra *Chagas*, el mayor soporte se obtiene de la relación *"{atención, Chagas, integral} => {válido}"* con un valor de 0,0619 y una confianza del 100%, así mismo, existe una probabilidad del 100%

de que la regla se cumpla cuando el antecedente contiene términos como *atención, integral,* y *Chagas* en aproximadamente el 6% de los tweets. Otro aspecto a destacar, es la relación del término *Bucaramanga* con términos como *masiva, declara, fumigación* o *guerra*, haciendo referencia a la masiva fumigación en la ciudad de Bucaramanga, como planes de mitigación de incidencia del Dengue, Zika y Chikungunya.

Tabla 36
Reglas de asociación para etiqueta válido en Chagas

Reglas de asociación	Soporte	Confianza	Frecuencia absoluta
{atención, Chagas, integral} => {válido}	0,061947	100%	7
{chile, europea} => {válido}	0,035398	100%	4
{europea, financiara} => {válido}	0,035398	100%	4
{Bucaramanga, masiva} => {válido}	0,035398	100%	4
{Bucaramanga, declara} => {válido}	0,035398	100%	4
{Chagas, municipios} => {válido}	0,035398	100%	4
{Chagas, vector} => {válido}	0,035398	100%	4
{enfermedad, investigador} => {válido}	0,035398	100%	4
{europea, investigación, unión} => {válido}	0,035398	100%	4
{Bucaramanga, fumigación, guerra} => {válido}	0,035398	100%	4
{atención, enfermedad, integral} => {válido}	0,035398	100%	4

La Tabla 37 muestra la asociación entre los términos que están relacionados con la etiqueta inválido, por ejemplo, *"sífilis", "donantes", "bancos", "hepatitis"* o *"VIH"* hacen parte del antecedente de las reglas. La regla con mayor soporte para la etiqueta inválido es *"{sífilis} => {inválido}*", con un valor de 0,02654 y una confianza del 75%, es decir que esta regla se cumple en el 2,6% de los tweets con una probabilidad del 75% de que al estar presente el término sífilis, está presente la etiqueta inválido. Así mismo, se evidencian reglas con menor soporte, que poseen igual confianza, tal como, *"{donantes, recibiendo, sífilis} => {inválido}"* y "*{confirmar, hepatitis, infecciones} => {inválido}"* ambos antecedentes se encuentran en el 1,7% de los tweets, cuyo significado no se encuentra relacionado con casos reales de ETV. Los términos identificados en el

antecedente de las presentes reglas hacen referencia a los bancos de sangre en portuguesa que no están recibiendo donantes de sangre porque no pueden confirmar que la sangre esté libre de infecciones como VIH, Hepatitis B y C, Chagas y sífilis, a pesar de que su contenido posee términos relacionados con salud pública, la información obtenida no hace parte de Colombia o Santander.

Tabla 37
Reglas de asociación para etiqueta inválido en Chagas

Reglas de asociación	Soporte	Confianza	Frecuencia absoluta
{sífilis} => {inválido}	0,026548673	75%	3
{donantes, portuguesa} => {inválido}	0,017699115	100%	2
{bancos, libre, recibiendo} => {inválido}	0,017699115	100%	2
{donantes, recibiendo, sífilis} => {inválido}	0,017699115	100%	2
{confirmar, hepatitis, infecciones} => {inválido}	0,017699115	100%	2
{Chagas, infecciones, VIH} => {inválido}	0,017699115	100%	2

6.6 Chikungunya

En la Tabla 38 se pueden observar los ítems más frecuentes para la enfermedad Chikungunya, seleccionando aquellos que tienen un valor de soporte mayor respecto de los demás términos, después de la enfermedad Chikungunya se puede evidenciar que los términos con mayor frecuencia son el conjunto Chikungunya y Zika, con un valor de soporte igual a 0,4175, es decir que los términos Chikungunya y Zika están presentes en aproximadamente el 42% de los tweets al mismo tiempo, en comparación con los términos Chikungunya y mosquito, quienes están presentes en un 14% de los tweets, por lo tanto, cuando se trata de la enfermedad Chikungunya se está hablando de la enfermedad Zika, haciendo referencia a la masiva fumigación en la ciudad de Bucaramanga, como planes de mitigación de incidencia del Dengue, Zika y Chikungunya a causa de las fuertes lluvias durante el periodo de estudio.

