Modelos de Predicción para la Estimación del Impacto Generado por un Evento Sísmico a través de Características Sociodemográficas y Económicas

Carlos Andrés Fontecha Ortiz, Sergei Felipe Beltrán Palencia

Trabajo de Grado para Optar el Título de Ingeniero Industrial

Director: Henry Lamos Díaz

PhD. Física-Matemática

Codirector: Daniel Orlando Martínez Quezada

Msc. Ingeniería Industrial

Universidad Industrial de Santander

Facultad de Ingenierías Físico-mecánicas

Escuela de Estudios Industriales y Empresariales

Ingeniería Industrial

Bucaramanga

2018

Tabla de Contenido

Lista de figuras

Lista de tablas

Lista de apéndices

(Ver apéndices adjuntos en el CD)

[1] SAP. Se refiere a los modelos sin ajuste de parámetros, es decir, con los parámetros por defecto.

[2] CAP. Se refiere a los modelos con parámetros óptimos hallados previamente.

RESUMEN

TITULO: MODELOS DE PREDICCIÓN PARA LA ESTIMACIÓN DEL IMPACTO GENERADO POR UN EVENTO SÍSMICO A TRAVÉS DE CARACTERÍSTICAS SOCIODEMOGRÁFICAS Y ECONÓMICAS.[*]

AUTORES: FONTECHA ORTIZ, Carlos Andrés[**]

BELTRÁN PALENCIA, Sergei Felipe[**]

PALABRAS CLAVE: Regresión, Máquinas de soporte vectorial, Random Forest, Predicción, Desastres, Sismos.

DESCRIPCION:

El propósito principal de este proyecto es determinar cuál de los modelos estadísticos propuestos (Regresión Lineal, Random Forest y Maquinas de Soporte Vectorial) presenta un mejor comportamiento frente a la predicción de los efectos de un evento sísmico.
A través del aprendizaje automático y diferentes modelos matemáticos de regresión se busca predecir los posibles efectos de este tipo de desastres como muertos, heridos y daños, teniendo en cuenta características socio-demográficas y económicas.
En los últimos 10 años han ocurrido más de 400 desastres naturales, por ejemplo, los terremotos de Haití, Chile, México y otros alrededor del mundo, debido a esto, ha crecido el interés por entender el comportamiento de los desastres naturales y desarrollar proyectos que reduzcan los efectos que estos pueden tener sobre la población.
Los diferentes entes educativos, gubernamentales y ONG's, a nivel mundial están trabajado en conjunto y usando técnicas de aprendizaje automático y logística humanitaria con el fin de estimar los efectos de eventos futuros y diseñar planes de acción que los mitiguen. Esto permite optimizar la gestión de la cadena de suministros y agilizar el apoyo a las poblaciones más vulnerables, el socorro de heridos y evitar la propagación de enfermedades.
Para el cumplimiento de los objetivos, este documento constara de un marco teórico referente al tema a desarrollar, revisión de antecedentes y el desarrollo de los algoritmos para cada uno de los modelos; se documentarán también la adecuación de las diferentes bases de datos con las que este se desarrollara, así como los resultados computacionales obtenidos de la aplicación de dichos modelos.

---

[*] Trabajo de grado

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director: Henry Lamos Díaz, PhD. Física-Matemática. Codirector: Daniel Orlando Martínez Quesada, Msc. Ingeniería Industrial.

ABSTRACT

TITLE: PREDICTION MODELS FOR ESTIMATION OF IMPACT GENERATED BY A SEISMIC EVENT THROUGH SOCIODEMOGRAPHIC AND ECONOMIC CHARACTERISTICS.[*]

AUTHORS: FONTECHA ORTIZ, Carlos Andrés[**]

BELTRÁN PALENCIA, Sergei Felipe[**]

KEYWORDS: Regression, Support vector machine, Random Forest, Prediction, Natural hazards, Natural disaster, Earthquakes, Seism.

DESCRIPTION:

The main purpose of this project is to determine which proposed statistical model (Linear Regression, Random Forest and Support Vector Machines) presents the best behavior at the prediction of the effects of a seismic event.
Through machine learning and different mathematical models of regression, the aim is to predict the possible effects of this kind of events, such as deaths, injuries and damages, considering socio-demographic and economic characteristics.
In the last 10 years there have been more than 400 natural hazards, for example, the earthquakes in Haiti, Chile, Mexico and other places in the world, due to this, the interest in understanding the behavior of these and the development of projects to reduce the effects they have over the population has increased.
The different educational entities, governments and NGOs, around the world are working together and using techniques of machine learning and humanitarian logistics in order to estimate the effects of future events and establishing action plans that mitigate them. This makes it possible to improve the supply chain management and speed up support to the most vulnerable populations, the relief of injured people and prevent the spread of diseases.
For the fulfillment of the objectives, this document will contain a theoretical framework related to the topic to develop, a review of antecedents and the development of the algorithms for each model; the adequacy of the different databases to use will be documented, as well as the computational results obtained from the application of said models.

---

[*] Bachelor Thesis

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director: Henry Lamos Díaz, PhD. Física-Matemática. Codirector: Daniel Orlando Martínez Quesada, Msc. Ingeniería Industrial.

Introducción

En los últimos 10 años han ocurrido más de 400 desastres naturales, como los terremotos de Haití, Chile, México, las inundaciones en Colombia, Filipinas, Tailandia, deslizamientos de tierra en Guatemala, Bolivia, Brasil; entre otros. Lo cual deja ver que, incluso en países del primer mundo, los esfuerzos por mitigarlos, evitarlos y/o intentar controlarlos aún son insuficientes.

Debido a esto, alrededor del mundo está creciendo de manera exponencial el interés por entender cómo se comporta la naturaleza y qué factores, características, condiciones o variables hacen que en el momento menos determinado ocurra un desastre. Los diferentes entes educativos, gubernamentales y ONG's, están en la búsqueda de comprender estos eventos a través del uso del 'machine learning' o aprendizaje automático. Esta es una rama de la inteligencia artificial (IA) y una de las áreas de investigación más activas actualmente, el cual implica el estudio y desarrollo de modelos computacionales de procesos de aprendizaje, cuyo objetivo principal de la investigación en este campo es construir máquinas capaces de mejorar su rendimiento con la práctica y de adquirir conocimiento por sí mismos. Michalski, Carbonell y Mitchell (2013).

Estimar los daños potenciales, puede ayudar a planear las actividades de la logística humanitaria frente a un desastre, ya que ésta soportada en dichas técnicas, busca cumplir su misión de apoyar a las poblaciones más vulnerables, socorrer a los heridos, evitar propagación de enfermedades, apoyándose en la gestión de la cadena de suministros, una herramienta que en las últimas décadas se ha aplicado a la gestión de desastres para maximizar los resultados positivos, reduciendo los costos y optimizando la utilización de los recursos.

Para apoyar dicha gestión, la presente investigación tiene como finalidad, desarrollar un modelo que permita predecir, con una alta confiabilidad, los efectos que podrían resultar al presentarse un terremoto en cualquier parte del mundo, basándose en modelo de aprendizaje supervisado, usando

características sociodemográficas como la densidad poblacional, el índice de desarrollo humano, entre otras.

El siguiente trabajo se dividirá de la siguiente manera: los capítulos uno, dos y tres describirán la esencia del proyecto, donde se plantea la problemática, justificación y objetivos respectivamente. En el capítulo cuatro se encuentra el marco teórico, que especifica y describe las técnicas y métodos con los que se pretende desarrollar la investigación; el capítulo cinco contiene el estado del arte, donde se realiza la revisión de diferentes publicaciones relacionadas con el tema central del proyecto, empleándolas como base de apoyo para este trabajo; el capítulo seis muestra cómo se construyen los algoritmos para los diferentes modelos de regresión junto con la compactación de la base de datos a utilizar; los capítulos siete y ocho recopilan la validación de los modelos y los resultados después de su entrenamiento; por último, en el capítulo nueve se muestran las conclusiones del proyecto de grado.

Inicialmente se realiza la obtención de los datos recopilando información de distintas fuentes, y haciendo uso de técnicas de imputación de datos, para mejorar la calidad de estos antes de realizar el análisis de correlación, lo cual nos permitirá seleccionar aquellas variables que positivamente ayuden al desarrollo del modelo de predicción. Las Máquinas de Soporte Vectorial y Random Forest son las técnicas propuestas para realizar el modelo, después de haber realizado el tratamiento de datos y se compararán el error (RMSE) de cada uno con los resultados del modelo corrido por regresión lineal, esto con el fin de determinar cuál de los tres es más efectivo y eficiente al momento de hacer una posible predicción de los variables respuesta propuestas.

Tabla 1.

Cumplimiento de Objetivos

Cumplimiento de Objetivos

Objetivo	Cumplimiento
Revisar literatura relacionada con los modelos de aprendizaje automático utilizados en la estimación del impacto generado por eventos sísmicos.	Este objetivo se da por cumplido en el numeral 5 que consiste en la revisión de literatura.
Revisar los fundamentos teóricos de aplicación de modelos de máquinas de soporte vectorial y Random Forest en problemas de regresión.	En el numeral 4 se encuentra el marco teorico que costa de la teoria de los diferentes modelos y otras herramientas a usar en el ejercicio.
Construir la base de trabajo de registros históricos de desastres generados por un evento sísmico relacionando características sociodemográficas mediante la consolidación de bases de datos de la web.	El capitulo 6 llamado tratameinto de datos, describe como se realiza la cosolidacion de las bases de datos utilizadas en el estudio.
Evaluar los modelos de predicción máquinas de soporte vectorial y algoritmo Random Forest, mediante el uso de software especializado.	En el capitulo 7 se describe el proceso de diseño de los codigos de los diferentes modelos y se evaluan en el capitulo 8.

Cumplimiento de Objetivos

Objetivo	Cumplimiento
Analizar los resultados obtenidos al aplicar el modelo seleccionado.	Se mencionan los resultados computacionales de los diferentes modelos en el capitulo 8 y en las conclusiones.
Elaborar un artículo académico de carácter publicable sobre los resultados del proyecto de investigación.	El articulo en cuestion se realiza a partir del documento del proyecto y se entrega en un documento por separado a la escuela de estudios industriales y empresariales.

Nota. Contiene los objetivos planteados para el proyecto y el capítulo donde se les da complimiento.

1. Justificación

Los sismos y terremotos han sido objetos de estudio a lo largo de la historia y en las últimas décadas ha crecido el interés por conocer con antelación dónde y cuándo sucederían y qué efectos podría tener en las regiones y poblaciones, tanto rurales como urbanas, estas últimas ocupadas por la mitad de la población mundial, siendo únicamente usada el 3% del área terrestre (Kondratyev, Krapivin, & Varostos, 2006). Lo cual intensifica el riesgo para los habitantes y las zonas donde existan asentamientos e invasiones ilegales de personas en situación de pobreza.

De acuerdo a las distintas fuentes de información, se encuentra que los impactos de un desastre natural son medidos principalmente por la pérdida de vidas humanas, cantidad de heridos y daños económicos, esta última representada en millones de dólares; debido a daños en infraestructura vial, casas y edificaciones destruidas y/o afectadas y cultivos afectados. Estas variables dependen de múltiples factores, por ejemplo: características del evento sísmico, profundidad, magnitud e

intensidad, densidad poblacional urbana y rural, índice de desarrollo humano (IDH), índice de Gini, entre otras.

En el presente trabajo se pretende seleccionar aquellos factores que tienen mayor incidencia en la predicción de los resultados de las variables mencionadas anteriormente, por medio de técnicas de imputación de datos, análisis de correlación entre otros, con el fin de mejorar los resultados en cada uno de los modelos planteados.

2. Planteamiento del problema

Los terremotos ocurren con frecuencia variable en diferentes partes del mundo a causa de las placas tectónicas que sostienen la superficie de la tierra y que se mueven con velocidades de hasta 12 centímetros por año, Mattioli y Jansma (2017). El estudio de estos, por parte de la comunidad científica ha incrementado en las últimas décadas, buscando determinar y/o predecir la ocurrencia de estos para evitar una catástrofe, ya que según el Centro para la Investigación sobre la Epidemiologia de los Desastres CRED en el periodo comprendido entre los años 2006 y 2016, estos desastres causaron la muerte de 358.428 personas y daños por un valor aproximado de 462 billones de dólares.

Estos desastres no solamente tendrían consecuencias a nivel poblacional y de infraestructura, sino también en la economía del país y/o la región en donde se presente. Como se muestra en Olsen, Carstensen y Hoyen (2003), el impacto de un desastre en una región, en el caso de no gestionarse adecuadamente, puede producir inestabilidad política y social, afectar a la seguridad y las relaciones internacionales (Rodríguez, Vitoriano, Montero y Kecman, 2011)

Para los países en desarrollo, es un factor clave reducir el gasto cuando se presenta algún desastre natural o social, lo cual se podría lograr mediante la implementación de políticas públicas

y una gestión del riesgo eficiente ya que, según el Banco Mundial, las repercusiones económicas, sociales y ambientales en estos países podría retrasar proyectos para continuar con su desarrollo.

Colombia, a pesar de ser pionera en América Latina en cuanto al tratamiento de los riesgos y desastres, aún se enfrenta a los daños a la propiedad, infraestructura y los medios de subsistencia. Campos et. al. (2012). Es por esto que ciudades como Bucaramanga, que hace parte del 86% de la población que se localiza en zonas de amenaza sísmica alta y media, debe contar con apoyo en gestión de desastres y logística humanitaria para mitigar el riesgo latente que presenta; esto a través de métodos y herramientas que sean eficientes pero cuya implementación sea barata y no afecta la economía de la región.

De acuerdo a lo anterior se evidencia la importancia de apoyar la gestión de desastres por medio de nuevas tecnologías, como la inteligencia artificial, que mediante máquinas de aprendizaje automático y modelos estadísticos, use datos para analizarlos y aprender de ellos, y así generar escenarios hipotéticos o de predicción que den a las ONG’s y gobiernos información sobre los posibles efectos en la población y región donde ocurra un terremoto, para así determinar la cantidad de ayuda humanitaria que se requeriría en la atención del mismo.

3. Objetivos

3.1. Objetivo general

Desarrollar modelos de predicción para la estimación del impacto generado por un evento sísmico, a través de características sociodemográficas y económicas.

3.2. Objetivos específicos

- Revisar literatura relacionada con los modelos de aprendizaje automático utilizados en la estimación del impacto generado por eventos sísmicos.
- Revisar los fundamentos teóricos de aplicación de modelos de máquinas de soporte vectorial y Random Forest en problemas de regresión.
- Construir la base de trabajo de registros históricos de desastres generados por un evento sísmico relacionando características sociodemográficas mediante la consolidación de bases de datos de la web.
- Evaluar los modelos de predicción máquinas de soporte vectorial y algoritmo Random Forest, mediante el uso de software especializado.
- Analizar los resultados obtenidos al aplicar el modelo seleccionado.
- Elaborar un artículo académico de carácter publicable sobre los resultados del proyecto de investigación.

4. Marco teórico

4.1. Logística humanitaria

En los últimos años, la logística humanitaria ha venido teniendo gran importancia en las sociedades, de esta forma, organizaciones tanto gubernamentales como ONG e investigadores han

desarrollado estudios en materia de logística para atención de desastres, con el fin de elaborar herramientas para el mejoramiento de la misma y así aliviar en parte el sufrimiento de la población.

Thomas (Como se citó en Apte, 2010 y Safeer, Anbuudayasankar, Balkumar y Ganesh, 2014) define la logística humanitaria como "el proceso de planificar, implementar y controlar el flujo eficiente y rentable y el almacenamiento de bienes y servicios, así como la información relacionada desde el punto de origen hasta el punto de consumo con el propósito de aliviar el sufrimiento de personas vulnerables".

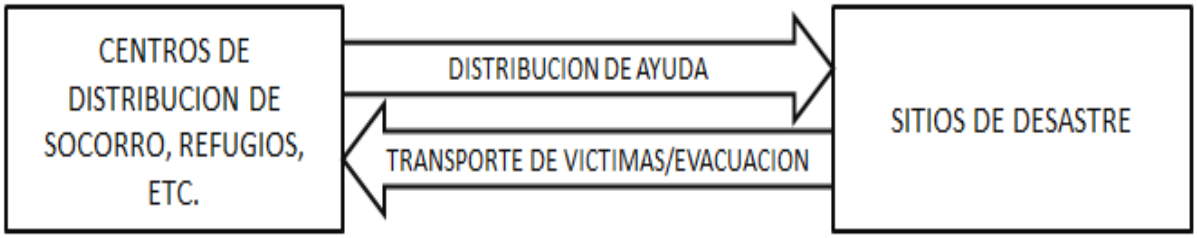


Figura 1. Operaciones logísticas humanitarias primarias de emergencia. Adaptado de (Safeer et al, 2014, p.2249).

Apte (2010) define logística humanitaria como una rama especial de la logística que gestiona la cadena de suministro de respuesta de suministros y servicios críticos con desafíos tales como aumentos de la demanda, suministros inciertos, ventanas de tiempo críticas frente a las vulnerabilidades de la infraestructura y gran alcance y tamaño de las operaciones (p.3).

4.2. Gestión de desastres

4.2.1. Desastre. El concepto de desastre tiene diferentes interpretaciones dependiendo el campo en que este se lleve a colación y del autor que lo interpret**e.**

La Organización de las Naciones Unidas (ONU) define un desastre como una perturbación grave del funcionamiento de una comunidad o de una sociedad. Los desastres involucran impactos humanos, materiales, económicos o ambientales generalizados, que exceden la capacidad de la

comunidad o sociedad afectada de hacer frente a sus propios recursos (World Confederation for Physical Therapy [WCPT], 2016).

Cortés (Como citó Villalibre, 2013), plantea que un desastre está ligado a los conceptos de riesgo, amenaza y vulnerabilidad. Según este autor un desastre es:

"Una situación extraordinaria causada por un fenómeno de origen natural, socio-natural o antrópico (la amenaza expresada en un evento real), que significa alteraciones intensas en las personas, los bienes, los servicios y el medio ambiente, excediendo la capacidad de respuesta. Es el resultado de un riesgo no manejado, y como tal entra a la ecuación añadiendo una flecha entre la R (riesgo) y una D de desastre: A (amenaza) x V (vulnerabilidad) = R (riesgo) – D (desastres)" (p.10).

Un aspecto común en las definiciones de desastre es que se centran más en los efectos sociales que en las características físicas de los hechos como tal. Los autores lo catalogan desastre cuando el impacto de la tragedia es sufrido por un número elevado de personas. Por ejemplo, un accidente aéreo con trescientas personas es determinado como desastre mientras que un accidente de un aeroplano con una sola persona involucrada no podría ser catalogado como tal (Villalibre, 2013).

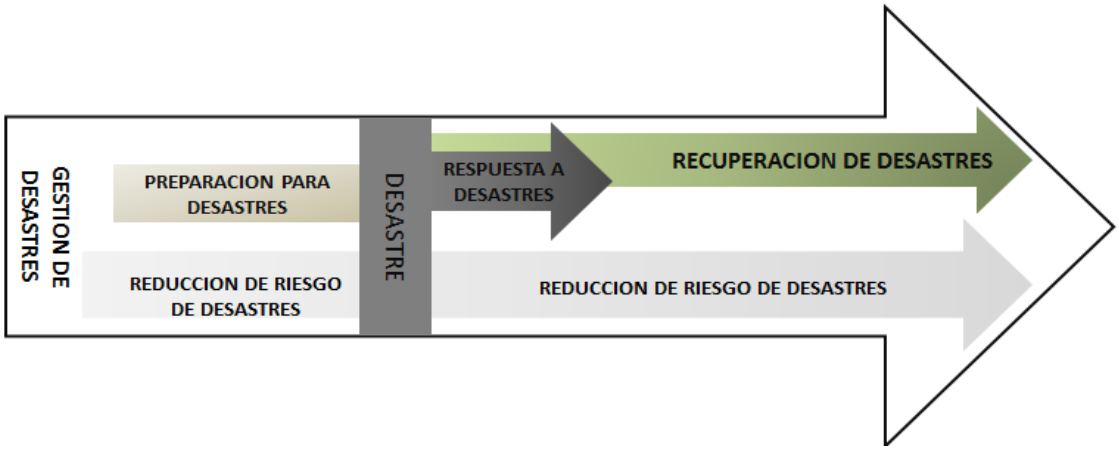


Figura 2. Etapas gestión de desastres. Adaptado de (Federación de la Cruz Roja Internacional **[**IFRC], s.f.)

Teniendo en cuenta la interpretación de desastre en sí y relacionándola con eventos naturales como lo son sismos y/o terremotos, se podría decir que es el impacto o efectos de un evento sísmico en una población vulnerable debido a la presencia de un riesgo, causando muertos, heridos y daños materiales a dicha comunidad.

4.2.2. Gestión de desastres. Se define la gestión de desastres como la organización y la gestión de recursos y responsabilidades para abordar todos los aspectos humanitarios de las emergencias, en particular la preparación, la respuesta y la recuperación a los desastres, a fin de reducir sus efectos (IFRC, s.f., parr.2).

"Las organizaciones locales, regionales, nacionales e internacionales participan en el desarrollo de una respuesta humanitaria a los desastres. Cada uno tendrá un plan de manejo de desastres preparado. Estos planes cubren la prevención, preparación, socorro y recuperación" (WCPT, 2016, parr.1).

4.2.3. Prevención de desastres. La prevención de desastres consiste en el desarrollo de actividades o la toma de medidas a fin de brindar protección contra los mismos. No todos los desastres se pueden evitar, en especial los desastres naturales, sin embargo, estas medidas pueden ayudar a mitigar la cantidad de víctimas y lesiones a presentarse. Estas medidas consisten en planes de evacuación, planificación ambiental y estándares de diseño. En 2005 se adoptó un registro denominado Marco de Hyogo por parte de 168 países, el cual contenía principios rectores, prioridades de acción y medios prácticos para lograr una mayor capacidad de superación ante las catástrofes en las poblaciones vulnerables (WCPT, 2016).

