Instituto Politécnico Nacional

Centro de Investigación en Computación

Monitoreo urbano de entidades y eventos geográficos basado en censado social

TESIS

Que para obtener el grado de:

Maestría en Ciencias de la Computación

Presenta:

Ing. Juan Carlos Salazar Carrillo

Directores:

Dr. Miguel Jesús Torres Ruiz

Dr. Marco Antonio Moreno Ibarra

Diciembre, 2015

SIP-14 bis

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D.F. siendo las 10:00 horas del día 07 del mes de diciembre de 2015 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis titulada:

"Monitoreo urbano de entidades y eventos geográficos basado en censado social"

Presentada por el alumno:

SALAZAR	**CARRILLO**	**JUAN CARLOS**
Apellido paterno	Apellido materno	Nombre(s)

Con registro:

B	1	3	0	1	2	6

aspirante de: **MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***APROBAR LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Directores de Tesis

Dr. Miguel Jesús Torres Ruiz

Dr. Marco Antonio Moreno Ibarra

Dr. Grigori Sidorov

Dr. Rolando Quintero Téllez

Dr. Clodoveu Augusto Davis Junior

M. en C. Sandra Dinora Orantes Jiménez

PRESIDENTE DEL COLEGIO DE PROFESORES

Dr. Luis Alfonso Villa Vargas

ESTADOS UNIDOS MEXICANOS
INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION EN COMPUTACION
DIRECCION

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de **México** el día **15** del mes **Diciembre** del año **2015**, el (la) que suscribe **Juan Carlos Salazar Carrillo** alumno (a) del Programa de **Maestría en Ciencias de la Computación** con número de registro **B130126**, adscrito a **Centro de Investigación en Computación,** manifiesta que es autor (a) intelectual del presente trabajo de Tesis bajo la dirección de **Dr. Miguel Jesús Torres Ruiz** y **Dr. Marco Antonio Moreno Ibarra** y cede los derechos del trabajo intitulado **Monitoreo urbano de entidades y eventos geográficos basado en censado social,** al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección **sacj.88@gmail.com**. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Juan Carlos Salazar Carrillo

Nombre y firma

Resumen

La proliferación de las redes sociales y la facilidad que otorgan a sus usuarios para compartir información alienta a publicar todo lo que ocurre a su alrededor. Las redes sociales ayudan a conocer diferentes acontecimientos. Conocer los eventos que ocurren en el entorno ayudan a tomar decisiones con mayor certeza. Por ejemplo, el tráfico vehicular, el conocer mediante redes sociales manifestaciones o un accidente en alguna avenida ayuda a eludir un congestionamiento vial desafortunado. Aprovechando la generación de contenido en Twitter y tomando como caso de estudio la Ciudad de México, se recolectó información de usuarios dedicados a publicar eventos viales. Por tanto, se propone una metodología para la *geocodificación* de textos cortos de Twitter y un método de aprendizaje automático basado en Máquinas de Soporte Vectorial para Regresión, con el cual se obtiene un modelo capaz de realizar un análisis espacio temporal de eventos viales. Los resultados experimentales muestran predicciones de eventos viales con una precisión aproximada de 74% dependiendo del número de *tweets* usados para generar el modelo de regresión y el umbral establecido para decir que ha ocurrido un acierto.

Este trabajo muestra eventos viales ocurridos, provenientes de Twitter y muestra un panorama de posibles congestionamientos dando características de tiempo y espacio. Como resultados se intenta optimizar la circulación vial, tomar decisiones para mejorar servicios de transporte público, revisión de semáforos, y administración de personal de tránsito. La metodología se clasifica en ocho etapas: (1) Recolección de información, (2) Análisis de la información y obtención de datos sobresalientes, (3) Creación de diccionarios y ejes equivalentes, (4) División del Gazetteer, (5) Estandarización, (6) Identificación y localización de eventos viales, (7) Generación de modelos de predicción y (8) Pronóstico de eventos viales. Finalmente, los resultados obtenidos por el modelo de predicción son evaluados mediante las medidas de *precision* y *recall*.

Abstract

Publishing information through social networks serves to know different kind of events, for example, social, cultural, economic, political and so on. To know events that are happening in the city help us to make decisions with clarity and certainty. As an illustration, be aware about traffic conditions aid us to avoid conflictive areas and make our travel pleasant. Social networks like Twitter besides of share ideas has many accounts that post certain information about distinct topics such as, art, health, business, entertainment, culture, sports, etc. Taking advantage from this user-generated content and taking Mexico City as a case of study, we collected information about traffic conditions. With this information, a machine learning method called Support Vector Regression (SVR) was trained; SVR generates a model capable to make spatiotemporal analysis of the traffic conditions. SVR has been widely used to make predictions of petrol, predictions of the stock market, and predictions of obesity rate, time travelling predictions or predictions of population growing. Preliminary results showed prediction of traffic conditions with precision near to 74%, depending of the number of tweets used to generate the prediction model and the threshold established to have a hit.

This work tries to be another source of information use to warning people about probable traffic conditions, therefore we pretend to reduce the concentration of cars, likewise serves as guideline for making decisions in order to improve public transportation, identification of broken traffic signals, management of traffic police, etc. Eight sections compose this work: (1) Retrieval information, (2) information from the tweet dataset, (3) dictionaries and equivalent road axis names, (4) division of the gazetteer, (5) standardization, (6) identification and location of traffic-related events, (7) prediction models (8) spatiotemporal analysis. Finally, this model is evaluated through precision and recall measures.

Agradecimientos

Agradezco al Instituto Politécnico Nacional y a sus profesores por la educación que me otorgaron, razón por la cual tengo los conocimientos teóricos y prácticos para poder realizar este trabajo. Agradezco a mi familia por ser siempre un apoyo económico, anímico y ético, principalmente a mi madre por ser una agradable compañía en esas noches de desvelo. Agradezco a mi novia Ximena que me brindó su consejo y apoyo, a mi lado siempre en los malos y buenos momentos, espero así sea siempre.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. Introducción

Más de la mitad de la población del mundo vive en áreas urbanizadas y el cambio de poblaciones rurales a urbanas sigue proyectándose en las siguientes décadas. Es inevitable que poblaciones de gran magnitud no lleguen a ser lugares confusos y desordenados [Chourabi, H. et al., 2012].

Ciudades y megalópolis generan nuevos tipos de problemas como administración de basura, carencia de recursos, contaminación de aire, problemas de salud en la población, desplazamiento de sus habitantes, congestionamiento de tráfico, y deterioro, inadecuación y envejecimiento de infraestructuras son los problemas básicos técnicos, físicos y materiales.

El Instituto Nacional de Estadística y Geografía (INEGI) es la dependencia encargada de generar estadística básica en el país, utiliza como fuentes de información censos, encuestas y registros administrativos; además se encarga de generar información geográfica del relieve, la vegetación, el clima, el suelo y las localidades. De acuerdo con el último censo de población publicado por el INEGI, se observa una tasa de crecimiento media anual del 1.8% del 2005 al 2010.

El desplazamiento de sus habitantes se ha convertido en una cuestión con un gran interés por controlar en las grandes áreas urbanas. Aunque existen esfuerzos por solventar este problema, como nuevas rutas de transporte público, construcción y ampliación de vialidades, creación de comisiones especializadas, entre otras, persisten grandes conflictos de movilidad en horas pico y continúa extendiéndose a horas donde no se percibía este fenómeno.

Los métodos de transporte no solventan la demanda de los usuarios, los largos traslados y la concentración de vehículos en calles y avenidas incrementa el número de accidentes, además generan otros problemas derivados.

Existen desarrollos de software que buscan minimizar el problema de movilidad urbana, aplicaciones relacionadas a la detección de accidentes en tiempo real como [*Waze*, 2015], aunque no realizan un análisis geoespacial es de gran ayuda para conocer la situación vial, también existen otro tipo de aplicaciones como [*Google maps*, 2015] que realiza ruteo, es

decir, muestra el mejor camino para llegar de un punto a otro, contemplando la situación vial en tiempo real, además también realiza análisis espacio temporal. *Google maps* utiliza el GPS de los dispositivos móviles, los cuales son precisos pero con desventajas por su naturaleza, por ejemplo una calle bloqueada se muestra como una calle sin tráfico, un usuario que decide detenerse por una razón externa es tomado en cuenta para promediar el avance, su información es restringida y para uso comercial tiene costo.

Por otra parte, el implementar enfoques como Ciudades Inteligentes (*Smart Cities*) cobra importancia en el desarrollo de las grandes ciudades. Por tanto, en este trabajo se pretende ser una fuente de información que lleve como resultado a otros análisis que contribuyan al monitoreo de las ciudades.

1.1. Planteamiento del problema

En los últimos años se ha desarrollado la llamada Web 2.0, en donde el usuario pasa de ser un simple consumidor de datos a ser un productor, de esto desprende lo que se denominan redes sociales como Facebook [Facebook, 2015] y Twitter [Twitter, 2015]. De ahí que se ha incrementado el interés de las personas de publicar diversos acontecimientos, los cuales pueden ser geocodificados y aprovechados para diversos fines [Aggarwal C., 2011] como por ejemplo el análisis del tránsito vehicular.

Existen diferentes proyectos enfocados al ámbito geográfico que han sido desarrollados en su mayoría por información proveniente de los usuarios, por ejemplo, OpenStreetMap [OpenStreetMap, 2015], en donde se otorga a los usuarios la capacidad de generar y editar mapas a través del GPS de los dispositivos móviles o a través de su sitio web. Wikimapia [Wikimapia, 2015] es un proyecto que permite asignar información a localidades dentro de los mapas de Google o Flickr [Flickr, 2015], la cual permite a sus usuarios enlazar fotos a lugares del mundo.

Hoy en día, la facilidad que brindan las redes sociales a los usuarios para compartir información los alienta a registrar gran cantidad de eventos diariamente [Lee R. et al., 2013]. Información proveniente de las redes sociales es utilizada como fuente en diversos proyectos, por ejemplo para la detección de incendios, detección de siniestros, aproximaciones de tiempo de traslado de un punto a otro, ayuda en desastres naturales, determinar la propagación de epidemias, caracterización de áreas urbanas, entre otros.

Por otra parte, el desarrollo y aplicación de métodos de Aprendizaje Automático también conocidos como *Machine Learning* han ayudado a la detección de correos electrónicos no deseados (SPAM), detección de rostros, reconocimiento de dígitos, minería de datos, análisis del mercado de valores, diagnósticos médicos, publicidad dirigida, etc. En los últimos años, los métodos de Aprendizaje Automático han destacado los resultados de las Máquinas de Soporte Vectorial (SVM por sus siglas en inglés), métodos utilizados para clasificación, agrupamiento y predicción de valores. El reconocimiento reciente de SVM a pesar de ser un método creado en la década de los 90's es debido a los grandes volúmenes de información agrupada y estandarizada con los que se trabaja en la actualidad.

Con la finalidad de generar predicciones de eventos viales en la Ciudad de México, se genera un modelo para realizar análisis espacio temporal, utilizando Maquinas de Soporte Vectorial que aprenden de información generada por los usuarios en la red social Twitter.

1.2. Justificación

Con la generación de un nuevo enfoque para obtener información vial y funcionar como otra herramienta de apoyo para los problemas de movilidad urbana, este proyecto se enfoca en la geocodificación de condiciones y accidentes viales para realizar análisis espacio temporal.

Se pretende ayudar a la toma de decisiones relacionadas con la movilidad urbana y que la población tenga un panorama de la concentración de las vialidades durante el transcurso del día, esta información es generada totalmente por colaboración de los usuarios de Twitter.

Los resultados del análisis espacio temporal son visualizados a través de un sitio web donde los usuarios pueden observar las zonas más afectadas dentro de la ciudad, con base en sus propios criterios de búsqueda. Información estadística es desplegada para mostrar comportamientos usuales basados en los eventos identificados por el sistema.

Asimismo, se definió como caso de estudio la Ciudad de México, ya que se cuenta con gran cantidad de perfiles en las redes sociales que contribuyen a publicar la condición de tráfico vehicular dentro de la ciudad, igualmente existe información geográfica abierta al público de

todas las vialidades que componen esta Ciudad. Este trabajo tiene implicaciones para diversas áreas:

Investigación. Es investigación útil y vigente. La UCGIS.org y el proyecto Varenious, involucraron el espíritu colaborativo en su agenda. Se han presentado este tipo de trabajos eventos de impacto en computación y GIS, como *GIScience* (www.giscience.org), en su edición más reciente se mostraron trabajos relacionados con VGI y áreas afines, representando aproximadamente el 25.6 %. En 2012 se hizo el Primer Taller Internacional en Información Geográfica Colaborativa y Voluntaria (GEOCROWD) dentro de la *ACM SIGSPATIAL GIS Conference*, agrupándose en Minería de datos de contenido generado por el usuario, datos cualitativos y semántica y herramientas, modelos y privacidad [Goodchild et al. 2013].

Sociedad. Es adecuado para este país contar con investigación en este sentido, ya que se pueden desarrollar herramientas para tener información oportuna, aprovechando a los ciudadanos; generando ventajas con respecto a enfoques tradicionales especialmente en cuanto a la rapidez, costo y el volumen de reportes que se pueden obtener.

Industria. Tiene un gran impacto, tal como lo cita el informe *Forrester Research*, en donde se estimó que empresas como McDonalds y General Motors gastaron $ 4.6 mil millones de dólares en 2013 [Singer 2008] para tecnología de Web 2.0. Por lo que representa que es de gran importancia abordar estos temas, para incorporar esta clase de trabajos, que resultan atractivos para la industria, debido a las tecnologías que se emplean.

1.3. Hipótesis

Las hipótesis de investigación relacionadas a este trabajo de tesis se enumeran a continuación:

1. La información obtenida es producida por cuentas especializadas en tráfico vehicular en rede sociales.
2. El incremento de accidentes y eventos viales identificados favorece la precisión del análisis espacio temporal.
3. Los pronósticos realizados muestran accidentes viales visualizados a través de un mapa de calor.
4. El tiempo para la generación de la función que modela situaciones viales aumenta con relación al número de accidentes y condiciones viales registradas.

5. La selección de las características correctas que conforman a un conjunto de entrenamiento ayuda a generar una función con mayor precisión.
6. Realizar una conexión de la infraestructura vial de la ciudad, la infraestructura social y la infraestructura de tecnología de la información.

1.4. Objetivos

1.4.1. Objetivo general

Desarrollar un modelo de predicción de eventos viales por medio de técnicas de aprendizaje automático, utilizando datos geocodificados de textos cortos de redes sociales.

1.4.2. Objetivos particulares

- Analizar el estado del arte.
- Generar una metodología para la geocodificación de tweets.
- Procesar tweets geocodificados para conformar un conjunto de entrenamiento.
- Utilizar un método de aprendizaje automático para realizar predicciones viales.
- Desarrollar un método de evaluación que permita medir la precisión de los resultados obtenidos.
- Contribuir al enfoque conocido como Ciudades Inteligentes.

1.5. Organización de la tesis

El resto del documento de tesis se organiza como sigue:

Capítulo 2. Estado del Arte, el cual describe los proyectos que han sido desarrollados con el enfoque de *Crowdsourcing* e información Geográfica Voluntaria y algunos proyectos que utilizan SVM que han dado buenos resultados para la clasificación y regresión de diferentes temas.

Capítulo 3. Marco Teórico, expone los términos necesarios para la compresión de los conceptos utilizados en la elaboración de este proyecto.

Capítulo 4. Metodología de la Investigación, describe las etapas necesarias para realizar el proceso de geocodificación de tweets, así como las etapas para generar la función que modela situaciones viales necesarias para realizar el análisis espacio temporal.

Capítulo 5. Pruebas y Resultados, se describe el método de evaluación y los resultados que fueron obtenidos analizando ciertas delegaciones que componen a la Ciudad de México.

Capítulo 6. Conclusiones y trabajo futuro, presenta las conclusiones basadas en los resultados obtenidos, así como elementos que pueden ser añadidos con el fin de mejorar y complementar este trabajo de investigación.

2. Estado del arte

En este capítulo se muestran metodologías y trabajos de investigación que realizan diversos procesos para obtener información geográfica a partir de información publicada en redes sociales, ya sea información generada sin la intención de hacer un aporte geográfico (*Crowdsourcing*), así como información que inicialmente se esperaba que fuera georeferenciada (Información Geográfica Voluntaria). Se describe como se realiza la recopilación, análisis, procesamiento y presentación de la información, así como los resultados obtenidos.

Por otra parte, se mencionan trabajos desarrollados bajo métodos de aprendizaje automático para clasificación, agrupamiento o regresión de información, con que datos son entrenados y que conclusiones se determinaron a partir de los resultados.

2.1. Uso de la Información Geográfica

Existen metodologías para obtener información georeferenciada de las redes sociales y aplicadas a fines distintos, por ejemplo para localizar usuarios, identificar eventos sobresalientes, observar el comportamiento de una población, determinar condiciones de tráfico vehicular, categorización de la población, confirmación de eventos, entre otros. A continuación se mencionan metodologías que aportaron fundamentos a este trabajo de tesis.

2.1.1. Información de redes sociales georeferenciada para localización de usuarios

Para inferir la localización de mensajes en Twitter [Davis Jr C. et al, 2011] se presenta una metodología basada en las relaciones entre usuarios, la localización del tweet es obtenida a partir de la información geográfica recolectada de los amigos del usuario. La lista de amigos del usuario se genera a través de la intersección entre los seguidores del usuario y los usuarios a los que sigue. Cuando se ha obtenido la lista de amigos, se busca el sitio más popular entre ellos analizando la localización de los sitios encontrados por GPS,

localización por IP o localización declarada textualmente. El resultado se asigna como la localización del usuario.

Otra metodología para identificar la localización geográfica de los usuarios [Backstrom L. et al, 2010] consiste en utilizar las direcciones proporcionadas por los usuarios de Facebook (alrededor del 6%). El estudio verificó que en promedio cada usuario tiene 10 amigos con localizaciones compartidas, por tanto, son generadas conexiones entre usuarios y localidades. Con esta información utilizando un enfoque de probabilidad máxima, el método puede encontrar al 69.1% de una muestra si los usuarios cuentan al menos con 16 amigos localizados dentro de los primeros 40 kilómetros. Este método es más eficiente que utilizar métodos basados en IP, los cuales encuentran un 57.2% de usuarios de la muestra bajo los mismos parámetros. Otro resultado de esta metodología fue que en distancias de medio y largo alcance, la probabilidad de amistad es inversamente proporcional a la distancia entre los usuarios. En distancias de corto alcance la probabilidad de amistad es constante al incremento de la distancia. Otro comportamiento observado fue que la probabilidad de amistad está relacionada con la densidad de población, a menor densidad de población la probabilidad de relación de amistad es mayor en los primeros 80 metros. Cuando la distancia es mayor a 80 metros se observó que el comportamiento se invierte, es decir, a mayor densidad de población la probabilidad de amistad aumenta.

2.1.2. Información de redes sociales georeferenciada para categorización de áreas urbanas

Para categorizar las áreas urbanas a través de su comportamiento en las redes sociales, [Lee R. et al, 2013] una metodología propuesta consiste primeramente en seccionar áreas geográficas basados en su actividad en Twitter, a continuación se obtienen los datos más relevantes de la colección total preservando la tendencia original, consecuentemente, usuarios son agrupados dentro de cada sección, en cada grupo se censa y analiza la cantidad de usuarios, la cantidad de *tweets* generados y los traslados realizados durante el día dividido en intervalos iguales de tiempo. Finalmente se comparan los resultados con los tipos de construcciones que existen en las zonas. Se verificó que las secciones con más construcciones recreativas como parques, aguas termales y miradores tiene más tweets relacionados al entretenimiento y las áreas que tienen construcciones como universidades, oficinas y residencias tienen más tweets relacionados a actividades de trabajo y estudios,

además se observó que las ciudades con gran densidad de población tienen más actividad en Twitter que las ciudades que tienen una baja densidad de población.

Otro trabajo de investigación desarrollado con relación a la distribución de los usuarios en Twitter [González R. et al, 2011] analizó la relación que existe entre usuarios y seguidores, así como elementos culturales que existen entre ellos. Los resultados obtenidos fueron que las características lingüísticas y culturales se encuentran vinculadas con el nivel de participación. Los usuarios de un país donde se habla el idioma inglés o es un segundo idioma muestran en sus relaciones usuario seguidor, un efecto de distribución externa a su país. Por otra parte, donde el idioma es diferente al inglés como el portugués el efecto de distribución en sus relaciones es interno a su territorio. Este efecto se debe a que la mayor parte de la comunidad de Twitter es de Estados Unidos y los países de habla inglesa tienen características culturales similares a éste, por otra parte países como Brasil no manifiestan este efecto con Portugal ya que a pesar de tener la misma lengua sus características culturales son diferentes.

2.1.3. Información de redes sociales georeferenciada para detección y confirmación de eventos

Una metodología para la detección de eventos en Twitter es a través de la explosividad de palabras identificadas [Albakour M. et al, 2013]. Se detecta un incremento repentino en el número de tweets dando por hecho que ha ocurrido un evento. Primeramente se obtiene un conjunto de ubicaciones que pertenecen a una sección geográfica que pueden ser de interés para los usuarios, el tamaño de cada sección puede ser obtenida por códigos postales, divisiones realizadas por autoridades locales, división de áreas simétricas, etc. después la herramienta de recuperación de eventos crea y califica tuplas, cada tupla relaciona una localización en un tiempo determinado. Para obtener la localización de cada tupla se analiza el tweet del usuario para verificar si insertó explícitamente una dirección o implícitamente puede tomarse la localización del usuario. La calificación de las tuplas se realiza verificando entre la localización obtenida y los eventos que pueden acontecer sobre esta localización. Un evento es relevante si está relacionado explícitamente con el tweet o implícitamente en el contexto del usuario. En otras palabras, puede decirse que la herramienta de recuperación de eventos define una función de clasificación, proporcionando una calificación a cada tupla, con base en el tweet realizado por el usuario.

Otra metodología para la detección de eventos en Twitter es utilizada en una herramienta de monitoreo en línea llamada Twitter Monitor [Mathioudakis M. et al, 2010]. Este proyecto detecta eventos para profesionales de la mercadotecnia en línea y compañías de rastreo de opinión. El primer paso es encontrar palabras que repentinamente tienen un gran número de apariciones, después se agrupan las palabras con un gran número de ocurrencias llamándole *tendencia*, finalmente se verifica que la tendencia se encuentre dentro de diferentes tweets. Además para cada tendencia se obtiene información adicional de otros *tweets* que están relacionados para añadir más detalles del evento identificado.

Una metodología para la confirmación de eventos a través de la Web [Becker H. et al, 2012], consiste en generar múltiples consultas a diferentes sitios en Internet para recuperar información de usuarios asociados al evento del cual se tiene interés, con el fin de obtener información específica. Esta metodología parte de un conocimiento previo del evento, se debe proporcionar el título, lugar, una descripción, hora y la fecha para generar las consultas. La información obtenida por las consultas sirve así mismo para generar nuevas consultas y verificar en otros sitios. Generalmente el primer paso es combinar características del evento conocido dentro de las consultas y obtener datos precisos; con esta información y el uso de técnicas de procesamiento de lenguaje natural se crean y mejoran nuevas consultas para obtener información más detallada.

2.1.4. Información de redes sociales georeferenciada para identificar condiciones del tráfico vehicular

El Observatorio de Tráfico [Ribeiro Jr. T. et al, 2012] es un sistema desarrollado en Belo Horizonte, Brasil. Esta metodología obtiene condiciones de tráfico vehicular. Primeramente, se decide seguir cuentas de Twitter que tienen como finalidad exclusivamente describir situaciones viales. Se realiza un procesamiento de estandarización de tweets eliminando acentos, hashtags y menciones a otras cuentas, así como la identificación de palabras comúnmente relacionadas a condiciones viales y accidentes. A continuación, se buscan calles de forma exacta y aproximada utilizando un diccionario geográfico compuesto por nombres de calles seccionadas por cruces y vecindarios acotados por calles.

2.2. Precisión de la Información georeferenciada

Como se mencionó anteriormente se han desarrollado diferentes metodologías para georeferenciar información proveniente de redes sociales y su aporte al desarrollo de Sistemas de Información Geográfica. Cuestionando la precisión de la información generada por usuarios, se mencionan elementos que garantizan la precisión de este tipo de información y cómo descartar información que no es verdadera.

2.2.1. Enfoques que aseguran la precisión de información generada por usuarios

Existen diversos estudios que han evaluado la precisión de la información geográfica generada por usuarios en Internet y la información geográfica generada por otras fuentes como *Google maps*, *Bing maps* o datos autorizados por agencias geográficas en diferentes países.