Tabla 38
Ítems más frecuentes para Chikungunya

Término (Ítem)	Soporte	Frecuencia absoluta
Chikungunya	0.7731959	150
Chikungunya, Zika	0.4175258	81
virus	0.1752577	34
Chikungunya, mosquito	0.1443299	28
Chikungunya, virus	0.1288660	25

La Tabla 39 muestra la asociación entre los términos que se relacionan con la etiqueta válido, por ejemplo, los términos *Dengue, enfermedades, Zika* y *Chikungunya* están relacionados con la etiqueta válido, con un valor de soporte igual a 0,092 y confianza del 100%, es decir que la regla se cumple, acorde a los datos, un 100% de las veces en aproximadamente el 10% de los tweets. Por otra parte, el mayor soporte se obtiene de la relación *"{Chikungunya, mosquito} => {válido}"* con un valor de 0,1391 y una confianza del 96%, lo que quiere decir que los antecedentes *Chikungunya* y *mosquito* están presentes en el 14% de los tweets y con una probabilidad del 96% se cumple que aparece el consecuente *válido;* lo que evidencia un plan de mitigación de los criaderos de mosquitos transmisores del Dengue, Zika y Chikungunya.

Tabla 39
Reglas de asociación para etiqueta válido en Chikungunya

Regla de asociación	Soporte	Confianza	Frecuencia absoluta
{transmisor} => {válido}	0,082474227	100%	16
{enfermedades} => {válido}	0,118556701	100%	23
{Chikungunya, transmisor} => {válido}	0,077319588	100%	15
{Chikungunya, salud} => {válido}	0,092783505	100%	18
{Chikungunya, enfermedades} => {válido}	0,113402062	100%	22
{Chikungunya, virus} => {válido}	0,128865979	100%	25
{Chikungunya, mosquito} => {válido}	0,139175258	96%	27
{Chikungunya, Dengue, enfermedades, Zika} => {válido}	0,092783505	100%	18

Continuación tabla 39

Reglas de asociación para etiqueta válido en Chikungunya

Regla de asociación	Soporte	Confianza	Frecuencia absoluta
{Aedes, aegypti, Chikungunya, Dengue, Zika} => {válido}	0,082474227	100%	16
{Aedes, Chikungunya, Dengue, mosquito, Zika} => {válido}	0,082474227	100%	16

La Tabla 40 muestra la asociación entre los términos que están relacionados con la etiqueta inválido, por ejemplo, *"miércoles"," animalito"," sorteo"* o *"pm",* hacen parte del antecedente de las reglas, los cuales no tienen relación con posibles casos de enfermedades y no corresponden a términos referentes en salud pública. La regla con mayor soporte para la etiqueta inválido es *"{Chikungunya, dio} => {inválido}",* con un valor de 0,0206 y una confianza del 100%, es decir que los términos *Chikungunya* y *dio* se encuentran en aproximadamente el 2% de los tweets y tienen un 100% de probabilidad de que su consecuente sea inválido. Así mismo, se evidencian reglas con menor soporte, sin embargo, poseen igual confianza.

Tabla 40

Reglas de asociación para etiqueta inválido en Chikungunya

Regla de asociación	Soporte	Confianza	Frecuencia absoluta
{Chikungunya, dio} => {inválido}	0,020618557	100%	4
{Chikungunya, menos} => {inválido}	0,010309278	100%	2
{miércoles, pm, sorteo} => {inválido}	0,010309278	100%	2
{animalito, miércoles, pm} => {inválido}	0,010309278	100%	2
{Chikungunya, miércoles, sorteo} => {inválido}	0,010309278	100%	2
{Chikungunya, miércoles, pm, sorteo} => {inválido}	0,010309278	100%	2

7. Caso de estudio

Bajo las mismas condiciones computacionales, se valida el modelo de SVM con kernel tipo Lineal y Polinomial, con el fin de evaluar el rendimiento del modelo de clasificación anterior ante la

presencia de nuevos datos de entrada y obtener sus métricas respectivas a partir de los 194 tweets recopilados del área de Santander, Colombia. Por consiguiente, se realiza el respectivo preprocesamiento de los datos, tal como la eliminación de tweets duplicados, de signos de puntuación, y de stopwords, conversión de texto a minúscula, reducción de palabras a su raíz, entre otros, y la representación del texto por medio de la matriz tf-idf, con el fin de realizar un análisis de contenido del texto identificando la importancia de cada término que lo conforma.

A continuación, se presenta en la Tabla 41 la matriz de confusión obtenida por el modelo Polinomial:

Tabla 41.
Matriz de confusión kernel Polinomial (Caso de estudio)

Observado	**Predicho**	
	Inválido	**Válido**
Inválido	8	0
Válido	1	38

Para el caso del kernel Lineal, la matriz de confusión arrojada por el modelo se presenta en la siguiente tabla:

Tabla 42.
Matriz de confusión kernel Lineal (Caso de estudio)

Observado	**Predicho**	
	Inválido	**Válido**
Inválido	8	0
Válido	1	38

A partir de la matriz de confusión arrojada por cada tipo de kernel, en la Tabla 43, se presentan las métricas previamente definidas para evaluar el modelo.