4.2.3.1. Preparación para desastres. Consiste en las actividades diseñadas para minimizar la pérdida de vidas y daños, actividades como lo son la reubicación de personas y bienes de una ubicación amenazada y facilitando el rescate oportuno y eficaz, el socorro y la rehabilitación. La preparación es la principal forma de reducir el impacto de los desastres. La preparación y la gestión comunitaria deben ser una prioridad en la gestión de la práctica física (WCPT, 2016).

4.2.3.2. Alivio de desastres. El alivio de desastres viene después de la ocurrencia del evento y se trata de dar una respuesta coordinada entre múltiples organismos con el fin de reducir el impacto de un desastre y sus resultados a corto y largo plazo. Las actividades de ayuda involucradas en esta etapa incluyen el rescate, la reubicación, la provisión de alimentos y agua, la prevención de enfermedades y discapacidades, la reparación de servicios vitales como telecomunicaciones y transporte, la provisión de refugio temporal y atención médica de emergencia. Es la atención que se brinda a las víctimas del desastre inmediatamente después del evento con el fin de no tener más víctimas mortales (WCPT, 2016).

4.2.3.3. Recuperación de desastres. Esta etapa se desarrolla una vez las necesidades de emergencia se han cumplido y la crisis inicial ha terminado, tras esto, las personas afectadas y las comunidades que las apoyan siguen siendo vulnerables. Las actividades de recuperación incluyen la reconstrucción de la infraestructura, el cuidado de la salud y la rehabilitación. Estos deben combinarse con actividades de desarrollo, como la creación de recursos humanos para la salud y el desarrollo de políticas y prácticas para evitar situaciones similares en el futuro. La gestión de los desastres está vinculada al desarrollo sostenible, en particular en relación con las poblaciones vulnerables (WCPT, 2016).

4.3. Sismos y terremotos.

Okulewicz (2017) define los terremotos son el movimiento rápido y las vibraciones causadas por el movimiento del suelo a lo largo de una fractura en una roca o a lo largo de una falla. El movimiento ocurre cuando las rocas no pueden almacenar más estrés, momento en el que alcanzan su punto de ruptura, liberan energía y crean un terremoto. El punto de origen de un terremoto por debajo de la superficie donde se libera su energía se conoce como el foco. El foco puede estar ubicado en cualquier superficie o profundidad. El punto en la superficie de la tierra directamente encima del foco se llama el epicentro (p.1).

Medir de manera cuantitativa los terremotos ha sido una tarea relevante para los científicos de esta área desde hace ya muchos años, se cuenta con diferentes métodos de medición de terremotos que se pueden realizar sobre dos características que poseen: magnitud e intensidad.

La magnitud es una clasificación numérica del tamaño o fuerza de un terremoto, basado en las lecturas del instrumento de medición de la misma, mientras que la intensidad es un tipo diferente de clasificación numérica que tiene que ver con los efectos reales de un terremoto en las personas, los edificios y el paisaje (Ansfield, 2017).

La escala de Mercalli es la escala que se ha convertido en la medida de intensidad estándar establecida en 1884, la cual, cuenta con una escala de doce valores. Los valores van desde I, que apenas se sentiría, hasta XII, que sería el más violento. La magnitud del terremoto se informa con un solo número y no varía de acuerdo a la distancia del epicentro y profundidad del mismo, este número se determina a partir de un registro instrumental de vibraciones terrestres. Cada terremoto lo suficientemente fuerte es detectado y registrado por un sismógrafo, asignándole un valor de magnitud de acuerdo a la escala utilizada en este caso que es la escala Richter. La magnitud en la escala de Richter de un terremoto se determina midiendo el mayor desplazamiento horizontal

desde el promedio del trazado de la pluma de registro en el sismograma de un instrumento estándar a una distancia estándar del epicentro del terremoto (Ansfield, 2017).

4.4. Minería de datos

La minería de datos (Data Mining en inglés) es el estudio de la recolección, limpieza, procesamiento, análisis y obtención de información útil a partir de datos. Existe una amplia variación en términos de los dominios de problemas, aplicaciones, formulaciones y representaciones de datos que se encuentran en aplicaciones reales. Por lo tanto, "minería de datos" es un amplio término paraguas que se utiliza para describir estos diferentes aspectos del procesamiento de datos (Aggarwal, 2015).

La minería de datos tiene diferentes etapas antes de la interpretación de datos como tal Aggarwal (2015) la describe como una tubería que contiene diferentes fases como: recolección y limpieza de datos, extracción de características y diseño algorítmico.

4.4.1. Recolección de datos. La recolección de datos puede requerir el uso de hardware especializado como una red de sensores, trabajo manual como la recopilación de encuestas de usuarios o herramientas de software como un motor de rastreo de documentos Web para recopilar documentos. También se requiere la buena elección de datos correctos debido a que errores pueden afectar significativamente el proceso (Aggarwal, 2015).

4.4.2. Extracción de características y limpieza de datos. Cuando se recopilan los datos, estos deben tratarse para que sean adecuados para su procesamiento. Por ejemplo, los datos pueden codificarse en registros complejos o en documentos de forma libre. En muchos casos, los diferentes tipos de datos pueden mezclarse arbitrariamente en un documento de forma libre.

Para que los datos sean adecuados para el procesamiento, es esencial transformarlos en un formato que sea amigable con los algoritmos de minería de datos, como multidimensional, series de tiempo o formato semiestructurado. En múltiples casos los datos son extraídos de diferentes fuentes y se deben estandarizar y transformarse a un conjunto de datos estructurado (Aggarwal, 2015).

4.4.3. Procesamiento analítico y algoritmos. La parte final del proceso de minería es diseñar métodos analíticos eficaces a partir de los datos procesados. En muchos casos, puede que no sea posible utilizar directamente un problema estándar de minería de datos, como los cuatro "superproblemas" (agrupación, clasificación, asociación de patrones de minería y detección de valores atípicos), para la aplicación en cuestión. Sin embargo, estos cuatro problemas tienen una cobertura tan amplia que muchas aplicaciones se pueden dividir en componentes que utilizan estos bloques de construcción diferentes.

El proceso general de minería de datos se ilustra en la figura 3. Obsérvese que el bloque analítico de la figura muestra varios bloques de construcción que representan el diseño de la solución para una aplicación particular. Esta parte del diseño algorítmico depende de la habilidad del analista y, a menudo, utiliza uno o más de los cuatro problemas principales como bloque de construcción (Aggarwal, 2015).

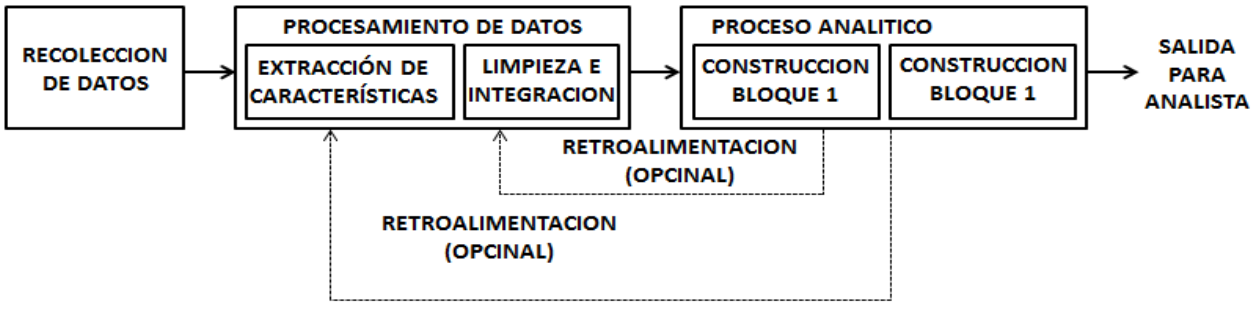


Figura 3. Tubería de procesamiento de datos. Adaptado de (Aggarwal, 2015).

4.4.4. Transformación de datos. Cuando los datos son de gran dimensión, muchos algoritmos de minería de datos no funcionan de manera efectiva. Además, muchas de las características de alta definición son ruidosas y pueden agregar errores al proceso de minería de datos. Por lo tanto, se utiliza una variedad de métodos para eliminar funciones irrelevantes o transformar el conjunto actual de características en un nuevo espacio de datos que es más fácil de analizar. Otro aspecto relacionado es la transformación de datos, donde un conjunto de datos con un conjunto particular de atributos puede transformarse en un conjunto de datos con otro conjunto de atributos del mismo tipo o de un tipo diferente. Por ejemplo, un atributo, como la edad, puede dividirse en rangos para crear valores discretos para la conveniencia analítica (Aggarwal, 2015).

En general, las trasformaciones se usan para tres propósitos: estabilizar la varianza de respuesta, hacer que la variable de distribución de la variable de respuesta esté más cerca de la distribución normal, y mejorar el ajuste del modelo de los datos (Montgomery, 1996).

Cuando se realiza la normalización, las magnitudes se escalan a valores apreciablemente bajos. Esto es importante para muchos algoritmos. Algunos de los métodos más comunes para este alcance son:

4.4.4.1. Estandarización de datos. Kotsiantis, Kanellopoulos, y Pintelas (2006), definen la normalización como una transformación de "escalamiento" de las características. Dentro de una característica, a menudo hay una gran diferencia entre los valores máximo y mínimo, por ejemplo: 0.01 y 1000.

Normalización Min-Max: Ejecuta una transformación lineal de los datos originales. Con base en los valores mínimo y máximo de un atributo, se calcula un valor de normalización v' con base en el valor v de acuerdo con la siguiente expresión:

$$v' = \frac{v - min_A}{max_A - min_A}(nuevo\_max_A - nuevo\_min_A) + nuevo\_min_A \qquad (1)$$

Este método conserva las relaciones entre los datos originales.

Normalización Z-core: Los valores para un atributo A son normalizados basados en la media y la desviación estándar de A. Un valor v de A es normalizado a v' con el cálculo de la siguiente expresión:

$$v' = \frac{v - media}{desv.estandar} \quad (2)$$

Este método es utilizado cuando el máximo y el mínimo del atributo A son desconocidos o cuando hay valores anómalos que predominan al utilizar la normalización min-máx.

Normalización de escala decimal: Normaliza moviendo los puntos decimales de los valores del atributo A. El número de puntos decimales movidos depende del máximo valor absoluto de A. Un valor v de A es normalizado a v' con el cálculo de la siguiente expresión:

$$v' = \frac{v}{10^j} \quad (3)$$

Donde j es el entero más pequeño de $Max(|v'|) < 1$.

4.4.4.2. Otras transformaciones. Las transformaciones susceptibles de mejorar el ajuste y la capacidad de predicción de un modelo son muy numerosas, a continuación, se mencionas algunas de ellas. La tabla 2, donde se presentan varias funciones que describen relaciones entre x y y que pueden producir una regresión lineal por medio de la transformación indicada. Además, en aras de que el análisis sea más exhaustivo, se presentan las variables dependiente e independiente que se utilizan en la regresión lineal simple resultante.

Tabla 2.

Algunas transformaciones para normalizar.

Forma funcional **Que relaciona y con x**	**Trasformación propia**	**Forma de hacer la regresión** **lineal simple**
Exponencial: $y = \beta_0 e^{\beta_1 x}$	$y^* = lny$	Hacer la regresión de y^*contra x
Potencia: $y = \beta_0 x^{\beta_1}$	$y^* = logy; x^* = logx$	Hacer la regresión de y^*contra x^*
Reciproca:$y = \beta_0 + \beta_1 \left(\frac{1}{x}\right)$	$x^* = \frac{1}{x}$	Hacer la regresión de y contra x^*
Hiperbólica: $y = \frac{x}{\beta_0 + \beta_1 x}$	$y^* = \frac{1}{y};\ x^* = \frac{1}{x}$	Hacer la regresión de y^*contra x^*

Nota: Características de la transformación de funciones. Adaptado de (Walpole et al., 2007).

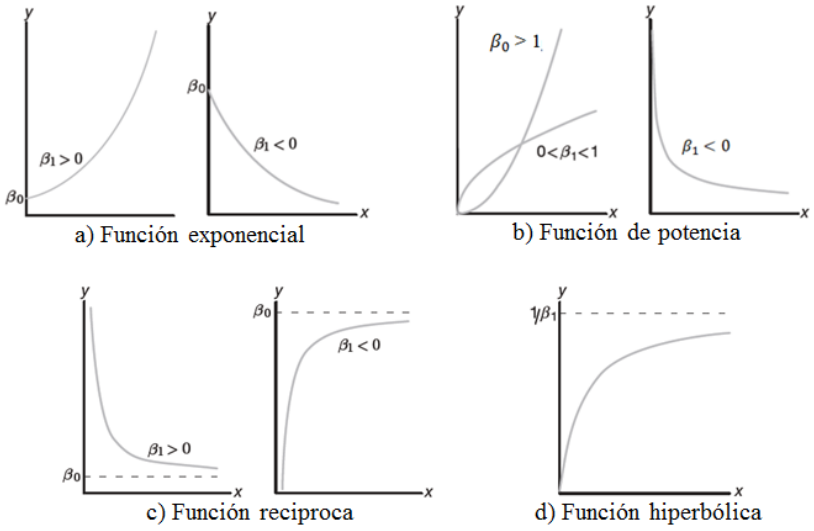


Figura 4. Diagramas de las funciones de la tabla 2. Adaptado de (Walpole et al., 2007).

La figura 4 ilustra las funciones que se listan en la tabla 2, las cuales sirven como guía en la elección de una transformación a partir de la observación de la gráfica de y contra x (Walpole, Myers, Myers, y Ye, 2007).

4.4.4.3. Transformación de Johnson. Cuando los datos que se tienen no se ajustan a una distribución normal como es supuesto, se presentan problemas para la aplicación de algunas técnicas en Control Estadístico de Procesos (Statistical Process Control), SPC. Sin embargo, cuando esto sucede es posible transformar los datos no-normales a datos normales, a través de técnicas como el Sistema de Familias de Distribuciones de Johnson (Lagos, y Vargas, 2003).

Para ajustar un conjunto de datos no-normales, es necesario establecer algunos criterios que permitan determinar la pertenencia de este conjunto a una de las tres familias Johnson existentes. Cada una de ellas tiene asociada una transformación de x a una variable normal estándar Z, así como condiciones especiales para los parámetros estimados y el rango de la variable x, que deben tenerse en cuenta cuando se va a escoger la familia con la que se quiere trabajar. Así mismo, Johnson en el año 1949 determina tres familias de dicha distribución (Lagos, y Vargas, 2003).

- SB : Se refiere a x acotada.

$$Z = \gamma + \eta \ln\left(\frac{x-\epsilon}{\lambda+\epsilon-x}\right) \quad (4)$$

Esta transformación está sujeta a las siguientes condiciones:

Condiciones de los parámetros: $\eta, \lambda > 0, -\infty < \gamma < \infty, -\infty < \epsilon < \infty$.

Condiciones de la variable x: $\epsilon < x < \epsilon + \lambda$.

- SL : Se refiere a x acotada por debajo o lognormal.

$$Z = \gamma + \eta \ln(x - \epsilon) \quad (5)$$

Esta transformación está sujeta a las siguientes condiciones:

Condiciones de los parámetros: $\eta > 0, -\infty < \gamma < \infty, -\infty < \epsilon < \infty$.

Condiciones de la variable x: $x > \epsilon$

- SU : Se refiere a x no acotada.

$$Z = \gamma + \eta \operatorname{sinh}^{-1}\left(\frac{x-\epsilon}{\lambda}\right) \quad (6)$$

Esta transformación está sujeta a las siguientes condiciones:

Condiciones de los parámetros: $\eta, \lambda > 0, -\infty < \gamma < \infty, -\infty < \epsilon < \infty$.

Condiciones de la variable x: $-\infty < x < \infty$.

4.4.4.4. Transformación de Box-Cox. En muchas aplicaciones de modelado estadístico, se requiere una transformación de la variable dependiente para lograr un modelo lineal teórico normal con una estructura media simple y errores homocedásticos. Cuando se conoce dicha transformación, los métodos habituales de inferencia de modelos de teoría normal pueden aplicarse directamente a las respuestas transformadas. Box y Cox (1964), como es citado en Yang (1999), sugiere estimar el parámetro de transformación cuando se desconoce dicha transformación y luego seleccionar el número simple más cercano correspondiente a una transformación de logaritmo, cuadrada, de raíz, etc., y luego llevar a cabo las inferencias habituales para los parámetros definidos e interpretados en la escala seleccionada (Yang, 1999).

La familia de transformaciones de potencia de Box-Cox es un método analítico conveniente que utiliza los datos para sugerir una transformación de potencia de la variable de respuesta continua y cuando y es positiva en el caso univariante. Su objetivo es encontrar un $\lambda \in R$ (típicamente $-2 \leq \lambda \leq 2$ en incrementos de 0.25) tal que se satisfagan los supuestos estadísticos subyacentes de la variable transformada. Esta familia de transformaciones es útil incluso cuando los datos no pasan una prueba exacta de normalidad tras la transformación. Contendrá una transformación satisfactoria o guiará al investigador a un análisis más detallado de la situación.

El método Box-Cox fue desarrollado para situaciones donde la respuesta y es continua y estrictamente positiva, es decir $y \in R^+$. Sin embargo, puede aplicarse en situaciones donde algunas observaciones de y son menores o iguales a cero. En tales casos, uno puede usar la familia

de potencia extendida al agregar una constante λ_2 a y (Típicamente $\lambda_2 = 1$) tal que $y^* = y + \lambda_2 > 0$, y luego aplicar el método Box-Cox a y^*(Malaeb, 1997).

Box-Cox suponen que se ha observado un vector de respuesta $n \times 1$ de observaciones $y = \{y_1, \dots, y_n\}$ y que existe un λ de tal manera que los datos pueden ser descritos apropiadamente por el modelo lineal.

$$E\{y^{(\lambda)} = \boldsymbol{x\beta}\}, \tag{7}$$

Donde $y^{(\lambda)\prime} = (y_1^{(\lambda)}, \dots, y_n^{(\lambda)})$ es el vector de las observaciones transformadas, x es una conocida matriz $n \times p$ de rango completo de valores de las variables independientes y β es un vector $p \times 1$ de parámetros desconocidos asociados con las observaciones transformadas. Para valores positivos de la variable de respuesta y, Box-Cox familia de transformaciones y a $y^{(\lambda)}$ Con el parámetro λ definir la transformación particular es:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, \lambda \neq 0 \\ lny, \ \lambda = 0 \end{cases} \tag{8}$$

Esta familia de "transformaciones de poder" se puede aplicar en muchos problemas con valores positivos de la variable de respuesta. Es muy útil para eliminar la asimetría y, por lo tanto, puede no funcionar bien cuando la respuesta se limita tanto arriba como abajo. Si algunos de los valores de y son cero o negativos, se puede usar una familia más flexible, llamada "familia de poder extendida".

$$y^{(\lambda)} = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1}, \lambda_1 \neq 0 \\ \ln(y+\lambda_2), \ \lambda_2 = 0 \end{cases} \tag{9}$$

Donde $y + \lambda_2 > 0$.

La familia (8), que contiene la conocida trasformación logarítmica, raíz cuadrada e inversa, es una modificación de la familia (10).

$$y^{(\lambda)} = \begin{cases} y^{\lambda}, & \lambda \neq 0 \\ lny, & \lambda = 0,\ y > 0 \end{cases} \quad (10)$$

Estudiada en detalle por Tukey (1957) para $|\lambda| \leq 1$. Esta modificación, evita una discontinuidad en $\lambda = 0$, ya que, según la regla de L'Hopital $\lim_{\lambda \to 0} y^{\lambda} = lny$. Las ecuaciones (8) y (10) son equivalentes ya que las estadísticas F en el análisis de varianza son invariantes bajo transformaciones lineales. La ecuación (8) es preferida por consideraciones teóricas para su continuidad en $\lambda = 0$. También se supone que para cada $\lambda, y^{(\lambda)}$ es una función monotónica de y sobre el rango admisible (Malaeb, 1997).

4.4.5. Datos faltantes. Durante la fase de preparación en la minería de datos, con frecuencia se deseará sustituir los valores perdidos en el conjunto de datos. Los valores perdidos son valores del conjunto de datos desconocidos, sin recopilar o incorrectamente introducidos. Por lo general, estos valores no son válidos en sus campos. Por ejemplo, el campo Sexo debe contener los valores M y F. Si se encuentran valores como Y o Z en el campo, puede asumir con seguridad que esos valores no son válidos y que se deben interpretar por lo tanto como espacios en blanco.

Del mismo modo, un valor negativo para el campo Edad no tendría sentido y, por tanto, también debería interpretarse como un valor vacío. En muchas ocasiones, estos valores obviamente erróneos se introducen deliberadamente o se dejan los campos vacíos durante un cuestionario para indicar la omisión de una respuesta. En ocasiones se deseará examinar estos elementos vacíos con mayor detenimiento para determinar si una respuesta omitida, como la negativa a proporcionar la edad de una persona, es un factor para predecir un resultado específico (IBM Knowledge Center, s.f).

Para tratar con datos faltantes se puede basar en uno de los muchos métodos de imputación. La imputación es el proceso de estimación de valores ausentes basado en valores válidos de otras

variables y/o casos de la muestra. El objetivo es emplear relaciones conocidas que puedan identificarse en los valores válidos de la muestra para ayudar en la estimación de valores faltantes. Sin embargo, el investigador debería considerar cuidadosamente el uso de la imputación en cada instancia, dados sus potenciales impactos sobre el análisis (Hair, Anderson, Tatham y Black, 1999).