OpenStreetMap, es un proyecto conocido ampliamente, funciona con Información Geográfica Voluntaria, aunque muestra un desplazamiento aproximado de seis metros, tiene una precisión superior que datos autorizados en diferentes partes del mundo. Goodchild M. et al [2012] asegura la precisión de la información geográfica generada por usuarios por medio de tres enfoques distintos que se mencionan a continuación:

Enfoque 1. Este enfoque está dado a través de la cantidad de personas que pueden ser partícipes de la generación de esta información conocido como *Crowdsourcing*, las personas pueden generar información geográfica, pero a la vez puede dedicar tiempo a validar que la información que ha puesto otra persona sea correcta. Este efecto también es visible en las redes sociales, cuando un evento es expuesto por un usuario generalmente viene acompañado de confirmaciones realizadas por otros usuarios sin existir una relación entre éstos. Entre mayor cantidad de usuarios que indican un evento de forma aislada en las redes sociales tiene mayor peso la credibilidad de que el evento que es publicado sea real.

Enfoque 2. Este enfoque se realiza a través de la creación de una jerarquía de usuarios y asignación de roles dentro del grupo de usuarios que generan la información geográfica. Estudios avalan que la generación de información por parte de los usuarios sigue una distribución respecto a la información generada y la cantidad de usuarios, es decir, un

grupo pequeño de personas generan gran cantidad de información y el resto de usuarios genera información moderada. A este grupo pequeño de personas, se les asigna un rol superior que a los demás y se les asignan responsabilidades adicionales como la validación de información, capacidad de edición, capacidad de eliminación de datos, etc.

El tercer enfoque está basado en la proximidad geográfica, básicamente es la primera Ley de la Geografía *"Todas las cosas están relacionadas entre sí, pero las cosas más próximas en el espacio tienen una relación mayor que las distantes"* [Tobler W, 1970]. El enfoque es utilizado para asignar un valor probabilístico si el evento expuesto puede ser falso o verdadero. Un ejemplo es el conocimiento de altos niveles de tráfico vehicular en una avenida determinada, es altamente probable que sea cierto algún registro de un accidente vehicular y aún más, si ya han existido registros de accidentes vehiculares anteriores en esta avenida.

Estos enfoques garantizan que la Información generada por usuarios sea congruente y con incrementos graduales en la precisión debido a las revisiones continuas por parte de la comunidad, la asignación de perfiles y la asignación de responsabilidades a los usuarios que administran y validan la información.

2.2.2. Estudios realizados que aseguran la precisión de la información generada por usuarios

Un estudio realizado para validar información sobre la red social Twitter [Mendoza M., et al., 2010] muestra que la información que es verdadera se comporta de una manera distinta a la información que es falsa. El estudio identificó siete eventos que eran verdaderos y siete eventos que eran falsos y se realizó un seguimiento del comportamiento de esta información en la red social. Los resultados muestran que la información que es verdadera se confirma en un 95.5% por la comunidad de la red social y se rechaza por solo un 0.3%. En cambio la información que es falsa se rechaza en un 50% además de ser cuestionada por un porcentaje mayor al de confirmación del evento. El estudio afirma que la comunidad de Twitter trabaja como un filtro colaborativo de la información.

Otro estudio desarrollado verificó que al utilizar Twitter durante una emergencia no se mencionan a otros usuarios en el texto del mensaje [Hughes A., et al., 2009]. El porcentaje varía entre 6% y 8%, contra un 22% en una situación normal; de manera inversa los tweets que incluyen direcciones de sitios en Internet son mayores en una emergencia que van

desde un 40% a un 50% contra un 25% en una situación cotidiana. Es posible inferir, que las emergencias siguen también un patrón que valida en gran medida la veracidad de la información.

2.3. Aplicaciones de los métodos de Aprendizaje Automático

Los métodos de aprendizaje automático dan solución a problemas diversos, inclusive existen soluciones que no se perciben y que están presentes, por ejemplo, la detección de correo electrónico no deseado, el objetivo es clasificar los correos de cada usuario en dos clases, correo deseado (HAM) o correo no deseado (SPAM), el problema radica en que las personas tienen gustos diferentes, por tanto, la clasificación debe ser personalizada por usuario. Este tipo de clasificación se llama clasificación binaria, es una tarea claramente identificada entre los métodos de aprendizaje automático, la solución es observar un conjunto de correos deseados y no deseados por el usuario, aprender de este conjunto de datos analizando las características de cada clase y a partir de estas características clasificar de forma automática los subsecuentes. Existe también la clasificación múltiple donde se clasifican elementos en más de dos clases, por ejemplo, el reconocimiento de dígitos, el problema es clasificar imágenes con dígitos escritos a mano del 0 al 9, por tanto se deben tener 10 clases para llevar a cabo esta tarea. A continuación se presentan algunos trabajos de investigación realizados con métodos de aprendizaje automático.

2.3.1. Aprendizaje automático, clasificación y agrupamiento

Una aplicación desarrollada con métodos de aprendizaje automático es el reconocimiento de expresiones basado en el Sistema de Codificación de Gestos Faciales (FACS) [Ekman P, 1978], en este trabajo se realizó una comparación entre métodos de aprendizaje automático para clasificar siete clases distintas, neutral, enojo, disgusto, miedo, alegría, tristeza y sorpresa. Los métodos con mejores resultados en la clasificación son utilizando en conjunto AdaBoost [Freund Y. et al, 1995] y Support Vector Machine (SVM) [Cortes C. et al, 1995], obteniendo resultados del 93% sobre el conjunto de datos DFAT-504, cabe resaltar que no se habían conseguido resultados tan altos con otras técnicas aplicadas a esta información.

Otro desarrollo utilizando métodos de aprendizaje automático es la detección de caídas y problemas de salud en personas mayores teniendo por objetivo asegurar la independencia de este sector de personas con la mínima intervención de terceros. Los datos recopilados fueron a través de diversos sensores que obtenían las posiciones del cuerpo humano así como los ángulos de inclinación producidos por partes adyacentes cuando es realizado un movimiento. La combinación de valores registrados por los sensores se catalogaron en las siguientes clases: caída, acostado, sentado, parado/caminando, sentarse y acostarse. Los métodos de aprendizaje con los mejores resultados fueron las Máquinas de Soporte Vectorial.

Un trabajo de investigación en esta línea fue el agrupamiento geográfico a través de imágenes. El método de aprendizaje automático realizó el agrupamiento de fotos de distintas playas estimando su localización geográfica [Wang Y. et al, 2013]. El método de aprendizaje automático utilizado fue Agrupación de Máximo Margen [Xu L. et al, 2004], al cual se añadió una restricción geográfica con el fin de mejorar el agrupamiento de los elementos.

2.3.2. Aprendizaje automático y predicción

Los métodos de aprendizaje automático también son utilizados para inferir nueva información, un ejemplo es el análisis de la bolsa de mercado donde es preciso saber cuándo se deben comprar acciones y cuando se deben vender, también para pronosticar los precios del petróleo así como sus reservas, el cálculo del producto interno bruto para el próximo sexenio o el porcentaje de la población con obesidad para el siguiente año.

Un trabajo de investigación utilizando métodos de aprendizaje automático para predicciones fue desarrollado para pronosticar el tiempo de traslado de un punto a otro [Wu C. et el, 2004], se analizaron diferentes factores, la velocidad del vehículo, el clima, el tráfico, si han ocurrido accidentes, entre otros. Los datos fueron recolectados durante un periodo de 5 semanas en 3 rutas distintas con diferentes horarios. Se demostró que el método realiza predicciones acertadas, superior a predicciones basadas solamente en históricos o métodos de predicción basados en historial reciente.

Otro desarrollo basado en métodos de aprendizaje automático fue la predicción de desbordamiento de ríos [Wu C. et al, 2008], el cual aplica diversos métodos y realizando

tareas de optimización, con lo anterior se logró desarrollar un modelo de predicción que muestra escenarios probables.

2.4. Discusión del estado del arte

Se mostraron diversos trabajos de investigación relacionados con el uso de la información generada por usuarios, enfoques que aseguran la veracidad de la información y se mencionó la precisión que tiene la información geográfica generada por usuarios. Además se describieron algunos trabajos desarrollados con métodos de Aprendizaje Automático para realizar clasificación, agrupamiento y predicción de la información.

Se presentaron algunas metodologías que obtienen información de las redes sociales enfocada a diferentes objetivos. De igual forma, se mencionaron trabajos de investigación que compararon métodos de aprendizaje automático y cuales obtuvieron los mejores resultados.

La investigación de estos trabajos y otros más [Pollino M. et al, 2012] [Utani A. et al, 2011] [De Longueville B. et al, 2009] [Abel F. et al, 2012] proporcionaron un panorama para llevar a cabo un análisis espacio temporal de eventos viales, utilizando métodos de aprendizaje automático que son entrenados con información generada por los usuarios en las redes sociales.

3. Marco Teórico

A continuación se describen todos los elementos necesarios para la comprensión y elaboración del trabajo de tesis, se mencionan los conceptos de *Crowdsourcing*, Información Geográfica Voluntaria (VGI), los Métodos de Aprendizaje Automático (*Machine Learning*), las Maquinas de Soporte Vectorial (SVM) y cómo es el funcionamiento de regresión (*Support Vector Regression*).

3.1. *Crowdsourcing*

Con el desarrollo de la Web 2.0 el mundo se ha convertido en una era de relaciones, cooperación y participación, basta con un equipo de cómputo y acceso a Internet para formar parte de esta red mundial. Internet facilita a cualquier persona el compartir sus habilidades, creatividad o trabajo, de igual forma relaciona personas que buscan consejos, puntos de vista o apoyo en un tema específico. Estas actividades generan conexiones entre los usuarios dando lugar a soluciones de forma colaborativa. Empresas y usuarios han decidido utilizar esta tendencia como un nuevo modelo de trabajo [Karabulut M, 2010]. Abrir un problema a usuarios en Internet viene acompañado de ideas innovadoras, de fácil implementación y creativas para resolverlo.

Existen en la actualidad diversas páginas en Internet donde se alienta a las personas en participar en la resolución de problemas, un ejemplo, es la página Kaggle [Kaggle, 2014], la cual realiza competiciones, algunos con premios en efectivo, con el fin de recibir múltiples soluciones a problemas relacionados generalmente a métodos de aprendizaje automático. Los participantes tiene acceso a toda la información necesaria para desarrollar una solución al problema, los usuarios mandan sus propuestas y verifican si su solución es la mejor.

El término *crowdsourcing* es derivado de las palabras *crowd* y o*utsourcing* que tiene como significado trabajo hecho por las personas. El término fue utilizado por primera vez en el 2006 [Howe J, 2006] para definir una nueva forma de colaboración e innovadora red de trabajo sobre Internet a través de herramientas colaborativas de la Web 2.0.

Unos años atrás se detectó que personas sin formación profesional pero entusiasta por aprender estaban contribuyendo a dar soluciones a problemas dando lugar a una tendencia

que cambió a la sociedad en Internet. En la actualidad, estas personas conocidas como Profesionales Amateurs tienen acceso a gran cantidad de información y comparten sus conocimientos a la comunidad en Internet, son caracterizadas por ser comprometidas, educadas y conectadas por medio de la tecnología [Leadbeater C. et al, 2004].

La Figura 1 muestra cómo funciona *Crowdsourcing* cuando una empresa tiene un problema por resolver [Brabham D, 2009]. El proceso inicia con la identificación del problema por parte de la empresa, después a través de una plataforma en línea la empresa hace público su problema a la comunidad, los usuarios comienzan a trabajar en el problema y envían soluciones, los patrocinadores escogen las mejores soluciones de acuerdo con sus necesidades y se adueñan y llevan a cabo la solución ganadora y premian al participante.

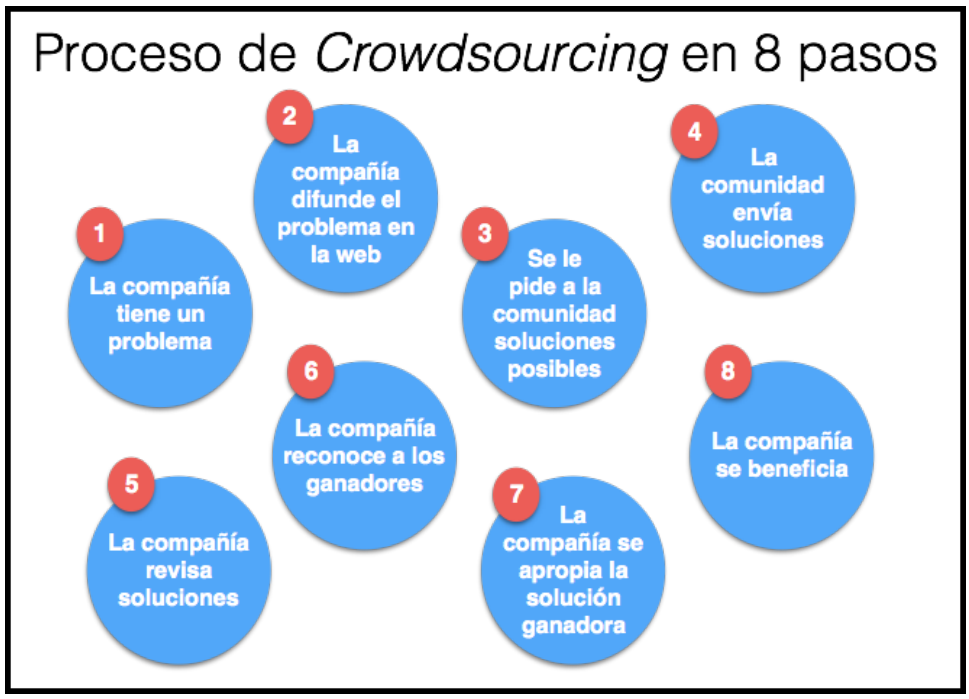

Figura 1: Proceso del *Crowdsourcing*.

3.2. Información Geográfica Voluntaria

La generación de contenido geográfico de forma voluntaria desarrollado por el usuario en la web es conocido como Información Geográfica Voluntaria "*Volunteered Geographic Information* – VGI". Desde que inició el interés en la Web para crear, reunir y difundir información geográfica realizada de forma voluntaria, diferentes sitios en Internet han optado por otorgarle poder a los usuarios para generar información que alimente sus

sistemas [Goodchild M., 2007]. Es cierto que los usuarios generalmente no tienen un entrenamiento previo para la generación de información geográfica, lo realizan de forma libre y sus resultados pueden ser imprecisos, pero cuando esta actividad se realiza de manera colectiva tiene un impacto relevante en los Sistemas de Información Geográfica, en la rama de la Geografía y con el público en general.

El desarrollo de VGI se debe gracias al progreso de diversos componentes, tales como, el desarrollo de páginas con altos niveles de interactividad, la creación de sistemas de referencia para localizar puntos en la superficie terrestre, desarrollo e instalación de canales de comunicación de banda ancha alrededor del mundo, la implementación de Sistemas de Posicionamiento Global (GPS) y su integración a diferentes dispositivos. Estos avances dan a los usuarios la capacidad de ofrecer o solicitar servicios en Internet, así como la creación de conexiones de forma directa para contribuir al mejoramiento y producción de los Sistemas de Información Geográficos y no Geográficos [Savelyev A., et al., 2011]. Diversos sitios en Internet actualmente utilizan VGI para abastecerlos de información y trabajar con ella.

Los humanos tienen la facultad de funcionar como sensores, ya que poseen la capacidad de degustar, oír, oler, sentir y observar lo que se encuentra en el entorno, igualmente interpretan lo que perciben y tienen la habilidad de visitar diferentes puntos de la superficie terrestre. Tomando en cuenta el número de personas que existen en todo el mundo se considera a VGI como un uso efectivo de una red de sensores conectados a través de la Web [Goodchild M., 2007]. Con la facilidad de acceso a sitos en Internet por medio de diferentes dispositivos se ha logrado que los usuarios registren todos los eventos de los que es espectador y/o participe. De las redes sociales se obtienen grandes volúmenes de datos generados por los usuarios, datos que sirven para generar nueva información.

A través de un análisis de la información obtenida de las redes sociales se producen interesantes perspectivas del comportamiento humano y de la interacción que existe entre las personas, por ejemplo, técnicas de minería de datos aplicados a las redes sociales se obtiene un mejor entendimiento de las opiniones de las personas acerca de un tema, identificar grupos de personas entre poblaciones enteras, analizar los cambios de la sociedad a través del tiempo, encontrar gente influyente o recomendar productos o servicios a gente en particular [Aggarwal C., 2011].

Los expertos en el área consideran una división de los perfiles de las personas que se dedican a la generación de información geográfica [Coleman D., et al., 2009]:

1. Principiante: Persona sin antecedentes formales en la materia pero cuenta con interés, tiempo y deseo de ofrecer una opinión de un tema de discusión.
2. Aficionado interesado: Persona que ha encontrado interés en un tema, ha comenzado a leer sobre él, consulta a otros interesados y ha expertos sobre tópicos específicos y gana experiencia apreciando el tema.
3. Experto interesado: Persona que tiene gran conocimiento del tema, práctica apasionadamente en ocasiones pero no confía en el tema para poder vivir de él.
4. Experto profesional: Persona que ha estudiado y practicado el tema, depende de este conocimiento para vivir y quizá tiene implicaciones legales si sus productos, conocimientos, recomendaciones u opiniones son inadecuadas, incorrectas o difamatorias.
5. Autoridad experta: Persona que es ampliamente estudiada y tiene un largo recorrido en el tema al punto de que es reconocido para tener un registro establecido de proveer productos de alta calidad y opiniones bien informadas. Puede llegar a perder su reputación y su forma de vida si pierde su credibilidad incluso solo por un periodo corto de tiempo.

VGI ha transformado como los datos geográficos, la información y el conocimiento son producidos y transmitidos [Sui D. et al, 2013]. VGI añade a los GIS todas las ventajas que la información voluntaria conlleva, haciendo partícipe a una gran cantidad de personas para generar, mantener, validar, eliminar y visualizar información geográfica.

3.3. Diferencias entre *Crowdsourcing* y VGI

Los temas vistos previamente están estrechamente relacionados aunque no sea clara la diferencia entre *Crowdsourcing* y VGI. *Crowdsourcing* es un término más amplio que abarca diferentes clases de información, no referente al contenido en sí mismo, sino a la forma en como es generado (Figura 2). Por otra parte, VGI es un tipo especial de *Crowdsourcing* que cumple con ciertas características en la forma en como es generado [James, F, 2012].

VGI es un tipo de información generada de forma voluntaria, es decir, las personas que aportan información geográfica lo hacen de una manera consiente y saben la aportación

que están realizando, por ejemplo colaboraciones para *OpenStreetMap* o el uso de aplicaciones móviles desarrolladas específicamente para referenciar puntos geográficos.

Por otra parte, *Crowdsourcing* además de incluir este tipo de información considera otras fuentes, por ejemplo, datos que son posibles procesar para obtener información requerida para un caso de estudio. Otro ejemplo, es información que ya es útil y no necesita de ningún proceso, sin embargo, esta información no fue generada con la intención de hacer una aportación a algún objetivo, por ejemplo, publicaciones en Twitter con o sin tener la localización activada en los dispositivos electrónicos, es un ejemplo de *Crowdsourcing*, los usuarios de Twitter no pretenden realizar una aportación a algún proyecto en específico, solo buscan manifestar un estado de ánimo, idea o pensamiento.

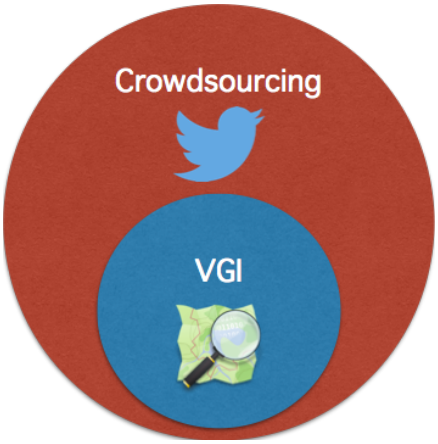

Figura 2: Esquema *Crowdsourcing* y VGI.

Como se mencionó anteriormente, aportaciones a *OpenStreetMap* son consideradas VGI, pero también es *Crowdsourcing*, esto es debido a que las aportaciones se realizan por diferentes usuarios alrededor del mundo y además se tiene conciencia de la aportación que se está haciendo. Por otra parte, la localización de un usuario cuando hace una transferencia en un cajero automático es *Crowdsourcing* ya que está produciendo información respecto a su ubicación, pero no es considerado VGI ya que no tienen la intención de contribuir a algún sistema donde su ubicación fue utilizada.

Una taxonomía formal propuesta por [Matevelli G., et al, 2015] clasifica técnicas de *Crowdsourcing* y *Crowdsensing* entendiendo *Crowdsensing* como la actividad de sensar eventos a nuestro alrededor por medio de dispositivos móviles (tabletas o teléfonos inteligentes), utilizando sensores ambientales embebidos. Esta clasificación se realiza bajo

dos dimensiones, la forma de captura de los datos y el nivel de colaboración del usuario (Figura 3).

- *Crowdsourcing* activo: El usuario facilita la información conscientemente, normalmente de forma voluntaria, sabiendo como será utilizada. Ejemplo, *OpenStreetMaps.*
- *Crowdsourcing* pasivo: La información es obtenida a partir de material publicado por el usuario para otras finalidades, es decir, los datos útiles eventualmente contenidos en el material publicado para otras finalidades son usados sin el conocimiento explícito del usuario. Ejemplo, recolección de tweets.
- *Crowdsensing* activo: El usuario proporciona activamente información que es capturada por sensores embebidos en los dispositivos móviles. Ejemplo, hacer *check-in* en un punto de interés con la posición determinada por el GPS.
- *Crowdsensing* pasivo: La información es capturada por sensores en dispositivos móviles, pero sin la necesidad de la interacción con el usuario. Ejemplo, incluyen la función de captura de trayectoria en herramientas como Waze.

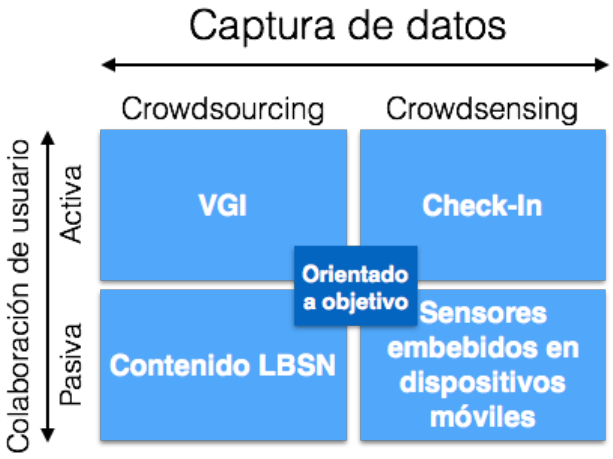

Figura 3: Taxonomía *Crowdsourcing* y *Crowdsensing*.

3.4. Ciudades inteligentes (*Smart Cities*)

Las ciudades inteligentes es un concepto que está emergiendo en los últimos años, conceptualizar este concepto sigue en progreso. Se han utilizado otras palabras para describir este concepto como ciudades digitales (*digital cities*) [Chourabi, H. et al., 2012]. Existen algunas definiciones aceptadas que se mencionan a continuación:

- Ciudad que monitorea e integra condiciones de todas sus infraestructuras críticas, incluyendo calles, puentes, túneles, vías, metro, aeropuertos, puertos, comunicaciones, agua y edificios con el fin de optimizar sus recursos, planear actividades preventivas, monitorear aspectos de seguridad mientras optimiza los servicios a sus habitantes.
- Ciudad que conecta su infraestructura física, infraestructura social, infraestructura operacional e infraestructura de tecnologías de la información para aprovechar la inteligencia colectiva de la ciudad.
- Uso de la inteligencia de tecnologías de la información para realizar los componentes de infraestructura y servicios importantes de una ciudad, la cual incluye administración, educación, salud, seguridad pública, trasporte, inmuebles y servicios públicos haciéndolos más inteligentes, interconectados y eficientes.

Las ciudades inteligentes se caracterizan por ser instrumentadas, interconectadas e inteligentes. La instrumentación permite el registro e integración de datos de la vida diaria a través de sensores, kioscos, cámaras, dispositivos personales, teléfonos inteligentes, la web y otros sistemas de adquisición de datos. La interconexión representa la integración de todos los datos recolectados en una plataforma de cómputo y la comunicación de esta información entre diversos servicios de la ciudad. Inteligencia se refiere a la inclusión de análisis complejos, modelación, optimización y visualización de procesos de negocio operacionales para realizar mejores decisiones.