Tabla 43.
Métricas del modelo (Caso de estudio)

Métricas	Kernel	
	Lineal	**Polinomial**
Accuracy	0,9787234	0,978723404
Precisión	1	1
Recall	0,97435897	0, 97435897
F-Measure	0,98701299	0, 98701299

Los kernel evaluados presentaron una exactitud (Accuracy) superior al 97%, una precisión del 100%, una sensibilidad (Recall) superior al 97% y una medida F (F-Measure) superior al 98%, lo cual valida la SVM con kernel de tipo Lineal y Polinomial como una herramienta eficiente de clasificación de tweets. A su vez, un clasificador SVM presenta menor costo computacional cuando utiliza un kernel de tipo Lineal y su desempeño no difiere en mayor proporción de un clasificador SVM con kernel de tipo Polinomial, situación que se presenta cuando se evalúa y valida el modelo con los datos obtenidos a nivel nacional. A continuación (Figura 30), se observan las métricas mediante la cuales se comparan los dos tipos de kernel:

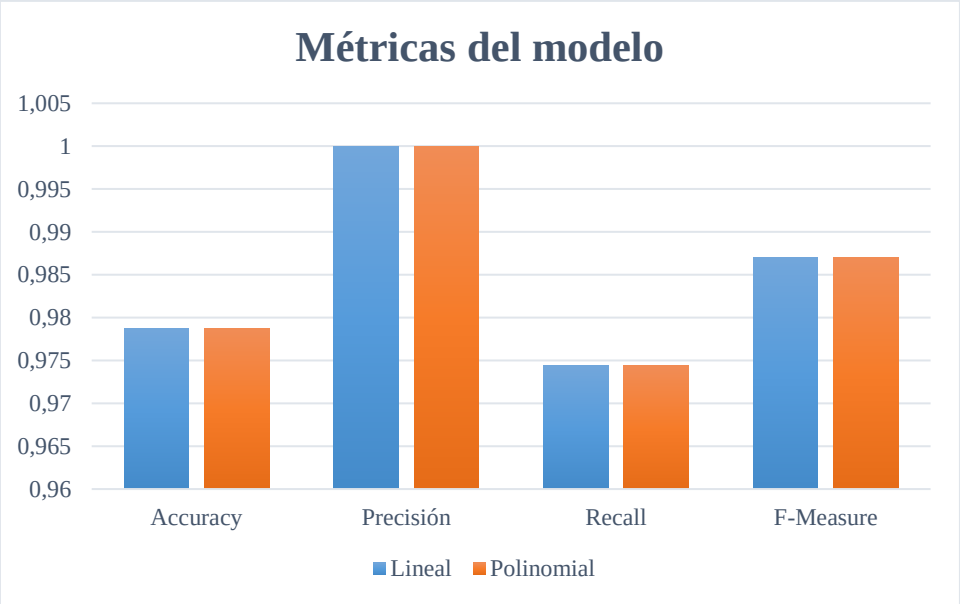

Figura 30. Métricas del modelo (Caso de estudio)

Con el fin de analizar los términos que aparecen tanto en los tweets etiquetados como válidos, como en los inválidos, se procede aplicar la técnica de reglas de asociación, teniendo en cuenta, las condiciones de los mejores valores para soporte y confianza obtenidos anteriormente. A continuación, se muestran los ítems más frecuentes y las reglas que cumplan con los mínimos establecidos tanto para soporte como para confianza. Se utiliza el algoritmo *apriori* de la librería "*arules"* con apoyo del software R, el cual permite encontrar de forma eficiente conjuntos de ítems frecuentes que sirven de base para generar reglas de asociación.

En la Tabla 44 se pueden observar los ítems más frecuentes para el Departamento de Santander, seleccionando aquellos que tienen un valor de soporte mayor respecto de los demás términos, se puede evidenciar que las enfermedades que mayor incidencia tienen dentro del Departamento son Dengue con un soporte de 0,8510 y Zika con un soporte de 0,31914, es decir, el término Dengue está presente en el 85% de los tweets, en comparación con el 31% con el que está presente la enfermedad Zika. Así mismo, se destacan términos como investigación, simposio y red, haciendo referencia al evento celebrado en el Departamento correspondiente al simposio de investigación de la red Aedes.