Hair et al. (1999), mencionan los siguientes métodos para sustitución o imputación de valores faltantes:

- **El uso de toda la información disponible como técnica de imputación**: Este método de imputación no reemplaza los datos ausentes, sino que imputa las características de distribución (por ejemplo, la desviación media o estándar) o las relaciones (por ejemplo, correlaciones) de todos los valores válidos disponibles.
- **Sustitución de datos ausentes**: Esta forma de imputación consiste en el método efectivo de sustitución de los datos ausentes por valores estimados sobre la base de otra información existente en la muestra. Esta medida puede llevarse a cabo de muchas maneras, que van desde una sustitución directa de valores, hasta procesos de estimación basados en relaciones entre variables.
- **Sustitución de caso**: En este método, las observaciones con datos ausentes se sustituyen con otras observaciones no muéstrales. Un ejemplo común es reemplazar un hogar que está en la muestra, pero con el que no se puede contactar o que tiene gran cantidad de datos ausentes con otro hogar que no está en la muestra, preferiblemente muy similar al de la observación original.
- **Solución por la media**: Uno de los métodos más empleados consiste en sustituir los valores ausentes por una variable cuyo valor medio se calcula sobre todas las respuestas

válidas. De esta forma, las respuestas de la muestra válida se usan para calcular el valor de sustitución. La lógica de esta aproximación es que la media es el mejor valor de sustitución.

- **Sustitución por valor constante**: En este método, se sustituyen los datos ausentes por un valor constante derivado de fuentes externas o investigación previa. Su naturaleza es similar al método de sustitución de la media, que difiere sólo en la fuente del valor de sustitución.
- **Imputación por regresión**: En este método se usa el análisis de regresión para predecir los valores ausentes de una variable basándose en su relación con otras variables del conjunto de datos.
- **Imputación múltiple**: Este método de imputación es en realidad una combinación de varios métodos. En esta aproximación, se usan dos o más métodos para derivar una estimación compuesta, usualmente la media de las diversas estimaciones para el dato ausente. La lógica de esta aproximación es que el uso de la aproximación múltiple minimiza los problemas específicos con cualquier método simple siendo su composición la mejor estimación.

4.5. Aprendizaje automático

El aprendizaje automático ha tenido gran crecimiento y desarrollo, ha recibido gran atención por parte de los científicos y debido a esto, ha tenido gran participación en un sin número de estudios, ayudando así a resolver tareas de la realidad con gran impacto en la sociedad. Expertos definen el aprendizaje automático de la siguiente forma:

El aprendizaje automático (Machine Learning en inglés), es la rama de la inteligencia artificial que se dedica al estudio de los programas que aprenden basados en su experiencia, para realizar una tarea determinada cada vez mejor. El objetivo principal de un modelo de aprendizaje

automático es utilizar la evidencia o información dada como datos de entrada a modo de ejemplos para poder crear una hipótesis y ser capaz de brindar una respuesta a una situación no conocida (Mitchell, 1997).

Los expertos sugieren diferentes enfoques del aprendizaje automático, entre ellos se encuentran:

4.5.1. Aprendizaje supervisado. Esta clase de aprendizaje consiste en que al algoritmo utilizado se le proporciona una serie de ejemplos con sus correspondientes etiquetas, es decir, todos estos ejemplos dados son calificados anteriormente. De esta forma en el proceso de aprendizaje, el algoritmo compara su salida actual con la etiqueta del ejemplo para luego realizar los cambios que considere necesarios (Espino, A. I. L., Mur, R. A., & de Miguel, M. A. S. 2004).

4.5.2. Aprendizaje no supervisado. A diferencia del aprendizaje supervisado en este tipo de aprendizaje no se conoce “a priori” el atributo dependiente, es decir, no existe una clasificación de los ejemplos como tal, el aprendizaje se guía por la similaridad y disimilaridad entre los ejemplos dados, al igual que el aprendizaje supervisado, pertenece al aprendizaje inductivo que busca obtener conceptos más generales a partir de los ejemplos antes mencionados (Cambronero, y Moreno, 2016).

4.6. Regresión lineal.

En la práctica se requiere resolver problemas que incluyen conjuntos de variables cuando se sabe que existen algunas relaciones inherentes entre ellas. La regresión lineal se refiere a encontrar la mejor estimación de la relación entre las variables involucradas.

La regresión lineal se convierte, entonces, en una herramienta de predicción, es decir, estimar el valor de una variable dependiente a partir de un dato ya conocido llamado variable

independiente. Regresión lineal supone una relación lineal entre las variables existentes en el ejercicio (Walpole et al., 1999).

4.6.1. Regresión lineal simple. En el caso de la regresión lineal simple, donde hay una sola variable de regresión independiente $\boldsymbol{x}$ y una sola variable dependiente $\boldsymbol{y}$. El modelo debe incluir al conjunto $[(\boldsymbol{xi}, \boldsymbol{yi});\ \boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}, \ldots, \boldsymbol{n}]$ de datos que implica n pares de valores $(\boldsymbol{x}, \boldsymbol{y})$. No se debe olvidar que el valor de $\boldsymbol{yi}$ depende de $\boldsymbol{xi}$ por medio de una estructura lineal que también incluye el componente aleatorio. La base para el uso de un modelo estadístico se relaciona con la manera en que la variable aleatoria $\boldsymbol{y}$ cambia con $\boldsymbol{x}$ y el componente aleatorio. La ecuación de regresión lineal simple se representa así: $\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\,\boldsymbol{x} + \boldsymbol{\varepsilon}$ en la cual $\boldsymbol{\beta_0}$ y $\boldsymbol{\beta_1}$ son los parámetros desconocidos de la intersección y la pendiente, respectivamente, y $\boldsymbol{\varepsilon}$ es una variable aleatoria que se supone está distribuida con $\boldsymbol{E}\,(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ y $\boldsymbol{Var}\,(\boldsymbol{\varepsilon}) = \boldsymbol{\sigma^2}$(Walpole et al., 2007).

Es frecuente que a la cantidad σ^2 se le denomine varianza del error o varianza residual. De este modelo se puede decir que la variable Y es aleatoria ya que ε es aleatoria y la variable x no es aleatoria y se mide con un error despreciable, la cantidad ε que a menudo recibe el nombre de error aleatorio tiene varianza constante, la presencia de este error evita que el modelo se convierta en tan solo una ecuación determinista, el hecho que $E\,(\varepsilon) = 0$ implica que para una x específica, los valores y se distribuyen alrededor de la recta verdadera o recta de regresión de la población $y = \beta_0 + \beta_1\,x$ (Walpole et al., 2007).

4.6.2. Regresión lineal múltiple. La regresión lineal múltiple, así como en la regresión lineal simple se utiliza con el fin de hallar una ecuación que describa el comportamiento de una variable a partir de datos independientes (en este caso dos o más), en este caso su complejidad es mayor debido a la existencia de dos más variables independientes.

Para el caso de k variables independientes, el modelo que da $x_1, x_2, \dots, x_k$ la media de $y\,|x_1, x_2, \dots, x_k$, es el modelo de regresión lineal múltiple $U_y\,|x_1, x_2, \dots, x_i = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2, \dots, \beta_k\, x_k$, y la respuesta estimada se obtiene a partir de la ecuación de regresión muestral $\hat{y} = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2, \dots, \beta_k\, x_k$, donde cada coeficiente de regresión β_i se estima a partir de los datos muestrales usando el método de mínimos cuadrados (Walpole et al., 2007).

Este método se basa en ajustar una línea de regresión a los puntos de un conjunto de datos que tiene la suma mínima de las desviaciones elevada al cuadrado (error de mínimos cuadrados). También se puede utilizar el método de estimación de máxima verosimilitud. Indica la probabilidad de que una muestra observada dependa de los posibles valores de los parámetros. Por lo tanto, cuando se maximiza la función de verosimilitud se determina los parámetros que tienen mayor probabilidad de producir los datos observados. (Minitab Inc., s.f.)

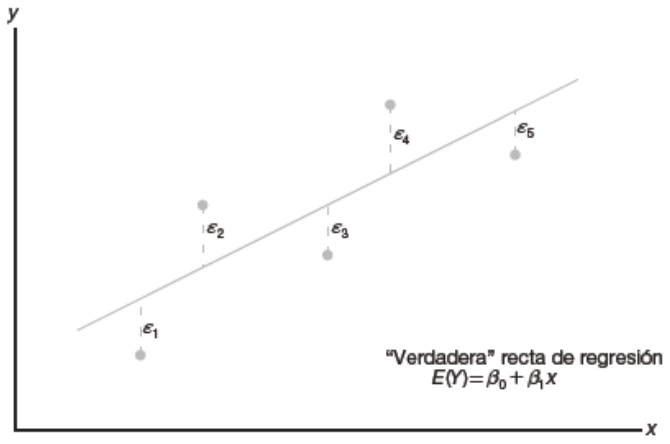


Figura 5. Datos (x, y) hipotéticos dispersos alrededor de la verdadera recta de regresión. Adaptado de (Walpole et al., 2007).

4.6.2.1. Método de mínimos cuadrados. En la mayoría de los problemas de investigación en los que se aplica el análisis de regresión se necesita más de una variable independiente para el modelo de regresión. La complejidad de la mayoría de mecanismos científicos es tal que, con el fin de predecir una respuesta importante, se requiere un modelo de regresión múltiple. Cuando un modelo es lineal en los coeficientes se denomina modelo de regresión lineal múltiple. Para el caso de k variables independientes, el modelo que da $\boldsymbol{x_1, x_2, \ldots, x_k}$, la media de $\boldsymbol{y|x_1, x_2, \ldots, x_k}$, es el modelo de regresión lineal múltiple (Walpole et. al, 2007).

$$y|x_1, x_2, \ldots, x_k = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k| \tag{11}$$

Y la respuesta estimada se obtiene a partir de la ecuación de regresión múltiple.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \tag{12}$$

Donde cada coeficiente de regresión β_i se estima por medio de b_i, a partir de los datos muestrales, usando el método de los mínimos cuadrados. Como ocurre en el caso de una sola variable independiente, a menudo el modelo de regresión lineal múltiple es una representación adecuada de una estructura más complicada dentro de ciertos rangos de las variables independientes (Walpole et. al, 2007)

Se calculan los estimadores de mínimos cuadrados de los parámetros $\beta_0, \beta_1, \ldots, \beta_k$ mediante el ajuste del modelo de regresión lineal múltiple

$$\mu_y|x_1, x_2, \ldots, x_k = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{13}$$

A los puntos de los datos

$$\{(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i); i = 1,2, \ldots, n \; y \; n = k\} \tag{14}$$

Donde y_i es la respuesta observada a los valores $x_{1i}, x_{2i}, \ldots, x_{ki}$ de las k variables independientes $x_1, x_2, \ldots, x_k$. Se supone que cada observación $x_{1i}, x_{2i}, \ldots, x_{ki}, y_i$ satisface la siguiente ecuación:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \quad (15)$$

O bien,

$$y_i = \widehat{y_i} + e_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i \quad (16)$$

Donde ε_i y e_i son el error aleatorio y residual, respectivamente, asociados con la respuesta y_i y con el valor ajustado $\widehat{y_i}$.

Como en el caso de la regresión lineal simple, se supone que los ε_i son independientes y están distribuidos en forma idéntica con media cero y varianza común σ^2. Al usar el concepto de mínimos cuadrados para obtener los estimados $b_0, b_1, b_2, \ldots, b_k$, se minimiza la expresión

$$SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_k x_{ki})^2 \quad (17)$$

Si, a su vez, diferenciamos la SCE respecto a $b_0,\ b_1,\ b_2, \ldots\ b_k$ y se iguala el resultado a cero, se genera un conjunto de $k + 1$ *ecuaciones normales para la regresión lineal múltiple.* Dichas ecuaciones se pueden resolver para $b_0,\ b_1,\ b_2, \ldots\ b_k$ utilizando cualquier método apropiado que permita resolver sistemas de ecuaciones lineales (Walpole et. al, 2007).

4.6.3. Criterios para la selección del modelo de regresión. A la hora de construir un modelo se tienen diferentes posibilidades, las cuales se ajustan mejor o peor a la realidad. Algunos de los criterios más usados para la selección de variables en modelos lineales son:

4.6.3.1. Coeficiente de determinación corregido o ajustado. Es un coeficiente que mide la intensidad de la relación lineal entre la variable objetivo y las predictoras.

4.6.3.2. Validación cruzada. La validación cruzada es una técnica utilizada para validar los métodos de regresión, al igual que otras técnicas como son la comparación de los parámetros derivados con los obtenidos mediante modelos físicos teóricos o con simulaciones o utilizar nuevos conjuntos de datos conocidos para comparar con los propuestos (Pérez, Delegido, Rivera y

Verrelst, 2015). Se hará énfasis en los métodos de validación cruzada como los llamados Hold-out y K-fold.

4.6.3.2.1. Método Hold-out. Consiste básicamente en dividir el conjunto de datos disponible en dos subconjuntos, un subconjunto se utilizaría como entrenamiento del modelo y el otro como test de validación del mismo, así, se crearía un modelo únicamente con los datos de entrenamiento, acto seguido, se obtienen datos de salida que se comparan con el conjunto de datos destinados para la validación. Los estadísticos obtenidos del subconjunto de validación son los que dan la validez del método empleado en términos de error, una alternativa en este mismo método consiste en repetir el proceso hold-out tomando distintos conjuntos de datos de entrenamiento aleatorios cierto número de veces de manera que se calculan los estadísticos de la regresión a partir de la media de los valores de cada una de las iteraciones (Pérez et al, 2015).

4.6.3.2.2. Método K-fold. El método k-fold tiene como base el hold-out, sin embargo, este tiene una utilidad mayor en un conjunto de datos pequeño. En este caso, el conjunto de datos se divide en k subconjuntos, de manera que se aplica el método hold-out k veces, utilizando cada vez un subconjunto diferente para la validación del modelo entrenado con los demás k-1 subconjuntos. El error medio que se obtiene de los k análisis desarrollados, proporciona el error cometido por el método, evaluando así, la validez del mismo.

El método k-fold tiene la ventaja de que todos los datos son tenidos en cuenta para entrenar y validar, por lo que se obtienen resultados más representativos a priori, en cambio el método hold-out, realiza el proceso n veces de manera aleatoria sin garantizar que todos los datos sean tenidos en cuenta (Pérez et al, 2015).

4.6.3.3. Criterio de Información de Akaike (AIC). Criterio propuesto por el estadístico japonés Hirotugu Akaike y que está basado en la teoría de la información. Está definida de forma

que bonifica la bondad de ajuste y penaliza la inclusión de parámetros a estimar, lo que ayuda a evitar el fenómeno del sobreajuste.

4.6.3.4. Criterio de Información Bayesiana (BIC). El profesor Gideon E. Schwarz propuso este criterio bajo un enfoque bayesiano que se basa en la probabilidad a posteriori de los modelos. Es, junto al AIC, el más usado (Guerra de la Corte, 2016).

No todos los coeficientes de las variables independientes o predictores se pueden estimar con una precisión razonable, por ello se debe considerar que variables incluir y de la misma forma excluir. Debería también tenerse en cuenta algunos procedimientos que se han para añadir y eliminar variables sistemáticamente, por ejemplo:

4.6.3.4.1. Procedimiento de selección hacia adelante. Se basa en el concepto de que las variables deben insertarse una por una hasta obtener una ecuación de regresión satisfactoria.

El procedimiento es como sigue:

Paso 1: Se elige la variable que proporciona la mayor suma de cuadrados de regresión cuando se ejecute la regresión lineal simple con y o, en forma equivalente aquella que proporcione el mayor valor de R^2. Esta variable se llamará x_I. Si x_1 es insignificante el proceso se suspende.

Paso 2: Se selecciona la variable que al ser integrada al modelo proporciona el mayor incremento de R^2, en presencia de x_I , sobre la R^2 encontrada en el primer paso. Ésta, por supuesto, es la variable x_j para la que

$$R(\beta_j|\beta_1) = R(\beta_1, \beta_j) - R(\beta_1) \tag{18}$$

Es más grande. Dicha variable se llamará x_2. Luego se ajusta el modelo de regresión con x_1 y x_2, y se observa R^2. Si x_2 es insignificante, el proceso se suspende.

Paso 3: se elige la variable x_j que proporciona el valor más grande de

$$R(\beta_j|\beta_1, \beta_2) = R(\beta_1, \beta_2, \beta_j) - R(\beta_1, \beta_2), \tag{19}$$

Otra vez da como resultado el incremento mayor de R^2 sobre el que se obtuvo en el paso 2. A esta variable se le denomina x_3, y ahora se tiene un modelo de regresión que incluye x_1, x_2 y x_3. Si x_3 es insignificante, el proceso se suspende, este proceso continúa hasta que la variable más reciente incluida ya no produce un incremento significativo en la regresión explicada.

4.6.3.4.2. Procedimiento de eliminación hacia atrás. Implica los mismos conceptos que la selección hacia delante, excepto que se comienza con todas las variables en el modelo. Por ejemplo, se supone que hay cinco variables en consideración. Los pasos son:

Paso 1: Se ajusta una ecuación de regresión con las cinco variables incluidas en el modelo. Se elige la variable que proporcione el valor más pequeño de la suma de cuadrados de regresión ajustada para las demás. Suponga que dicha variable es x_2. Elimine x_2 del modelo si

$$f = \frac{R(\beta_2|\beta_1,\beta_3,\beta_4,\beta_5)}{s^2} \tag{20}$$

es insignificante.

Paso 2: Se ajusta una ecuación de regresión utilizando las variables restantes x_1, x_3, x_4 y x_5, y repita el paso 1. Suponga que esta vez se elige la variable x_5. Nuevamente, si

$$f = \frac{R(\beta_5|\beta_1,\beta_3,\beta_4)}{s^2} \tag{21}$$

es insignificante, se retira del modelo la variable x_5. En cada paso la s^2 que se usa en la prueba F es el cuadrado medio de error para el modelo de regresión en esa etapa.

Regresión por etapas: Se lleva a cabo con una modificación ligera pero importante del procedimiento de selección hacia delante. La modificación requiere efectuar más pruebas en cada etapa para garantizar la eficacia continuada de las variables que se hubieran incluido en el modelo durante alguna etapa anterior. Esto representa una mejoría sobre la selección hacia delante, ya que es muy posible que una variable que haya entrado a la ecuación de regresión en una etapa temprana resulte poco importante o redundante debido a las relaciones que existen entre ella y las otras

variables que se incluyeron en etapas posteriores. Por lo tanto, en la etapa en que se incluyó una variable nueva a la ecuación de regresión mediante un incremento significativo de R_2, según lo determina la prueba F, todas las variables que ya estén en el modelo se someten a pruebas F (o bien, a pruebas t) a la luz de esta nueva variable, y si no muestran un valor f significativo, se eliminan. El procedimiento continúa hasta que se alcance una etapa donde ya no sea posible insertar ni eliminar variables adicionales (Walpole et. al, 2007).

4.6.3.5. Medición de errores de predicción. Una medida de verificación en ejercicios de regresión es el error cuadrático medio (RMSE por sus siglas en ingles), que se define como la raíz cuadrada de la media de las diferencias al cuadrado entre los elementos correspondientes del pronóstico y las observaciones. Es la desviación estándar de los residuos (errores de predicción).

Los residuos son una medida de cuán lejos están los puntos de datos de la línea de regresión; el RMSE es una medida de la dispersión de estos residuos. En decir, dice qué tan concentrado está la información en la línea de mejor ajuste. El error cuadrático medio se usa comúnmente en la predicción y el análisis de regresión para verificar los resultados experimentales (Barnston, 1992).

Para ello se debe conocer el error cuadrado medio (MSE por sus siglas en ingles), "esta estadística no es muy informativa por sí misma, pero se puede usar para comparar los ajustes obtenidos mediante el uso de diferentes métodos" (Minitab Inc., s.f).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2 \tag{22}$$

Donde Y son las observaciones, $\hat{Y}$ son los valores predichos y n es la cantidad de residuos del ejercicio.

$$RMSE = \sqrt{MSE} \tag{23}$$

A menudo, se prefiere el RMSE al MSE ya que está en la misma escala que los datos. Históricamente, el RMSE y el MSE han sido populares, en gran parte debido a su relevancia teórica

en el modelado estadístico. Sin embargo, son más sensibles a valores atípicos que otros estimadores en la misma escala, lo que ha llevado a algunos autores a recomendar su uso en la evaluación de la precisión del pronóstico (Hyndman, y Koehler, 2006).

4.7. Algoritmo Random Forest - RF

Uno de los métodos o marcos más populares utilizado por los científicos de datos en la práctica profesional es el algoritmo Random Forest. El cual es considerado uno de los mejores algoritmos de clasificación, capaz de clasificar grandes cantidades de datos con precisión según muchos expertos. RF es un método de aprendizaje conjunto (también considerado como una forma de predictor más cercano).

El bagging o la agregación de bootstrap es una técnica para reducir la varianza de una función de predicción estimada. El bagging parece funcionar especialmente bien para procedimientos de baja varianza y alta varianza, como los árboles. Para la regresión, simplemente se ajusta el mismo árbol de regresión muchas veces para arrancar versiones muestreadas de los datos de entrenamiento y promediar el resultado. Para la clasificación, un conjunto de árboles emite una decisión para la clase predicha. Los bosques aleatorios (Breiman, 2001) son una modificación sustancial del bagging que forma una gran colección de árboles sin correlación, y luego los promedia. En muchos problemas, el rendimiento de los bosques aleatorios es muy similar al aumento de estos, y son más sencillos de entrenar y sintonizar. Como consecuencia, los bosques aleatorios son populares y se implementan en una gran variedad de paquetes (Hastie, Tibshirani, y Friedman, 2001).