3.5. Métodos de aprendizaje automático

Primeramente para entender que son los métodos de aprendizaje automático se debe definir que es aprendizaje en los seres humanos y cuál es el proceso en cómo los humanos realizan esta actividad.

3.5.1. Aprendizaje

En este documento se define el concepto de aprendizaje como un proceso donde se adquieren o modifican habilidades, destrezas, conocimientos, conductas o valores como resultado del estudio, la experiencia, la instrucción, el razonamiento y la observación [Wikipedia, 2015].

El proceso de aprendizaje más básico es la imitación, la repetición de un proceso de observación, por ejemplo, saltar. La imitación necesita tiempo para ver los detalles, el espacio adecuado, desarrollar las habilidades necesarias y otros recursos. Mediante la observación los niños aprenden a desarrollar las habilidades necesarias para sobrevivir, como comer, beber, caminar, hablar, etc.

La Taxonomía de Bloom originalmente creada como un método para clasificar objetivos educativos en la evaluación del rendimiento de los estudiantes ahora es utilizada como una taxonomía para adquirir conocimiento en el dominio cognitivo, esta taxonomía se divide en seis categorías: conocimiento, comprensión, aplicación, análisis, síntesis y evaluación (Figura 4).

1. Conocimiento se refiere a los materiales previos al aprendizaje como hechos, términos, conceptos y respuestas.
2. La comprensión es el entendimiento de los hechos e ideas por organización, comparación, traducción, interpretación y descripción.
3. Aplicación es el uso del nuevo conocimiento para resolver problemas.
4. Análisis es la exploración y división de información en partes mediante la identificación de motivos y causas.
5. Síntesis es la unión de información en una nueva forma, combinando elementos en patrones o proponiendo soluciones alternativas.
6. Evaluación es la presentación y defensa de opiniones haciendo juicios acerca de la información, validando ideas o calidad de trabajo basado en un conjunto de criterios.

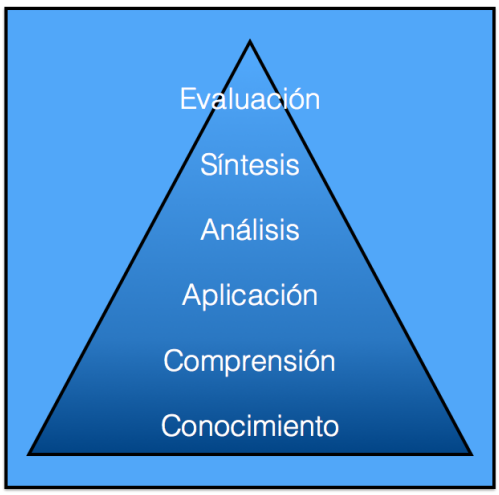

Figura 4: Taxonomía de Bloom.

3.5.2. Aprendizaje Automático (Machine Learning)

Conociendo acerca del proceso de aprendizaje se deduce que es una tarea complicada intentar reproducir o simular este proceso humano mediante una computadora. La computadora a pesar de realizar algunas tareas en un tiempo menor a las personas, existen otras tareas que una computadora no realiza con la misma velocidad, un humano identifica de manera sencilla reconocimiento de números, lo cual para una computadora conlleva una gran cantidad de operaciones para tomar una decisión.

Al igual que los humanos necesitan observar un cierto número de veces un suceso o actividad para realizarlo o distinguirlo, los métodos de aprendizaje automático, requieren de una fase de entrenamiento para asociar un significado a cada suceso o elemento para después realizar alguna conjetura. Cuando un niño ve a un animal y sabe que se trata de un perro, la pregunta es, ¿Cómo el niño sabe que es un perro en lugar de un gato o un tigre? La respuesta viene de un proceso de aprendizaje donde el niño por medio de otra fuente de información (padres, escuela, televisión, etc.) asoció a ese animal el nombre de perro, seguido de más confirmaciones de lo que es un perro. Los métodos de aprendizaje automático al igual realizan un análisis de colecciones de información aunque la procesan de forma distinta (Figura 5).

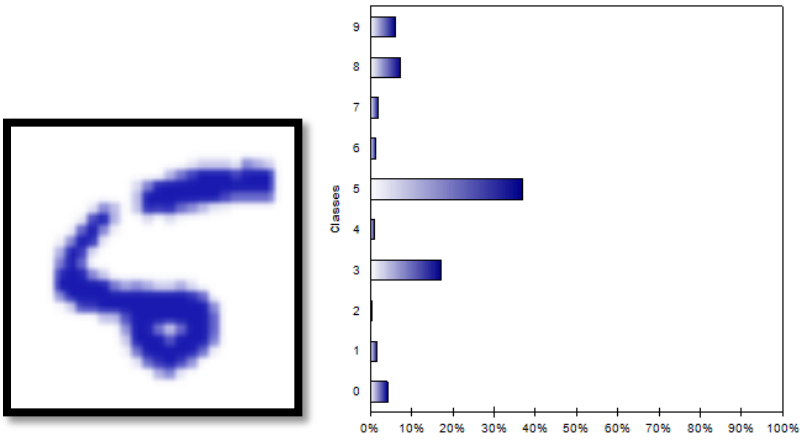

Figura 5: Como reconoce un número una computadora.

El eslogan del Aprendizaje Automático podría definirse como "*Usar las características correctas para construir los modelos correctos que logren las tareas correctas*." [Flatch P, 2012] representado por la Figura 6.

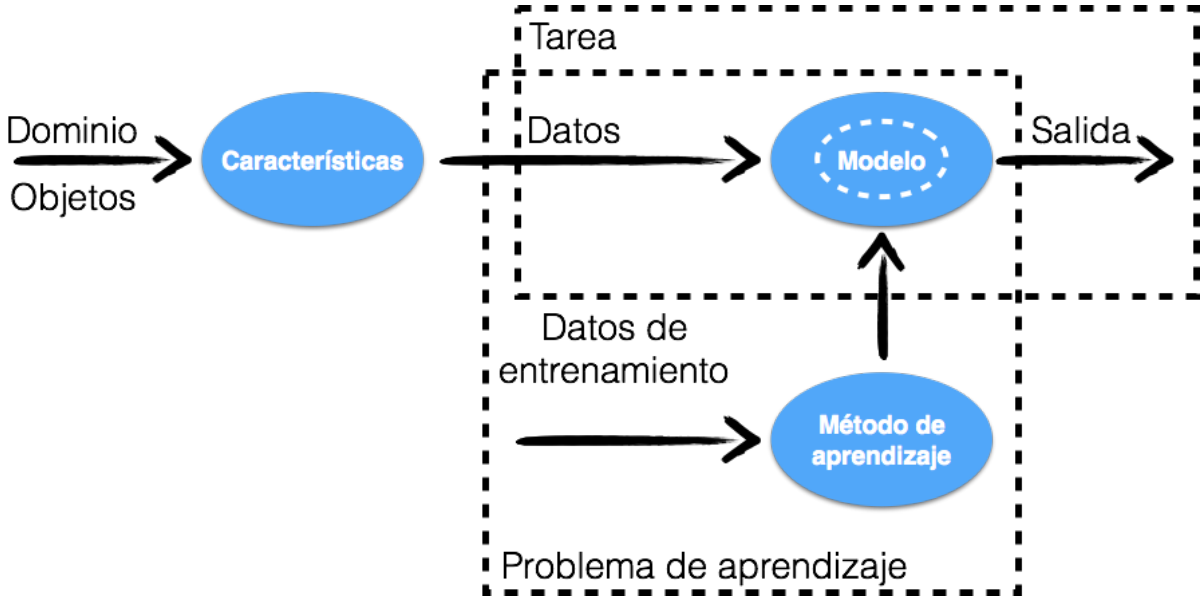

Figura 6: Como el Aprendizaje Automático aborda una tarea.

Las características definen el lenguaje en donde se describen los objetos relevantes de nuestro dominio, por ejemplo, perros, gatos, números, correos electrónicos o complejos modelos moleculares. No es necesario regresar al dominio de los objetos mismos cuando se han conseguido las características adecuadas de representación, que es donde juegan un papel importante en el aprendizaje automático. Una tarea es una representación abstracta de un problema que se quiere resolver con respecto a los objetos del dominio, la forma más común es realizar una clasificación de dos o más clases. Muchas de estas tareas son representadas mediante un mapeo de puntos de datos a las salidas. Este mapeo o modelo por sí mismo es la salida de un algoritmo de aprendizaje automático utilizando los datos de entrenamiento.

Los métodos de aprendizaje realizan trabajo de clasificación ya sea binario o multiclase. Cuando se requiere trabajar sobre un espacio no discreto y predecir números reales se le llama regresión, generalmente conlleva aprender una función de valores reales con muestras de entrenamiento que tienen valores de función verdaderos. Por otra parte, la tarea de conjuntar datos sin previa información de los grupos es llamado agrupamiento,

estos algoritmos buscan la similitud entre las instancias, poniendo instancias similares en el mismo grupo e instancias que no son similares en grupos distintos. Los algoritmos de aprendizaje automático son divididos en supervisados y no supervisados.

El aprendizaje automático supervisado consiste en crear una función a partir de los datos de entrenamiento, los datos de entrenamiento consisten en pares entrada y salida, las entradas son comúnmente vectores *n*-dimensionales y la salida puede ser un valor continuo en problemas de regresión o una clase a la que pertenece la instancia en problemas de clasificación. El problema es predecir el valor de la función para cualquier entrada válida después de haber analizado un conjunto de muestras de entrenamiento. Los métodos de aprendizaje automático supervisado con resultados notables para clasificación y regresión son las Máquinas de Soporte Vectorial, Redes Neuronales y Árboles de Decisión (Figura 7).

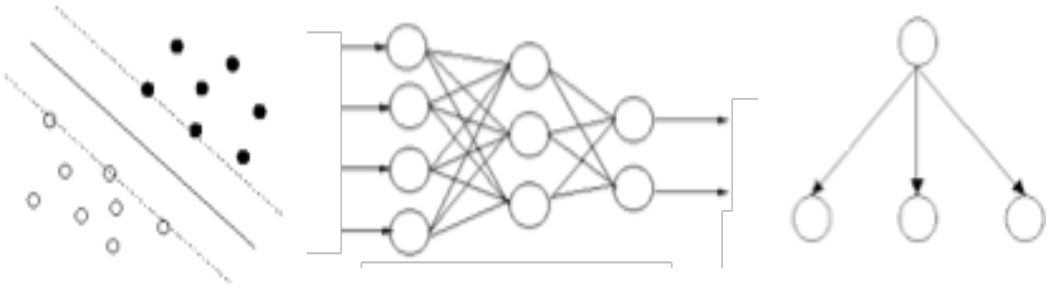

Figura 7: Métodos de Aprendizaje Automático Supervisados.

El aprendizaje automático no supervisado consiste en realizar predicciones solamente con el vector de entrada sin ningún tipo de entrenamiento previo, este tipo de problema no tiene una definición matemática bien definida, por ejemplo, en problemas de agrupamiento no existe una forma estandarizada de medir la afinidad entre los elementos, en muchas ocasiones es determinado de forma subjetiva (Figura 8).

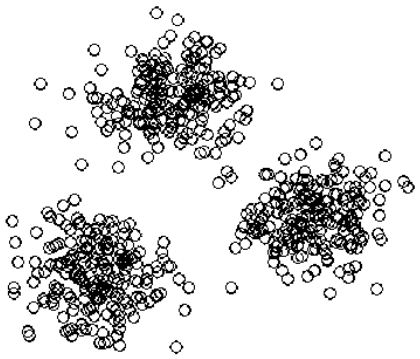

Figura 8: Métodos de Aprendizaje Automático No Supervisados.

3.6. Máquinas de Soporte Vectorial

Desde la década pasada hasta ahora las Maquinas de Soporte Vectorial o SVM por su acrónimo en inglés, han generado diversos artículos de investigación aplicados a diferentes temas como la detección de partículas, reconocimiento facial, categorización de textos, bioinformática, comercialización de bases de datos, etc. [Bennett K. et al, 2000]. De las Máquinas de Soporte Vectorial se están obteniendo resultados superiores en problemas de clasificación y en problemas de regresión no lineal, por esta razón, se seleccionó como método de aprendizaje para este trabajo de tesis. Se explica a continuación el funcionamiento de ambos:

El caso básico en problemas de clasificación es suponer que existen muestras de dos clases (clase azul - clase verde), las cuales son linealmente separables, en la Figura 9 se observa que existen diferentes soluciones al problema, pero ¿cuál sería la mejor solución de estas tres?

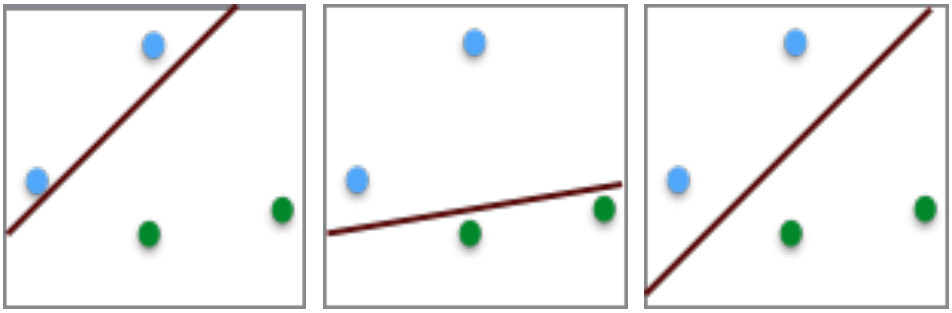

Figura 9: Soluciones para separar linealmente las muestras.

Todas las soluciones separan linealmente las muestras de las dos clases correctamente pero el margen de error obtenido es diferente para cada solución (Figura 10). La mejor solución es la línea que genera el mayor margen de error, si se opta por la imagen izquierda o la imagen en el centro una nueva muestra generada de alguna clase puede ser clasificada de forma incorrecta inclusive si se encuentra cercana a elementos de su propia clase debido a su cercanía con la línea que las separa, en cambio, si se tiene una línea que se encuentra lo más separada posible de las dos clases se reduce la posibilidad de que una muestra quede mal clasificada. Por tanto, el problema es encontrar la línea que genere el mayor margen de error que separa linealmente ambas clases.

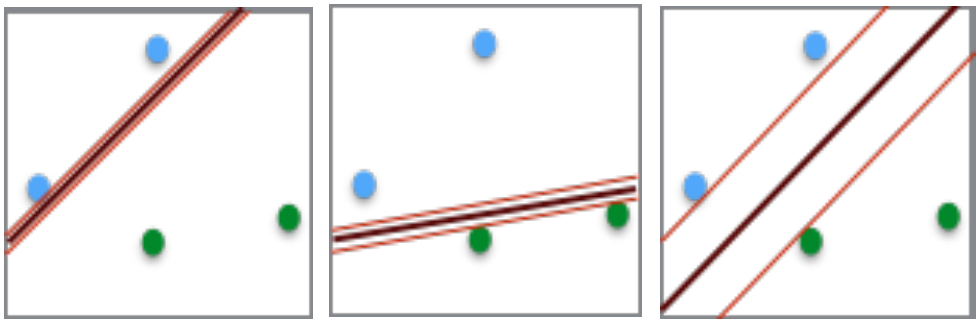

Figura 10: Margen de error de las soluciones propuestas.

El problema parece ser a primera instancia fácil de resolver pero conlleva una cantidad de operaciones considerables.

Supóngase que existe un vector $\vec{w}$ el cual es perpendicular al plano que divide las muestras positivas (color verde) y negativas (color azul) y un vector $\vec{u}$ el cual debe encontrarse si está por debajo o por encima del plano (Figura 11).

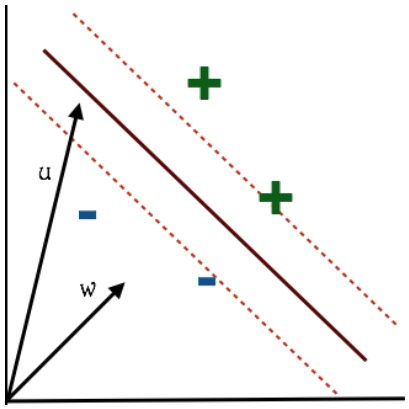

Figura 11: Funcionamiento de SVM.

Al realizar la proyección de $\vec{u}$ sobre $\vec{w}$, es decir, el producto punto de estos dos vectores, estableciendo un parámetro de decisión c, se dice que si la proyección de $\vec{u}$ sobre $\vec{w}$ es mayor o igual que c el vector $\vec{u}$ se encuentra por encima del plano y se trata de una muestra positiva, por el contrario, si es menor que c el vector $\vec{u}$ se encuentra por debajo del plano y se trata de una muestra negativa.

$$\vec{w} \cdot \vec{u} \geq c \quad (1)$$

Despejando la ecuación y reemplazando $c = -b$ la regla de decisión queda de la siguiente manera:

$$\vec{w} \cdot \vec{u} + b \geq 0 \quad \textit{entonces muestra positiva.} \quad (2)$$

Por tanto, el problema es encontrar los valores de $\vec{w}$ y b, dado un vector $\vec{u}$ de cualquier longitud y dirección, para esto deben agregarse algunas restricciones. Se va a restringir que una muestra para considerarse como positiva tiene que tener un valor mayor o igual que 1 y las muestras negativas deben tener un valor menor o igual a -1, entonces las ecuaciones quedan de la siguiente manera:

$$\vec{w} \cdot \vec{x}_+ + b \geq 1 \quad \textit{muestras positivas} \quad (3)$$

$$\vec{w} \cdot \vec{x}_- + b \leq -1 \quad \textit{muestras negativas} \quad (4)$$

Se va a introducir una variable y con el fin de hacer los cálculos más simples, esta variable y_i tendrá el valor de 1 cuando se trate de muestras positivas y -1 para muestras negativas.

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 \quad \textit{muestras positivas} \quad (5)$$

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 \quad \textit{muestras negativas} \quad (6)$$

Se observa que las ecuaciones resultantes son idénticas, es una de las razones por las cuales se introduce la variable y_i. El siguiente paso es despejar una de las dos ecuaciones anteriores a cero para obtener los valores de las muestras cuando se encuentra exactamente sobre las orillas de los márgenes de error.

$$y_i(\vec{w} \cdot \vec{x}_i + b) - 1 = 0 \quad (7)$$

Conociendo los valores en las orillas del margen de error se obtiene la ecuación del ancho del margen mediante la resta de dos muestras que se encuentren en las orillas y proyectarlas sobre el vector $\vec{w}$ que es perpendicular al plano (Figura 12).

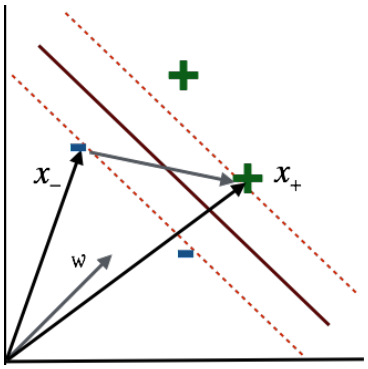

Figura 12: Cálculo del ancho del margen.

La ecuación para calcular el ancho del margen de error queda de la siguiente forma:

$$(x_+ - x_-) \cdot \frac{\vec{w}}{||\vec{w}||} \quad (8)$$

Reemplazando (7) en (8) se tiene como resultado la ecuación del ancho del margen de error.

$$\frac{2}{||\vec{w}||} \quad (9)$$

Por tanto es la ecuación que se desea maximizar,

$$\max \frac{2}{||\vec{w}||} \quad (10)$$

Si se pretende maximizar la ecuación (10) puede eliminarse el dos de la parte superior de la división por ser irrelevante o es posible decir que sería igual a minimizar solamente la magnitud del vector $\vec{w}$, el cual se divide entre 2 y se eleva al cuadrado para hacer los cálculos siguientes de forma más fácil.

$$\max \frac{2}{||\vec{w}||} = \max \frac{1}{||\vec{w}||} = \min ||\vec{w}|| \rightarrow \min \frac{1}{2} ||\vec{w}||^2 \quad (11)$$

Para resolver este problema se recurre a los Multiplicadores de LaGrange, este procedimiento sirve para problemas de optimización encontrando mínimos y máximos de funciones de múltiples variables sujetos a restricciones, que es el problema que se tiene, se busca la minimización de $\vec{w}$ y *b* y la maximización de α [Abu-Mostafa Y, 2012]. Los Multiplicadores de LaGrange transforman un problema de *n* variables y *k* restricciones a un problema de *n* + *k* variables que se puede resolver de forma más fácil. La ecuación a resolver para nuestro problema queda de la siguiente manera:

$$L(\vec{w},b,\alpha)=\frac{1}{2}\left|\left|\vec{w}\right|\right|^2-\sum\alpha_i\left[y_i(\vec{w}\cdot\vec{x}_i+b)-1\right] \quad (12)$$

El procedimiento de los Multiplicadores de LaGrange toma la ecuación del ancho del margen de error restando las restricciones, en este caso solamente es la restricción relacionada a las muestras en las orillas del margen, que es multiplicada por un Multiplicador de LaGrange. El paso siguiente es realizar las derivadas parciales con respecto a $\vec{w}$ y *b*, obteniendo los siguientes resultados:

$$\frac{\delta L}{\delta\vec{w}}=\vec{w}-\sum\alpha_i y_i\vec{x}_i=0 \longrightarrow \vec{w}=\sum\alpha_i y_i\vec{x}_i \quad (13)$$

$$\frac{\delta L}{\delta b}=-\sum\alpha_i y_i=0 \longrightarrow \sum\alpha_i y_i=0 \quad (14)$$

Sustituyendo (13) y (14) en (11) y simplificando se tiene la ecuación final:

$$L=\sum\alpha_i-\frac{1}{2}\sum_i\sum_j\alpha_i\alpha_j y_i y_j(x_i\cdot x_j) \quad (15)$$

La cual debe estar sujeta a las restricciones:

$$\sum_1^i y_i\alpha_i=0 \quad (16)$$

$$\alpha_i\geq 0\rightarrow n=1,2,3,...,n \quad (17)$$

De igual forma para la regla de decisión se sustituye (13) en (2).

$$\sum\alpha_i y_i(\vec{x}_i\cdot\vec{u})+b\geq 0 \quad \textit{entonces muestra positiva.} \quad (18)$$

Para obtener el valor de b se utiliza la siguiente fórmula:

$$b = \frac{1}{N_S} \sum_{i \in S} \left(y_i - \sum_{j \in S} \alpha_j y_j k(x_i, x_j) \right) \quad (19)$$

Donde N_S es el número de Multiplicadores de LaGrange que son mayores que 0, a los cuales se les llama Vectores de Soporte y $k(x_i, x_j)$ es el producto punto entre los vectores de entrada en cualquier espacio donde es posible esta operación. Este concepto es conocido como “kernel trick”, que es utilizado cuando los puntos no son linealmente separables.

El problema de optimización depende solamente del producto punto de los vectores $\vec{x}_i$ y $\vec{x}_j$ en (15) lo cual es de gran ayuda para los casos en que las muestras no son linealmente separables como el que se tiene en la Figura 13.

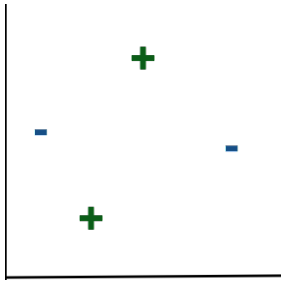

Figura 13: Muestras linealmente no separables.

Kernel trick es una de las características por las cuales SVM ha tenido gran éxito. Kernel trick realiza una transformación del espacio en donde las muestras son separables (Figura 14).

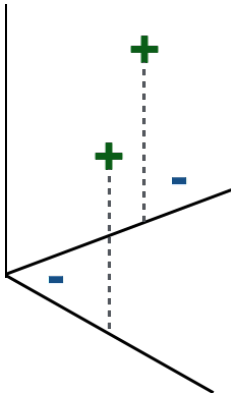

Figura 14: Muestras en un espacio diferente donde si son separables.

Como se mencionó anteriormente la ecuación (15) el problema de optimización depende solamente del producto punto entre las muestras, por tanto, puede ser reemplazado por otra función que busque separar las muestras en un espacio diferente.