Tabla 44
Ítems frecuentes para Santander, Colombia

Ítems	Soporte	Frecuencia absoluta
Dengue	0,8510638	40
Zika	0,3191489	15
investigación	0,2765957	13
Dengue, investigación	0,2765957	13
simposio	0,2553191	12
Dengue, simposio	0,2553191	12
red	0,2340426	11
Aedes	0,2340426	11
red, simposio	0,2340426	11

En la Tabla 45 se muestran algunos ejemplos de las reglas de asociación generadas para el Departamento de Santander con un nivel de confianza del 70% y un soporte de 0,04255319. En los antecedentes de las reglas de asociación *"{Dengue, investigación, salud} => {válido}"*, *"{Dengue, prevención, Zika} => {válido}"*, *"{Aedes, investigación, Zika} => {válido}"*, se puede evidenciar la relación que existe entre las enfermedades Zika y Dengue para la obtención de un consecuente válido, lo que significa que en Santander son las dos ETV que predominan y a su vez, la alcaldía de Bucaramanga está realizando campañas de sensibilización para la prevención de dichas enfermedades. Las reglas con mayor soporte son *"{casos} => {válido}"* y *"{Dengue, prevención, Zika} => {válido}"* con un valor de 0,127659574 y un nivel de confianza del 100%, es decir, que los términos casos o Dengue, prevención y Zika están presentes en aproximadamente el 13% de los tweets y existe una probabilidad del 100% de que la regla se cumpla. Por otra parte, son de gran interés las reglas cuyo antecedente contienen el término *epidemia*, indicando una alerta de posibles altas incidencias de enfermedades como Zika y Dengue en el Departamento.

Tabla 45

Reglas de asociación para etiqueta válido - Santander, Colombia

Reglas de asociación	Soporte	Confianza	Frecuencia absoluta
{epidemia} => {válido}	0,085106383	100%	4
{casos} => {válido}	0,127659574	100%	6
{enfermedades} => {válido}	0,106382979	100%	5
{virus} => {válido}	0,085106383	100%	4
{Dengue, medidas} => {válido}	0,063829787	100%	3
{Dengue, Santander} => {válido}	0,063829787	100%	3
{Aedes, epidemia} => {válido}	0,063829787	100%	3
{epidemia, investigación} => {válido}	0,063829787	100%	3
{Dengue, epidemia} => {válido}	0,085106383	100%	4
{Dengue, enfermedades} => {válido}	0,085106383	100%	4
{diagnostico, investigación} => {válido}	0,085106383	100%	4
{virus, Zika} => {válido}	0,085106383	100%	4
{Dengue, investigación, salud} => {válido}	0,063829787	100%	3

Continuación tabla 45

Reglas de Asociación para etiqueta válido - Santander, Colombia

Reglas de asociación	Soporte	Confianza	Frecuencia absoluta
{Dengue, prevención, Zika} => {válido}	0,127659574	100%	6
{Aedes, investigación, Zika} => {válido}	0,085106383	100%	4

En la Tabla 46 se muestran las reglas de asociación generadas para la etiqueta inválido, el número de reglas generadas es muy limitado debido a que no existen relación entre los datos cuyo consecuente es inválido, además el número de tweets para dicha etiqueta es muy reducido. Sin embargo, se observan dos reglas cuyo antecedente contiene términos como "*vida*" o "*Dengue, vida*" con un soporte de 0,042 y una confianza del 100%, es decir, que los términos Dengue y vida están presentes en aproximadamente el 4% de los tweets al mismo tiempo y existe una probabilidad del 100% de que la regla se cumpla.

Tabla 46

Reglas de asociación para etiqueta inválido - Santander, Colombia

Reglas de asociación	Soporte	Confianza	Frecuencia absoluta
{vida} => {inválido}	0,04255319	100%	2
{Dengue, vida} => {inválido}	0,04255319	100%	2

8. Conclusiones

La captura de información para conformar el repositorio de datos se basó en la metodología utilizada por Allen et al., (2016), de manera que en este proyecto se utiliza el API de twitter con apoyo del software R, y a partir de la construcción de radios de búsqueda que tenían como centroide coordenadas de latitud y longitud, se considera que este método de extracción de información es adecuado para involucrar dentro de la base de datos objeto de estudio cualquier

opinión de los usuarios que interactúan en la red social twitter alrededor de una tema en específico, en este caso en particular las opiniones relacionadas con ETV.

En los documentos soporte de la revisión de literatura, se destacan las SVM, como métodos de clasificación para analizar información difundida sobre enfermedades en la red social Twitter como lo menciona en su estudio Aslam et al., (2014), así, este método de aprendizaje automático aplicado en el proyecto de investigación fue adecuado porque permitió relacionar los atributos y las etiquetas de clasificación (válido o inválido) de cada tweet, con el fin de descubrir aquella información que está relacionada o no con posible presencia de casos de ETV.

Las SVM con kernel Polinomial tienen mejor desempeño, pero su entrenamiento requiere un tiempo de procesamiento (32 horas) significativamente mayor, en comparación con el kernel lineal (21 minutos). Por tanto, un kernel de tipo lineal resultará más adecuado para apoyar a los tomadores de decisiones en caso de requerir resultados inmediatos y su rendimiento no difiere considerablemente en relación con el de un kernel Polinomial.