Los parámetros usados en Random Forest son:

- Número de árboles (ntree)
- Numero de variables predictoras elegidas al azar por cada corrida (mtry)

- Número mínimo de nodos (nodesize)
- Numero de variables predictoras (p)

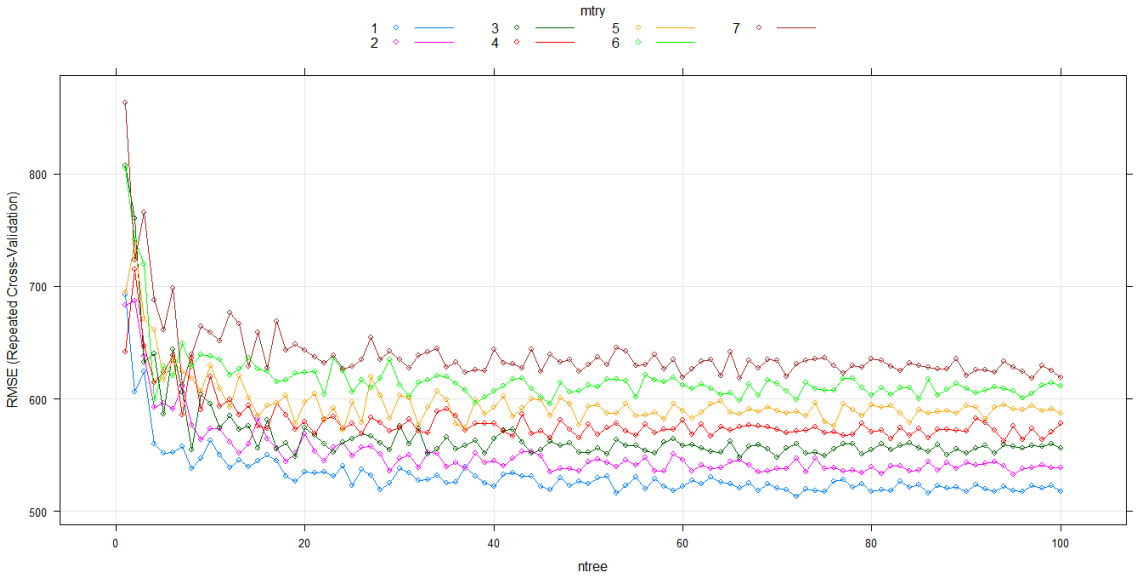


Figura 6. Comportamiento de los árboles de un bosque de regresión para diferentes mtry.

En la figura 6 se observa el comportamiento de los árboles de regresión para valores de mtry desde 1 hasta 7 especificadas por color en la parte superior, figura obtenida con la ayuda del software RStudio por medio de la librería "caret".

Como parámetros en la construcción del bosque, el inventor (Breiman, 2001) hace las siguientes recomendaciones:

• Para la clasificación, el valor predeterminado para mtry es $\sqrt{p}$ y el mínimo el tamaño del nodo es uno.

• Para la regresión, el valor predeterminado para mtry es $p / 3$ y el mínimo el tamaño del nodo es cinco.

El valor mtry debe ser el número entero inmediatamente menor al resultado de la operación tanto para clasificación como para regresión, en la práctica, los mejores valores para estos parámetros dependerán del problema, y deben ser tratados como parámetros de ajuste. En la figura

6, $mtry = 1$ funciona mucho mejor que el valor predeterminado $8/3 = 2{,}66$, siendo $mtry = 2$, es decir, presenta menor error en el modelo.

El algoritmo Random Forest es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio muestreado independientemente y con la misma distribución de todos los árboles del bosque. El error de generalización de un bosque depende de la fuerza de los árboles individuales y la correlación entre ellos. Las estimaciones internas supervisan el error, fuerza, y la correlación, se utilizan para demostrar la respuesta al aumento del número de características utilizadas en la división. También se utilizan estimaciones internas para medir la importancia variable (Breiman, 2001).

RF para la regresión está formado por árboles en crecimiento según un vector aleatorio Θ, de modo que el predictor de árbol h (**x**, Θ) adopta valores numéricos en oposición a las etiquetas de clase. Los valores de salida son numéricos y se supone que el conjunto de entrenamiento se extrae independientemente de la distribución del vector aleatorio $(Y, \boldsymbol{X})$. El error de generalización de la media cuadrática para cualquier predictor numérico $h\ (\boldsymbol{X})$ es:

$$E_{\boldsymbol{X},Y}(Y - h(\boldsymbol{X}))^2 \tag{24}$$

El predictor de Random Forest se forma tomando el promedio sobre k de los árboles $\{h\ (\boldsymbol{X}, \Theta k)\}$. De manera similar al caso de clasificación, se cumple lo siguiente:

Teorema1: Como la cantidad de árboles en el bosque va al infinito, casi seguramente:

$$E_{\boldsymbol{X},Y}\ (Y - av_k h\ (\boldsymbol{X}, \Theta k))^2 \rightarrow\ E_{\boldsymbol{X},Y}\ (Y - E_{\Theta} h\ (\boldsymbol{X}, \Theta))^2 \tag{25}$$

Denotar el lado derecho de (25) como PE * (bosque) - el error de generalización del bosque.

Define el error de generalización promedio de un árbol como:

$$PE^*(\acute{A}rbol) = E_{\Theta}\ E_{\boldsymbol{X},Y}(Y - h\ (\boldsymbol{X}, \Theta))^2 \tag{26}$$

Teorema 2: Asuma que para todo $\Theta,\ EY = E_{\boldsymbol{X}h}(\boldsymbol{X}, \Theta)$. Entonces,

$$PE^{*}(Bosque) \leq \rho' PE * (\acute{A}rbol) \quad (27)$$

Donde ρ' es la correlación ponderada entre los residuos $[Y - h(\boldsymbol{X}, \Theta)]$ e $[Y - h(\boldsymbol{X}, \Theta')]$ donde $[\Theta, \Theta']$ son independientes. Prueba:

$$PE * (Bosque) = E_{\boldsymbol{X},Y}[E_{\Theta}(Y - h(\boldsymbol{X}, \Theta)]^2 = E_{\Theta}E_{\Theta'}E_{\boldsymbol{X},Y}(Y - h(\boldsymbol{X}, \Theta))(Y - h(\boldsymbol{X}, \Theta')) \quad (28)$$

El término de la derecha en (28) es una covarianza y puede ser escrita como:

$$E_{\Theta}E_{\Theta'}(\rho(\Theta, \Theta')\, sd(\Theta)\, sd(\Theta')) \quad (29)$$

Donde $sd(\Theta) = \sqrt{E_{\boldsymbol{X},Y}(Y - h(\boldsymbol{X}, \Theta))^2}$. Define la correlación ponderada como:

$$\bar{\rho} = E_{\Theta}E_{\Theta'}(\rho(\Theta, \Theta')\, sd(\Theta)\, sd(\Theta')) / (E_{\Theta} sd(\Theta))^2 \quad (30)$$

Entonces,

$$PE * (Bosque) = \bar{\rho}(E_{\Theta}\, sd(\Theta))\, 2 \leq \bar{\rho}\, PE * (\acute{a}rbol) \quad (31)$$

El teorema 2 identifica los requisitos para bosques de regresión precisos: baja correlación entre residuos y árboles de bajo error. El bosque aleatorio disminuye el error promedio de los árboles empleados por el factor ρ. La aleatorización empleada debe apuntar a una baja correlación (Breiman, 2001).

4.8. Máquinas de soporte vectorial – MSV

Las máquinas de vectores de soporte iniciaron como una máquina de aprendizaje para solucionar problemas de clasificación de dos grupos donde los vectores de entrada se asignaban de forma no lineal a un espacio de funciones de más alta dimensión (Cortes y Vapnik, 1995).

Dichos vectores son observaciones que se encuentran en el borde de un área en el espacio que presenta un límite entre una de estas clases de observaciones, por ejemplo, los cuadrados y otra clase de observaciones, los círculos (ver figura 7), estos vectores se usan para identificar el hiperplano que separa las clases. En la terminología de MSV se habla del espacio entre estas dos

regiones como el margen entre las clases. Cada región contiene observaciones con el mismo valor para la variable objetivo, es decir, la clase. (Williams, 2011).

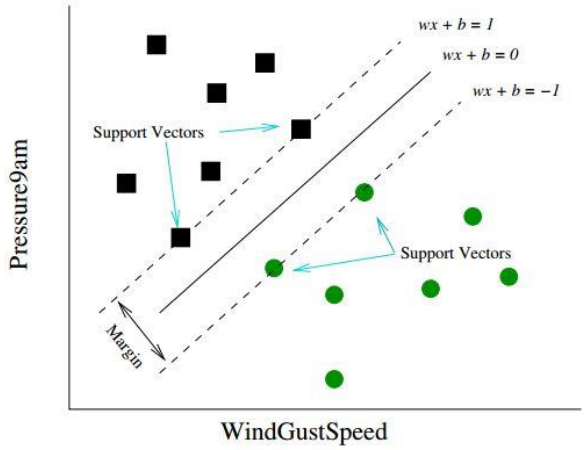


Figura 7. Modelo básico de MSV. Adaptado de (Williams, 2011)

Se plantea brevemente el planteamiento general de las MSV desde la perspectiva de un aprendizaje supervisado, es decir, el conocimiento de las salidas de un conjunto de entradas permite cuantificar (supervisar) la bondad de los resultados del modelo (González, 2003).

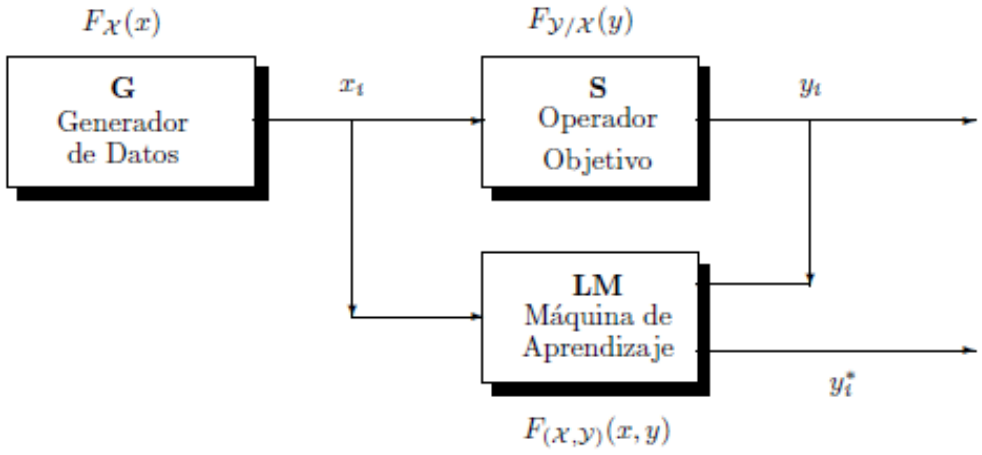


Figura 8. Esquema de configuración de una máquina de aprendizaje a partir de ejemplos. Adaptado de (González, 2003)

El objetivo fundamental de este tipo de estudios es aprender a partir de los datos y para ello busca la existencia de alguna dependencia funcional entre un conjunto de vectores inputs (o, de entrada)

$$\{x_i, i = 1, \dots, n\} \subseteq X \subseteq \mathbb{R}d \tag{32}$$

Y los valores outputs (o salidas)

$$\{y_i \; i = 1, \dots, n\} \subseteq Y \subseteq \mathbb{R} \tag{33}$$

El modelo representado por la figura 8, recoge de manera clara el objetivo que se persigue. En este esquema, G representa un modelo generador de datos que nos proporciona los vectores $x_i \in X$, independientes e idénticamente distribuidos de acuerdo con una función de distribución $FX(x)$ desconocida pero que suponemos no varía a lo largo del proceso de aprendizaje. Cada vector x_i es la entrada del operador objetivo S, el cual lo transforma en un valor y_i según una función de distribución condicional $F_{Y|X=xi}(y)$. Así la máquina de aprendizaje, que denotamos LM (learning machine) recoge el siguiente conjunto de entrenamiento,

$$Z = \{(x_i, y_i), \dots, (x_n, y_n)\} \subseteq X * y = \mathbb{Z} \tag{34}$$

el cual es obtenido independiente e idénticamente distribuido siguiendo la función de distribución conjunta:

$$F_{(X,Y)}(x, y) = F_X(x).F_{Y|X=x}(y) \tag{35}$$

A partir del conjunto de entrenamiento Z, la máquina de aprendizaje "construye" una aproximación al operador desconocido la cual proporcione para un generador dado G, la mejor aproximación (en algún sentido) a las salidas proporcionadas por el supervisor. Formalmente construir un operador significa que la máquina de aprendizaje implementa un conjunto de

funciones, de tal forma que, durante el proceso de aprendizaje, elige de este conjunto una función apropiada siguiendo una determinada regla de decisión.

Se puede estimar la función de esperanza matemática condicional:

$$E[Y/Z = x] \stackrel{\text{def}}{=} \int y dF_{Y|x}(y) \tag{36}$$

El objetivo del problema es la construcción de una función $f(x;\ y)$ dentro de una determinada clase de funciones F'. elegida a priori, la cual debe cumplir un determinado criterio de la mejor manera posible. Formalmente el problema se plantea así:

Dado un subespacio vectorial $\mathbb{Z}$ *de* $\mathbb{R}^{d+1}$ donde se tiene definida una medida de probabilidad $F_{\mathbb{Z}}(z)$, un conjunto $F' = \{f(z), z \in \mathbb{Z}\}$ de funciones reales y un funcional R: $F' \rightarrow \mathbb{R}$.

Buscar una función $f* \in F'$ tal que:

$$R\,[f*] = \min_{(f \in F')} R[f] \tag{37}$$

Dada una clase $F' = \{f\,(z), z \in \mathbb{Z}\}$ de funciones reales y una medida de probabilidad $F_{\mathbb{Z}}(z)$ se define el riesgo, R: $F' \rightarrow \mathbb{R}$, como:

$$R\,[f] = \int_z^{\cdot} c\,(z, f\,(z))\,d\,F_{\mathbb{Z}}\,(z) \tag{38}$$

donde c (-,-) se denomina función de perdida (o de coste) y tomará valores no negativos.

A la vista de la figura (8) se llega a la conclusión que los valores y_i e $y^*{}_i$ no necesariamente coinciden. Cuando esto sea así, la máquina de aprendizaje habría cometido un error que se debe cuantificar de alguna forma y este es precisamente el sentido que tiene la función de pérdida.

En este planteamiento, dado un conjunto $\{(x_i, y_i), \ldots, (x_n, y_n)\}$, el principal problema consiste es formular un criterio constructivo para elegir una función de F' puesto que la ecuación (38) por

sí mismo no sirve como criterio de selección, ya que la función $F_{\mathrm{Z}}(z)$ incluida en él es desconocida.

Dado un riego definido por (38), un conjunto de funciones F' y una muestra $\{z_1, \dots, z_n\}$. Al funcional R_{emp}: $F' \rightarrow \mathbb{R}$ definido como:

$$R_{emp}[f] = \frac{1}{n}\sum_{i=1}^{n} c(z_1, f(z_1)),\ f \in F' \tag{39}$$

se le denomina riesgo empírico.

La forma clásica de abordar estos problemas es: si el valor mínimo del riesgo se alcanza con una función f_0 y el mínimo del riesgo empírico con f_n para una muestra dada de tamaño n, entonces se considera que f_n es una aproximación a f_0 en un determinado espacio métrico. El principio que resuelve este problema se denomina principio de minimización del riesgo empírico.

Este es el principio utilizado en los desarrollos clásicos por ejemplo cuando se plantea a partir de un conjunto de datos la regresión lineal mínima cuadrática. Se debe imponer algunas condiciones de regularidad a sus elementos, con un funcional $Q[f]$. Así en la elaboración del problema se debe buscar una adecuada relación entre la precisión alcanzada con un particular conjunto de entrenamiento, medido a través de $R_{emp}[f]$, y la capacidad de la máquina medida por $Q[f]$. Ello lleva a considera el problema de minimizar un riesgo regularizado, donde este se define para algún $\lambda > 0$ en la forma:

$$R_{reg}[f] = R_{emp}[f] + \lambda Q[f] \tag{40}$$

Indicar que, en las MSV el espacio de trabajo es

$$F' = \{f(x) = w.x + b, w \in \mathbb{R}^d, b \in \mathbb{R}\} \tag{41}$$

(Funciones lineales) y la restricción se impone sobre el parámetro w (González, 2003).

4.8.1. Vectores de soporte para regresión – SVR. Los resultados efectivos de las máquinas de soporte vectorial para problemas de clasificación se trasladaron para abordar problemas de regresión, donde existen dos clases: $\boldsymbol{\nu}$-regresión y $\boldsymbol{\varepsilon}$-regresión.

En este caso, la idea es seleccionar el hiperplano regresor que mejor se ajuste al conjunto de datos de entrenamiento. Ahora se dispone de clases para separar. La idea se basa en considerar una distancia margen ε, de modo que esperamos que todos los ejemplos se encuentren en una banda o tubo entorno a nuestro hiperplano, es decir, que disten una cantidad menor de ε del hiperplano. A la hora de definir el hiperplano sólo se consideran los ejemplos que disten más de ε de nuestro hiperplano. En este caso esos ejemplos serán los considerados como vectores soporte. (Martín, 2016, p. 15).

El algoritmo SVR se explica de la siguiente manera: según Thomas, Pillai, y Pal (2017):

Considere un conjunto de datos de entrenamiento $\{(x_1, x_1), (x_2, x_2), \dots, (x_n, x_n)\}$, donde $x_i \in R^d$, $y_i \in R$, $i = 1 \dots n$. Los datos de entrenamiento x_i del espacio de entrada X se asignan a un espacio de características Q tal como $\theta: x_i \to \theta(x_i)$ usando una función no lineal predefinida $\vartheta(x)$. Sea f la función lineal que tiene la forma como en la ecuación (42)

$$f(x) = w^t x + b = \langle w, x \rangle + b, \text{ donde } w \in X, b \in R, \langle .,. \rangle \textit{ denota producto punto} \quad (42)$$

4.8.1.1. ε-Regresión (ε-SVR). En la regresión e-SVR (Vapnik, 1995), la función $\boldsymbol{f(x)}$ se calcula de manera que es plana, pero al mismo tiempo tiene una desviación máxima de $\boldsymbol{\varepsilon}$. Por lo tanto, la banda de error permisible para la función es $[-\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}]$. La función $\boldsymbol{f(x)}$ alcanza la planitud cuando el valor de $\boldsymbol{w}$ es pequeño y para obtener el valor mínimo para $\boldsymbol{w}$ se obtiene la solución de norma mínima que es $\|\boldsymbol{w}\|^2 = \langle \boldsymbol{w}, \boldsymbol{w} \rangle$.De ahí que al reformular el problema como un problema de optimización convexa factible, obtenemos la ecuación (43) (Thomas et al. 2017).

$$\text{Minimizar: } \frac{1}{2}\|w\|^2 \text{ sujeto a } \begin{Bmatrix} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{Bmatrix} \tag{43}$$

La ecuación anterior puede reformularse de nuevo (Vapnik, 1995) para incluir restricciones no factibles del problema mediante la introducción de variables de holgura

$$\text{Minimizar: } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \text{ sujeto a } \begin{Bmatrix} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{Bmatrix} \tag{44}$$

donde C es la compensación entre error ε permisible y $\|w\|$. El objetivo es minimizar el valor de riesgo empírico dado como $E = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|_\varepsilon$ donde $|y_i - y_i'|_\varepsilon$ ε es una función de pérdida insensible como se muestra en la figura 9, tal que $|y_i - y_i'|_\varepsilon = \begin{Bmatrix} 0, |y_i - y_i'| \leq \varepsilon \\ |y_i - y_i'| - \varepsilon, de\ otra\ manera \end{Bmatrix}$, donde y_i y y_i' son valores objetivo y predichos, respectivamente. Esta fórmula es también conocida como regresión linear de pérdida insensible-ε (Cortes y Vapnik 1995, Vapnik 1998).

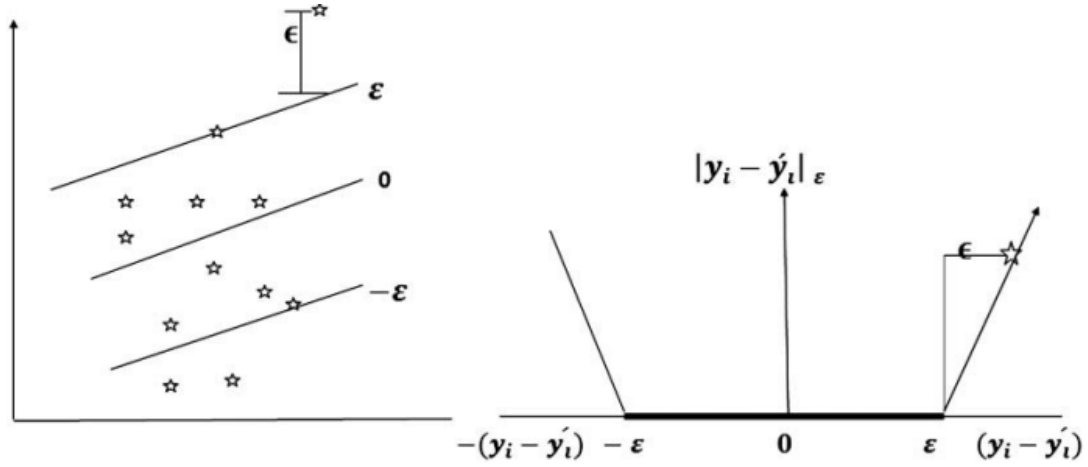


Figura 9. SVR con función de pérdida insensible- ε. Adaptado de (Thomas et al., 2017).

4.8.1.1. Regresión (ν-SVR). Otra versión de SVR, $\boldsymbol{\nu}$-SVR, fue propuesta por Schölkopf, Smola, Williamson y Bartlett (2000), que usa el parámetro $\boldsymbol{\nu}$ del rango [0, 1].