Mejor aún es que no se necesita saber la transformación en el otro espacio de $\phi(x_i)\cdot\phi(x_j)$ solamente se necesita la función *kernel* que realiza el producto punto en otro espacio $k(x_i,x_j)$ [Winston P, 2014].

$$L=\sum\alpha_i-\frac{1}{2}\sum_i\sum_j\alpha_i\alpha_j y_i y_j k(x_i,x_j) \quad (20)$$

Sujeto a (16) y (17).

Algunos de los *kernel* más utilizados son:

- *Kernel* lineal.
- *Kernel* gaussiano.
- *Kernel* polinomial.
- *Kernel* sigmoide.
- *Kernel* de función de base radial (RBF).

3.7. Máquinas de Soporte Vectorial para Regresión

Las Máquinas de Soporte Vectorial para Regresión (SVR) por su acrónimo en inglés, al igual que la clasificación, han tenido gran éxito, se ha demostrado minimizar el error cuadrático medio de mejor forma en comparación con otros métodos de predicción [Wu C. et al, 2004]. SVR ha sido utilizado satisfactoriamente en predicción del mercado financiero, estimación del consumo de energía, reconstrucción de sistemas caóticos y predicción del tráfico vehicular.

SVR contiene todas la ventajas de SVM como el uso de *kernel*, trabajar con ruido en datos y trabajar en un espacio convexo, por tanto se tiene la certeza de buscar mínimos y máximos globales, situación que no pasa con las redes neuronales que en ocasiones quedan establecidas en mínimos y máximos locales.

SVR considera de igual forma que en el SVM los datos de entrenamiento $\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$ donde cada x_i puede ser un vector de *n* dimensiones y y_i es un valor que pertenece al conjunto de los reales. La idea de SVR es determinar una función

basada en los valores de entrenamiento que pueda aproximar valores futuros de forma precisa. La forma general de la función es la misma que SVM para clasificación.

$$f(x) = (w \cdot \phi(x)) + b \quad (1)$$

Donde $\phi(x)$ es una muestra que puede estar en un espacio no linealmente separable y w y b están relacionados al plano que divide las muestras. El objetivo es encontrar de igual forma que en SVM los valores de w y b tal que los valores de x puedan ser determinados minimizando del riesgo de regresión.

$$R_{reg}(f) = C\sum_{i=0}^{l} \Gamma(f(x_i) - y_i) + \frac{1}{2}\|w\|^2 \quad (2)$$

Se observa que es la ecuación de minimización del margen más la restricción para reducir el riesgo de regresión y esto multiplicado por una constante de penalización C.

SVR trabaja con una variable de insensibilidad con el fin de tener soluciones dispersas, reemplaza la función de error cuadrático por una función de insensibilidad. El valor es cero si la diferencia absoluta entre la predicción $f(x)$ y la salida y es menor que ε donde $\varepsilon > 0$. Todos los puntos menores a ε no son considerados para la generación de la función de regresión (Figura 15).

$$\Gamma(f(x) - y) = \begin{cases} |f(x) - y| - \varepsilon, & |f(x) - y| \geq \varepsilon \\ 0, & cualquier\ otro \end{cases}$$

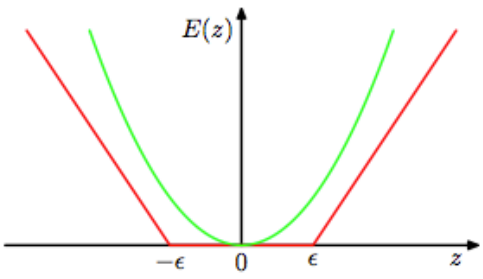

Figura 15: Gráfica de la función de insensibilidad ε.

Puede visualizarse ε como un tubo que rodea a nuestra función, donde los valores dentro del tubo no son considerados para la generación de la función de regresión (Figura 16).

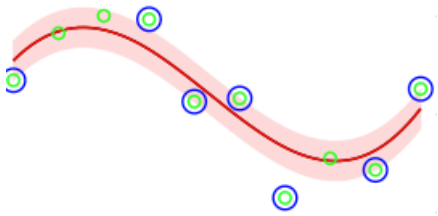

Figura 16: Los valores dentro del tubo ε no son considerados.

El valor de ε es introducido por el usuario cuando realiza el proceso de regresión, pero hay que tener ciertas consideraciones cuando éste es definido. Valores grandes de ε producen una función que no realiza de forma correcta la predicción de puntos y una ε demasiado pequeña genera sobre entrenamiento y da como resultado un sistema bastante complejo de generar, por tanto, es preciso experimentar con este valor cuando el modelo de predicción de SVR es generado. Existen algunas bibliotecas que realizan esta tarea de forma automática.

Para hacer flexible la generación de la función de regresión se introducen variables de holgura. Para cada valor x_i se necesitan dos variables de holgura $\xi_i \geq 0$ y $\hat{\xi}_i \geq 0$, donde $\xi_i > 0$ corresponde al punto $y_i > f(x_i) + \varepsilon$ y $\hat{\xi}_i > 0$ corresponde a un punto donde $y_i < f(x_i) - \varepsilon$ (Figura 17).

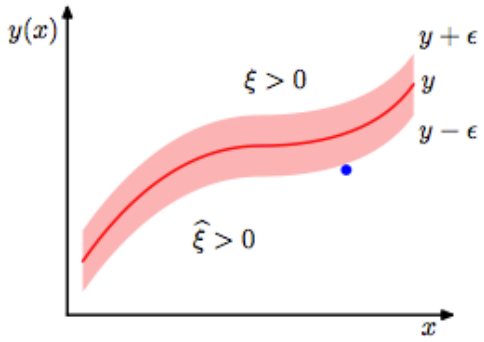

Figura 17: Función generada con insensibilidad ε y variables de holgura.

Tomando en cuenta las variables de holgura las restricciones son las siguientes:

$$y_i \leq f(x_i) + \varepsilon + \xi_i \quad (3)$$

$$y_i \geq f(x_i) - \varepsilon - \hat{\xi}_i \quad (4)$$

Utilizando nuevamente los multiplicadores de LaGrange tomando en cuenta las restricciones $\xi_i \geq 0$, $\hat{\xi}_i \geq 0$, así como las restricciones (3) y (4) mencionadas anteriormente, queda la siguiente ecuación.

$$L(w,b,\xi,\hat{\xi}) = C\sum_{i=0}^{l}(\xi_i + \hat{\xi}_i) + \frac{1}{2}\|w\|^2 - \sum_{i=0}^{l}(\mu_i\xi_i + \hat{\mu}_i\hat{\xi}_i)$$

$$-\sum_{i=0}^{l}\alpha_i(\varepsilon + \xi_n + f(x_i) - y_i) - \sum_{i=0}^{l}\hat{\alpha}_i(\varepsilon + \hat{\xi}_i - f(x_i) + y_i) \quad (5)$$

Después de realizar las derivadas parciales correspondientes a las variables que se desean maximizar y minimizar, se obtiene lo siguiente.

$$\frac{\delta L}{\delta w} = 0 \rightarrow w = \sum_{i=1}^{l}(\alpha_i - \hat{\alpha}_i)\phi(x_i) \quad (6)$$

$$\frac{\delta L}{\delta b} = 0 \rightarrow \sum_{i=1}^{l}(\alpha_i - \hat{\alpha}_i) = 0 \quad (7)$$

$$\frac{\delta L}{\delta \xi_i} = 0 \rightarrow \alpha_i + \mu_i = C \quad (8)$$

$$\frac{\delta L}{\delta \hat{\xi}_i} = 0 \rightarrow \hat{\alpha}_i + \hat{\mu}_i = C \quad (9)$$

Los Multiplicadores de LaGrange α_i y $\hat{\alpha}_i$ son la solución al problema de optimización, estos multiplicadores pueden verse como elementos que dan forma a la función, situándola cerca de los valores y_i que fueron introducidos como salida en el conjunto de entrenamiento. Los valores distintos de cero son conocidos como vectores de soporte y son los que deben ser tomados en cuenta para la función de regresión.

La constante C que se encuentra en la ecuación (2) sirve para determinar la penalización por muestras que no son bien clasificadas, es una constante con la cual se debe tener mucho cuidado, ya que valores altos pueden causar un sistema sin tolerancia a errores, generando una función con poca generalización y demasiado complicada de generar, por el contrario, un valor cercano a 0 permite gran cantidad de errores con una función muy generalizada aunque fácil de obtener. De igual forma, que con ε se debe experimentar el modelo de predicción de SVR generado con diferentes valores de C.

Sustituyendo los valores de (6), (7), (8) y (9) en (5) y agregando la función *kernel* se tiene como resultado:

$$L(\alpha,\hat{\alpha}) = -\frac{1}{2}\sum_{i=0}^{l}\sum_{j=0}^{l}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)k(x_i,x_j)$$

$$-\varepsilon\sum_{i=0}^{l}(\alpha_i + \hat{\alpha}_i) + \sum_{i=0}^{l}(\alpha_j - \hat{\alpha}_j)y_i \quad (10)$$

Sujeto a las siguientes restricciones:

$$0 \le a_i \le C \quad (11)$$

$$0 \le \hat{a}_i \le C \quad (12)$$

$$\sum_{i=1} a_i - \hat{a}_i = 0 \quad (13)$$

Para obtener predicciones para nuevos valores se debe sustituir (6) en (1).

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \hat{\alpha}_i)k(x,x_i) + b \quad (13)$$

El valor de b está dado por la siguiente ecuación:

$$b = y_i - \varepsilon - w^T\phi(x_i) \quad (14)$$

$$b = y_i - \varepsilon - \sum_{i=1}^{l}(\alpha_i - \hat{\alpha}_i)k(x,x_i) \quad (15)$$

El *kernel* más utilizado para regresión en las Máquinas de Soporte Vectorial es el *kernel* de función de base radial (RBF) [Wu C. et al, 2004].

$$k(x_i, x) = \exp\{-\gamma \left| x - x_i \right|^2\} \quad (16)$$

El *kernel rbf* tiene la característica de contemplar la distancia entre las muestras, es decir, las muestras son afectadas en diferente magnitud con relación a la distancia. Las muestras más cercanas a un evento inusual son las más afectadas y va decreciendo su impacto para las muestras más distantes siguiendo una distribución gaussiana.

4. Metodología

La metodología para realizar análisis espacio temporal de eventos de tránsito por medio de SVM utilizando *Crowdsourcing* y VGI, se compone de ocho etapas que están divididas en dos categorías:

Geocodificación:

1. Recolección de información.
2. Análisis de la información y obtención de datos sobresalientes.
3. Creación de diccionarios y ejes equivalentes.
4. División del *Gazetteer*.
5. Estandarización.
6. Identificación y localización de eventos viales.

Análisis espacio temporal:

7. Generación de modelos de predicción.
8. Pronóstico de eventos viales.

4.1. Marco de trabajo

Dadas las características de tiempo y espacio la metodología propuesta muestra zonas conflictivas donde existen eventos viales que afectan la circulación. El proceso de la metodología de forma general se describe a continuación (ver Figura 18).

En la primera etapa se desarrolló una metodología para llevar a cabo el proceso de Geocodificación, el cual consiste en asignar coordenadas a los eventos de tránsito mencionados en Twitter, además de llevar a cabo una interpretación geográfica adecuada al tipo de evento ocurrido. Incluye las siguientes etapas:

1. ***Recolección de información.*** Se desarrolla un proceso de almacenamiento de tweets para su futuro procesamiento.
2. ***Análisis de la información y obtención de datos sobresalientes.*** Se analiza la información recolectada y se obtienen datos que pueden ser utilizados para enriquecer el *Gazetteer*.

3. ***Creación de diccionarios y ejes equivalentes.*** Se desarrollan diccionarios de entidades mencionadas frecuentemente dentro de la colección de tweets, por ejemplo, abreviaciones comunes, apodos a lugares conocidos, monumentos famosos, etc.
4. ***División del Gazetteer.*** Existen calles frecuentemente mencionadas dentro de Twitter, estas calles conforman un grupo reducido al cual es más fácil de acceder.
5. ***Estandarización.*** Con el fin de mejorar el proceso de identificación de calles, tanto el *Gazetteer* como cada tweet deben tener un formato parecido, por ejemplo, minúsculas, cambiar abreviaciones por palabras completas, reemplazar hashtags, etc.
6. ***Identificación y localización de eventos viales.*** Se utilizan en conjunto el *Gazetteer* y los diccionarios para la identificación de elementos geográficos y clasificación de eventos viales.

   En la segunda categoría se lleva a cabo el análisis espacio temporal abordando los procesos necesarios para realizar predicciones de tráfico basado en información obtenida por el proceso de Geocodificación. Incluye las siguientes etapas:
7. ***Generación de modelos de predicción.*** Creación de datos de entrenamiento a partir de la información obtenida del proceso de Geocodificación y creación de los modelos de predicción.
8. ***Pronóstico de eventos viales.*** Uso de los modelos de predicción en datos de prueba.

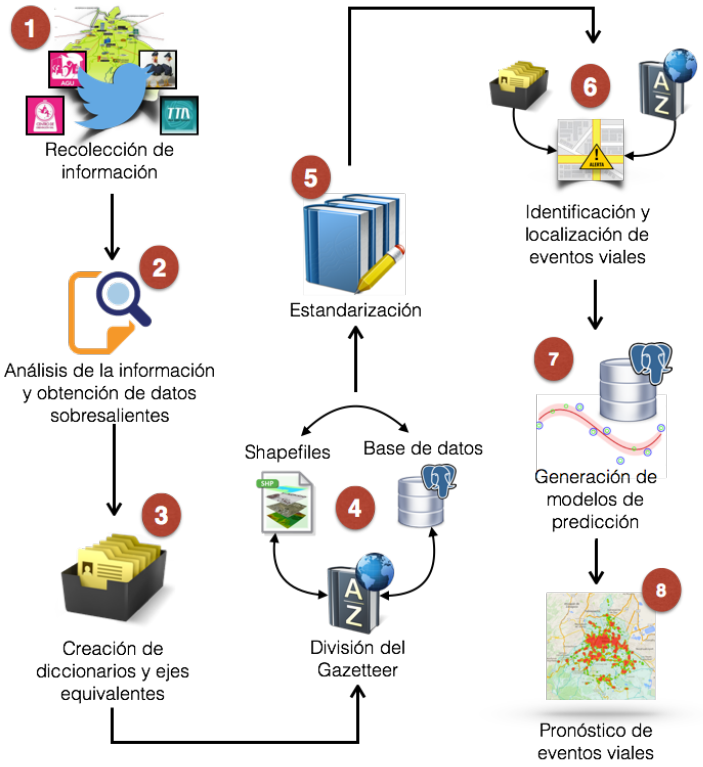

Figura 18: Esquema general de la metodología.

4.2. Recolección de información

Para el proceso de recolección de información se realizó un análisis de cuentas de *Twitter* relacionadas al tránsito vehicular en la Ciudad de México. Se analizaron ciertos parámetros para decidir que perfiles considerar y cuáles deberían ser descartados. Los parámetros a considerar son los siguientes:

1. **Localidad**: Dónde ha sido creada la cuenta. (Existen cuentas que no son de la Ciudad de México; sin embargo, los datos que son publicados pertenecen a la Ciudad y son correctos).
2. **Fecha de creación**: Analizar la antigüedad de la cuenta.
3. **Número de seguidores**: El número de seguidores afirma la utilidad de los datos que publica y verifica que sus publicaciones sean correctas.
4. **Promedio de tweets por día**: Considerar si la cuenta publica suficiente información para ser considerada.

5. **Pertenece al gobierno**. Las cuentas que pertenecen al gobierno están obligadas a dar información verdadera y mantienen sus cuentas abiertas por periodos indefinidos.
6. **Página Web**: Las cuentas que tienen sitio web, generalmente pertenecen al gobierno o a una empresa privada, como noticieros de radio y televisión.

Las cuentas que son consideradas en este trabajo se muestran en la Tabla 1.

Tabla 1: Listado de cuentas seleccionadas.

Cuenta Twitter	Localidad	Fecha de creación	Seguidores	Número de tweets	Pertenece al gobierno.	Sitio Web
SSPDFVIAL	México	07.14. 2010	369,115	154.65	Si	sup.df.gob.mx
PolloVial	México	01.31.2013	667	71.91	No	Sin sitio web
Trafico889	México	05.14.2009	137,099	90.54	No	siempre889.com/trafico
Alertux	México	10.16.2012	179,574	35.59	No	www.alertux.com
072AvialCDMX	México	10.20.2010	83,535	134.71	Si	www.agu.df.gob.mx
RedVial	México	03.09.2010	63,702	44.81	No	rvial.mx

4.3. Análisis de la información y obtención de datos sobresalientes

De la colección de tweets recolectada se identificaron los detalles que son mencionados por los usuarios, se tiene conocimiento que la información publicada está relacionada a eventos viales debido a la naturaleza de las cuentas, pero también son identificables patrones repetitivos, por ejemplo calles populares, abreviaciones comunes, lugares emblemáticos de la ciudad, monumentos históricos, etc. Por tanto, se desarrolló un script para determinar los patrones más repetitivos dentro de la colección de tweets obtenida. El script busca los n-grama más comunes. Un n-grama es un subconjunto de palabras de un conjunto más grande. Da cada tweet se obtuvieron los n-gramas y se contabilizó el número de ocurrencias en toda la colección. Se seleccionaron los n-gramas más repetitivos, en este caso mayor a 100 repeticiones. A pesar de que un n-grama por definición considera cualquier combinación de caracteres, palabras en cualquier secuencia, en este caso el

script solo considera elementos contiguos de n-palabras. La Figura 19 muestra un ejemplo del funcionamiento del script.

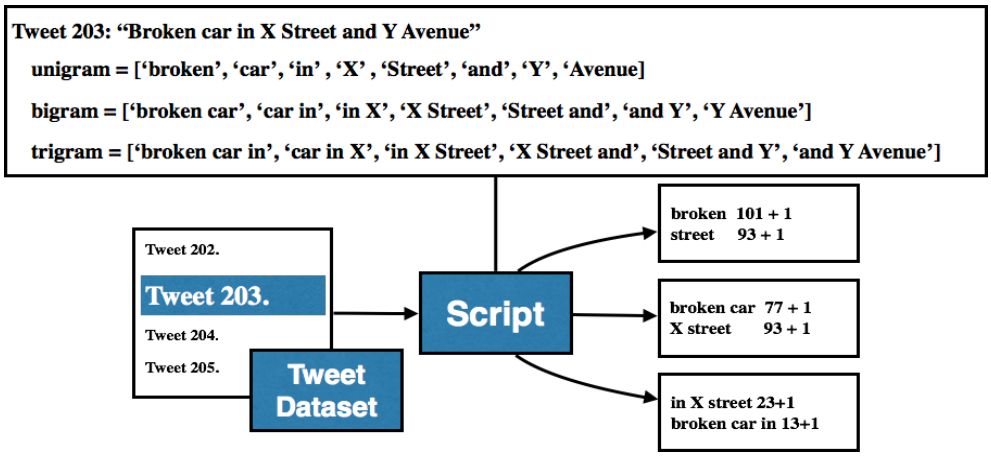

Figura 19: Obtención de n-gramas y contabilización de ocurrencias.

De los n-gramas más frecuentes se identificaron 456 calles, 150 eventos de tráfico, 135 hashtags, 69 apodos, 65 edificios, lugares y monumentos, 34 abreviaciones y 26 preposiciones y relaciones espaciales. La Figura 20 muestra la información recolectada de la colección de tweets.

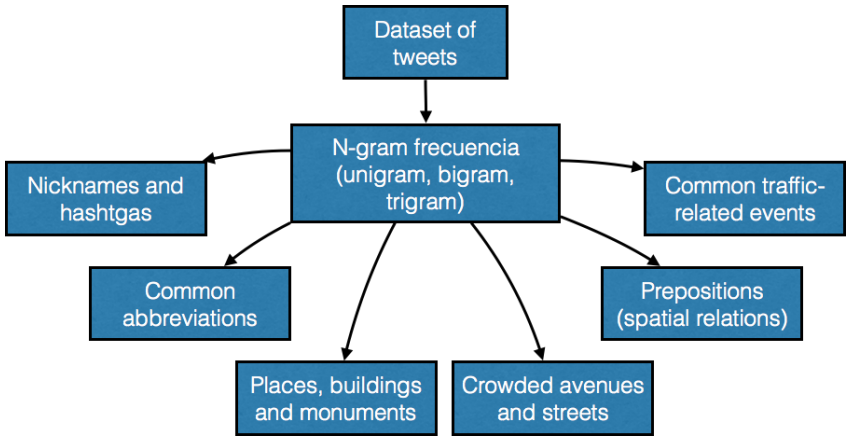

Figura 20: Información recolectada de la colección de tweets.

4.4. Creación de diccionarios y ejes equivalentes

De los resultados obtenidos en el paso anterior, datos de *OpenStreetMap* y del Instituto Nacional de Estadística y Geografía (INEGI) fueron utilizados para generar diccionarios con el fin de enriquecer la información del *Gazetteer*. Los diccionarios generados son los siguientes:

- Diccionario de abreviaciones.
- Diccionario de hashtags.
- Diccionario de apodos.
- Diccionario de eventos viales.
  - Diccionario de transporte público.
  - Diccionario de calles principales.
  - Diccionario de edificios, lugares y monumentos.
  - Diccionario de vecindarios.

Los últimos cuatro diccionarios tienen una componente geográfica, la cual es usada en el proceso de geocodificación explicado más adelante (Figura 21). A estos diccionarios los llamados diccionarios de elementos geográficos. Un diccionario de preposiciones y relaciones espaciales pudo ser creado, pero este trabajo realiza una clasificación de eventos viales de acuerdo con el número de elementos geográficos identificados en cada tweet.

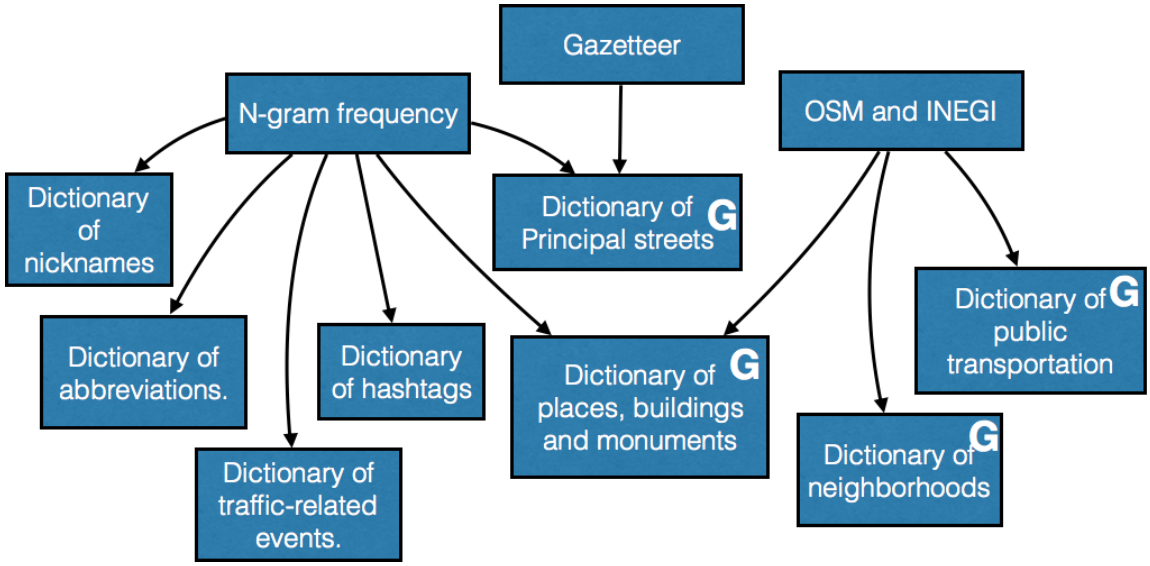

Figura 21: Creación de diccionarios geográficos y no geográficos.

Acerca de los ejes equivalentes, es frecuente que las calles tengan más de un nombre oficial. La ciudad de México tienen 31 ejes viales y 2 circuitos que cubren más de 10 mil kilómetros de longitud, Estos ejes y circuitos viales cambian de nombre a lo largo de su camino cuando cruzan con otras calles. Por esta razón, la gente suele llamarlas por su nombre principal o por el nombre de la calle en un segmento determinado (Figura 22).

Debido a que las dos opciones son válidas, todas las alternativas posibles deben ser buscadas dentro de cada tweet, como resultado un diccionario de ejes equivalentes es añadido a la colección de diccionarios listados anteriormente.