La red social Twitter es una fuente complementaria de datos que permite apoyar a los tomadores de decisiones para revisar posibles estrategias de control y prevención en cuanto a situaciones específicas relacionadas con salud pública, y también permitiría realizar intervenciones y monitoreo en tiempo real. En este sentido, se destaca la oportunidad de utilizar métodos descriptivos como reglas de asociación para identificar la relación entre los términos obtenidos en un mensaje corto con el fin de descubrir posibles situaciones que puedan afectar la salud de una comunidad objeto de estudio.

En el caso de estudio se valida que los modelos aplicados con fines similares al de este proyecto, tienen la capacidad de que los datos analizados están representados por el modelo y que este modelo se desempeña adecuadamente frente a nuevos datos, por tanto, estos métodos de analítica podrían ser una oportunidad para ser adoptados por las entidades tomadoras de decisiones en cuanto a aspectos de salud pública no solo enfocados al análisis de ETV sino también al análisis de otros eventos de interés de salud pública tanto a nivel nacional como departamental, esto con el fin de apoyar los métodos de vigilancia tradicionales.

9. Recomendaciones

Implementar modelos de aprendizaje semi-supervisado, los cuales permiten aprovechar los datos no etiquetados y obtener un modelo predictivo, en donde el reto es utilizar una pequeña muestra de datos etiquetados y obtener el mismo nivel de resultados que un método supervisado, es decir, se reduce el esfuerzo en etiquetar, lo que disminuye los costos asociados a asignar etiquetas. Actualmente existe una gran cantidad de datos disponibles, sin embargo, no todos tienen asignada una clase o etiqueta, con la cual realizar la clasificación. Los ejemplos más claros de esto son los datos no estructurados como texto e imágenes, en donde existe una gran cantidad de información no etiquetada.

Se recomienda tener en cuenta diferentes fuentes de datos para la recopilación de la información en las plataformas virtuales, tales como, blogs, consultas de motores de búsqueda, Facebook, Instagram, entre otras, que permitan ampliar el repositorio objeto de estudio y aprovechar la

revolución digital de las plataformas de redes sociales, puesto que, la red social Twitter a través de su API posee una limitante de tiempo al momento de realizar la captura de la información.

Los diferentes usuarios de Twitter publican información variada durante cada día, ya sean tweets, conversaciones con amigos o retweets, lo cual genera gran cantidad de datos en esta red social, convirtiéndose en un fuente de datos destacable para la captura de la información en tiempo real, con el fin de proporcionar oportunidades de detección y monitoreo ante un evento en tiempo real.

Se recomienda dar a conocer a las entidades interesadas en salud pública la oportunidad de realizar vigilancia digital y de adoptar los modelos desarrollados en la presente investigación con el fin de complementar la vigilancia tradicional, y así, generar valor agregado a dichas entidades.

Referencias Bibliográficas

AL-Rubaiee, H., Qiu, R., & Li, D. (2017). Visualising Arabic Sentiments and Association Rules in Financial Text. *International Journal of Advanced Computer Science and Applications*, *8*(2), 1–7. https://doi.org/http://dx.doi.org/10.14569/IJACSA.2017.080201

Allen, C., Tsou, M. H., Aslam, A., Nagel, A., & Gawron, J. M. (2016). Applying GIS and machine learning methods to twitter data for multiscale surveillance of influenza. *PLoS ONE*, *11*(7), 1–10. https://doi.org/10.1371/journal.pone.0157734

Alpaydin, E. (2010). *Introduction to Machine Learning Second Edition*. *Introduction to Machine Learning*. https://doi.org/10.1007/978-1-62703-748-8_7

Andrea, J., Hernández, R., Milena, D., Herrera, R., & Rodríguez Rodríguez, J. E. (2016). A Research Comparative Among Association Rules Algorithms, *10*(2), 210–217.

Aslam, A. A., Tsou, M. H., Spitzberg, B. H., An, L., Gawron, J. M., Gupta, D. K., … Lindsay, S. (2014). The reliability of tweets as a supplementary method of seasonal influenza surveillance. *Journal of Medical Internet Research*, *16*(11). https://doi.org/10.2196/jmir.3532

Beltrán, M. B. (2014). Minería de datos. *Benemérita Universidad Autónoma de Puebla*, *1*(5), 67. Retrieved from http://bbeltran.cs.buap.mx/NotasMD.pdf

Bernardo, T. M., Rajic, A., Young, I., Robiadek, K., Pham, M. T., & Funk, J. A. (2013). Scoping review on search queries and social media for disease surveillance: A chronology of innovation. *Journal of Medical Internet Research*, *15*(7), 1–13. https://doi.org/10.2196/jmir.2740

Betancour, G. (2005). Las máquinas de soporte vectorial (SVMs). *Scientia Et Technica*, (27), 67–