Es similar a $\boldsymbol{\varepsilon}$-SVR con $\boldsymbol{\varepsilon}$ considerada como un parámetro para tener un control sobre el conteo del vector de soporte. La formulación para $\boldsymbol{\nu}$-SVR es similar a ε-SVR, con un ligero cambio. Por lo tanto, la ecuación (44) se reformula para $\boldsymbol{\nu}$-SVR como:

$$\text{Minimizar: } \frac{1}{2}\|w\|^2 + C\nu\varepsilon + C\sum_{i=1}^{n}(\varepsilon_i + \varepsilon_i^*) \text{ sujeto a } \begin{Bmatrix} y_i - \langle w, x_i\rangle - b \leq \varepsilon + \varepsilon_i \\ \langle w, x_i\rangle + b - y_i \leq \varepsilon + \varepsilon_i^* \\ \varepsilon_i, \varepsilon_i^* \geq 0, \varepsilon \geq 0 \end{Bmatrix} \quad (45)$$

4.8.1.2. Núcleo o Kernel. Gunn S. R. (1998) explica.

El método que se puede usar para construir un mapeo en un espacio de características de alta dimensión mediante el uso de kernels reproductores. La idea de la función kernel es permitir que las operaciones se realicen en el espacio de entrada en lugar del espacio de características potencialmente de gran dimensión. Por lo tanto, el producto interno no necesita ser evaluado en el espacio de características. Esto proporciona una forma de abordar la maldición de la dimensionalidad. Sin embargo, el cálculo sigue siendo críticamente dependiente del número de patrones de entrenamiento y para proporcionar una buena distribución de datos para un problema de dimensiones elevadas generalmente se requerirá un gran conjunto de entrenamiento. Los kernel mencionados por David (2017), deben ser usado tanto en el entrenamiento de los datos como en la predicción.

- Kernel lineal: es el kernel más simple y está definido así: $K(x, x') = x * x'$ Cuando x y x' son dos vectores. Funciona mejor para problemas de clasificación
- Kernel polinomial: $K(x, x') = (\gamma(x * x') + coef0)^{degree}$
- Kernel Radial: $K(x, x') = e^{(-\gamma|x-x'|^2)}$, el parámetro gamma debe ser mayor que 0
- Kernel sigmoidal: $K(x, x') = (\tanh(\gamma\langle x, x'\rangle + coef0))$

4.8.1.3. Parámetros de ajuste en MSV. Thomas et al. (2017) refiere que $\boldsymbol{\varepsilon}$-SVR usa los parámetros C [0, ∞) y $\boldsymbol{\varepsilon}$ [0, ∞) para aplicar una penalización a la optimización para los puntos que no se predijeron correctamente. No hay penalidad asociada con los puntos que se predicen dentro de la distancia $\boldsymbol{\varepsilon}$ del valor real. Al disminuir $\boldsymbol{\varepsilon}$, se obtiene un mejor ajuste en los datos. En cuanto a $\boldsymbol{\nu}$-SVR, usa los parámetros C [0, ∞) y $\boldsymbol{\nu}$ [0, ∞).

El parámetro de penalización $\boldsymbol{\varepsilon}$ se reemplazó por $\boldsymbol{\nu}$. $\boldsymbol{\nu}$ representa un límite superior en la fracción de muestras de entrenamiento que son errores poco predecibles y un límite inferior en la fracción de muestras que son vectores de soporte. $\boldsymbol{\varepsilon}$ y $\boldsymbol{\nu}$ son versiones del parámetro de penalización.

Los otros dos parámetros utilizados son C (costo) y gamma. El costo representa la penalización asociada con errores mayores que épsilon. El aumento en el valor de costo les da una mejor adaptación a los datos. El parámetro gamma controla la forma del hiperplano de separación. El aumento de gamma generalmente aumenta el número de vectores de soporte.

David (2017) describen los parámetros usados en máquinas de soporte vectorial en el paquete "e1071" para el software R:

Costo: parámetro de penalización C del término de error, controla la compensación entre el límite de decisión uniforme y la clasificación correcta de los puntos de entrenamiento, aplica para los 4 kernel mencionados.

Épsilon: Épsilon en el modelo épsilon-SVR. Especifica el tubo épsilon dentro del cual no se asocia ninguna penalización en la función de pérdida de entrenamiento con los puntos predichos dentro de una épsilon de distancia del valor real, debe ajustarse en los 4 kernel anteriores.

Núcleo: Especifica el tipo de kernel que se usará en el algoritmo. Debe ser uno de lineal, 'polinomial', 'radial', 'sigmoide'. Si no se proporciona ninguno, se usará radial. Si se proporciona un invocable, se usa para calcular previamente la matriz del kernel.

Grado: Grado de la función del kernel polinomial. Ignorado por todos los demás núcleos.

Gamma: Coeficiente para Kernel radial, polinomial y sigmoidal. Si gamma es automático, entonces se usarán 1/ (dimensión de los datos), define hasta dónde llega la influencia de un único ejemplar de capacitación

Coef0: Término independiente para proyección en función del kernel. Solo es significativo en polinomial y sigmoidal.

4.9. Ventajas y desventajas de MSV y RF

4.9.1. Máquinas de Soporte Vectorial.

Chang, y Lin (2011) describen algunas de las ventajas y desventajas de las máquinas de soporte vectorial.

Ventajas

- Efectivas en espacios de alta dimensión.
- Sigue siendo efectivo en casos donde el número de dimensiones es mayor que el número de muestras.
- Utiliza un subconjunto de puntos de entrenamiento en la función de decisión (llamados vectores de soporte), por lo que también es eficiente desde el punto de vista de la memoria.
- Es versátil, es decir, se pueden especificar diferentes funciones del Kernel para la función de decisión. Se proporcionan núcleos comunes, pero también es posible especificar Kernels personalizados.

Desventajas

- Si la cantidad de características es mucho mayor que la cantidad de muestras, evite el sobre ajuste al elegir las funciones Kernel y el término de regularización es crucial.
- Las MSV no proporcionan estimaciones de probabilidad directamente, estas se calculan utilizando una costosa validación cruzada de cinco subconjuntos.

4.9.2. Random Forest. University of California, Departament of Statistics (s.f.), menciona algunas características que se clasifican como ventajas y desventajas del modelo matemático Random Forest.

Ventajas

- Es insuperable en precisión entre los algoritmos actuales.
- Se ejecuta de manera eficiente en grandes bases de datos.
- Puede manejar miles de variables de entrada sin eliminación de variables.
- Da estimaciones de qué variables son importantes en la clasificación y regresión.
- Genera una estimación interna no sesgada del error de generalización a medida que avanza la construcción del bosque.
- Tiene un método efectivo para estimar datos faltantes y mantiene la precisión cuando falta una gran proporción de los datos.
- Tiene métodos para equilibrar el error en los conjuntos de datos desequilibrados de la población de la clase.
- Los bosques generados se pueden guardar para su uso futuro en otros datos.
- Calcula las proximidades entre pares de casos que se pueden usar en la agrupación, la localización de valores atípicos o (por escalado) ofrece vistas interesantes de los datos.

- Las capacidades de lo anterior se pueden extender a datos no etiquetados, lo que lleva a agrupaciones no supervisadas, vistas de datos y detección de valores atípicos.
- Ofrece un método experimental para detectar interacciones variables.

Desventajas

- La correlación entre dos árboles en el bosque. Aumentar la correlación aumenta la tasa de error del bosque.
- La fuerza de cada árbol individual en el bosque. Un árbol con una baja tasa de error es un clasificador fuerte. Aumentar la fuerza de los árboles individuales disminuye la tasa de error del bosque.
- La reducción de mtry reduce tanto la correlación como la fuerza. Aumentarlo aumenta ambos. En algún punto intermedio hay un rango "óptimo" de mtry, generalmente bastante ancho. Este es el único parámetro ajustable al cual los bosques aleatorios son algo sensibles.

5. Estado del arte

Las ciudades y zonas urbanas en general son vulnerables a los desastres naturales debido a la concentración de la población y toda la infraestructura correspondiente. Yoon y Jeong (2016) comentan que a pesar de que las magnitudes de riesgo natural sean similares, las pérdidas económicas y muertes debido a esto se distribuyen de manera desigual entre las naciones, las regiones, las comunidades y los mimos individuos; donde las comunidades social, económica y ambientalmente vulnerables son más propensas a sufrir desproporcionadamente los desastres.

Agregan que identificar los factores de vulnerabilidad a los desastres, es información crítica que los administradores y planificadores de desastres deben tener clara para la elaboración de políticas y estrategias para mitigar el impacto negativo de estos eventos.

Montoya y Masser (2005), hacen referencia al aumento poblacional en las urbes de los países en desarrollo y a la necesidad de incorporar la gestión de desastres en sus agendas, creando redes efectivas y planes de mitigación específicos en cada caso, ya que un solo desastre puede destruir décadas de desarrollo.

Es por eso que las cadenas de suministro humanitarias buscan que las actividades logísticas llevadas a cabo antes y después de un desastre se realicen de tal manera que las poblaciones afectadas puedan satisfacer sus necesidades más básicas en primera instancia. Ngwenya y Naude (2016), el objetivo de la gestión de estas cadenas es buscar la excelencia logística coordinada. El flujo de información y la gestión de la demanda son dos factores claves que nombran los autores para lograr el éxito de cualquier cadena de suministro, y es precisamente allí donde hay deficiencias, ya que, por no contar con información veraz, estudios estadísticos previos y cifras correctas, se generan dificultades para determinar con precisión el número de personas afectadas, su ubicación geográfica y el tipo y cantidad de suministros de socorro necesarios (p. 2).

Debido a esa falta de información es que los desastres naturales y los causados por el hombre han afectado a miles de personas, cientos han desaparecido y otras perdieron sus vidas. En Colombia, por ejemplo, en el 2017 se presentó una avalancha en Mocoa (departamento del Putumayo), que dejó como resultado 333 personas muertas, 76 desaparecidas y 24.000 familias damnificadas (el colombiano, 2017). Esto deja ver la vulnerabilidad en la que están países emergentes y aún los desarrollados, ante los desastres naturales. Es por eso que en muchas ocasiones los desastres requieren de apoyo internacional como las fuerzas militares, grupos de rescate, etc., pero también, como lo dicen Kunz y Reiner (2012), de organizaciones no gubernamentales de socorro, que tienen los conocimientos y recursos para ayudar a las poblaciones afectadas por estas crisis.

Otro ejemplo reciente, es la cadena de terremotos ocurrida en México, donde tras tres eventos de este tipo (7, 19 y 23 de septiembre), 468 personas perdieron la vida, directa o indirectamente y más de 48.000 han recibido atención médica (Revista Proceso, 2017) y los daños de infraestructura supera los US$ 2.100 millones. (El Periódico, 2017). Hasta el momento 27 países, incluidas ONG's y empresas privadas, han brindado apoyo técnico, financiero o en especie, completando 440 toneladas en ayuda humanitaria (El Universal, 2017).

Por años se ha considerado que los terremotos son uno de los desastres naturales con mayor cantidad de víctimas, ya que por su naturaleza no pueden predecirse, esto sumado a la gran cantidad de daños estructurales que repercute en las poblaciones y a su vez, la mayor cantidad de muertes, se dan por la caída de escombros. (Kondratyev et al., 2006).

"En los últimos cuatro mil años, alrededor de 13 millones de personas han muerto como resultado de la actividad sísmica, y una cantidad desconocida de destrucción de la propiedad también ha ocurrido." (Okulewicz, 2017)

Según Van Wassenhove (2006, p. 475, 476), la logística puede marcar la diferencia al momento de mitigar los desastres, ésta comprende el 80% de la gestión de los mismos, por tal motivo se debe hacer una efectiva y eficiente gestión de la cadena de suministro para lograr una exitosa operación.

Cozzolino (2012), aborda las cuatro fases que conforman el ciclo de gestión de desastres y describe la incertidumbre y complejidad que caracteriza los esfuerzos de los organismos de socorro, por lo cual es importante contar con una gestión de desastres que contenga claros procesos estratégicos.

- Mitigación: Básicamente describe la responsabilidad del gobierno de organizar y estructurar las leyes y los mecanismos para reducir la vulnerabilidad de la población

Figura 10. Ciclo de gestión de desastres. Adaptado de (Cozzolino, 2012).

- Preparación: Basado en experiencias pasadas se hacen adaptaciones y mejoras a las estrategias que al implementarse su respuesta operativa será exitosa. Principalmente su objetivo es evitar las consecuencias más graves.
- Respuesta: Cozzolino (2012) divide esta fase en dos objetivos, siendo el primero responder de manera inmediata a la situación; y en el segundo se restablecen los servicios básicos y se hace entrega de bienes al mayor número de afectados.
- Reconstrucción: Cozzolino (2012) comenta que esta fase se trata de la rehabilitación de la población y de la región, dado que los efectos del desastre pueden tener consecuencias al largo plazo.

Thomas et al. (2017) explican los inconvenientes de utilizar el análisis de regresión para el desarrollo de ecuaciones predictivas, ya que la no linealidad y no homogeneidad entre las variables independientes afectan directamente los coeficientes de las variables independientes en la ecuación de regresión desarrollada. Ya que, en el análisis de regresión, el modelo se desarrolla en base a una ecuación lineal o no lineal predefinida, con la hipótesis de la normalidad de los residuos para probar el modelo desarrollado. Por lo tanto proponen utilizar técnicas más nuevas para reducir los errores existentes en la estimación del parámetro de movimiento en el suelo, así que decidieron utilizar las máquinas de soporte vectorial para regresión (SVR), usando tres algoritmos de aprendizaje ε-SVR, ν-SVR y Ls-SVR para pronosticar la aceleración pico de terreno (PGA), un parámetro asociado con el movimiento de una señal sísmica. Obteniendo con modelo de predicción de Ls-SVR y kernel radial propuesto mejores resultados en comparación con todos los modelos de predicción de parámetros de movimiento terrestre convencionales.

Por otro lado, Rodríguez et al. (2011), proponen un "Sistema para la Evaluación y Diagnóstico de Desastres" (SEDD), basado en metodologías de aprendizaje automático, las cuales, con conjuntos de datos de fácil acceso, realizan una evaluación de las consecuencias en un escenario de desastre dado, ya sean heridos, muertos y victimas en general; lo cual permite a las organizaciones tomar decisiones rápidas y en tiempo real, con el fin de evitar que dichas consecuencias se incrementen a medida que pasa el tiempo.

Inicialmente este sistema fue diseñado para implementarse en la fase de respuesta, ayudando a los responsables de las distintas organizaciones a tomar decisiones acertadas, enfocadas en la estimación y evaluación de las consecuencias, clasificando la gravedad del escenario de desastre tan preciso como fuese posible, con el fin de aumentar la probabilidad de éxito en las operaciones.

Naturalmente este tipo de decisiones se toman bajo una presión muy alta y por sobre todas las cosas se debe priorizar la vida y seguridad de las personas que se encuentran en la zona del evento.

Según los autores, es posible utilizar este tipo de sistemas para apoyar el diseño de políticas para la mitigación de desastres, es decir, que las fases de preparación y mitigación pueden ser cubiertas sin mayor contratiempo debido a que se pueden evaluar diferentes escenarios.

Una de las grandes ventajas de este tipo de metodologías (SEDD), es que permite utilizar datos de eventos históricos con el fin de poder deducir posibles patrones en caso de desastres y así acercarse más a la realidad de un evento futuro, teniendo en cuenta características socioeconómicas y demográficas para robustecer las bases de datos de sismos y apuntarle a la realidad actual de la región donde pueda ocurrir un desastre natural.

Los autores resaltan con las pruebas realizadas que, es fácil la interpretación de los datos que un modelo alterno basado en reglas difusas provee, si se compara con técnicas estadísticas ordinarias, por ejemplo: regresiones lineales, intervalos de predicción, análisis discriminante, entre otros. Por tal motivo estos modelos al ser más flexibles y precisos, generarían una confianza mayor en las organizaciones no gubernamentales para hacer evaluaciones rápidas y correctas de las consecuencias de los desastres.

Las máquinas de soporte vectorial han sido utilizadas por años en gran cantidad de campos de acción como la medicina, la ciencia, la tecnología, la seguridad, la meteorología, entre otros; lo cual muestra la versatilidad y según los resultados reportados en cada uno de ellos, el resultado con las MSV suele sobresalir comparándolo con otros métodos. Por ejemplo, en Pedroza (2007) es aplicado para reconocimiento de voz, aprovechando su robustez y eficiencia lograron obtener un resultado superior a los Modelos de Mezclas Gaussianas y a las Redes Artificiales Neuronales.

Es el caso también de Pérez (2014), quien opta por este tipo de método, ya que al ser una clasificación de datos supervisada trabaja con datos ya existentes y en el caso de su tesis doctoral, utiliza los datos de descarga de un dispositivo de fusión nuclear, pero al ser tan soberbia la cantidad de datos por descarga decide inclinarse por este modelo que permiten agrupar muestras ya que se conocen las clases de clasificación. También enfatiza en su gran eficiencia en modelos de predicción y nombra características como el uso de un kernel o función de transformación, el cual busca la minimización del riesgo estructural. Señala que el modelo generado va a depender de los datos con mayor información, lo cual es conveniente en el proceso de entrenamiento de la máquina así se trabaje con un número elevado de atributos, ya que la parametrización es mínima.

Gutiérrez (2007), utiliza las máquinas de soporte vectorial para clasificar imágenes utilizando el algoritmo de optimización mínima secuencial (SMO por sus siglas en inglés), ya que este descompone el problema en tareas más pequeñas con el fin de facilitar las operaciones matriciales reduciendo las exigencias de recursos como memoria y procesador. Este algoritmo también entrena con mayor rapidez la máquina mediante programación dinámica de forma secuencial, dice el autor.

Lin et al. (2017) usaron datos de deslizamientos en la región sudeste de Taiwán entre los años 2008 y 2011, para que a través de regresiones lineales y entrenamiento de máquinas de soporte vectorial se preparan mapas lo más preciso posibles para determinar la susceptibilidad a los eventos, buscando así, cuantificar los daños y víctimas y reducirlos a través de planes de mitigación con la información suministrada con los modelos. En las pruebas encontraron que factores claves como: la pendiente de la región, el índice de humedad topográfica, distancia al río, entre otras, son capaces de determinar cuántos pixeles se deslizamiento de tierra habría cuando se presentan lluvias en máximo 48 horas seguidas. Como resultado se demostró que las MSV aplicada con los Kernel

polinomial y radial, tuvieron precisiones superiores a los otros modelos, lo cual demuestra la sensibilidad de las MSV para proveer información y medias más completas.

Dentro de las máquinas de aprendizaje inteligente se encuentra Random Forest (RF), la cual se desenvuelve eficientemente en grandes bases de datos y provee estimaciones importantes de variables específicas en la clasificación y resuelve la relación multivariable y no lineal. Wang et al. (2015).

Blanco, Alonso y Gomariz (2014) plantean la creación de mapas de suelos a través de métodos estadísticos o de aprendizaje automático, como el algoritmo Random Forest y las máquinas de soporte vectorial, usando como variables cualitativas el tipo y uso del suelo, y como variables cuantitativas, la posición topográfica, el clima, el índice de rugosidad del terreno, entre otras. Debido al gran número de variables optaron por usar el método del factor de inflación de la varianza (VIF), el cual calcula un estadístico para cada variable de tal manera que resume el vector correspondiente, buscando el de mayor colinealidad. Luego de usar 1150 datos para calibrar los modelos, el resultado mostro que con los métodos RF y MSV se obtienen mejores resultados ya que muestran una menor incertidumbre, 0.57 y 0.89, respectivamente. Concluyen pues que "… los métodos más flexibles basados en aprendizaje automático han obtenido resultados considerablemente mejores que los métodos estadísticos más clásicos"

Sabemos que los desastres naturales son inevitables pero los seres humanos se adaptan a los cambios, buscando la forma de que los daños y pérdidas se reduzcan y eso es lo que pretenden Wang et. al. (2015) al usar una máquina Random Forest para estimar, a través de amplias bases de datos, el riesgo de inundaciones en la cuenca del río Dongjiang en China, el cual se mide generalmente por la probabilidad de ocurrencia; ya que este es de los riesgos más comunes alrededor del mundo, el cual ha causado más de 7 millones de muertes en el último siglo y más de

US $600 millones en pérdidas. Uno de los problemas críticos con los que se enfrentaron fue la no linealidad entre índices y niveles de riesgo por lo cual un método fuese tolerante a los valores atípicos y al ruido sin dejar de ser preciso en sus pronósticos, por lo cual, usaron RF para su análisis, encontrando así, que las zonas con mayor riesgo eran aquellas en las que habían mayor precipitación y por lo general tierras bajas y planas donde se concentraba toda el agua lluvia; pero para confirmar y validar el análisis, se realizó un mapa con datos históricos de inundaciones en la región y se sobrepuso con el mapa arrojado por el estudio inicial, lo cual demostró que se ubicaban muy cerca de las zonas críticas.

Chen, Yu y Li (2017), en consecuencia, hablan de otro de los desastres que se presenta en todo el globo, son los deslizamientos de tierra, causado tanto por la misma naturaleza y geografía de las distintas regiones, como por las alteraciones de esta por parte del hombre; el hecho de deforestar regiones, abrir carreteras, crear industrias y afectar la hidrografía, son aspectos que están agravando los deslizamientos. Es por eso que la supervisión oportuna de deslizamientos puede ayudar a evitar efectos no deseados en las poblaciones vulnerables y a comprender el mecanismo de las cadenas peligrosas. Gracias al desarrollo y los avances de la teledetección se han podido superar obstáculos para el desarrollo de este tipo de estudios, ya que los costos para fotografiar las regiones son muy elevados y el uso de modelos de aprendizaje automático se hace más factible, como el modelo Random Forest, que gracias a que es una de los más rápidos para entrenar y que tiene un rendimiento robusto, es de gran ayuda para este tipo de casos. Por ende, los autores lo usaron para entrenarse con 26.000 contornos conectivos, luego de una selección de 500 árboles, generando un modelo de detección de deslizamientos que funciona automáticamente y que es aplicable a gran escala enfatizándose esencialmente en la textura de cada contorno

Para gestionar este tipo de riesgo, Zhang, Wu, Niu, Yang y Zhao (2017), afirman que, para prevenir y reducir los deslizamientos, se debe contar con cartografía de susceptibilidad de deslizamiento lo más detallada posible. Su estudio se realizó en el área de Zigui-Badong en la región de las Tres Gargantas de China. Donde existen altas montañas y valles muy profundos y donde priman las tierras cultivadas, boscosas, edificadas y además cuenta con un embalse, estos dos últimos usos han hecho que la actividad, frecuencia y tamaño de los desastres aumenten en esta región.