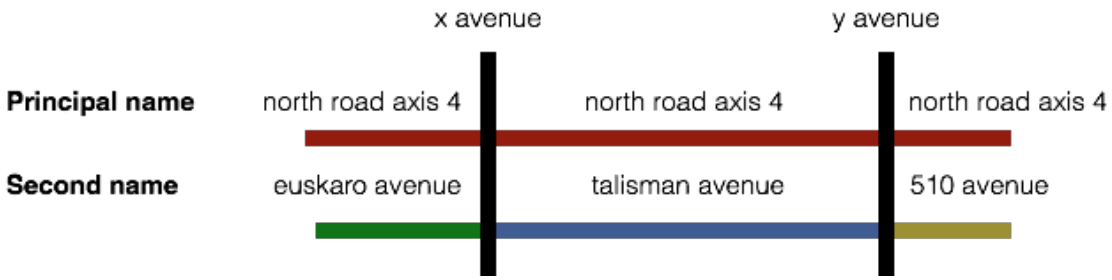

Figura 22: Representación de ejes equivalentes.

4.5. División del *Gazetteer*

Se detectó un reducido grupo de calles donde se concentran los eventos viales, basados en el diccionario de calles principales, se encontró que solo el 19% del *Gazetteer* completo aparece dentro de los tweets. Por tanto, se decidió dividir el *Gazetteer* en dos partes, la primera está conformada por las calles que son frecuentemente mencionadas (calles dentro del diccionario de calles principales que se encuentran dentro del *Gazetteer*) y la segunda compuesta por el 81% restante. Aunque la división del *Gazetteer* no mejora los resultados del proceso de Geocodificación, el rendimiento de la identificación y localización es mejorado.

4.6. Estandarización

Para mejorar el proceso de estandarización, los diccionarios de elementos no geográficos son utilizados (diccionario de abreviaciones, diccionario de apodos, diccionario de hashtags). Dentro del *Gazetteer* se encuentra que las calles incluían abreviaciones, calles en mayúsculas, contienen acentos y elementos sin nombre asignado o con valores por defecto. Por tanto, en esta etapa el nombre de las calles es transformado a minúsculas y se eliminaron los acentos (CLL TALISMÁN - cll talisman). Además, utilizando el diccionario de abreviaciones comunes, fueron reemplazadas por su nombre completo (cll talisman - calle talisman). Finalmente las calles sin nombre asignado y con valores por defecto fueron eliminadas.

Otros problemas detectados dentro de los tweets además de los solucionados previamente fueron enlaces a otras cuentas, apodos, faltas de ortografía y hashtags (http://t.co/hAN0K0WS, @OVIALCDMX, 'El ángel', 'circuito iterior', #avenidainsurgentes). Para solventar estos errores se utilizaron los diccionarios de apodos y *hashtags*, ('El ángel' - 'ángel de la Independencia' e #avenidainsurgentes - avenida insurgentes). Enlaces y mención a otras cuentas fueron suprimidos de cada tweet, faltas de ortografía no son resueltas en esta metodología, pero se propone utilizar diccionario de faltas de ortografía o algoritmos de coincidencias aproximadas.

Tanto en el *Gazetteer* como en la colección de *tweets* se filtraron *stop words*, que son palabras que no agregan ningún sentido al enunciado y son más comunes en un lenguaje (artículos, pronombres y preposiciones). No existe una lista universal de *stop words* usadas en el procesamiento de lenguaje natural. Por esta razón, la lista de *stop words* utilizada en esta metodología es la definida por la biblioteca Natural Language Toolkit [Bird, 2006].

4.7. Identificación y localización de eventos viales

La identificación de elementos geográficos en *tweets* se llevó a cabo utilizando todos los diccionarios de elementos geográficos: diccionario de calles principales, diccionario de calles no comunes, diccionario de vecindarios, diccionario de transporte público y diccionario de lugares, edificios y monumentos.

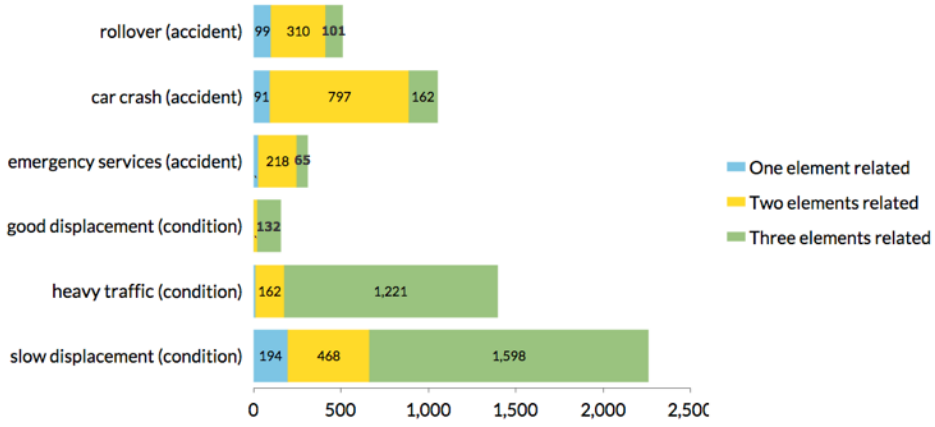

Figura 23: Eventos viales y su relación con el número de elementos geográficos identificados.

Los eventos viales reportados por las cuentas de Twitter seleccionadas comúnmente hablan de accidentes, malas y buenas condiciones. Por ejemplo, un evento vial considerado como accidente es mencionado dentro de tweets como **choque, volcadura,**

servicios de emergencia, un evento vial considerado como mala condición es mencionado como **lento desplazamiento, asentamientos, vuelta de rueda**, por otra parte, una buena condición es mencionada como **buen desplazamiento, sigue avanzando, sin problemas**, entre otras. Aunque existe una gran cantidad de propaganda comercial, preguntas y consejos de seguridad vial, son fácilmente filtrados, debido a la ausencia de elementos geográficos detectados y su reducida longitud del mensaje. Con base en la colección de tweets obtenida y al análisis de frecuencia de n-gramas, un accidente es considerado como un evento que ocurre en un punto específico que tiene relacionado uno o dos elementos geográficos, por ejemplo, **choque en avenida X con avenida Y, semáforo descompuesto en la intersección de calle Z y calle W, volcadura sobre la estación del metro J**. Una buena o mala condición es considerada como la situación actual de un segmento de una calle, comúnmente con uno, dos o tres elementos geográficos relacionados, por ejemplo, **asentamientos sobre calle X de calle Y hasta calle W, buen desplazamiento entre calle A y calle B sobre calle C, tráfico pesado en calle T en vecindario S**. Parte del diccionario de eventos viales clasificado por accidentes, buenas y malas condiciones es mostrado en la Tabla 2. Desde que un *tweet* está restringido a 140 caracteres, es difícil publicar un enlace, una mención a una cuenta, un evento vial y más de 3 elementos geográficos, por tanto, se identifica que el número de elementos geográficos mencionados dentro de los *tweets* tienen una fuerte relación con el tipo de evento vial (Figura 23).

Tabla 2: Listado de cuentas seleccionadas.

Accidente	Frecuencia	Condición	Frecuencia
Servicio de emergencia	378	Carretera bloqueada	4377
Darse vuelta	612	Aún cerca	1053
Accidente	1162	Tráfico pesado	1423
Inundación	432	Desplazamiento lento	2779
Coque de carros	1312	Carretera en reparacoón	1225
Emergencia en lugar	508	Carretera cerrada	2521
Carro descompuesto	1002	Embotellamiento	1423
señales de tráfico fuera de servicio	241	Estancamiento	1101

Cada diccionario de elementos geográficos tiene una primitiva geográfica de representación (punto, línea, polígono). El diccionario de transporte público es representado por una

colección de puntos, el diccionario de calles es representado por una colección de segmentos de línea y los diccionarios de vecindarios y lugares, edificios y monumentos son representados por polígonos. Como resultado de la búsqueda de elementos geográficos de los diccionarios y el *Gazetteer* dentro de los *tweets* se obtiene una colección de primitivas geográficas (ver Figura 24).

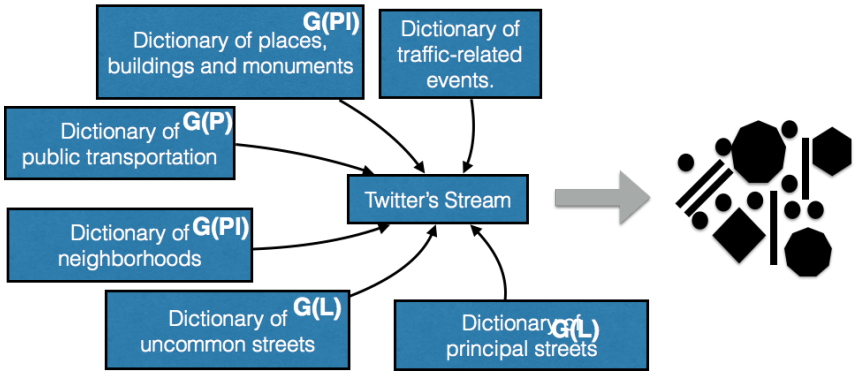

Figura 24: Resultado del proceso de identificación de la colección de tweets.

Asumiendo que pueden existir 1, 2 ó 3 referencias geográficas, el número de posibles relaciones que pueden existir entre ellas corresponde con la fórmula de combinaciones con reemplazo (Ecuación 1).

$$CR_m^n = \left(\begin{array}{c} m+n-1 \\ n \end{array} \right) = \frac{(m+n-1)!}{n!(m-1)!},$$

Ecuación 1. Ecuación de combinaciones con reemplazo.

Donde *m* es el número de elementos posibles a seleccionar, en este caso punto, línea o polígono y *n* es el número de elementos encontrados. Así, para un elemento se tiene [(punto), (línea), (polígono)], para dos elementos identificados es [(punto, punto), (punto ,línea), (punto, polígono), (línea, línea), (línea, polígono), (polígono, polígono)] y para tres elementos [(punto, punto, punto), (punto, punto, line), (punto, punto, polígono), (punto, línea, línea), (punto, línea, polígono), (punto, polígono, polígono), (línea, línea, línea), (línea, línea, polígono), (línea, polígono, polígono), (polígono, polígono, polígono)].

Varias de estas relaciones geográficas no son precisas con respecto a la localización de un evento vial o la frecuencia dentro de la colección de tweets no es numerosa, así que algunas son descartadas. Para esta metodología solo las siguientes relaciones fueron consideradas: [(punto), (punto, línea), (línea, línea), (punto, punto, línea), (punto, línea,

línea), (línea, línea, línea), (línea, línea, polígono)]. Estas relaciones fueron claramente identificadas dentro de la colección de tweets asumiendo las situaciones listadas a continuación:

- (punto) representa un accidente en una estación de transporte público.
- (punto, línea) representa una condición de un segmento de calle frente a una estación de transporte público.
- (línea, línea) representa un accidente en una intersección.
- (punto, punto, línea) representa una condición de un segmento de calle delimitado por dos estaciones de transporte público.
- (punto, línea, línea) representa una condición de un segmento de calle delimitado por otra calle y una estación de transporte público.
- (línea, línea, línea) representa una condición de un segmento de calle delimitado por dos calles.
- (línea, línea, polígono) representa un segmento de calle delimitado por una calle y un lugar, edificio o monumento.

Aunque es probable que existan otras situaciones viales que consideren estos grupos de elementos geográficos, las situaciones listadas son más frecuentes. Con el fin de obtener el resultado de estas suposiciones, tres operaciones geográficas fueron consideradas:

1. Encontrar la intersección entre calles.
2. Encontrar el punto más cercano a otro elemento geográfico.
3. Encontrar el *bounding box* o envolvente convexa de un segmento de línea.

Las operaciones espaciales fueron ejecutadas utilizando las funciones de *PostGIS* como *ST_Intersection*, *ST_ClosestPoint*, *ST_Envelope* y *ST_ConvexHull*. El resultado es un elemento geográfico, donde el evento vial ocurrió. Por ejemplo, para encontrar la relación (línea, línea, línea) el pseudocódigo es el siguiente:

- *Mientras todas las posibles combinaciones no han sido probadas:*
  - *Existe intersección del elemento A y el elemento B (Operación 1)*
    - *Salvar*
- *¿Existen más de dos intersecciones?*
  - *Si: ¿Existe un elemento en común en las dos intersecciones?*
    - *Si: Encuentra el bounding box (o envolvente convexa) del elemento en común delimitado por las dos intersecciones (Operación 3)*

- *No: Verifica relación (línea, línea)*

- *No: Verifica relación (línea, línea)*

4.8. Generación de modelos de predicción

Los modelos de predicción son generados con Máquinas de Soporte Vectorial para Regresión (SVR), para generar los modelos de predicción es necesario un conjunto de entrenamiento, el cual consiste de un vector n-dimensional de características representadas con valores categóricos y asociados a un valor de salida.

4.8.1. Generación del conjunto de entrenamiento

La selección de características para los métodos de aprendizaje automático determinan el éxito o el fracaso de un método de aprendizaje, sin embargo, la mejor forma de seleccionar los atributos más relevantes es de forma manual, basado en el conocimiento profundo del problema de aprendizaje y el conocimiento de que representa cada valor [Witten, I. et al., 2005].

La información de cada tweet se encuentra restringida a 140 caracteres como se mencionó anteriormente; por tanto, obtener información para conformar un vector de múltiples características está severamente restringido. El resultado del proceso de geolocalización de los tweets es la salida que tiene asociado cada vector de características, por tanto, las características que fueron consideradas para conformar el vector están relacionadas al tiempo cuando éste ocurrió (ver Tabla 3).

Tabla 3: Características para el vector característico.

Característica	Valores asignados	Descripción
Mes	[1,…,12]	Mes del año.
Día del mes	[1,2,…,31]	Día cuando fue publicado el *tweet*.
Día de la semana	[0,1,2,3,4,5,6]	Día de la semana cuando fue publicado el *tweet*. El número 0 corresponde al día Domingo, 2 al día Lunes y así sucesivamente hasta el día Domingo.
Hora del día	[0,1,…,24]	Hora del día cuando fue publicado el *tweet*.

Existen algunas características que fueron descartadas debido a que generan ruido en la generación del modelo de predicción, en el siguiente capítulo se mencionan algunas características probadas y cuáles fueron los resultados obtenidos.

Las características espaciales relacionadas a cada tweet son escasas ya que en pocas ocasiones se hace mención de colonias, códigos postales, delegaciones o zonas urbanas. Aunque la delegación, colonia o código postal pueden ser definidos a partir del proceso de geocodificación de cada tweet, teóricamente en los métodos de aprendizaje automático supervisado, el vector de características no puede ser definido a partir del valor de salida asociado.

Existen parámetros adicionales que fueron obtenidos del proceso de geocodificación que no son utilizados para conformar el conjunto de entrenamiento, sin embargo, son necesarios para filtrar la información y obtener información que no produzca ruido para la generación del modelo de predicción.

Se descartaron para el conjunto de entrenamiento tweets considerados como buenas condiciones y fueron agrupados por delegaciones. Aunque se genera un modelo de predicción por cada delegación, se incrementa considerablemente su precisión y exactitud debido a que se reduce el número de vectores que tienen un valor aproximado en tiempo pero en lugares distantes (Figura 25).

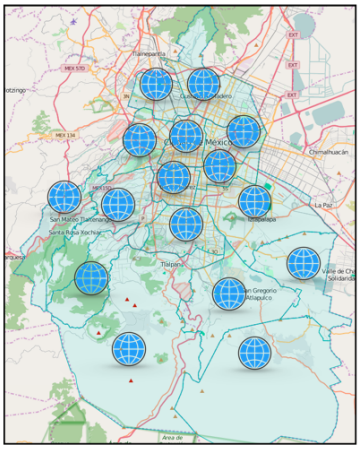

Figura 25: Modelo de predicción por cada delegación.

La matriz resultante para la generación del modelo de predicción está conformada por todos los elementos que fueron geocodificados agrupados por delegación. Cada vector de

entrada contiene características de tiempo asociados a una salida, la cual es la coordenada geográfica donde ocurrió el evento vial (Tabla 4).

Tabla 4: Conjunto de entrenamiento para generar el modelo de predicción de eventos viales.

hora	día	d_sem	mes	coordenada
22	26	5	9	POINT(-99.13388 19.420084)
21	22	3	10	POINT(-99.14466 19.407152)
17	5	3	11	POINT(-99.156445 19.4381015)
12	5	3	11	POINT(-99.140053 19.403607)
13	23	2	12	POINT(-99.146662 19.431037)
18	29	1	9	POINT(-99.1425965 19.4247005)
10	30	4	10	POINT(-99.15687025 19.43211625)
6	10	5	10	POINT(-99.137952 19.436513)
22	26	3	11	POINT(-99.162119 19.425651)

4.8.2. Herramienta SVM para Regresión

El modelo de predicción es generado con la biblioteca Scikti-Learn [Scikit-learn, 2015], la cual contiene diferentes métodos de aprendizaje automático implementados en Python, para hacer uso de las Máquinas de Soporte Vectorial para Regresión Scikit-Learn encapsula las bibliotecas de *libsvm* y *liblinear* [Chang, C. et al., 2011], las cuales son bibliotecas desarrolladas especialmente para utilizar Máquinas de Soporte Vectorial de forma fácil en cualquier aplicación.

El uso de Scikti-Learn para generar los modelos de predicción involucra dos etapas. La primera consiste en realizar el entrenamiento del sistema de aprendizaje, y la segunda consiste en validar el modelo de predicción con información proveniente de datos de prueba.

EL proceso de entrenamiento de la Máquina de Soporte Vectorial para Regresión se da bajo los siguientes parámetros:

- ***C***. Factor de penalización. Valor por defecto 1.0.
- ***Epsilon/e-tube***. En SVR se refiere al ε-tube, es decir, el radio de insensibilidad de la función, en el cual no está asociada ninguna penalización. Valor por defecto 0.1.
- ***Kernel***. Es el tipo de *kernel* utilizado en el algoritmo, puede ser lineal, rbf, polinomial, sigmoide o precalculado. Valor por defecto rbf.

- ***Degree***. Grado de la función *kernel*, es importante en los *kernel rbf*, polinomial y sigmoide. Valor por defecto 3.0.
- ***Gama***. Coeficiente del *kernel* para rbf y polinomial, si gamma es 0.0 se toma 1/# de características como valor. Valor por defecto 0.0.
- ***Coef0***. Término independiente en la función *kernel*. Importante para el *kernel* polinomial y sigmoide. Valor por defecto 0.0.
- ***Shrinking***. Para aplicar la heurística *shirinking*. Valor por defecto verdadero.
- ***Tol***. Tolerancia para criterio de parada. Valor por defecto 1e-3.
- ***Cache size***. Especifica el tamaño de la cache del *kernel* en megabytes. Valor por defecto 200.
- ***Verbose***. Habilita modo detallado. Toma tiempo de procesamiento y no funciona correctamente en un contexto multi-hilo. Valor por defecto falso.
- ***Max iter***. Límite de iteraciones dentro de la solución, si el valor es -1 no hay límite. Valor por defecto -1.

Los valores asignados al método de SVR para la generación del modelo de predicción de eventos vehiculares son los siguientes:

- *Kernel*: rbf
- C: 1e3
- Gamma: 4e-2
- *Epsilon/e-tube*: 1e-4

Los parámetros restantes quedaron con su valor por defecto. La selección de los valores se realizó de forma manual mediante pruebas utilizando validación cruzada, método que garantiza que los resultados son independientes del conjunto de entrenamiento y el conjunto de prueba.

```
svr_rbf = SVR(kernel='rbf', C=1e3, gamma=0.04, epsilon=0.0001)
```

El conjunto de entrenamiento es seccionado en la entrada compuesta por los vectores de características y la salida conformada por los puntos de los eventos geocodificados. Ambas colecciones de entradas y salidas son los parámetros que recibe SVR para realizar el proceso de aprendizaje del modelo de predicción.

```
svr_rbf.fit(input_train, output_train)
```

Con el modelo entrenado se realizan predicciones con la sección de entradas del conjunto de prueba, este conjunto de prueba es tomado de una sección del conjunto de entrenamiento y no forma parte para el proceso de aprendizaje del modelo de predicción.

```
svr_rbf.predict(input_test)
```

Los resultados obtenidos son comparados con la sección de salida del conjunto de prueba. A continuación en la figura 26 se muestra una comparación de la sección de salida del conjunto de prueba (ícono Twitter azul) y los puntos pronosticados (ícono precaución rojo).

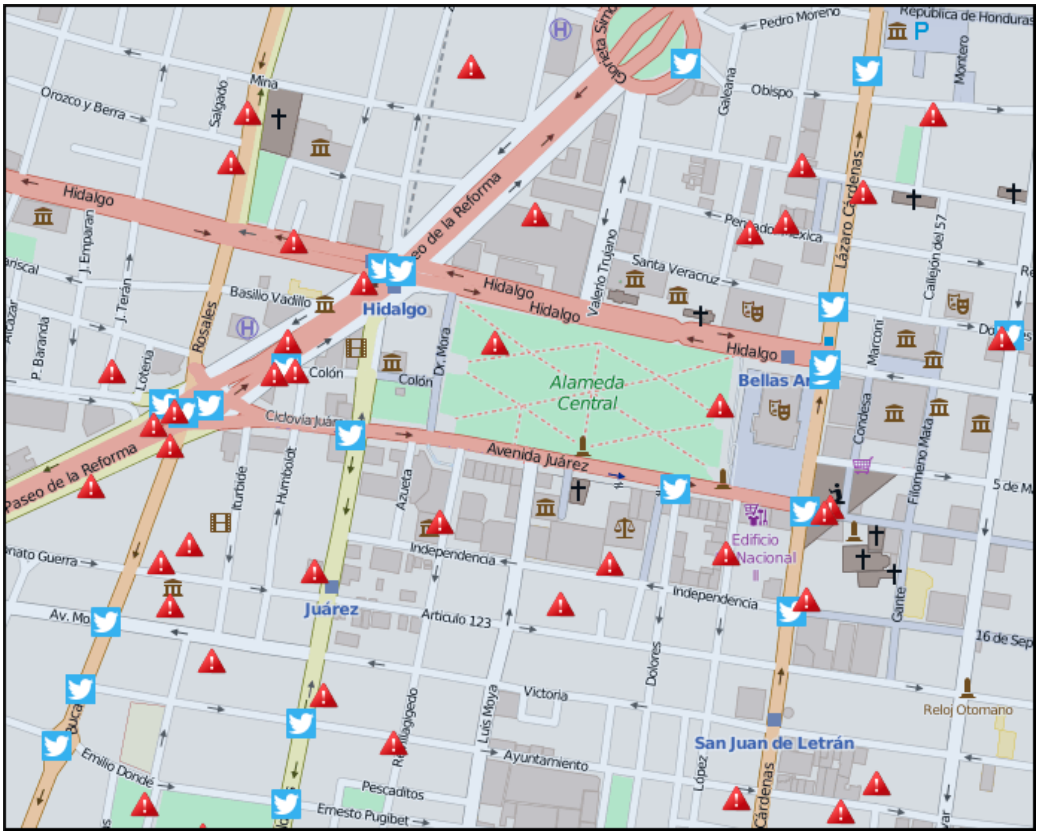

Figura 26: Visualización de puntos pronosticados y los puntos de prueba.

4.9. Análisis espacio temporal de eventos

El análisis espacio temporal de eventos de tránsito se realiza mediante una combinación posible de un conjunto de características de tiempo y espacio, con las cuales se genera una colección de vectores que funcionan como entrada al modelo de predicción, las características de entrada son las siguientes:

- **Delegación(es)**. Área en la cual se requiere conocer la situación vial.
- **Mes**. Mes en el cual se requiere conocer la situación vial.
- **Día**. Días en los cuales es posible que acontezca una situación vial.
- **Día semana**. Día de la semana en el cual se requiere conocer la situación vial.
- **Hora**. Hora en la cual se requiere conocer la situación vial.

Las características son vectorizadas con la misma estructura que se compone la entrada del conjunto de entrenamiento. El número de vectores enviado al conjunto de entrenamiento varía con respecto a un análisis de regresión, utilizando nuevamente SVR donde la entrada del conjunto de entrenamiento está compuesta por las características de tiempo hora y día de la semana, además se le suma la característica espacial delegación, representada de forma categorizada y como salida relacionada se tiene el número de accidentes ocurridos (Tabla 5).