72. https://doi.org/10.22517/23447214.6895

Blanco, E.-J., & Sanz, H. (2016). Algoritmos de clustering y aprendizaje automático aplicados a Twitter. *Universidad Politécnica de Catalunya*. Retrieved from https://upcommons.upc.edu/bitstream/handle/2117/82434/113257.pdf?sequence=1&isAllowed=y

Brownstein, J. S., Freifeld, C. C., & Madoff, L. C. (2010). Digital disease detection--harnessing the Web for public health surveillance. *Search*, *360*(November 2002), 2153–2157. https://doi.org/10.1056/NEJMp0900702.Digital

Burton, S. H., Tanner, K. W., Giraud-Carrier, C. G., West, J. H., & Barnes, M. D. (2012). Right time, right place" health communication on twitter: Value and accuracy of location information. *Journal of Medical Internet Research*, *14*(6). https://doi.org/10.2196/jmir.2121

Carneiro, H. A., & Mylonakis, E. (2009). Google Trends: A Web- Based Tool for Real- Time Surveillance of Disease Outbreaks. *Clinical Infectious Diseases*, *49*(10), 1557–1564. https://doi.org/10.1086/630200

Chen, L. S., Lin, Z. C., & Chang, J. R. (2015). FIR: An Effective Scheme for Extracting Useful Metadata from Social Media. *Journal of Medical Systems*, *39*(11). https://doi.org/10.1007/s10916-015-0333-0

Chen, M.-S., Han, J., & Yu, P. S. (1996). Data Mining: An Overview from Database Perspective. *IEEE Transactions on Knowledge and Data Engineeringand Data Engineering*, *8*, 866–883.

D.Manning, C., Raghavan, P., & Schütze, H. (2009). Introduction to Information Retrieval. *Information Retrieval*, (c), 1–18. https://doi.org/10.1109/LPT.2009.2020494

De Moya Amaris, M. E., & Rodríguez Rodríguez, J. E. (2003). LA CONTRIBUCIÓN DE LAS REGLAS DE ASOCIACIÓN A LA MINERÍA DE DATOS, 94–109.

https://doi.org/10.1016/S1076-6332(98)80562-X

Dunn, A. G., Leask, J., Zhou, X., Mandl, K. D., & Coiera, E. (2015). Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: An observational study. *Journal of Medical Internet Research*, *17*(6), e144. https://doi.org/10.2196/jmir.4343

Elkan, C. (2011). Evaluating Classifiers. *University of San Diego, California, Retrieved [01-11-2012] from Http://Cseweb. Ucsd. Edu/~ Elkan B*, *250*, 1–11. https://doi.org/10.1145/775107.775137

Erik G. (2014). Introduction to Supervised Learning, 1–5. Retrieved from https://people.cs.umass.edu/~elm/Teaching/Docs/supervised2014a.pdf%0Ahttp://people.cs.umass.edu/~elm/Teaching/Docs/supervised2014a.pdf

Felgaer, P., Britos, P., Sicre, J., Servetto, A., García-Martínez, R., & Perichinsky, G. (2003). Optimización de Redes Bayesianas Basada en Técnicas de Aprendizaje por Instrucción, 1687.

FH Joanneum. (2005). Cross-Validation Explained. Retrieved from http://genome.tugraz.at/proclassify/help/pages/XV.html

Fox, S., & Jones, S. (2009). The social life of health information. *Washington, DC: Pew Internet & American Life …*, 1–72. Retrieved from file://r/Article PDFs/Inflexxion Research Master/Inflexxion Research Master/1174/social life of health information.pdf%5Cnhttp://classweb.gmu.edu/gkreps/721/20. Fox, The Social Life of Health Information.pdf

Gómez, F. D. (2014). Aprendizaje automático: métodos y aplicaciones. Retrieved from http://www.eui.uva.es/doc/anuncios/curso.2013.2014/aprendizajeAutomatico.pdf

Gómez, M. M. y. (2005). Minería de texto empleando la semejanza entre estructuras semánticas.

Computación y Sistemas, *9*(1), 63–81.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of Medical Internet Research*, *15*(11), 1–9. https://doi.org/10.2196/jmir.2721

Greene, M. (2010). Epidemiological monitoring for emerging infectious diseases. *SPIE Defense, Security, and Sensing*, *7666*, 76661B--76661B. https://doi.org/10.1117/12.849351

Hamad, E. O., Savundranayagam, M. Y., Holmes, J. D., Kinsella, E. A., & Johnson, A. M. (2016). Toward a mixed-methods research approach to content analysis in the digital age: The combined content-analysis model and its applications to health care twitter feeds. *Journal of Medical Internet Research*, *18*(3), 1–17. https://doi.org/10.2196/jmir.5391

Hernández Orallo, J., & Ferri Ramírez, C. (2006). Práctica de Minería de Datos Introducción al weka. Retrieved from http://users.dsic.upv.es/~jorallo/docent/doctorat/weka.pdf

Hussain, A., & Cambria, E. (2018). Semi-supervised learning for big social data analysis. *Neurocomputing*, *275*, 1662–1673. https://doi.org/10.1016/j.neucom.2017.10.010

Jain, V. K., & Kumar, S. (2018). Rough set based intelligent approach for identification of H1N1 suspect using social media, *45*(2), 8–14.