Usaron el modelo Random Forest mediante la integración de datos de varias fuentes que estudiaban el mismo tema y los resultados del experimento se compararon con deslizamientos existentes, donde el 70% del conjunto de datos fue usado para el entrenamiento del modelo y el 30% se usó para las pruebas de validación, arrojando finalmente que un 77.12% de las celdas de red de deslizamiento están contenidas en regiones peligrosas y altamente peligrosas, con una precisión de predicción del 86.1% los datos se pueden considerar creíbles.

Asim, Martinez, Basit y Iqbal (2017) cuentan cómo en la región de Hindukush, una de las más propensa a terremotos, usan técnicas de aprendizaje automático para intentar predecir alguna de las características de terremotos como la ubicación exacta, la magnitud, el lapso de ocurrencia y la probabilidad de ocurrencia; dado que en la actualidad no se cuenta con ningún modelo capaz de predecir desastres, aun cuando hay autores que consideran esta idea como imposible.

Para su estudio usan cuatro modelos: Redes Neuronales, Random Forest, Red Neuronal de Reconocimiento de Patrones y Programación Lineal Aumentada; y consideran los eventos sísmicos registrados en el Centro de Estudios de Terremotos y el Servicio Geológico de los Estados Unidos, desde 1979 hasta 2013 para un total de 1.137 eventos con una magnitud igual o superior a 4.0. También usan parámetros como el tiempo, la magnitud media, la energía liberada, entre

otros; para entrenar los modelos antes de la predicción en tiempo real ya que son modelos supervisados.

El resultado de los cuatro modelos arroja una media de 76.5% de exactitud en la predicción de un evento sísmico, aunque por ejemplo el modelo PRNN muestra una menor sensibilidad hacia las concurrencias sísmicas también toma ventaja en especificidad (tasa de negativos reales previsto) y el LPBoost muestra mayor precisión y sensibilidad, pero menor especificidad.

A pesar de los resultados, los autores no consideran que el estudio es 100 % preciso pero anuncia que es un paso prometedor con resultados motivadores para continuar con el mejoramiento del modelo ya que al final, lo que se quiere buscar es evitar la mayor cantidad de efectos negativos sobre las poblaciones que puedan sufrir un evento catastrófico.

Otro de los usos del aprendizaje automático a través de modelos de predicción como el algoritmo Random Forest se ve en la estimación que las empresas de servicios públicos y lo entes gubernamentales hacen para precisar la duración de los cortes de energía causados por huracanes antes de tocar tierra, ya que el funcionamiento de estos depende en gran parte del fluido eléctrico. Es por eso que Nategui, Guikema y Quiring (2014), buscan identificar las variables clave para predecir estos apagones y su grado de influencia en la restauración del servicio.

Lo primero que definieron, fueron aquellas variables que se predicen antes de que ocurra una tormenta y aquellas que no. Como variables predecibles se consideran las características de los huracanes, es decir la velocidad del viento y la duración del mismo y que sean superiores a 44.7 millas/h, la información climática del área geográfica, niveles de humedad del suelo a diferentes profundidades, entre otras. Luego de definir dichas variables, crearon un conjunto de entrenamiento con los datos seleccionados para luego ajustar un árbol de regresión entrenando el conjunto de datos previamente seleccionados al azar, probando así el error predictivo y por último

definieron para el análisis 500, como la cantidad óptima de árboles para garantizar un resultado más preciso, ya que se cuenta con una cantidad de datos complejos y la técnica de baja polarización del algoritmo RF proporciona una mayor precisión y una mayor robustez para el ruido y los valores atípicos, adicionalmente es más rápido computacionalmente hablando, que otros métodos, lo cual ayuda a reaccionar de manera ágil en caso que se presente una tormenta o un huracán.

Lee, Sameen, Predham y Park (2017), evaluaron 5 modelos para seleccionar un enfoque adecuado para la modelización de la susceptibilidad de deslizamiento de tierra en entornos donde no existieran suficientes datos, en este caso solo se analizaron 418 datos de deslizamientos con 18 factores de acondicionamiento, entre ellos: la altitud, ángulo de inclinación, aspecto de pendiente, curvatura de planta, rugosidad del terreno; entre otros. Los modelos fueron evaluados para el área de Sanju, localizada al norte de la provincia de Gyeongsang en Corea del Sur.

Las lluvias intensas y los terremotos son los principales responsables de los deslizamientos de tierra, cuyas áreas se pueden identificar y evaluar con base en análisis científico y/o modelado de susceptibilidad a deslizamiento, siendo un tipo básico de análisis empleado para la evaluación del riesgo.

En primer lugar, optaron por corregir los problemas de multicolinealidad en los factores de deslizamiento, eliminando, mediante el estadístico chi-cuadrado, las observaciones problemáticas y factores no significativos. Usaron el método Bland-Altman para comprender la repetibilidad de los modelos de predicción a diferentes escalas, con varios conjuntos de datos; para establecer el acuerdo entre dos mediciones cuantitativas mediante la construcción de límites de acuerdo, los cuales se calculan utilizando la media y la desviación estándar (s). Realizaron el método con el conjunto completo de datos y luego con una porción de ellos para cada uno de los modelos de predicción.

Como resultado y con una alta precisión (85%), lograron identificar que MSV llegaban a este valor utilizando todo el conjunto de datos para su entrenamiento. Y para los modelos Random Forest y Peso de Evidencia (WoE), - por sus siglas en inglés – encontraron que el acuerdo era más débil en comparación de los otros modelos en función de los valores de diferencia de medias y las desviaciones; todo esto con un nivel de confianza del 95%.

6. Tratamiento de datos

6.1. Base de datos

La base de datos utilizada en este proyecto se fundamentó en los datos suministrados por "Los centros nacionales de información ambientar" (NOAA), por sus siglas en inglés, quienes preservan, monitorean, evalúan y brindan acceso a datos geofísicos a través del intercambio internacional, los datos recopilados hasta octubre del 2017 en el Banco Mundial de Desarrollo y las ubicaciones de los tres nidos sísmico existentes, lo cual nos arroja casi 6000 datos brutos.

6.1.1. Variables predictoras.

Las variables a tratar son:

Latitud: Distancia angular que hay desde un punto de la superficie de la tierra hasta el paralelo del ecuador, valores válidos 0 a 90 (hemisferio norte) -90 a 0 (hemisferio sur).

Longitud: Distancia angular de un punto de la superficie terrestre al meridiano de Greenwich, valores válidos 0 a 180 (hemisferio oriental) -180 a 0 (hemisferio occidental) .

Profundidad: Profundidad del terremoto medida en kilómetros.

Magnitud: Valor de la magnitud del terremoto primario, valores válidos de 0.0 a 9.9.

Intensidad: Basada en la escala sismológica de Mercalli, (The Modified Mercalli Intensity Scale, s.f.) es el efecto de un terremoto en la superficie de la tierra y cuenta con 12 grados para

evaluar la intensidad de los terremotos a través de los efectos y daños causados a distintas estructuras (Escala sismológica de Mercalli, s.f.).

Distancia a Nido 1: Bucaramanga, Colombia (Lat. 6.7545 - Long. -73.0281).

Distancia a Nido 2: Vrancea, Rumanía (Lat. 45.917 - Long. 26.54).

Distancia a Nido 3: Hindu Kush, Afganistán (Lat. 36.00 - Long. 73.00).

Densidad Poblacional: La densidad de población se define como la población a mitad de año dividida por la superficie territorial en kilómetros cuadrados. La población se basa en la definición de facto de la población, que incluye a todos los residentes independientemente de su estado legal o de ciudadanía, con excepción de los refugiados no asentados permanentemente en el país de asilo, que suelen considerarse parte de la población del país de origen. (Banco Mundial, s.f.).

Población Urbana (% del total): La población urbana se refiere a las personas que viven en áreas urbanas según lo definido por las oficinas nacionales de estadística. Los datos son recopilados y suavizados por la División de Población de las Naciones Unidas. (Banco Mundial, s.f.).

IDH: Indicador desarrollado por la Naciones Unidas con el fin de medir el desarrollo de un país, para este fin se analiza lo siguiente:

Salud (esperanza de vida al nacer): La esperanza de vida al nacer se mide en el IDH utilizando un valor mínimo de 20 años y un valor máximo de 83.57. De forma que, por ejemplo, el componente de longevidad para un país cuya esperanza de vida al nacer sea de 55 años vendrá a ser de 0,551.

Educación: Se mide a través de los años de escolarización para adultos y los años de escolarización previstos para niños y niñas en edad escolar.

El componente de ***riqueza (o estándares de vida digna)*** se mide a través del INB per cápita ($PPP) en lugar del PIB per cápita ($PP) como se hacía anteriormente. Los límites mínimo y máximo son 100$ (PPP) y 87,478$ (PPP). (Expansión, s.f.).

Índice de GINI: El índice de Gini mide hasta qué punto la distribución del ingreso (o, en algunos casos, el gasto de consumo) entre individuos u hogares dentro de una economía se aleja de una distribución perfectamente equitativa. Una curva de Lorenz muestra los porcentajes acumulados de ingreso recibido total contra la cantidad acumulada de receptores, empezando a partir de la persona o el hogar más pobre. El índice de Gini mide la superficie entre la curva de Lorenz y una línea hipotética de equidad absoluta, expresada como porcentaje de la superficie máxima debajo de la línea. Así, un índice de Gini de 0 representa una equidad perfecta, mientras que un índice de 100 representa una inequidad perfecta. (Banco Mundial, s.f.).

Inflación, precios al consumidor (% anual): La inflación medida por el índice de precios al consumidor refleja la variación porcentual anual en el costo para el consumidor medio de adquirir una canasta de bienes y servicios que puede ser fija o variable a intervalos determinados, por ejemplo, anualmente. Por lo general se utiliza la fórmula de Laspeyres. (Banco Mundial, s.f.).

Acceso a mejoras sanitarias: El acceso a mejoras en las instalaciones sanitarias se refiere al porcentaje de la población con un acceso al menos adecuado a instalaciones de desecho de excreciones que puedan evitar eficazmente el contacto de humanos, animales e insectos con las excreciones. Las mejoras en las instalaciones van de letrinas sencillas pero protegidas hasta baños con descarga y conexión cloacal. Para que sean eficaces, las instalaciones deben construirse correctamente y someterse a un mantenimiento adecuado. (Indexmundi, s.f.).

Gasto en investigación y desarrollo: Los gastos en investigación y desarrollo son gastos corrientes y de capital (público y privado) en trabajo creativo realizado sistemáticamente para

incrementar los conocimientos, incluso los conocimientos sobre la humanidad, la cultura y la sociedad, y el uso de los conocimientos para nuevas aplicaciones. El área de investigación y desarrollo abarca la investigación básica, la investigación aplicada y el desarrollo experimental. (Banco Mundial, s.f.).

6.1.2. Variables respuesta.

Muertos: valores de 0 a 1,100,000 (siempre que sea posible se enumera las cifras de muertos)

Heridos: valores de 0 a 30,000 (siempre que sea posible se lista el número de heridos)

Daños: estos valores se multiplican por 1,000,000 para obtener el monto en dólares (de EEUU).

6.2. Imputación de datos

Partiendo de las diferentes variables predictoras se realizó un análisis de datos con el complemento XLSTAT (software estadístico para Excel), para determinar cuáles variables se deben imputar y con qué método, revisando también la cantidad de datos que había por cada variable; lo cual permitió definir dos posibles escenarios con el fin de comparar el comportamiento de los modelos con cada una de ellas.

Con el fin de tener una mayor precisión en la imputación de los datos se realizó el procedimiento para cada uno de los países por separado, teniendo en cuenta, solamente los variables con más de 3 datos (parámetro exigido por el software).

Se realizó una imputación de cuatro variables predictoras teniendo en cuenta la cantidad de datos dando prioridad a las que presentaban la menor cantidad de datos (base de datos 1), también se imputaron tres variables más además de las anteriores (base de datos 2), esto se explicara a continuación:

6.2.1. Base de datos con cuatro variables imputadas (base de datos 1). En esta base de datos se realizó la imputación del índice de desarrollo humano y acceso a mejoras sanitarias, usando el método de imputación por regresión, ya que hay un patrón definido de datos faltantes debido a que la información recopilada por el banco mundial de desarrollo es limitada desde la década de los 60 hasta la actualidad.

Para las siguientes dos variables imputadas, densidad poblacional y población urbana, se usó la regresión exponencial debido a que la explicación de la variable Y no era lineal con respecto a la X, por lo cual dicha regresión arrojó valores más acordes a los reales.

La cantidad de observaciones sin datos vacíos para cada variable respuesta fue la siguiente:

- Muertos = 261
- Herido = 282
- Daños = 129

6.2.2. Base de datos con siete variables imputadas (base de datos 2). Esta base de datos complementa a la anterior con la imputación de las variables profundidad, magnitud e intensidad, características típicas de los eventos sísmicos, donde por países realizamos la imputación teniendo en cuenta la media de cada variable, apreciación empírica optada por los autores.

En este caso las observaciones sin datos vacíos aumentaron así:

- Muertos = 722
- Herido = 757
- Daños = 324

6.3. Correlación de variables predictoras

A partir de la base de datos después de la imputación de datos se procedió a realizar un estudio de correlaciones entre variables predictores con ayuda del software RStudio, con el fin de observar la correlación gráfica y fácilmente tanto para las diferentes bases de datos se hace uso de mapas de calor, es importante especificar que en dichas graficas un color rojo significa una correlación nula, es decir, cero, mientras que un color amarillo muestra una correlación elevada de uno.

6.3.1. Correlación base de datos 1. Como se ha mencionado anteriormente, se cuenta con una base de datos, la cual tiene catorce variables predictoras, en la figura 11 se observa la correlación entre dichas variables.

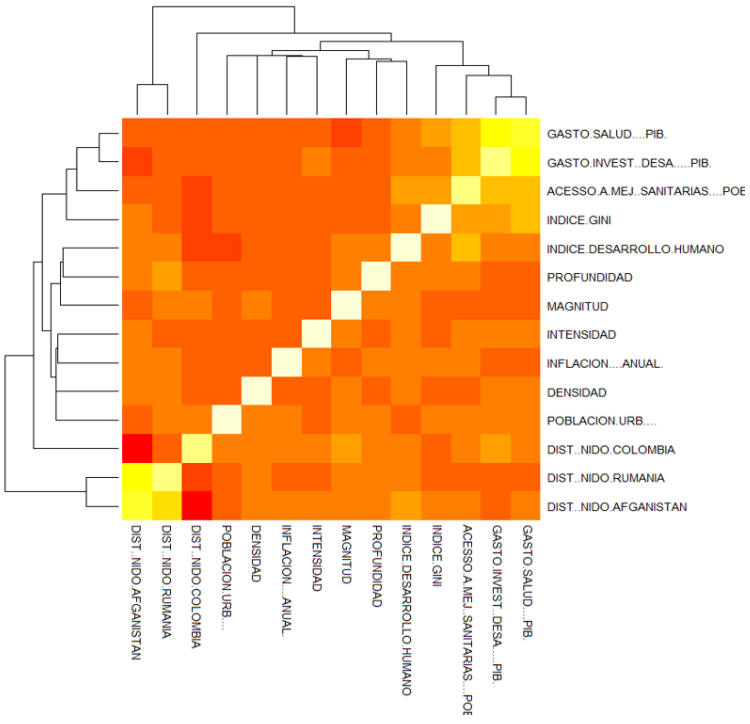


Figura 11. Correlación 1 para la base de datos 1.

Se observa en la figura 11 una alta correlacion entre las distancias de los nidos sismicos de Rumania y Afganistan, por lo tanto se planteó reducir estas tres variables (distancia a cada uno de

los 3 nidos sísmicos) a una sola que realacione las tres anteriores como la menor distancia a un nido sismico, es decir, tomar la menor distancia desde cualquier punto hasta cualquiera de los tres nidos simicos existentes; de acuerdo a esto se realiza el cambio y se decide eliminar la variable Indice de Gini debido a su limitante de cantidad de datos con el fin de evitar una restriccion futura, asi se decide realizar de nuevo un mapa de calor con estas variables y se obtiene la figura 12.

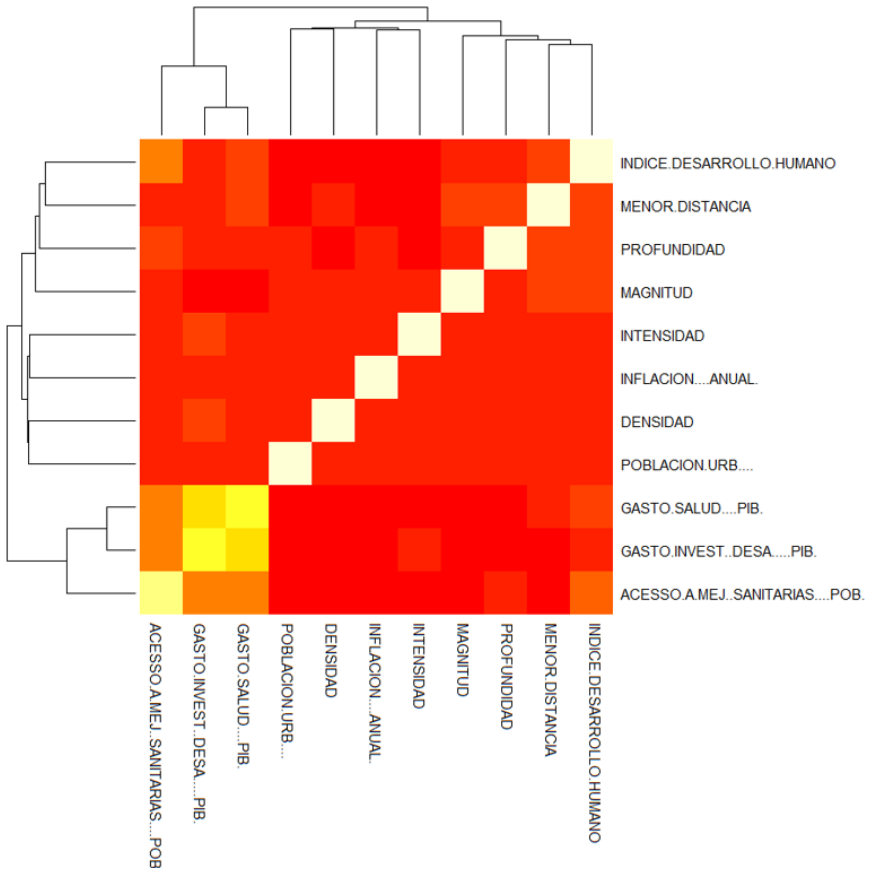


Figura 12. Correlación 2 para la base de datos 1.

A partir de la figura 12, se denota una correlación importante entre las variables gasto en investigación y desarrollo - gasto en salud pública, acceso a mejoras sanitarias – gasto en investigación y desarrollo y en acceso a mejoras sanitarias – gasto en salud pública que tenían una correlación de 0.7461, 0.4650 y 0.408 respectivamente, por lo tanto, se decide realizar de nuevo

un análisis, esta vez sin las variables gasto en salud pública y gastos en investigación y desarrollo, se prefiere mantener acceso a mejoras sanitarias por encima de gasto en salud pública debido a la cantidad limitante de datos, obteniendo como resultado la figura 13.

En esta figura se obvia la correlación entre las variables, existiendo una ligera correlación entre el índice de desarrollo humano y el acceso a mejoras sanitarias por debajo de 0,5, con este proceso implementado se procede a la construcción de los modelos con las variables representadas anteriormente sin problemas de correlación alta. Dejando así definitivamente las 9 variables predictoras: Profundidad, magnitud, intensidad, menor distancia a un nido sísmico, densidad poblacional, población urbana (%), índice de desarrollo humano, inflación y acceso a mejoras sanitarias.

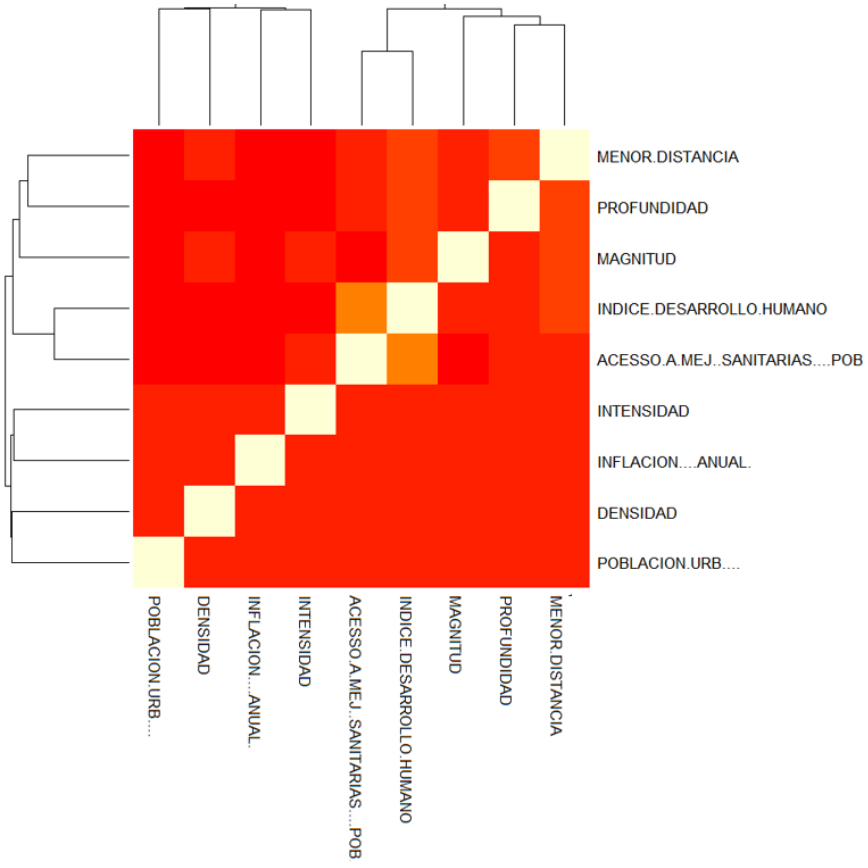


Figura 13. Correlación 3 para la base de datos 1.