Tabla 5: Conjunto de entrenamiento para generar el modelo de predicción de número de eventos viales.

hora	d_sem	id_del	occ
22	*5*	*8-[0,...1,]*	*120.00*
21	*3*	*3-[0,0,1,]*	*116.00*
17	*3*	*3-[0,0,1,]*	*113.00*
12	*3*	*3-[0,0,1,]*	*122.00*
13	*2*	*4-[..,0,1]*	*119.67*
18	*1*	*12-[..,1,..]*	*18.00*
10	*4*	*12-[..,1,..]*	*67.60*
6	*5*	*3-[0,0,1,]*	*83.00*
22	*3*	*3-[0,0,1,]*	*82.00*

Del modelo de predicción se obtiene el número de vectores que deben ser generados para enviarlos al modelo de predicción, este número debe ser suficiente para representar el comportamiento usual relacionado con las características de tiempo y espacio requeridas.

Cada vector que será enviado al modelo de predicción está conformado por características de tiempo y espacio seleccionadas, por ejemplo, *Realizar el análisis espacial del próximo jueves a las 2 de la tarde en la delegación Cuauhtémoc* (ver Figura 27).

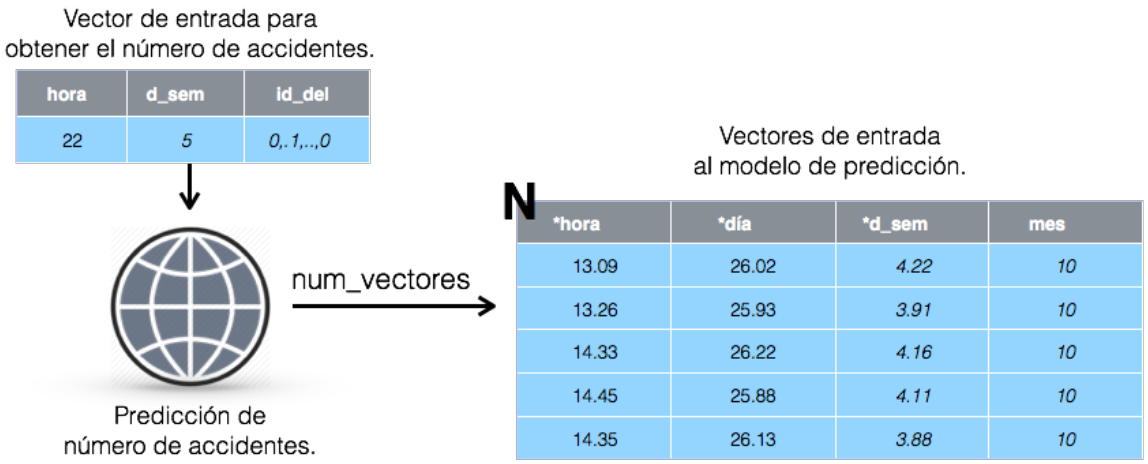

Figura 27: Proceso para generar la colección de vectores para el análisis espacio temporal.

Los parámetros de hora, día y día semana son desfasados en umbrales de [-1,1] con el fin de representar el comportamiento vial próximo a los valores requeridos. Como resultado se obtiene un análisis espacio temporal conformado por un conjunto de eventos viales relacionados a parámetros requeridos (ver Figura 28).

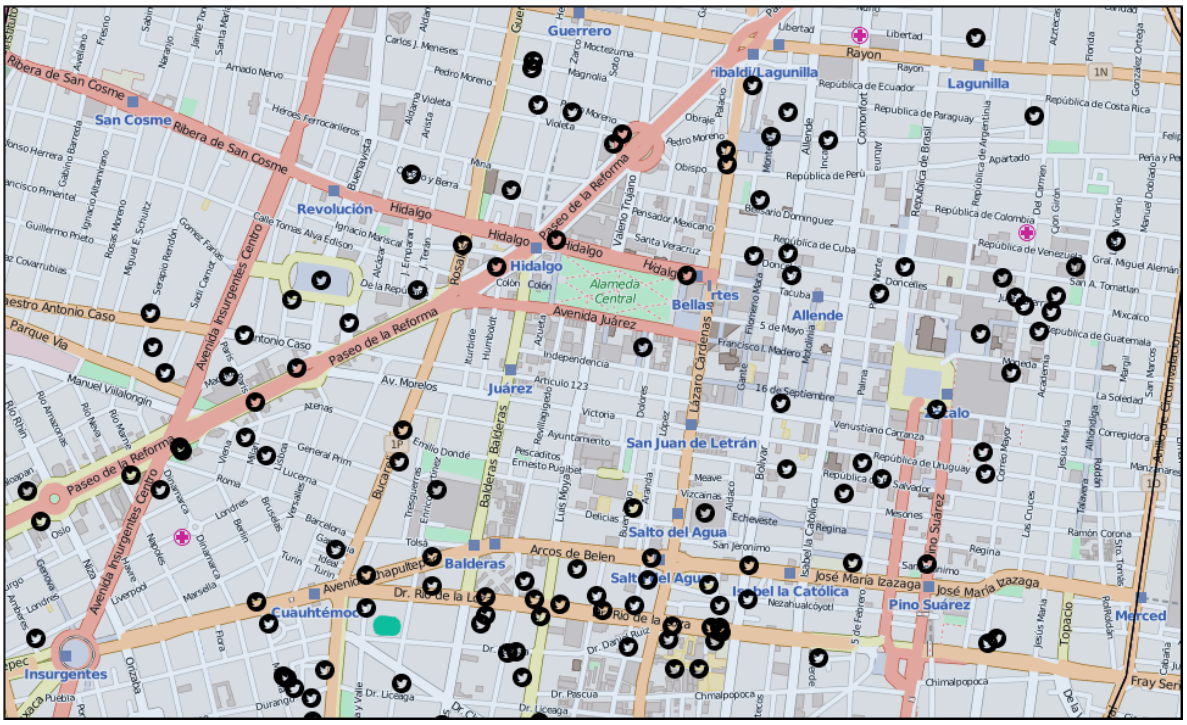

Figura 28: Resultado del análisis espacio temporal.

Los resultados anteriores son visualizados como un mapa de calor con el fin de representar la concentración de los valores recurrentes (Figura 29)

Figura 29: Vectores generados para el modelo de predicción como mapa de calor.

5. Evaluación, experimentos y resultados

Este capítulo muestra los métodos de evaluación propuestos para validar la precisión de la geocodificación y de los modelos de predicción, los experimentos realizados, como fueron llevados a cabo y los resultados obtenidos.

Primeramente se describe como se realiza la evaluación, experimentos y resultados en geocodificación. Se describe cuando se considera un verdadero positivo, un falso positivo y un falso negativo, con el fin de obtener medidas de precisión y recall. En experimentos y resultados se realizan comparaciones entre el gazetteer inicial y el gazetteer con los procesos añadidos.

Para el proceso de predicción de igual forma se describe cuando se considera un verdadero positivo, un falso positivo y un falso negativo para obtener medidas de precisión, recall y medida F. Para experimentos y resultados se realizan cambios en el vector de características, cambios en la representación de las características, en el tamaño del conjunto de entrenamiento y en los parámetros del modelo de predicción y por último se muestran comparaciones entre los resultados obtenidos.

De forma general un verdadero positivo, falso negativo, falso positivo son definidos de la siguiente manera:

Verdadero positivo: muestra encontrada por el sistema que pertenece al conjunto de la solución.

- **Falso positivo**: muestra encontrada por el sistema que NO pertenece al conjunto de la solución.
- **Falso Negativo**: muestra NO encontrada por el sistema que pertenece al conjunto de la solución.

La precisión, recall y medida-F se definen y calculan de la siguiente forma:

La **precisión** esta definida como la fracción de instancias que son relevantes con relación al conjunto de instancias recuperadas. Se calcula de la siguiente manera:

$$P = \frac{VP}{VP + FP}$$

La **recall** esta definida como la fracción de instancias que son relevantes con relación al conjunto de instancias que debieron ser recuperadas. Se calcula de la siguiente manera:

$$R = \frac{VP}{VP + FN}$$

La **medida F** es la medida de precisión que tiene una prueba, por tanto tiene relación con precisión y recall. Se calcula de la siguiente manera:

$$F = \frac{2PE}{P + E}$$

5.1. Evaluación para geocodificación.

El proceso de evaluación considera la calidad de los resultados en los experimentos realizados. Los parámetros seleccionados para considerar la calidad de los resultados para la metodología de geocodificación son *precisión* y recall. Estos parámetros son obtenidos a partir de verdaderos positivos, falsos positivos y falsos negativos.

Con el fin de medir la precisión de la metodología de geocodificación, una colección de prueba de 652 tweets fue geocodificada a mano. Se identificaron calles, estaciones de transporte publico, vecindarios, lugares, edificios y monumentos.

Para considerar un verdadero positivo debe ser un elemento geográfico identificado por la metodología y que se encuentre dentro de la colección de prueba. Un falso positivo es un elemento geográfico identificado por la metodología que no pertenece a los elementos del conjunto de prueba. Un falso negativo es un elemento geográfico que pertenece al conjunto de prueba que no fue identificado por la metodología.

La colección de prueba fue comparada con la metodología descompuesta en etapas, primero por el proceso de estandarización, seguido de la adición de ejes equivalentes y finalmente con los diccionarios geográficos.

5.2. Experimentos y resultados para geocodificación.

Se obtuvo la precisión y recall del proceso de geocodificación partiendo de una línea base la cual es el gazetteer sin ningún proceso añadido, solo fue pasado a letras minúsculas junto con cada tweet procesado. Después se obtuvo la precisión y recall añadiendo el proceso de estandarización descrito en la metodología, subsecuentemente se añadieron los ejes equivalentes y por último los diccionarios geográficos.

Se considero cuando todos los elementos que perecen al conjunto de prueba son encontrados, cuando al menos un elemento que pertenece al conjunto de prueba fue encontrado y los errores de identificación que se cometieron. Los resultados son mostrados en la Tabla 6.

Tabla 6: Comparación de resultados entre línea base y metodología de geocodificación por etapas

	Línea base	Estandarización	Estandarización + ejes equivalentes	Estandarización + ejes equivalentes + diccionarios	Colección de prueba
Todos los elementos encontrados	152	152	427	456	652
Al menos un elemento encontrado	289	388	599	608	652
Errores	363	264	53	44	0
Precisión	0.39	0.43	0.83	0.85	1.0
Recall	0.31	0.39	0.80	0.83	1.0

Se observa que la metodología propuesta tiene una precisión y recall de 85% y 83% respectivamente la cual es superior al 39% y 31% obtenidos de la línea base.

Se observo que el mayor aporte a la identificación de elementos geográficos es a través del diccionario de ejes equivalentes.

Durante el desarrollo de la metodología de geocodificación se identifico que el comportmiento de Twitter tiene relación con el la actividad en el mundo real. El número de tweets publicados de las 18 a las 20 horas es mayor que el resto. De igual forma otro

periodo de participación en Twitter es alrededor de las 8 horas, lo cual esta ligado a las horas pico dentro de la Ciudad de México (Figura 30).

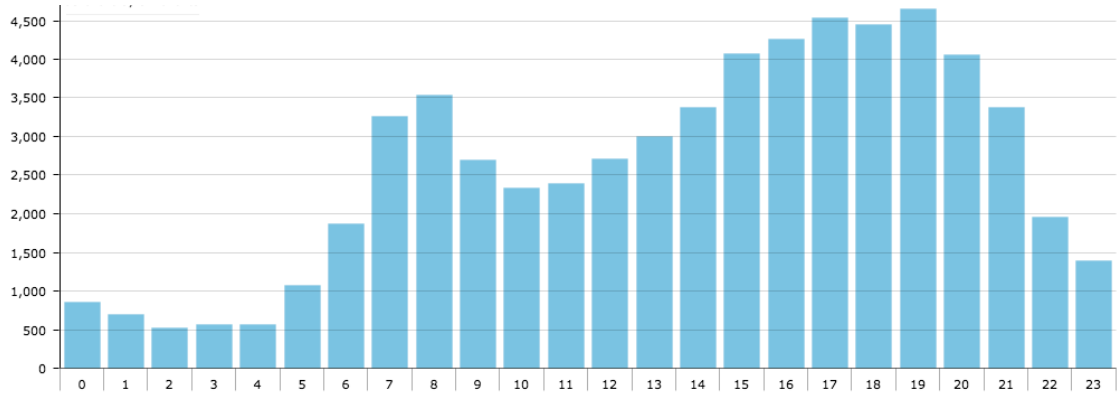

Figura 30: Participación en Twitter durante el día.

5.3. Evaluación para predicción.

Los problemas de regresión trabajan en el conjunto de los números reales, por tanto, los resultados obtenidos por un modelo de regresión cuentan con un rango de distancia entre el valor de la función y el valor real. Los umbrales establecidos en este trabajo son de *50 metros y 100 metros* entre el conjunto de elementos pronosticado por el modelo de predicción y el conjunto de los valores de prueba (Figura 31).

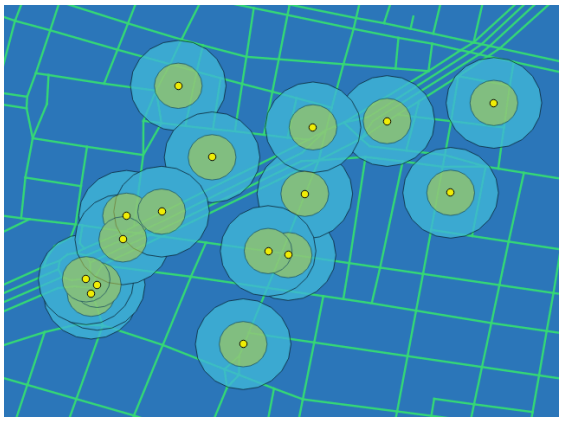

Figura 31: Pronóstico con buffer de 50 y 100 metros.

Para eventos vehiculares que trabajan con coordenadas latitud y longitud, los verdaderos positivos, falsos positivos y falsos negativos son definidos de la siguiente manera:

Un **verdadero positivo** es considerado cuando un elemento del conjunto de prueba (punto verde) se encuentra dentro del umbral de un evento pronosticado (punto amarillo). La

Figura 32 muestra un verdadero positivo con umbral de 50 metros y uno con umbral de 100 metros.

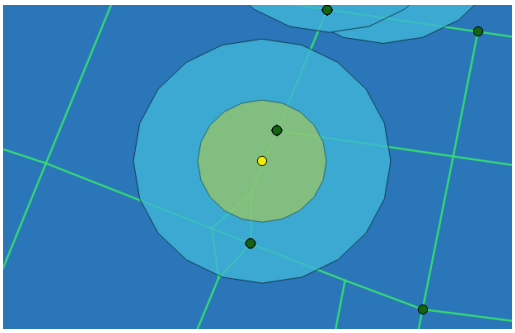

Figura 32: Verdadero positivo.

Los **falsos positivos** son eventos (punto amarillo) que fueron pronosticados por el modelo y que dentro del umbral establecido no se encuentra ningún elemento del conjunto de prueba (punto verde). En la Figura 33 se tiene un falso positivo con umbral a 50 metros y un verdadero positivo con umbral a 100 metros.

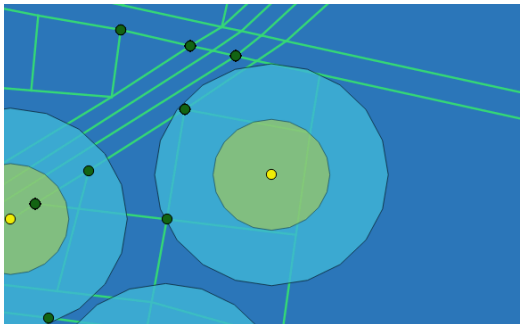

Figura 33: Falso positivo.

Los **falsos negativos** son elementos del conjunto de prueba (punto verde) que no se encontraron dentro del umbral establecido de ningún elemento del conjunto pronosticado (punto amarillo). En la Figura 34 existen 2 falsos negativos.

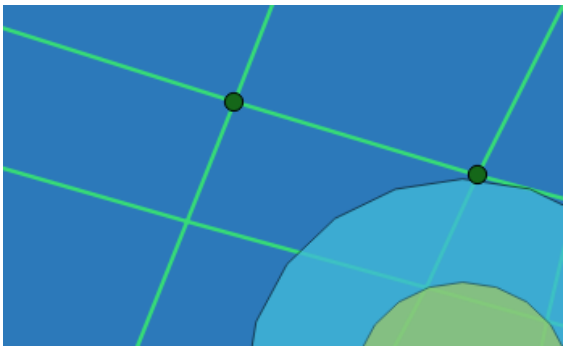

Figura 34: Falso negativo.

El cálculo de *precision*, *recall* y *medida-f* para los conjuntos obtenidos por el modelo de predicción se llevan a cabo de la siguiente manera:

El proceso de evaluación se lleva a cabo mediante un conjunto de prueba, el cual es una partición del conjunto de entrenamiento y no es utilizado para generar el modelo de predicción. Los vectores del conjunto de prueba son enviados al modelo de predicción y la salida del modelo generado se compara con las coordenadas del conjunto de prueba (ver Figura 35), obteniendo verdaderos positivos, falsos positivos y falsos negativos.

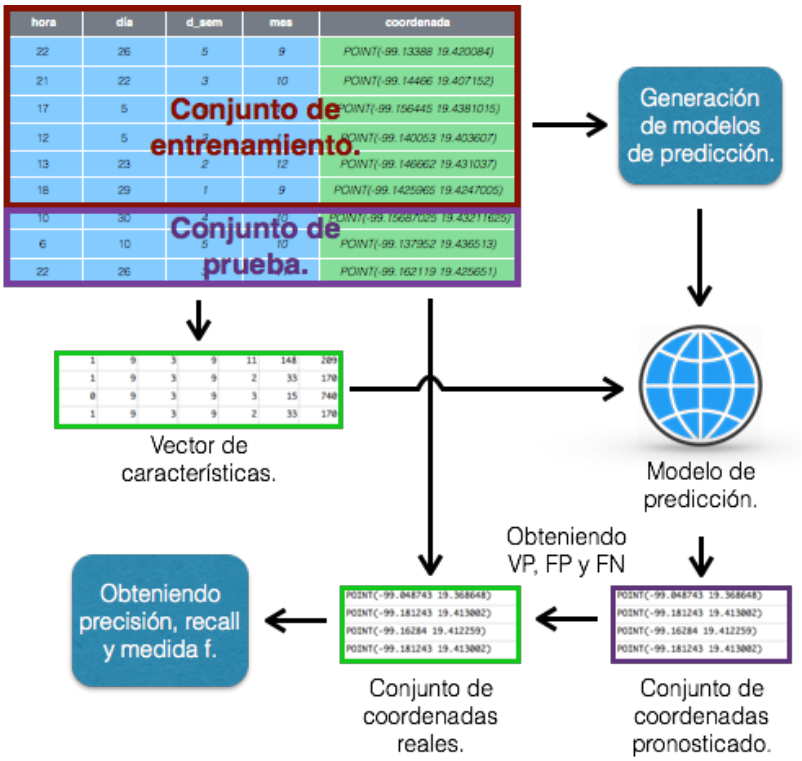

Figura 35: Proceso de evaluación.

Para garantizar que los resultados son independientes del conjunto de entrenamiento y del conjunto de prueba se utiliza validación cruzada. Este método de validación realiza *n* particiones del conjunto de entrenamiento. En la primera etapa se toma la primera partición del conjunto de entrenamiento y se utiliza como conjunto de prueba, las particiones restantes son utilizadas para la generación del modelo de predicción y se realiza el proceso de evaluación mencionado anteriormente, en la segunda etapa se toma la segunda partición como conjunto de prueba y las restantes como conjunto de entrenamiento y se

repite el proceso *n* veces, uno por cada partición, se obtiene el *precision*, *recall* y *medida f* en cada iteración y se realiza el promedio de los resultados (ver Figura 36).

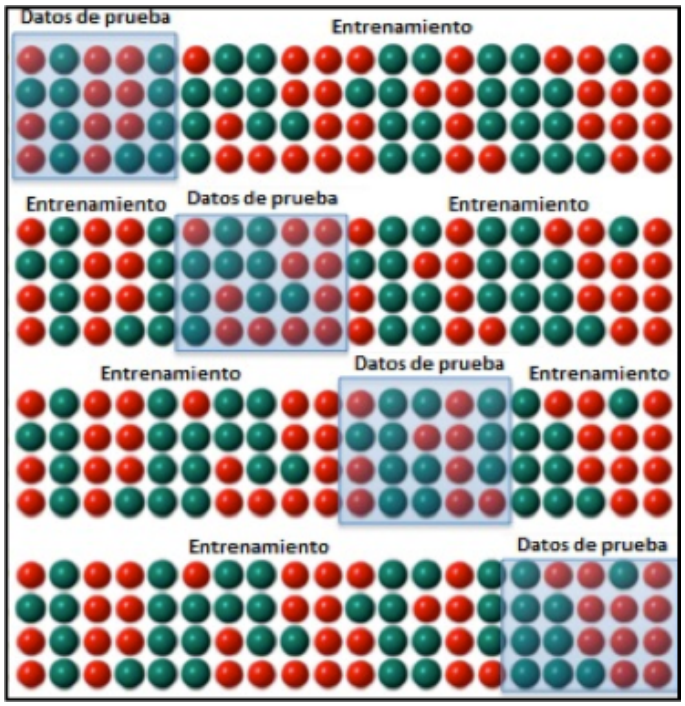

Figura 36: Validación cruzada de 4 particiones.

5.4. Experimentos y resultados para predicción

Los experimentos se realizaron con un subconjunto de la información recolectada, se utilizó la información de la delegación Cuauhtémoc. Esta delegación cuenta con eventos suficientes para generar un modelo de predicción con resultados aceptables y con la cartografía necesaria para llevar a cabo los experimentos.

5.4.1. Relación del número de eventos con *Precision*, *Recall* y Medida-*F*

El primer experimento consiste en realizar incrementos del tamaño del conjunto de entrenamiento y observar la relación que tiene con *precision*, *recall* y la *medida-f* del modelo de predicción. Se utilizó la delegación Cuauhtémoc con un umbral de 100 metros, realizando incrementos en el tamaño del conjunto de entrenamiento para cada proceso de evaluación. El conjunto de entrenamiento cuenta con las características mencionadas en el capítulo de Metodología (mes, día, día de la semana y hora) y el método de predicción SVR

cuenta con los parámetros indicados en la Metodología (kernel rbf, gamma 0.04, constante de penalización de 1e3 y épsilon 1e-4). Los resultados obtenidos se muestran en la Figura 37.

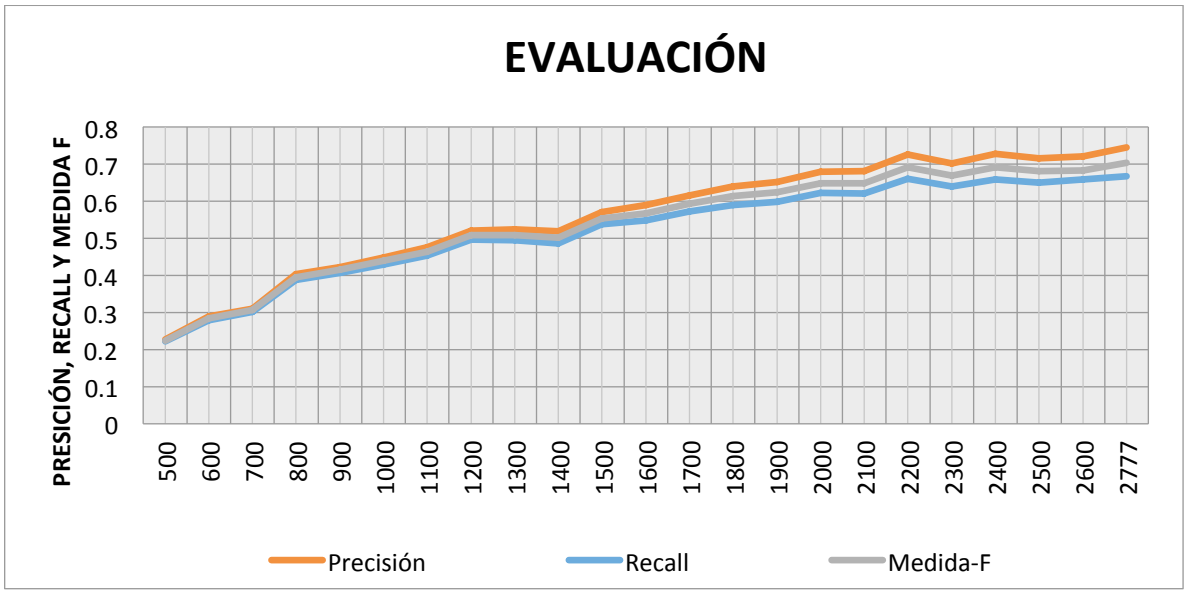

Figura 37: Relación del tamaño del conjunto de entrenamiento con los resultados de la evaluación.