Ji, X., Chun, S. A., Wei, Z., & Geller, J. (2015). Twitter sentiment classification for measuring public health concerns. *Social Network Analysis and Mining*, *5*(1), 1–25. https://doi.org/10.1007/s13278-015-0253-5

Kagashe, I., Yan, Z., & Suheryani, I. (2017). Enhancing seasonal influenza surveillance: Topic analysis of widely used medicinal drugs using twitter data. *Journal of Medical Internet Research*, *19*(9), 1–14. https://doi.org/10.2196/jmir.7393

Kotsiantis, S., & Kanellopoulos, D. (2006). Association Rules Mining: A Recent Overview.

Science, *32*(1), 71–82.

Lampos, V., Yom-Tov, E., Pebody, R., & Cox, I. J. (2015). Assessing the impact of a health intervention via user-generated Internet content. *Data Mining and Knowledge Discovery*, *29*(5), 1434–1457. https://doi.org/10.1007/s10618-015-0427-9

Liu, B., Ma, Y., & Wong, C. K. (2000). Improving an Association Rule Based Classifier, 504–509. https://doi.org/10.1007/3-540-45372-5_58

Lopes, L. F., Silva, F., Couto, F., Zamite, J., Ferreira, H., Sousa, C., & Silva, M. J. (2010). Epidemic Marketplace: An Information Management System for Epidemiological Data. *Information Technology in Bio- and Medical Informatics, ITBAM 2010, 6266*, 31–44. Retrieved from https://link.springer.com/chapter/10.1007%2F978-3-642-15020-3_3#citeas

Matich, D. J. (2001). Redes Neuronales: Conceptos Básicos y Aplicaciones. *Historia*, 55. Retrieved from ftp://decsai.ugr.es/pub/usuarios/castro/Material-Redes-Neuronales/Libros/matich-redesneuronales.pdf

Ministerio de Salud. (n.d.). Salud Pública. Retrieved from https://www.minsalud.gov.co/salud/Paginas/salud-publica.aspx

Ministerio de Salud y Proteccion Social, & Federacion Medica Colombiana. (2013). Memorias © 2012 - 2013. *ILADIBA Educación En Salud*, 1–31.

Moore, A. (2001). Cross-validation for detecting and preventing overfitting. *Science*, 1–27. Retrieved from http://www.cs.cmu.edu/~cga/ai-course/overfit.pdf

Nagar, R., Yuan, Q., Freifeld, C. C., Santillana, M., Nojima, A., Chunara, R., & Brownstein, J. S. (2014). A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. *Journal of Medical Internet Research*, *16*(10), e236. https://doi.org/10.2196/jmir.3416

Nair, L. R., Shetty, S. D., & Shetty, S. D. (2018). Applying spark based machine learning model on streaming big data for health status prediction. *Computers and Electrical Engineering*, *65*, 393–399. https://doi.org/10.1016/j.compeleceng.2017.03.009

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., & Gonzalez, G. (2015). Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, *22*(3), 671–681. https://doi.org/10.1093/jamia/ocu041

Observatorio de Salud Pública de Santander. (2017). Indicadores Básicos Situación de Salud en Santander. *Salud En Números/Indicadores Básicos*, 48. Retrieved from http://www.dgis.salud.gob.mx/contenidos/sinais/indica_basicos.html

Ochoa, M. E. (2014). Enfermedades de transmisión vectorial (ETV). Santander, 2013. *Informe Epidemiológico de Santander*, 322908.

Orallo, J. H., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004). *Introducción a la minería de datos* (Pearson).

Organización Mundial de la Salud. (n.d.). Información sobre las enfermedades transmitidas por vectores. Retrieved from http://www.who.int/campaigns/world-health-day/2014/vector-borne-diseases/es/

Organización Mundial de la Salud. (2016). Enfermedad por el virus de Zika. Retrieved from http://www.who.int/mediacentre/factsheets/zika/es/

Organización Mundial de la Salud. (2018a). Leishmaniasis. Retrieved from http://www.who.int/mediacentre/factsheets/fs375/es/

Organización Mundial de la Salud. (2018b). Paludismo.