Cabe tener en cuenta que el análisis anterior se hizo aun con datos faltantes, es decir con entradas de datos con el valor cero, por lo tanto, existe la posibilidad de que la correlación cambie al contar únicamente con las entradas de datos completos, en razón a esto, se decide analizar nuevamente la correlación y si es el caso prescindir de alguna variable, este proceso se realiza tanto para muertos, heridos y daños, ya que sus datos no son los mismos.

En las figuras 14a, 14b y 14c se denota una correlación alta que involucra las variables población urbana, índice de desarrollo humano y acceso a mejoras sanitarias con coeficientes de correlación por encima de 0.7, de decide eliminar las variables población urbana y acceso a mejoras sanitarias debido a la importancia del índice de desarrollo humano.

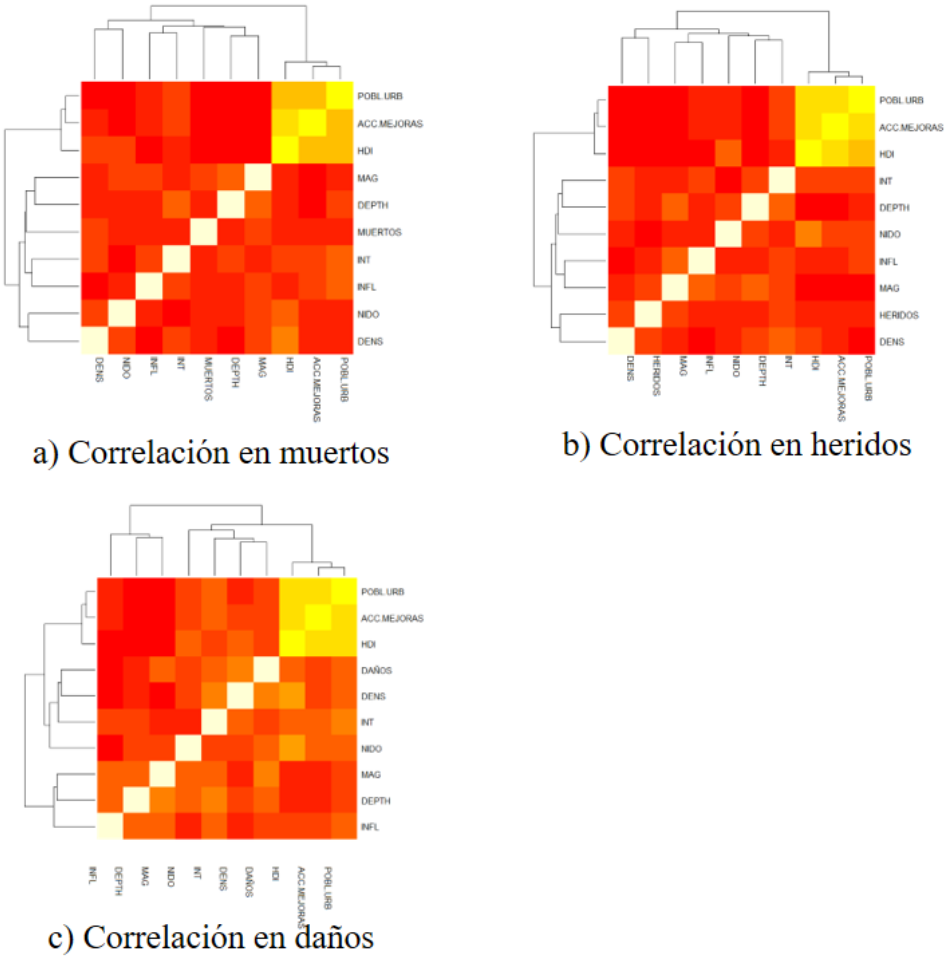
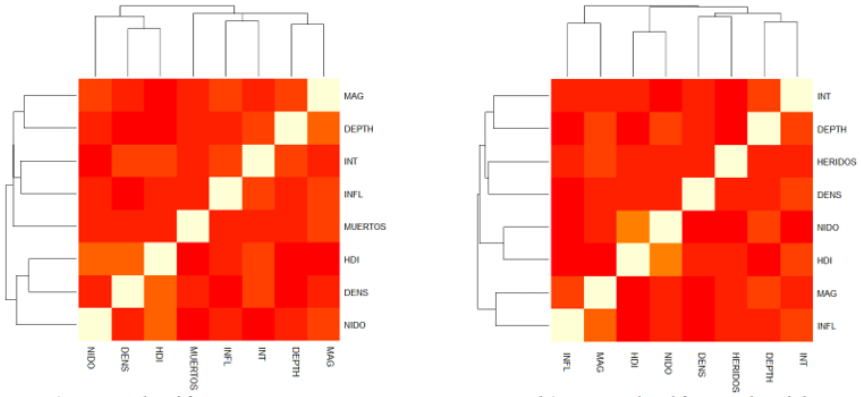
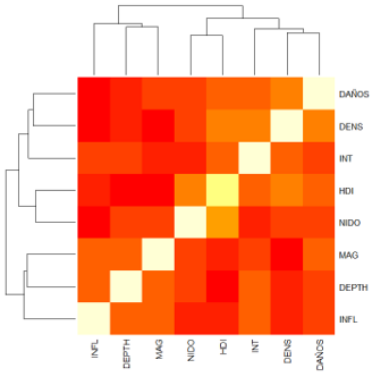


Figura 14. Correlación 4 para la base de datos 1 para cada variable respuesta.



a) Correlación en muertos

b) Correlación en heridos



c) Correlación en daños

Figura 15. Correlación definitiva para cada variable respuesta.

Las figuras 15a, 15b y 15c representan el mapa de calor de la correlación de las variables predictoras definitivas para la construcción de los diferentes modelos de predicción en cuanto a la base de datos que cuenta con cuatro variables imputadas.

6.3.2. Correlación base de datos 2. El proceso realizado anteriormente se realiza también para esta base de datos, sin embargo, no se tiene en cuenta el Índice de Gini como variable inicial por la cantidad de datos y las distancias de los nidos sísmicos representados en una sola variable como la menor distancia, se procede a calcular la correlación entre las variables resultantes y se obtiene la figura 16.

En la figura 16 se visualiza una correlación importante entre las variables acceso a mejoras sanitarias – índice de desarrollo humano y gasto en salud pública - gastos en investigación y

desarrollo con coeficientes de 0,8785 y 0,7460 respectivamente, de decide descartar la variable acceso a mejoras sanitarias debido a dicha correlación y la importancia del índice de desarrollo humano, eliminan también gasto en investigación y desarrollo por su coeficiente y gasto en salud debido a la limitación de datos, para así obtener la figura 17.

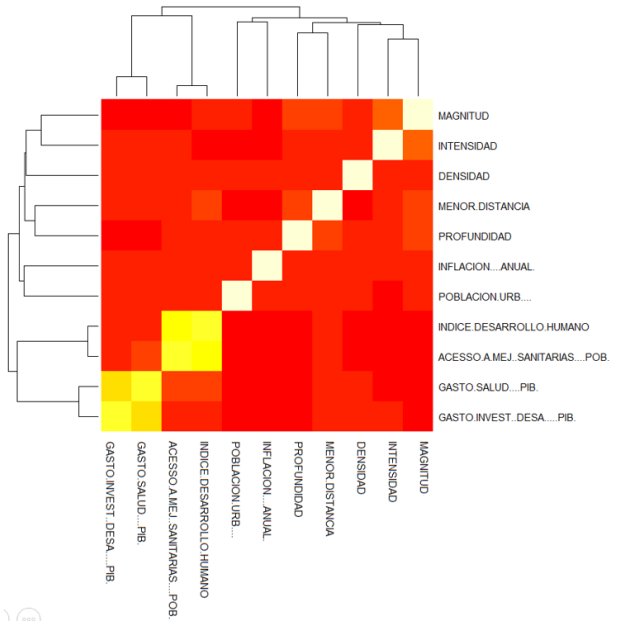


Figura 16. Correlación 1 para la base de datos 2.

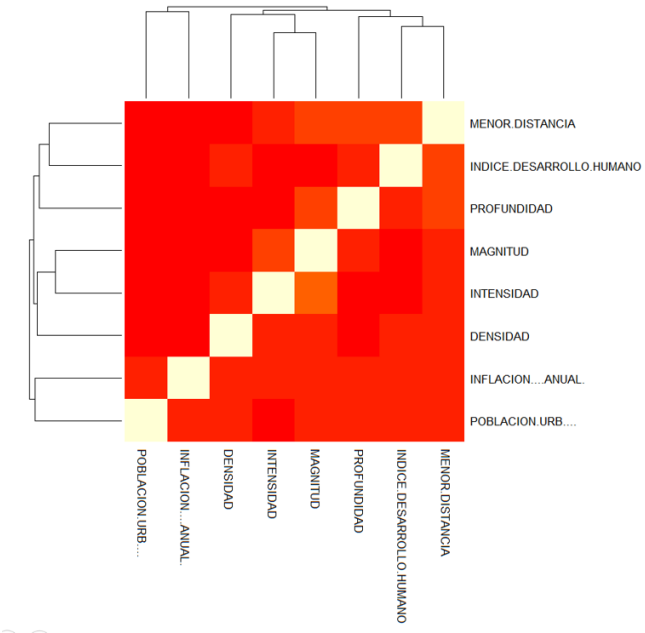


Figura 17. Correlación 2 para la base de datos 2.

En la anterior figura se observa que ninguna de las variables tiene alguna incidencia o correlación con las demás, por lo tanto, se precede a realizar los modelos correspondientes. Lo cual permite obtener en definitiva 8 variables predictoras que son: Profundidad, magnitud, intensidad, menor distancia a un nido sísmico, densidad poblacional, población urbana (%), índice de desarrollo humano e inflación.

Al igual que en el apartado anterior, es necesario realizar un nuevo análisis sobre las correlaciones al quitar las entradas que eran ceros o valores nulos, tal hecho arrojó como resultado la figura 18.

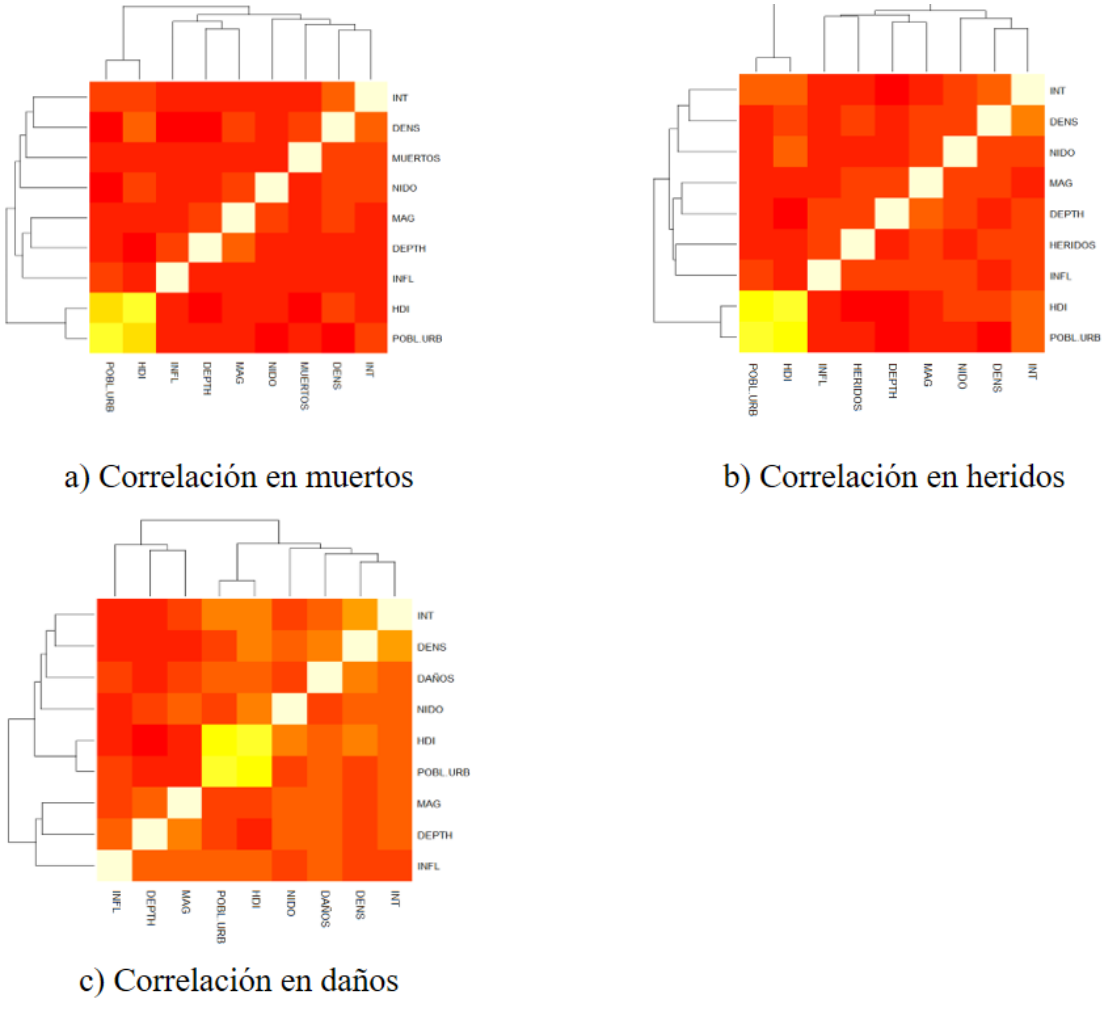


Figura 18. Correlación 4 para la base de datos 2 para cada variable respuesta.

En las figuras 18a, 18b y 18c es notoria la correlación entre las variables índice de desarrollo humano y población urbana, por lo tanto, se elimina la segunda variable.

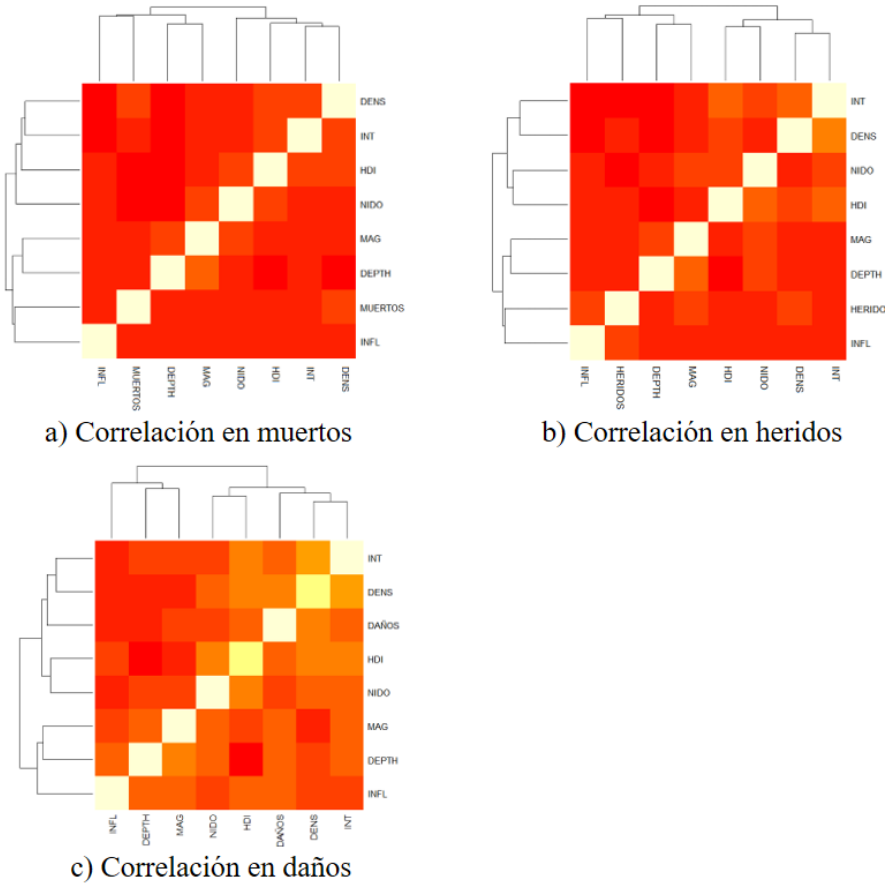


Figura 19. Correlación definitiva para cada variable respuesta.

En la figura 19 se muestra como la correlación disminuye al eliminar la variable de población urbana, se observa también que hay una correlación más cercana a cero en las bases de datos para muertos y para heridos que en la de daños, sin embargo, la mayor correlación en la última base posee un coeficiente de 0.33. Dando como definitiva esta combinación de variables para el diseño de los modelos.

En la figura 20, se describe como se desarrolló el proceso de los modelos matemáticos de forma general.

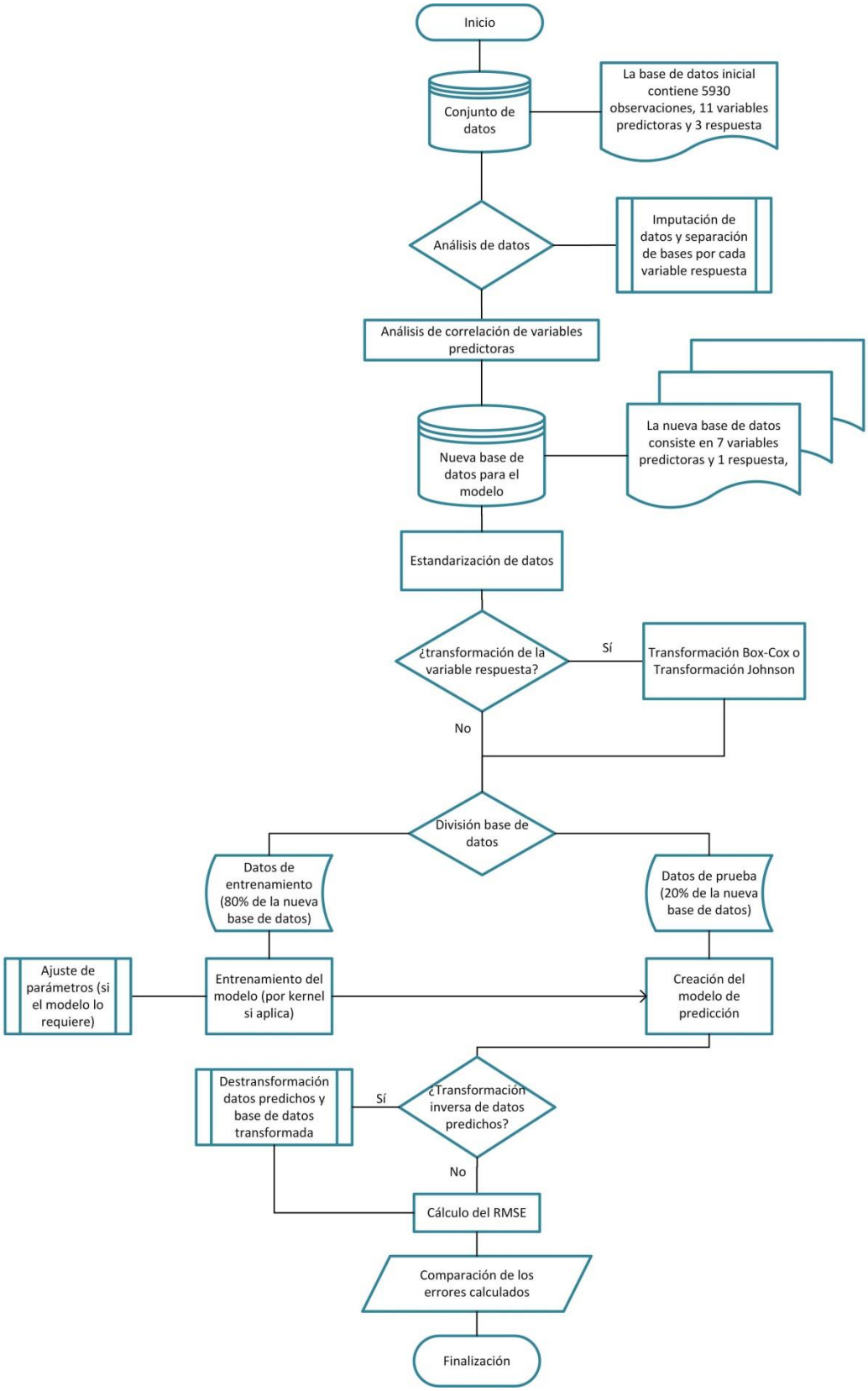


Figura 20. Diagrama de flujo de procesamiento de datos, algoritmo de aprendizaje y predicción.