El incremento en el tamaño del conjunto de entrenamiento tiene relación con el incremento en *precision*, *recall* y *medida f* obtenidos por los modelos de predicción generados. SVR realiza un proceso de aprendizaje del conjunto de entrenamiento, por tanto, entre más grande sea el conjunto de entrenamiento se obtienen modelos más precisos.

La generación del modelo de predicción es un proceso que conlleva un tiempo considerable, por tanto, se realizaron pruebas que muestran el tiempo consumido en la generación de los modelos de predicción con relación al tamaño del conjunto de entrenamiento, lo cual se puede observar en la Figura 38.

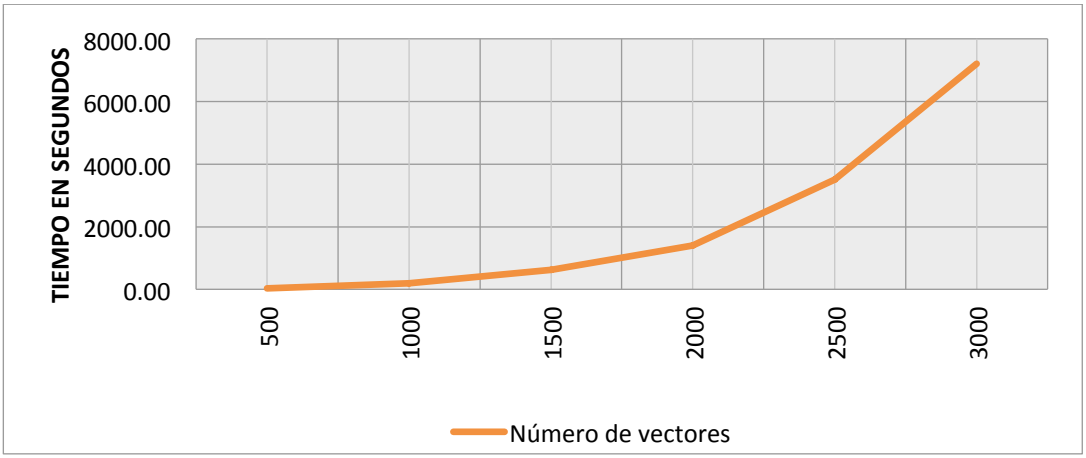

Figura 38: Relación del tamaño del conjunto de entrenamiento con el tiempo para generar el modelo de predicción.

El modelo de predicción no es generado cada vez que se realiza una predicción, los modelos son generados cuando se cuenta con un nuevo conjunto de eventos obtenidos por periodos de 7 días, 15 días o 31 días; dependiendo del número de eventos que han sido geocodificados. El obtener una predicción dado un vector de características es de tiempo constante, es decir, el modelo de predicción es una función generada que dado un valor 'x', se obtiene su valor 'y' (ver Figura 39).

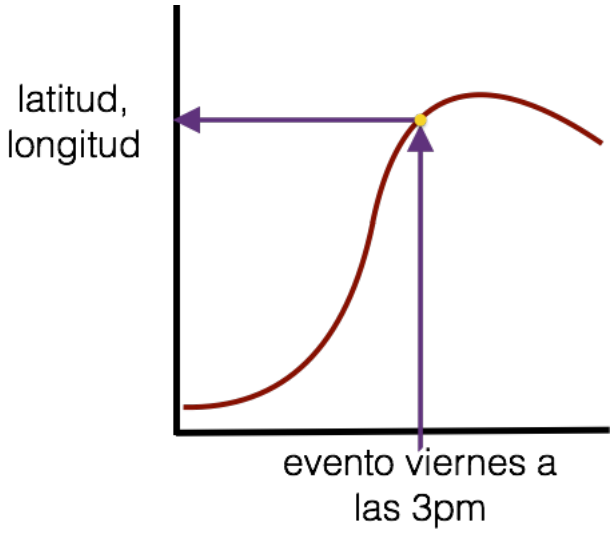

Figura 39: Obtener coordenadas del modelo de predicción dado un vector.

5.4.2. Selección de características

La selección de características que conforman al conjunto de entrenamiento es un punto fundamental para los resultados que se obtienen del modelo de predicción, para comprobar esta hipótesis se agregó el parámetro minuto donde ocurrió el evento al vector de características:

- Mes del evento ocurrido.
- Día del mes evento ocurrido.
- Día de la semana del evento ocurrido.
- Hora del evento ocurrido.
- Minuto del evento ocurrido.

Los resultados fueron comparados con el conjunto de entrenamiento inicial, se realizaron incrementos en el conjunto para medir *precision*, *recall* y *medida-F*, los resultados se muestran en las Figuras 40, 41 y 42.

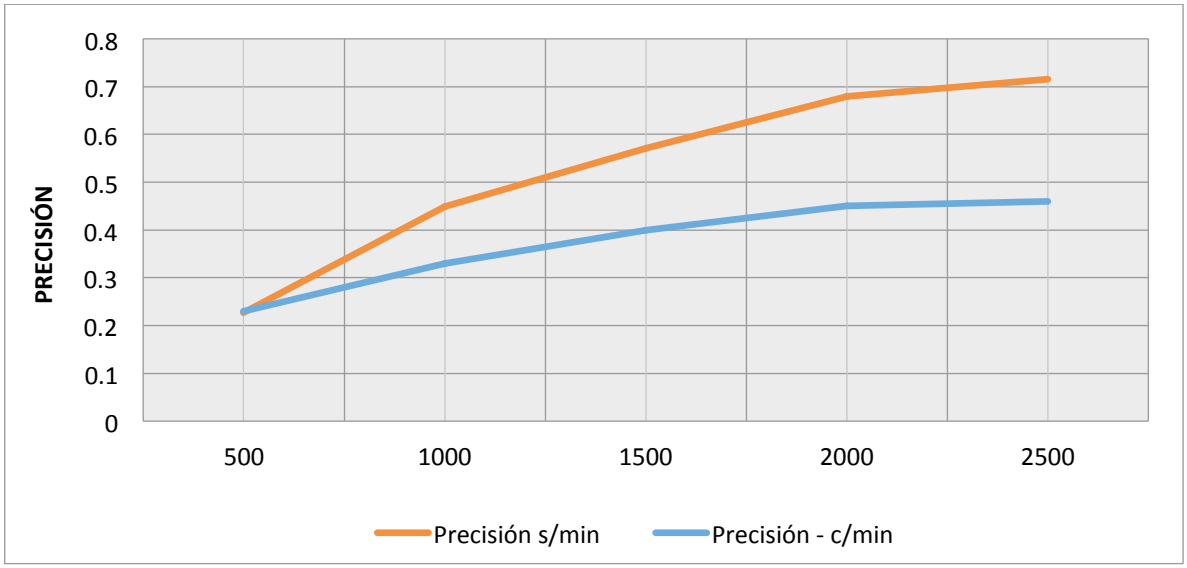

Figura 40: Comparación en precisión incluyendo el minuto en el vector de características.

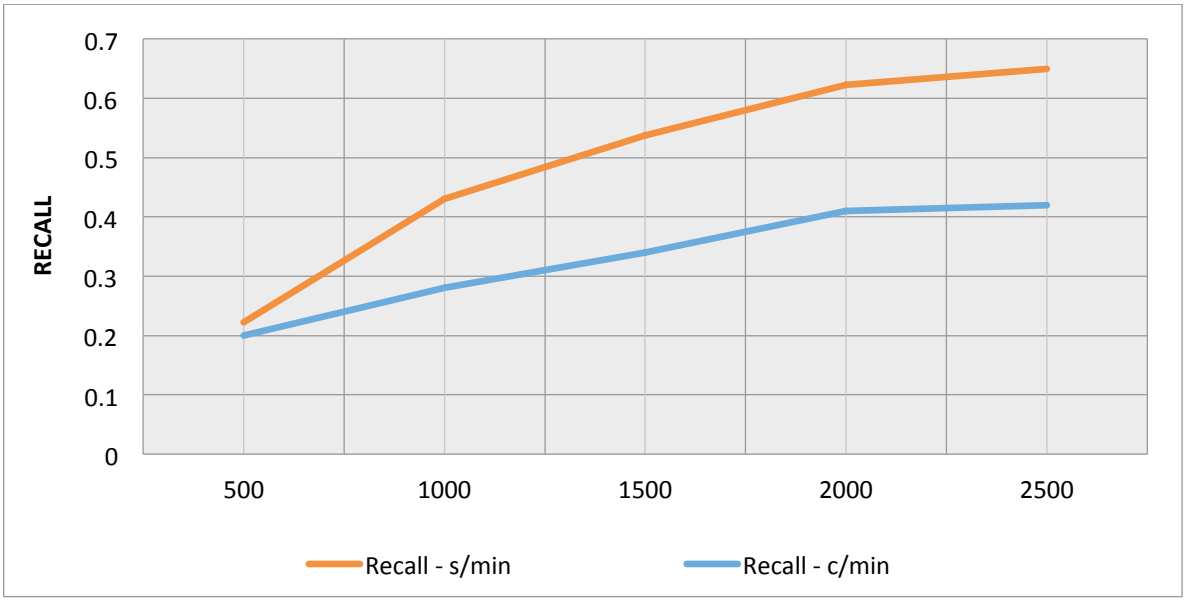

Figura 41: Comparación en *recall* incluyendo el minuto en el vector de características.

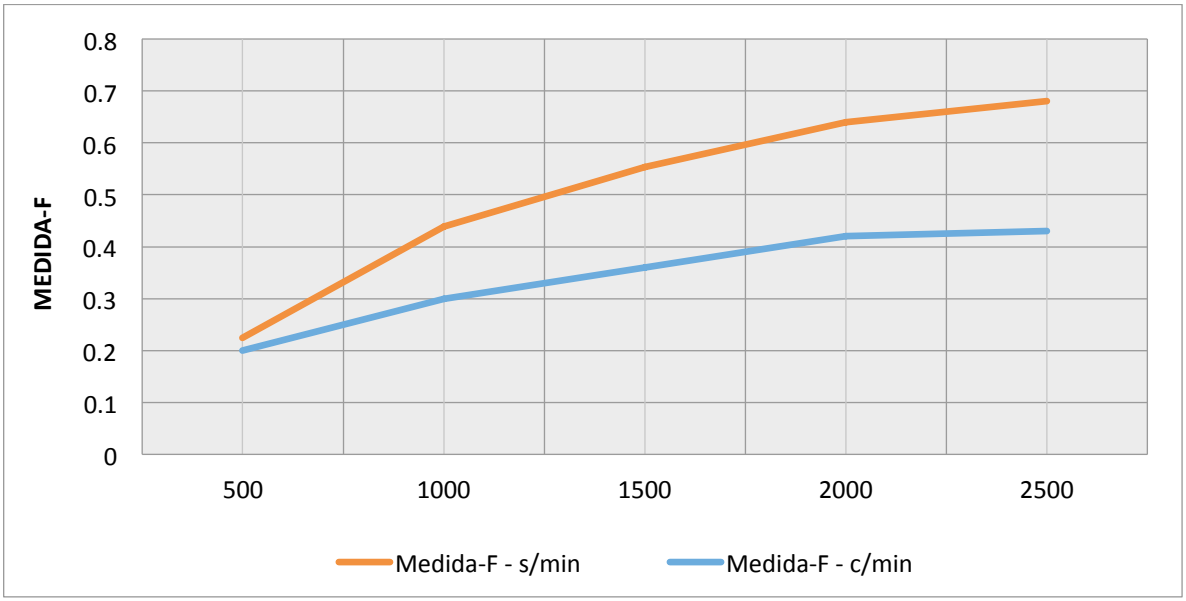

Figura 42: Comparación en medida-F incluyendo el minuto en el vector de características.

Se agregó al vector de características, el segundo cuando ocurrió el accidente y se comparó con los resultados obtenidos inicialmente, el vector quedó conformado con los siguientes parámetros:

- Mes del evento ocurrido.
- Día del mes evento ocurrido.

- Día de la semana del evento ocurrido.
- Hora del evento ocurrido.
- Minuto del evento ocurrido.
- Segundo del evento ocurrido.

Los resultados se muestran en las Figuras 43, 44 y 45.

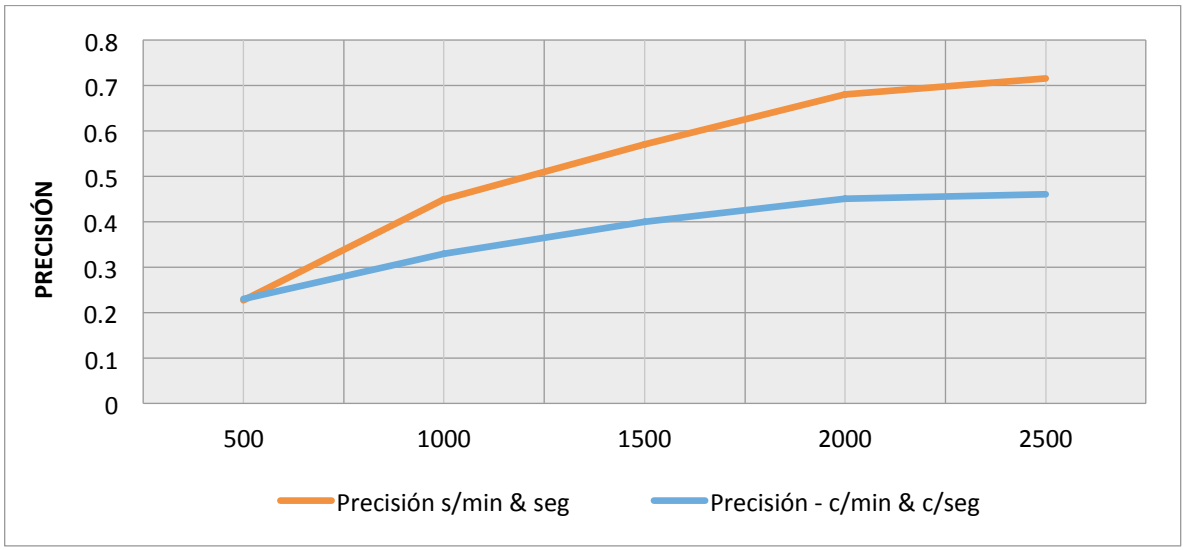

Figura 43: Comparación en precisión incluyendo minuto y segundo en el vector de características.

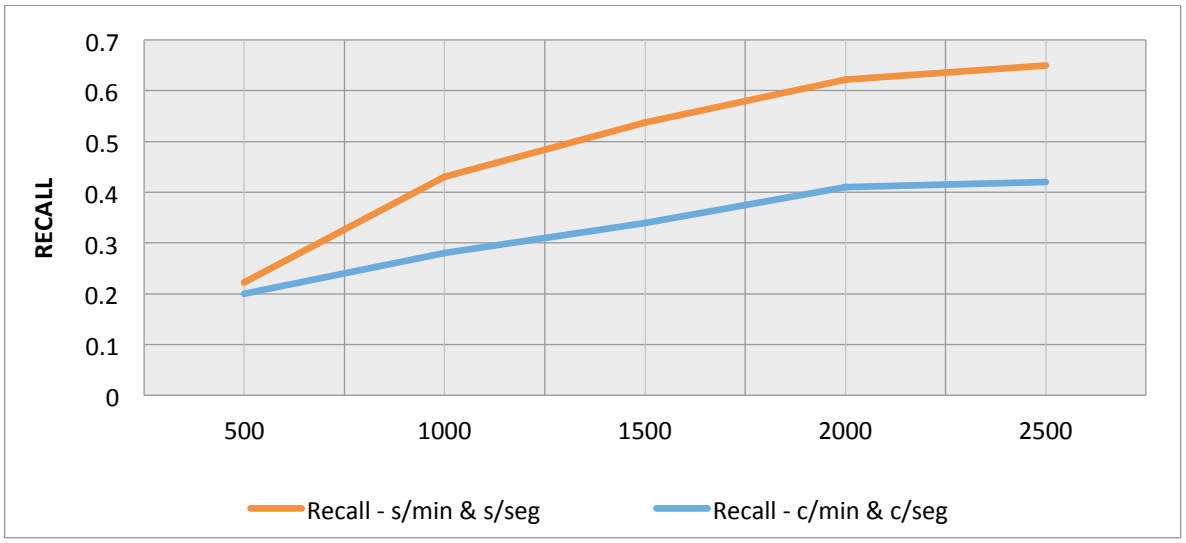

Figura 44: Comparación en *recall* incluyendo minuto y segundo en el vector de características.

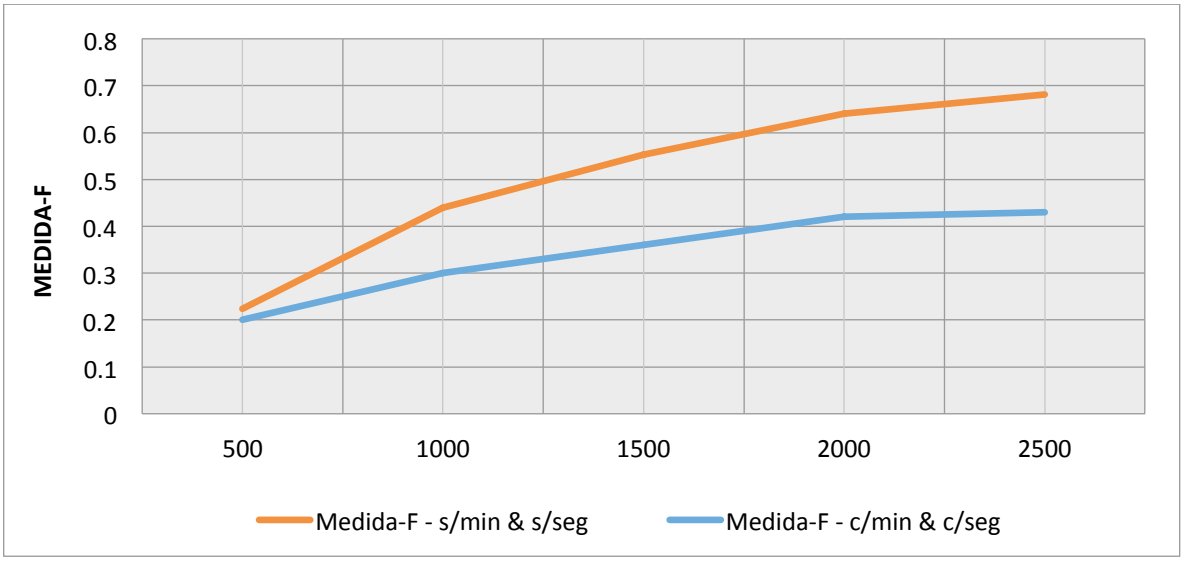

Figura 45: Comparación en Medida-F incluyendo minuto y segundo en el vector de características.

Los resultados no fueron lo esperado debido a que el minuto y segundo introducen ruido al método de aprendizaje automático SVR, ya que el conjunto de entrenamiento produce un efecto de dispersión geográfica, el cual SVR no puede modelar de forma correcta. Por esta razón fueron descartadas estas características en el conjunto de entrenamiento.

5.4.3. Representación de características

La representación de las características en los vectores fue un factor considerado cuando se generó el conjunto de entrenamiento. Existe otra forma de representar las características que forman parte del vector cuando ocurrió el evento. Esto es mediante la presencia y la ausencia del elemento, es decir, se reemplaza cada una de las características por un conjunto de tamaño *n*, donde *n*, es el número de valores posibles que puede tomar cada característica. Un ejemplo es la característica hora, la cual puede tomar valores de 0 a 23, por tanto, se agregan 24 características nuevas al vector con valores de 0 ó 1, se establece el valor de 1 a la hora cuando ocurrió el evento y se dejan las restantes en 0. Se realizó la prueba modificando el tipo de representación de los parámetros hora, día, día semana y mes y fue comparado con los resultados de la forma de representación inicial. Las Figuras 46, 47 y 48 muestran las comparaciones de *precision*, *recall* y *medida-f*.

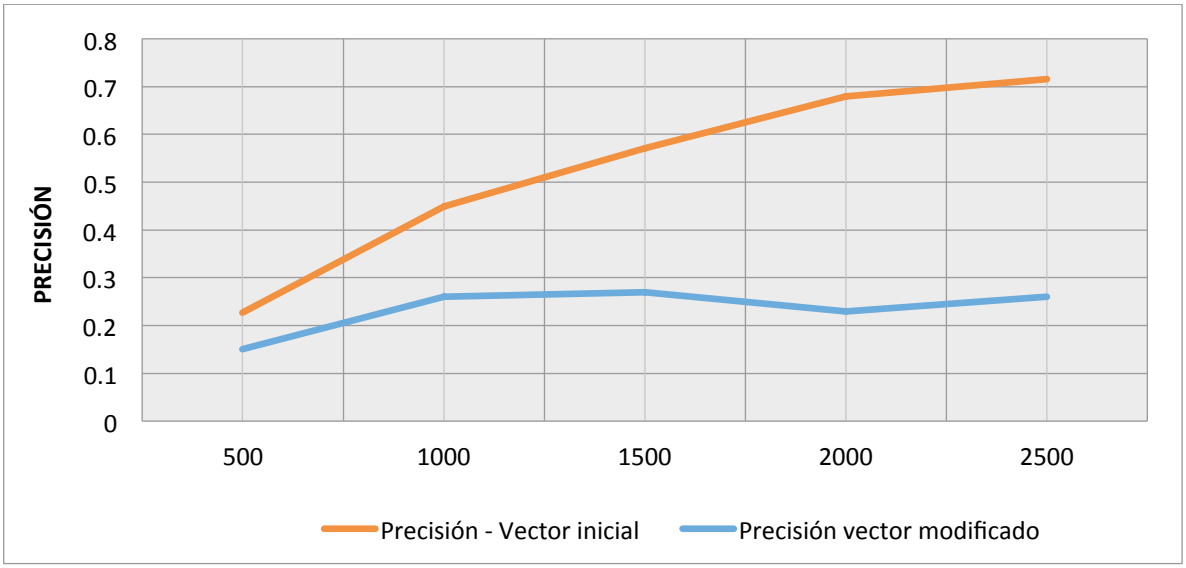

Figura 46: Comparación de precisión entre la representación de características alternativa e inicial.

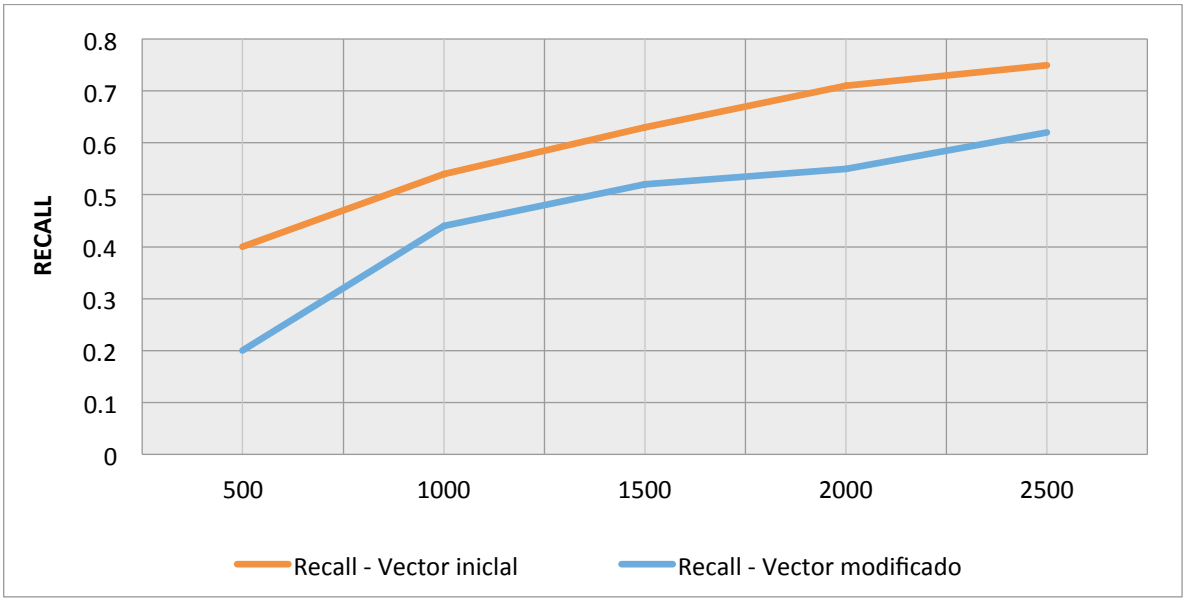

Figura 47: Comparación de *recall* entre la representación de características alternativa e inicial.