Organización Mundial de la Salud. (n.d.-a). Dengue. Retrieved from

http://www.who.int/topics/dengue/es/

Organización Mundial de la Salud. (n.d.-b). Enfermedad de Chagas. Retrieved from http://www.who.int/topics/chagas_disease/es/

Organización Mundial de la Salud. (2017). Chikungunya. Retrieved from http://www.who.int/mediacentre/factsheets/fs327/es/

Palma, M., Tomás, J., & Morales, R. M. (2008). *Inteligencia artificial: métodos, técnicas y aplicaciones* (McGraw-Hil). España. Retrieved from https://ebookcentral.proquest.com/lib/bibliouissp/detail.action?docID=3194970.

Papalexakis, E. E., Faloutsos, C., & Sidiropoulos, N. D. (2016). Tensors for Data Mining and Data Fusion. *ACM Transactions on Intelligent Systems and Technology*, *8*(2), 1–44. https://doi.org/10.1145/2915921

Payam, R., Lei, T., & Huan, L. (2008). Cross-Validation. *Arizona State University*. https://doi.org/10.1159/000331070

Pérez-Planells, L., Delegido, J., Rivera-Caicedo, J. P., & Verrelst, J. (2015). Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. *Revista de Teledeteccion*, *2015*(44), 55–65. https://doi.org/10.4995/raet.2015.4153

Perichinsky, G., Servente, M., Servetto, A., García-Martínez, R., Orellana, R., & Plastino, A. (2003). Taxonomic Evidence and Robustness of the Classification Appying Intelligent Data Mining, 1797–1808.

Prieto, V. M., Matos, S., Álvarez, M., Cacheda, F., & Oliveira, J. L. (2014). Twitter: A Good Place to Detect Health Conditions. *PLoS ONE*. https://doi.org/10.1371/Citation

Rodríguez González, A. Y., Martínez Trinidad, J. F., Carrasco Ochoa, J. A., & Ruiz Shulcloper, J. (2009). Minería de Reglas de Asociación sobre Datos Mezclados. *Inaoe*, 35. Retrieved from

http://ccc.inaoep.mx/portalfiles/file/CCC-09-001.pdf

Rodríguez, Y., & Díaz, A. (2009). Herramientas de Minería de Datos. *Rcci*, *3*, 73–80. Retrieved from https://msdn.microsoft.com/es-es/library/ms174467(v=sql.120).aspx

Rueda Restrepo, J. A., & Cotes Torres, J. M. (2009). EVALUACIÓN DE DOS MÉTODOS DE ESTABILIDAD FENOTÍPICA A TRAVÉS DE VALIDACIÓN CRUZADA. *Revista Facultad Nacional de Agronomía Medellín.*, *62*, 5111–5123. Retrieved from https://revistas.unal.edu.co/index.php/refame/article/view/24923/36933

Ruiz Sánchez, R. (2006). Heurísticas de selección de atributos para datos de gran dimensionalidad, 189. Retrieved from http://www.tdx.cat/handle/10803/114710

Salathé, M. (2016). Digital pharmacovigilance and disease surveillance: Combining traditional and big-data systems for better public health. *Journal of Infectious Diseases*, *214*(Suppl 4), S399–S403. https://doi.org/10.1093/infdis/jiw281

Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., … Gonzalez, G. (2015). Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, *54*, 202–212. https://doi.org/10.1016/j.jbi.2015.02.004

Sparks, R. S., Robinson, B., Power, R., Cameron, M., & Woolford, S. (2017). An investigation into social media syndromic monitoring. *Communications in Statistics: Simulation and Computation*, *46*(8), 5901–5923. https://doi.org/10.1080/03610918.2016.1186182

Trstenjak, B., Mikac, S., & Donko, D. (2009). KNN with TF-IDF Based Framework for Text Categorization. *International Journal of Pharmaceutics*, *381*(2), 205–213. https://doi.org/10.1016/j.proeng.2014.03.129

Wang, Y. C., Kraut, R. E., & Levine, J. M. (2015). Eliciting and receiving online support: Using computer-aided content analysis to examine the dynamics of online social support. *Journal*

of Medical Internet Research, *17*(4), e99. https://doi.org/10.2196/jmir.3558

World Health Organization, & Regional Office for South-East Asia. (2014). Vector-borne diseases. Report of an informal expert consultation, (April), 7–8. Retrieved from http://apps.who.int/iris/handle/10665/206531

Xia, H., & Huan, L. (2013). Text Analytics in Social Media. *Springer Science*, *9781461432*, 1–522. https://doi.org/10.1007/978-1-4614-3223-4

Yin, Z., Fabbri, D., Rosenbloom, S. T., & Malin, B. (2015). A scalable framework to detect personal health mentions on Twitter. *Journal of Medical Internet Research*, *17*(6), e138. https://doi.org/10.2196/jmir.4305

Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997). New Algorithms for Fast Discovery of Association Rules. *Data Mining and Knowledge Discovery*, *1*(4), 343–373. https://doi.org/10.1023/A:1009773317876