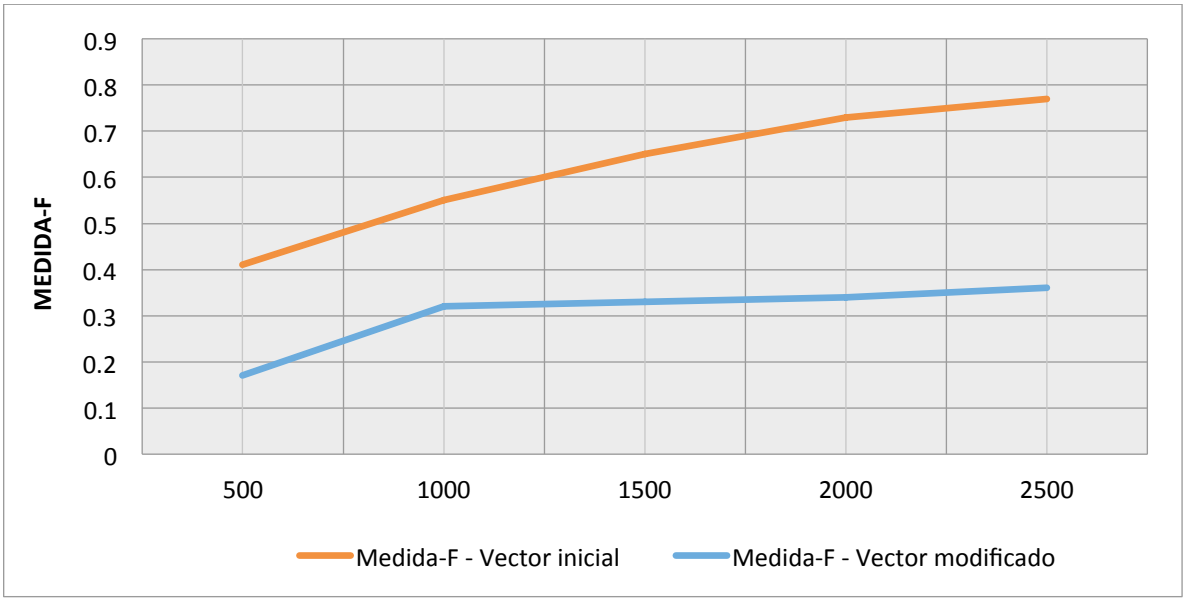

Figura 48: Comparación de medida-f entre la representación de características alternativa e inicial.

El cambio en la representación de los vectores no genera resultados con mayor *precision*, *recall* o *medida-f*, esto se debe a la forma de representación tanto de presencia y ausencia pierden la continuidad del tiempo y el método de aprendizaje automático no realiza distinciones entre la proximidad de lapsos de tiempo, es decir, no distingue la proximidad entre las 18 y 19 horas de las 10 y las 20 horas. Este es un aspecto fundamental en el comportamiento de la circulación vial, el cual debe ser modelado por SVR, por tanto, el conjunto de entrenamiento que sirve para generar los modelos de predicción más acertados se realiza con el conjunto de entrenamiento inicial.

5.4.4. Parámetros del modelo de predicción

Se realizaron cambios en los parámetros con los cuales se generó el modelo de predicción, utilizando el conjunto de entrenamiento inicial. Los valores con los que se genera actualmente el modelo de predicción son los siguientes:

- *Kernel*: rbf.
- C: 1000
- Gamma: 0.04
- *Epsilon/e-tube*: 1e-4

Los valores restantes tienen el valor por defecto asignado por la biblioteca y son mencionados en el Capítulo de Metodología. La selección de estos valores es de forma manual, los cuales fueron seleccionados mediante el entendimiento de qué significa cada valor y de los resultados obtenidos aplicando la validación cruzada. Los valores con mayor impacto en los resultados son los siguientes:

- El *kernel* en SVR se refiere al espacio donde se genera la función que pueda aproximar los valores futuros, reduciendo el riesgo en la regresión.
- C es el costo asignado por cada error cometido en la generación de la función.
- Gamma. Es un valor requerido por el *kernel rbf*, el cual es necesario para controlar el impacto que tiene una muestra con sus vecinos más cercanos. Un valor de gamma alto representa un impacto centrado que no afecta a las muestras vecinas y un valor de gamma pequeño muestra un impacto de amplio alcance entre las muestras.
- Épsilon es el radio del tubo de insensibilidad (ε-tube), donde la función no realiza ningún proceso de aprendizaje para los elementos que se encuentran dentro de este tubo.

Se realizaron diferentes pruebas para escoger los valores que obtienen los mejores resultados en *precision*, *recall* y *medida-f*. A continuación se muestran los valores que se insertaron y los resultados obtenidos con muestras de 500, 1000, 1500 y 2000 eventos para el conjunto de entrenamiento. Los parámetros que no son mostrados tienen los valores por defecto.

Prueba 1. Cambios de kernel

Utilizando kernel = lineal.

1000	Prueba	Inicial
Precisión	Sin solución lineal.	0.449
Exactitud	Sin solución lineal.	0.43
Medida-F	Sin solución lineal.	0.439

El *kernel lineal* no es el adecuado para generar la función de regresión, ya que si no existe una solución lineal, la generación del modelo de predicción queda ciclada.

Utilizando kernel = polinomial.

100	Prueba	Inicial
Precisión	Sin resultado	0.011
Exactitud	Sin resultado	0.011
Medida-F	Sin resultado	0.011

El *kernel polinomial* no soporta conjuntos de entrenamiento superiores a los 10 elementos. El costo computacional requerido para el *kernel polinomial* es demasiado grande, ya que realiza potencias en la generación de la función de regresión.

El *kernel rbf* tiene la característica de contemplar la distancia entre las muestras, es decir, las muestras son afectadas en diferente magnitud con relación a la distancia. Las muestras más cercanas a un evento inusual son las más afectadas y va decreciendo su impacto siguiendo una distribución gaussiana para las muestras más distantes.

Prueba 2. Reduciendo el costo de penalización C. Utilizando kernel = rbf, Gamma = 0.04 y Épsilon = 1e-4

500, C=10	Prueba	Inicial
Precisión	0.158	0.227
Exactitud	0.155	0.222
Medida-F	0.157	0.224

1000, C=10	Prueba	Inicial
Precisión	0.260	0.449

Exactitud	0.248	0.43
Medida-F	0.254	0.439

2000, C=10	**Prueba**	**Inicial**
Precisión	0.368	0.680
Exactitud	0.337	0.622
Medida-F	0.352	0.649

La reducción en la penalización por errores en la generación de la función de regresión permite que exista gran cantidad de errores y por tanto la función disminuya su precisión, exactitud y medida-f. Por el contrario, aumentar el costo de penalización genera un sistema intolerante a errores y genera sobre entrenamiento, afectando de igual forma la precisión, exactitud y medida-f. Además aumenta el tiempo en la generación de la función de regresión de forma considerable.

Prueba 3. Aumentando el ε-tube. Utilizando kernel = rbf, C = 1000 y Gamma = 0.04

500, ε = 1e-2	**Prueba**	**Inicial**
Precisión	0.038	0.227
Exactitud	0.037	0.222
Medida-F	0.038	0.224

1000, ε = 1e-3	**Prueba**	**Inicial**
Precisión	0.154	0.449
Exactitud	0.15	0.43

Medida-F	0.152	0.439

2000, ε = 1e-2	**Prueba**	**Inicial**
Precisión	0.178	0.680
Exactitud	0.165	0.622
Medida-F	0.171	0.649

El aumento de radio en el ε-tube genera una función plana que no realiza ningún tipo de aprendizaje, ya que la mayoría de las muestras quedan dentro del ε-tube, las cuales no contribuyen a la generación de la función de regresión, por tanto, se reduce su *precision*, *recall* y *medida-f*. Por el contrario, una disminución en el radio del ε-tube genera sobre entrenamiento, ya que la mayoría de las muestras contribuyen al aprendizaje de la función de regresión y para datos nuevos produce malos resultados afectando el *precision*, *recall* y *medida-f*.

6. Conclusiones

Este capítulo presenta las conclusiones, aportaciones, limitaciones y trabajo futuro para esta tesis.

6.1. Conclusiones

El trabajo de tesis *"Monitoreo urbano de entidades y eventos geográficos basado en monitoreo social",* fue desarrollado con el propósito de hacer predicción de eventos viales. Esta actividad se lleva a cabo utilizando métodos de aprendizaje automático supervisado, entrenados con *Crowdsourcing* e Información Geográfica Voluntaria (VGI). La elaboración de esta tesis se lleva a cabo mediante dos etapas: la geocodificación y predicción.

En geocodificación se propone una metodología para asignar coordenadas geográficas a información proveniente de Twitter. Se mejoró considerablemente la geocodificación enriqueciendo un *Gazetteer* con diccionarios auxiliares. Se realizó una representación más precisa del tipo de evento vial geocodificado y se comprobó que el número de elementos geográficos identificados en cada tweet tiene una relación con el tipo de evento. Además se comprobó que la participación en Twitter tiene una relación directa con las horas pico en la Ciudad de México.

En predicción se propone una forma de crear un conjunto de entrenamiento a partir de datos geocodificados con el fin de entrenar un método de aprendizaje automático supervisado para regresión. Se propuso una forma de llevar a cabo análisis espacio temporal de eventos viales, así como una forma de evaluar los resultados en un ámbito geográfico. Se comprobó que el número de eventos geocodificados utilizados como conjunto de entrenamiento está relacionado con la precisión del modelo de predicción generado. Entre mayor sea el conjunto de entrenamiento mayor es la precisión y la exactitud del modelo de predicción. Se comprobó que la selección de características es un aspecto fundamental para generar un modelo de predicción acertado. La representación de las características de igual forma, es fundamental para generar un modelo de predicción acertado, las características deben ser presentadas de forma categorizada con el fin de que no pierdan o agreguen sentido a la característica que se quiere modelar.

La selección de parámetros en la generación del modelo afecta significativamente la precisión, parámetros amplios generan falta de aprendizaje y parámetros justos generan sobre entrenamiento, estos valores pueden ser definidos a prueba y error utilizando validación cruzada.

Por lo tanto. se realizó un aporte a la migración de *Ciudades Inteligentes*, realizando una conexión entre una infraestructura vial, una infraestructura social y una infraestructura de tecnologías de la información.

6.2. Aportaciones

Las aportaciones de este trabajo de tesis se enlistan a continuación:

- Metodología para la geocodificación, clasificación y representación de tweets.
- Conjunto de datos *(Dataset)* de 100,000+ tweets geocodficados para su uso general en el Laboratorio PIIG.
- Artículo de geocodificación de tweets titulado "*Geocoding of traffic-related events from Twitter*" presentado en GEOINFO 2015, *XVI Brazilian Symposium on GeoInformatics*.
- Metodología para entrenar un algoritmo de aprendizaje automático supervisado con tweets geocodificados.
- Diseño de criterios de evaluación para predicciones geoespaciales en contexto de eventos viales.

6.3. Limitaciones

Tanto la etapa de geocodificación como la etapa de predicción tienen limitaciones. Ambos están restringidos por la naturaleza de su funcionamiento, a continuación se describen las principales limitaciones de cada etapa:

Para geocodificación las limitaciones parten por la constitución del *Gazetteer* y por la naturaleza de los tweets publicados. El enfoque de esta metodología no contempla el sentido en que el evento ha ocurrido debido a que el *Gazetteer* no contiene información relacionada al sentido de la calle; además, gran parte de los tweets no mencionan el sentido donde ha ocurrido el accidente.

Para predicción, la naturaleza del método de aprendizaje automático es poco preciso para considerar condiciones viales, debido a que una condición está compuesta por una secuencia de coordenadas, las cuales tendrían asignado un mismo vector de características temporales, lo cual afecta considerablemente la generación del modelo de regresión. Otro aspecto no considerado es el impacto posible de cada accidente sobre la calle, es decir, cuánto tiempo puede afectar una vialidad un evento, como línea de base (*baseline*) se asigna una duración de tiempo *t* a cada accidente cuando es pronosticado o visualizado. Otro aspecto es la cantidad de información para generar los modelos de predicción, delegaciones que contienen pocos eventos registrados producen malas predicciones.

6.4. Trabajo futuro

Como trabajo futuro se propone recopilar gran cantidad de eventos geocodificados para conformar conjuntos de entrenamiento que ayuden a generar mejores modelos de predicción e inclusive generar conjuntos de entrenamiento especializados a los requerimientos de tiempo y espacio.

Además, se propone desarrollar una metodología para realizar análisis temporal de condiciones viales y obtener predicciones más precisas en su forma de representación.

A su vez, s pretende desarrollar un enfoque para obtener el sentido donde ocurrió el evento vial, así como adaptar nuevas fuentes de información (por ejemplo Waze) para generar conjuntos de entrenamiento más grandes y en menor tiempo.

También se propone generar una metodología para establecer el tiempo en que un accidente puede afectar una vialidad y retroalimentar a la red social Twitter para validar los resultados obtenidos por los modelos de predicción.

En consecuencia, este trabajo es un punto de partida hacia el concepto de Ciudades Inteligentes, generando información relevante de eventos viales en la Ciudad de México de los cuales se puede realizar análisis de información, no solo para generar predicciones sino también puede ser utilizada en conexión con otros datos como transporte público, infraestructura vial, otras redes sociales, etc.

La conexión con otras fuentes de información puede ser utilizada con el fin de reafirmar los eventos o descartarlos, generando información altamente confiable, desarrollando una

conexión estrecha entre lo que acontece en la ciudad y los usuarios de las redes sociales o en general dentro de la web.

La conexión con fuentes gubernamentales puede hacer más eficientes sus actividades como administración de personal de tránsito, administración de horarios de transporte público, mantenimiento de infraestructura vial, detección de cruces o vialidades peligrosas, clasificación y planeación de vialidades, etc.

Referencias

[Abel F. et al., 2012]
Abel, F., Hauff, C., Houben, G. J., Stronkman, R., & Tao, K. (2012, June). Semantics+ filtering+ search= twitcident. exploring information in social web streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media* (pp. 285-294). ACM.

[Abu-Mostafa Y, 2012]
Abu-Mostafa Y. [caltech]. (2012, Mayo 18) *Lecture 14 – Support Vector Machines* [Archivo de video]. Obtenido de https://www.youtube.com/watch?v=eHsErlPJWUU Consultado en: 2014.

[Aggarwal C, 2011]
Aggarwal, C. C. (2011). *An introduction to social network data analytics* (pp. 1-15). Springer US.

[Albakour M, et al., 2013]
Albakour, M., Macdonald, C., & Ounis, I. (2013, May). Identifying local events by using microblogs as social sensors. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval* (pp. 173-180). LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

[Backstrom L. et al, 2010]
Backstrom, L., Sun, E., & Marlow, C. (2010, April). Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web* (pp. 61-70). ACM.

[Bartlett M. et al, 2004]
Bartlett, M. S., Littlewort, G., Lainscsek, C., Fasel, I., & Movellan, J. (2004, October). Machine learning methods for fully automatic recognition of facial expressions and facial actions. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on* (Vol. 1, pp. 592-597). IEEE.

[Becker H. et al, 2012]
Becker, H., Iter, D., Naaman, M., & Gravano, L. (2012, February). Identifying content for planned events across social media sites. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 533-542). ACM.

[Bennett K. et al, 2000]
Bennett, K. P., & Campbell, C. (2000). Support vector machines: hype or hallelujah? *ACM SIGKDD Explorations Newsletter*, *2*(2), 1-13.

[Bird, 2006]
Bird, S. (2006) "NLTK: the natural language toolkit." In: *Proceedings of the COLING/ACL on Interactive presentation sessions* (pp. 69-72). Association for Computational Linguistics.

[Brabham D, 2009]
Brabham, Daren C. "Crowdsourcing process2.jpg" Wikipedia. 2009. http://en.wikipedia.org/wiki/File:Crowdsourcing_process2.jpg (Feb 13, 2010)

[Chang, C. et al., 2011]
Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 27.

[Chourabi, H. et al., 2012]
Chourabi, H., Nam, T., Walker, S., Gil-Garcia, J. R., Mellouli, S., Nahon, K., ... & Scholl, H. J. (2012, January). Understanding smart cities: An integrative framework. In *System Science (HICSS), 2012 45th Hawaii International Conference on* (pp. 2289-2297). IEEE.

[Coleman D. et al, 2009]
Coleman, D. J., Georgiadou, Y., & Labonte, J. (2009). Volunteered geographic information: The nature and motivation of producers. *International Journal of Spatial Data Infrastructures Research*, *4*(1), 332-358.

[Cortes C. et al, 1995]
Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

[Davis Jr C. et al, 2011]
Davis Jr, C. A., Pappa, G. L., de Oliveira, D. R. R., & de L Arcanjo, F. (2011). Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, *15*(6), 735-751.

[De Longueville B. et al, 2009]
De Longueville, B., Smith, R. S., & Luraschi, G. (2009, November). Omg, from here, I can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 international workshop on location based social networks* (pp. 73-80). ACM.

[Ekman P. et al, 1978]
Ekman P. & Friesen W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA.

[Facebook, 2015]
https://www.facebook.com/ (Consultado el: 20/06/2015)

[Flatch P, 2012]
Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press.

[Flickr, 2014]
https://www.flickr.com/ (Consultado el: 20/06/2014)

[Freud Y. et al, 1995]
Freund, Y., & Schapire, R. E. (1995, January). A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory* (pp. 23-37). Springer Berlin Heidelberg.

[Goldberg D. et al, 2007]
Goldberg, D. W., Wilson, J. P., & Knoblock, C. A. (2007). From text to geographic coordinates: the current state of geocoding. *URISA journal*, *19*(1), 33-46.

[González R. et al, 2011]
González, R., Cuevas, R., Cuevas, A., & Guerrero, C. (2011). Where are my followers? Understanding the Locality Effect in Twitter. *arXiv preprint arXiv: 1105.3682*.

[Goodchild M, 2007]
Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211-221.

[Goodchild M. et al, 2012]
Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial statistics*, *1*, 110-120.

[Goodchild M. et al, 2013]
Goodchild, M., Pfoser, D., & Sui, D. (2013). GEOCROWD 2012 workshop report: the First Int'l Workshop on Crowdsourced and Volunteered Geographic Information 2012 (Redondo Beach, CA-Nov. 6, 2012). SIGSPATIAL Special, 5(1), 7-8.

[Google Maps, 2015]
https://www.google.com.mx/maps (Consultado el: 16/05/2015)

[Howe J, 2006]
Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, *14*(6), 1-4.

[Hughes A. et al, 2009]
Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, *6*(3), 248-260.

[James, F, 2012]
James, F. [WeoGeo Support]. (2012, Noviembre 7*). Hangouts with James Fee Volunteered, Crowdsourced and Awesome* [Archivo de video]. Obtenido de https://www.youtube.com/watch?v=e95BH7gEpbY Consultado en: 2014.

[Kaggle, 2014]
https://www.kaggle.com/competitions (Consultado el: 12/02/2015)

[Karabulut N, 2010]
Karabulut, N. (2010). Crowdsourcing: The Power of The Crowd. In *7th International Symposium of Interactive Media Design, http://newmedia. yeditepe. edu. tr/pdfs/isimd_10/nejla-karabulut. pdf*.

[Leadbeater C. et al, 2004]
Leadbeater, C., & Miller, P. (2004). *The Pro-Am revolution: How enthusiasts are changing our society and economy*. Demos.

[Lee R. et al, 2013]
Lee, R., Wakamiya, S., & Sumiya, K. (2013). Urban area characterization based on crowd behavioral lifelogs over Twitter. *Personal and Ubiquitous Computing*, *17*(4), 605-620.

[Matevelli G., et al, 2015]
Matevelli, G. V. ; Machado, N. G. ; Moro, M. M. ; DAVIS JUNIOR, C. A. . Taxonomia e Desafios de Recomendação para Coleta de Dados Geográficos por Cidadãos. In: XXX Simpósio Brasileiro de Bancos de Dados (SBBD), 2015, Petrópolis (RJ). Anais do XXX Simpósio Brasileiro de Bancos de Dados. Porto Alegre (RS): Sociedade Brasileira de Computação (SBC), 2015.

[Mathioudakis M. et al, 2010]
Mathioudakis, M., & Koudas, N. (2010, June). Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 1155-1158). ACM.

[Mendoza M., et al., 2010]
Mendoza, M., Poblete, B., & Castillo, C. (2010, July). Twitter Under Crisis: Can we trust what we RT? *Proceedings of the first workshop on social media analytics* (pp. 71-79). ACM.
[OpenStreetMap, 2015]
http://www.openstreetmap.org/ (Consultado el: 14/04/2015)

[Pollino M. et al, 2012]
Pollino, M., Fattoruso, G., La Porta, L., Della Rocca, A. B., & James, V. (2012). Collaborative open source geospatial tools and maps supporting the response planning to disastrous earthquake events. *Future Internet*, *4*(2), 451-468.

[Ribeiro Jr. T. et al, 2012]
Ribeiro Jr, S. S., Davis Jr, C. A., Oliveira, D. R. R., Meira Jr, W., Gonçalves, T. S., & Pappa, G. L. (2012, November). Traffic observatory: a system to detect and locate traffic events and conditions using Twitter. In *Proceedings of the 5th International Workshop on Location-Based Social Networks* (pp. 5-11). ACM.

[Savelyev A. et al, 2011]
Savelyev, A., Xu, S., Janowicz, K., Mülligann, C., Thatcher, J., & Luo, W. (2011, November). Volunteered geographic services: Developing a linked data driven location-based service. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Spatial Semantics and Ontologies* (pp. 25-31). ACM.

[Singer, M, 2008]
Singer, M. (2008) Web 2.0: Companies Will Spend $4.6 Billion By 2013, Forrester, InformationWeek, Abril 21, Consultado Septiembre 2013, http://www.informationweek.com/news/internet/web2.0/showArticle.jhtml?articleID=207400790

[Scikit-learn, 2015]
scikit-learn (2015), sckit-learn, Machine Learning in Python [scikit-learn.org] de http://scikit-learn.org/stable/index.html

[Sui D. et al, 2013]

Sui, D. Z., Elwood, S., & Goodchild, M. (2013). *Crowdsourcing geographic knowledge*. Springer.

[Tobler W, 1970]
Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, 234-240.

[Tweepy, 2015]
Tweepy (2015), Tweepy, An easy-to-use Python library for accesing the Twitter API [www.tweepy.org] de http://www.tweepy.org/

[Twitter, 2015]
https://twitter.com/ (Consultado el: 06/06/2015)

[Utani A. et al, 2011]
Utani, A., Mizumoto, T., & Okumura, T. (2011, December). How geeks responded to a catastrophic disaster of a high-tech country: rapid development of counter-disaster systems for the great east Japan earthquake of March 2011. In *Proceedings of the Special Workshop on Internet and Disasters* (p. 9). ACM.

[Wang Y. et al, 2013]
Wang, Y., & Cao, L. (2013). Discovering latent clusters from geotagged beach images. In *Advances in Multimedia Modeling* (pp. 133-142). Springer Berlin Heidelberg.

[Waze, 2015]
https://www.waze.com/es/ (Consultado el: 14/04/2015)

[Wikimapia, 2015]
Wikimapia CC-BY-SA ©. (2015), Wikimapia, Wikimapia [wikimapia.org] de https://wikimapia.org/

[Wikipedia, 2015]
http://en.wikipedia.org/wiki/Main_Page (Consultado el: 05/02/2015)

[Winston P, 2014]
Winston, P. [MIT OpenCourseWare]. (2014, Enero 10) *16. Learning: Support Vector Machines* [Archivo de video] Obtenido de https://www.youtube.com/watch?v=_PwhiWxHK8o Consultado en: 2014.

[Witten, I. et al., 2005]
Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[Wu C. et al, 2004]
Wu, C. H., Ho, J. M., & Lee, D. T. (2004). Travel-time prediction with support vector regression. *Intelligent Transportation Systems, IEEE Transactions on*, *5*(4), 276-281.

[Wu C. et al, 2008]
Wu, C. L., Chau, K. W., & Li, Y. S. (2008). River stage prediction based on a distributed support vector regression. *Journal of Hydrology*, *358*(1), 96-111.

[Xu L. et al, 2004]
Xu, L., Neufeld, J., Larson, B., & Schuurmans, D. (2004). Maximum margin clustering. In *Advances in neural information processing systems* (pp. 1537-1544).