7. Diseño de modelos estadístico en el software RStudio

7.1. Modelo de regresión lineal

Al contar con las variables definidas tanto con cuatro variables imputadas como con siete, se crean varias bases de datos, una para cada variable respuesta, con el fin de tratar las variables como casos independientes, en este caso muertos**,** heridos y daños.

Basados en el software RStudio se realiza el modelo inicial de regresión, en el cual, al cargar los datos, estos se normalizan con la forma Z-core, se procede dividir aleatoriamente los datos en dos grupos, uno para el entrenamiento de los modelos y el segundo para la prueba del mismo de la forma 70% y 30% respectivamente.

Se realiza un experimento con la base de datos de cuatro variables imputadas con la variable respuesta muertos, se ajusta un modelo lineal inicial y se hallan los residuos con el fin de revisar los supuestos de normalidad, homocedasticidad, no colinealidad e independencia de los mismos. Al realizar diferentes pruebas de normalidad de residuos y se tiene:

Tabla 3.

Resultados de las pruebas de normalidad.

	Kolmogorov/ Smirnov	**Shapiro/Wilk**	**Kurtosis**	**Weisberg/ Bingham**
Valor p	2.2e-16	2.2e-16	2.2e-16	2.2e-16

Nota: Pruebas de normalidad de residuos con sus estadísticos y valores p para el modelo lineal inicial, según los resultados, no existe normalidad de residuos.

Para la prueba Breush – Pagan para heterocedasticidad el valor p es de 0,8955 y está por encima del nivel de significancia 0,05, lo cual refleja que no se puede rechazar la hipótesis de homocedasticidad.

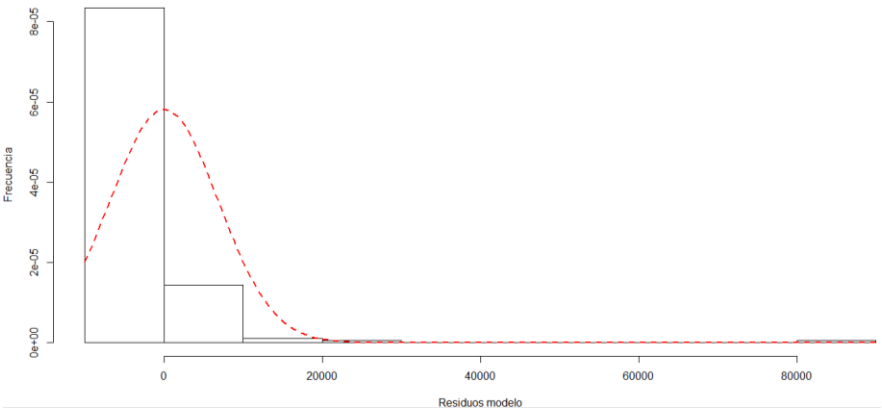


Figura 21. Histograma de residuos del modelo 1 sin transformar variable respuesta.

Se implementó la prueba Durbin–Watson para la independencia de residuos, arrojando un estadístico DW de 2,0461 y un valor p de 0,6233 dando a conocer un resultado favorable en cuanto a este supuesto.

En cuanto a la no colinealidad se acudió a los factores de inflación de varianza (VIF), el cual no arroja ningún valor por encima de 2, lo que significa que no existe colinealidad entre las variables independientes.

Se decidió realizar dos transformaciones a la variable respuesta con el fin de dar solución a la no normalidad de residuos, y se contó con la transformación de Box-Cox y la transformación de Johnson.

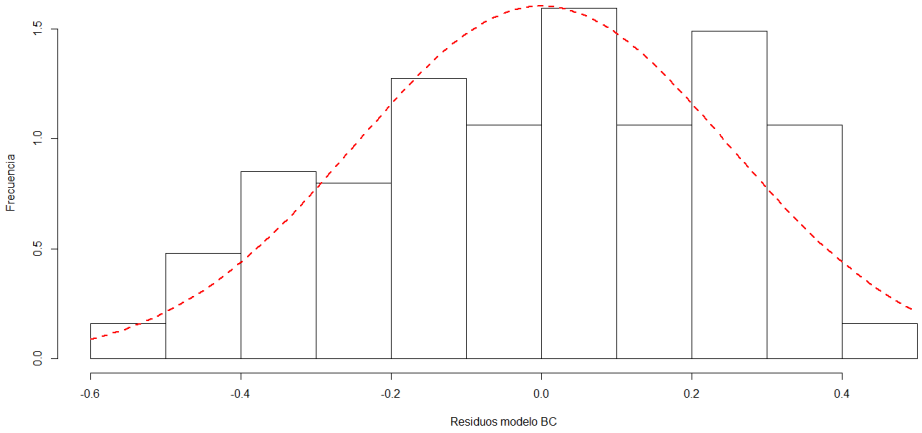


Figura 22. Histograma de residuos del modelo 1 con la transformación de Box-Cox.

En la figura 22 se observa una mayor normalidad en los datos de la variable respuesta con respecto a la anterior distribución.

En esta transformación de Johnson es importante tener en cuenta que puede cambiar la familia (S_U, S_L o S_B) con la que esta se realiza, esto se debe a que la cantidad de datos es diferente en la base de datos 1 a la base de datos 2. Lo mismo sucede cuando se aplica el modelo Random Forest y MSV.

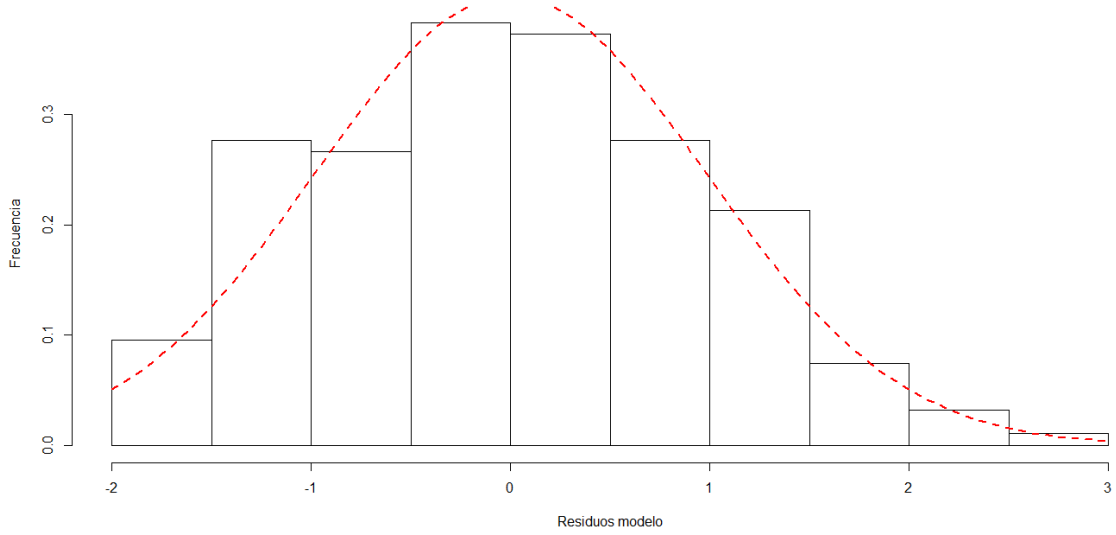


Figura 23. Histograma de residuos del modelo 1 con la transformación de Johnson.

La figura 23 muestra el histograma de los residuos del modelo creado con la variable respuesta transformada usando las familias Johnson, a su vez, esta posee una mayor normalidad que los datos sin transformar y aparentemente mejor que la transformada con las familias Box-Cox. A continuación, se muestra una tabla con las pruebas de normalidad para las anteriores transformaciones y sus respectivos resultados.

Se identifica en la tabla 4 que la transformación de Johnson trae consigo una mayor normalidad para este caso. Al tener estas pruebas un nivel de significancia de 0.05, se determina que, según las 4 pruebas realizadas, que no se cumple el supuesto de normalidad ya que al menos tres de estas cuatro deberían ser favorables y esto no se da.

Tabla 4.

Resultados de las pruebas de normalidad con datos transformados.

Transformación		**Kolmogorov – Smirnov**	**Shapiro - Wilk**	**Kurtosis**	**Weisberg - Bingham**
Box - Cox	Valor p	0,2302	0.0003	0,0125	0,001
Johnson	Valor p	0,3113	0,015	0,261	0,0315

Nota: Pruebas de normalidad de residuos con sus estadísticos y valores p para el modelo lineal inicial.

Este proceso se realiza análogamente para cada una de las variables respuestas tanto con cuatro variables imputadas como con siete, así, tendríamos la siguiente tabla de normalidades.

Tabla 5.

Prueba de normalidad 1.

Variable	**Transformación**	**Kolmogorov-Smirnov**	**Shapiro-Wilk**	**Kurtosis**	**Weisberg-Bingham**
Muertos	Sin Transformación	2,2e-16	2,2e-16	2,2e-16	2,2e-16
Heridos		2,2e-16	2,2e-16	2,2e-16	2,2e-16
Daños		2,2e-16	2,2e-16	2,2e-16	2,2e-16
Muertos	Box-Cox	0,2302	0.0003	0,0125	0,001
Heridos		0,8613	0,9035	0,6215	0,905
Daños		0,8727	0,8124	0,329	0,922
Muertos	Johnson	0,3113	0,015	0,261	0,0315
Heridos		0,743	0,0714	0,0585	0,045
Daños		0,7497	0,5711	0,368	0,433

Nota: Incluye los valores P y los estadísticos de las pruebas de normalidad para la base de datos con 4 variables imputadas para la variable respuesta muertos.

Se observa que en el caso donde no se trasforma la variable respuesta el valor p es insignificante y por lo tanto no se cumple normalidad. Sin embargo, las transformaciones de Box-Cox y de Johnson proveen resultados satisfactorios en algunas pruebas.

Tabla 6.

Prueba de normalidad 2.

Variable	Transformación	Kolmogorov-Smirnov	Shapiro-Wilk	Kurtosis	Weisberg-Bingham
Muertos	Lineal	2,2e-16	2,2e-16	2,2e-16	2,2e-16
Heridos		2,2e-16	2,2e-16	2,2e-16	2,2e-16
Daños		2,2e-16	2,2e-16	2,2e-16	2,2e-16
Muertos	Box-Cox	6,817e-6	2,543e-9	0,0015	2,2e-16
Heridos		0,9514	0,1014	0,0955	0,231
Daños		0,5275	0,4109	0,119	0,6465
Muertos	Johnson	4,634e-9	6,317e-8	0,603	2,2e-16
Heridos		0,0008	5,38e-8	0,02	2,2e-16
Daños		0,4714	0,768	0,619	0,8635

Nota: Incluye los valores P y los estadísticos de las pruebas de normalidad para la base de datos con 7 variables imputadas para la variable respuesta muertos.

También cabe mencionar que el valor p es menor que el nivel de significancia 0,5% por lo tanto no se tiene normalidad cuando no se realiza transformación alguna, se observa que la variable muertos no presenta normalidad salvo bajo la prueba de Kurtosis, y heridos bajo la transformación de Box-Cox, así, se puede concluir que en este caso funcionó mejor la transformación de Box-Cox.

Luego de revisar los supuestos de normalidad, se procede a plantear el modelo de regresión, se crean diversos modelos para cada variable respuesta, además se establece un modelo para cada transformación de datos y así, al entrenar y probar los diferentes códigos, se obtendrán diferentes resultados que al evaluarlos se decide la mejor opción usando como criterio de selección el menor valor de la raíz del error cuadrado medio (RMSE) de los datos predichos.

7.2. Algoritmo Random Forest

Tomando como referencia el modelo lineal construido previamente, se decide construir modelos diferentes para cada una de las trasformaciones para la variable respuesta, de esta manera se tendrán tres modelos por cada variable respuesta.

Este modelo de predicción trabaja con algunos parámetros por defecto como lo son;

- Número de árboles = 500
- Mtry = (Numero de variables predictoras) / 3 = 2
- Número mínimo de nodos = 5

Con el fin de hallar los parámetros óptimos en el modelo Random Forest se hace uso de la función “tune.randomForest” extraída de la librería “e1071”, esta función realiza un análisis secuencial de diferentes combinaciones de los parámetros pertenecientes al modelo, dicha función permite hallar el número óptimo de árboles (ntree), cantidad de nodos de dichos árboles (nodesize) y el número de variables seleccionadas aleatoriamente en cada corrida (mtry), todo esto para que el modelo tenga el menor error posible en los datos predichos.

Se hace uso del paquete “randomForest**”** necesario para crear el bosque, se ajusta con los valores de los parámetros hallados con ayuda de la función tune y, con esto se realiza la predicción con los datos de prueba establecidos anteriormente en el modelo.

En este estudio, se realiza una prueba con los valores por defecto de los modelos, y uno con los parámetros óptimos, como objetivo de analizar la diferencia entre estos y si la parametrización es efectiva o no.

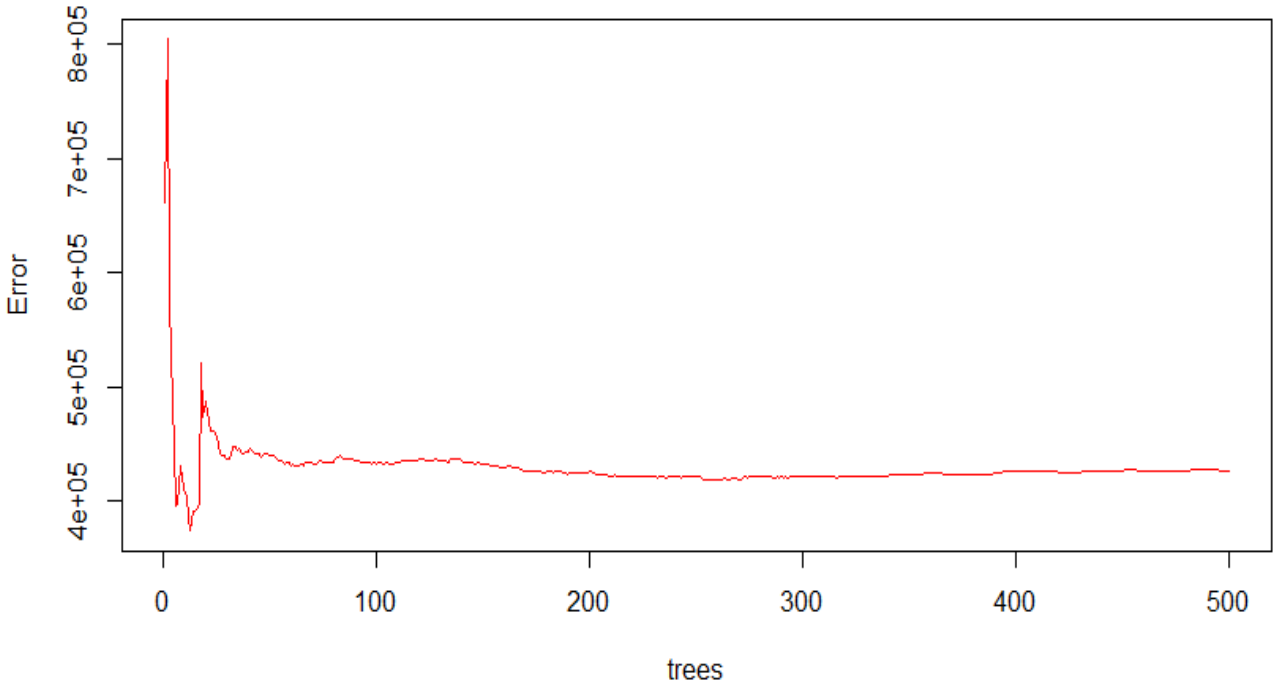


Figura 24. Comportamiento de un bosque creado con los parámetros por defecto.

En la figura 24 se muestra el comportamiento de un bosque creado con los parámetros por defecto, en este caso, para la variable muertos y sin transformación de esta.

7.3. Máquinas de soporte vectorial

Para el desarrollo de este método de aprendizaje supervisado usamos el paquete del "e1071" disponible para el software R, del cual extraemos las funciones necesarias tanto para entrenar el modelo, crear la función de predicción e incluso hacer el tuneo de parámetros y también hacer uso de los diferentes tipos de kernel que éste pone a disposición y los compararemos entre sí basados en el RMSE. El código para las 2 bases de datos (4 y 7 variables imputadas) y para cada una de

las variables respuesta (Muertos, Daños y Heridos), solo difiere en el tipo de transformación que se realiza a estas últimas, por ende, se hace una descripción general del procedimiento de la siguiente manera:

- El procedimiento da inicio cargando la base de datos y normalizando los datos de las variables predictoras usando la técnica Z-core
- Seguidamente y si es el caso se realiza la transformación, ya sea Box-Cox o Johnson
- Estás técnicas de aprendizaje automático recomiendan hacer una división del conjunto de datos, uno para entrenamiento y otro para las respectivas predicciones, en este caso, el 70% de ellos se dedicaron a la fase de entrenamiento, mientras el 30% restante a la fase de prueba. Estos datos son seleccionados de manera aleatoria
- La fase de entrenamiento se realizó de dos maneras: la primera sin ajuste de parámetros, y la segunda usando la función “tune.MSV” de la librería “e1071”, esto con el fin de observar e identificar el aporte que ese ajuste puede representar en el desarrollo del modelo predictivo. Cada kernel debe ajustarse según los parámetros que tenga cada uno, el costo y épsilon son los únicos que se repiten en todos los kernel
- Luego de haber realizado el entrenamiento, la función “predict” realiza la predicción de los datos de la variable respuesta, usando como argumentos, el modelo entrenado anteriormente y la base de datos de prueba, es decir el 30% de los datos originales
- Por último, se calculó el error (RMSE) para cada uno de los kernel y para las diferentes bases de datos.

8. Resultados computacionales

En esta sección presentamos los resultados de cada uno de los modelos propuestos, donde destacamos un mejor comportamiento de la base de datos con 4 variables imputadas, ya que los errores en cada uno de los modelos son mucho menores que los de la base de datos con 7 variables imputadas.

En cuanto al modelo lineal, encontramos que para estos datos no se cumple el supuesto de normalidad, pero si se cumplen: el supuesto de homocedasticidad, independencia de residuos y no colinealidad, aun así, se tomó la decisión de correr el código y realizar la comparación con los otros modelos de predicción.

El modelo Random Forest presenta una gran ventaja con respecto a los otros dos, ya que el ajuste de los parámetros de este, es mucho más rápido y menos costoso computacionalmente, pero es menos preciso que las máquinas de soporte vectorial, estás presentan un costo al momento de hallar los parámetros, ya que maneja la opción de evaluar cuatro núcleos diferentes, pero sus resultados son significativamente mejores que los demás. A continuación, presentamos los resultados de cada algoritmo:

8.1. Resultados modelo de regresión lineal

Por medio de la tabla 7 se muestra el comportamiento de la predicción con regresión lineal, donde el modelo sin alguna transformación se comporta de mejor manera en cuanto al error arrojado en las tres variables respuesta.

Cabe resaltar el extraño comportamiento de la variable daños al realizar la trasformación de Johnson, cuyo error se eleva repentinamente de forma inesperada, al revisar el porqué del suceso, se encontró que un solo dato está presentando este error ya que su predicción es extremadamente

distante de su dato observado, de esto se puede concluir que en ocasiones la transformación de datos puede alterar los resultados de manera negativa y mostrar un dato erróneo.

Tabla 7.

Valores RMSE para los modelos corridos con cuatro variables imputadas.

	Regresión lineal	**Regresión lineal Box-Cox**	**Regresión lineal Johnson**
Muertos	460,6	479,29	479,09
Heridos	1472,79	1531,29	1529,98
Daños	1348,00	1426,76	105993,83

Nota. Incluye los resultados por cada variable respuesta y cada transformación por medio de la regresión lineal con la imputación de cuatro variables predictoras.

Tabla 8.

Valores RMSE para los modelos corridos con siete variables imputadas.

	Regresión lineal	**Regresión lineal Box-Cox**	**Regresión lineal Johnson**
Muertos	1337,73	1293,25	1293,17
Heridos	1605,69	1467,51	1465,73
Daños	1941,72	1983.28	1967,52

Nota. Incluye los resultados por cada variable respuesta y cada transformación por medio de la regresión lineal con la imputación de siete variables predictoras.

En el caso de la tabla 8 se observa el comportamiento similar en las diferentes transformaciones por cada respuesta, así, Así se determina que en cuanto a las variables muertos y heridos el mejor resultado se da con la transformación de Johnson, mientras que para daños su mejor resultado es sin transformación.

Además, a partir de las tablas 7 y 8 se concluye que, en cuanto a regresión lineal, el mejor comportamiento es cuando solo se imputan cuatro variables, es decir, no siempre que se imputen datos el resultado tiende a mejorar, también se observa que en cuanto a la variable daños**,** las transformaciones incrementan su error de predicción.

Tabla 9.

Mejores resultados del modelo lineal.

Mejores resultados modelo lineal			
Variable respuesta	Muertos	Heridos	Daños
Transformación	Sin transformar	Sin transformar	Sin transformar
RMSE	460.6	1472,79	1348

Nota. Incluye los valores de RMSE para los mejores resultados con el modelo de Regresión Lineal.

Con excepción de la variable heridos que presentó un RMSE de 1465,73 con la transformación de Johnson de la base de datos con siete variables imputadas, se decide mantener la base de datos 1 ya que presenta mayoría en los modelos con mejor resultado. Estos RMSE significan que al realizar una predicción se tiene la posibilidad de tener un error en 461 en cuestión de muertos. Los modelos mostrados en la tabla 9 serán tenidos en cuenta para la comparación con los resultados de los demás modelos Random Forest y Maquinas de Soporte Vectorial.

En la figura 25a, 25b y 25c se observa el comportamiento del error con respecto al aumento en los valores predichos para muertos, heridos y daños respectivamente, se nota una ligera disminución, es decir, a medida que aumenta la estimación, el error disminuye. En la figura 25d se observa el error de los modelos predictores, siendo muertos la variable mejor ajustada.

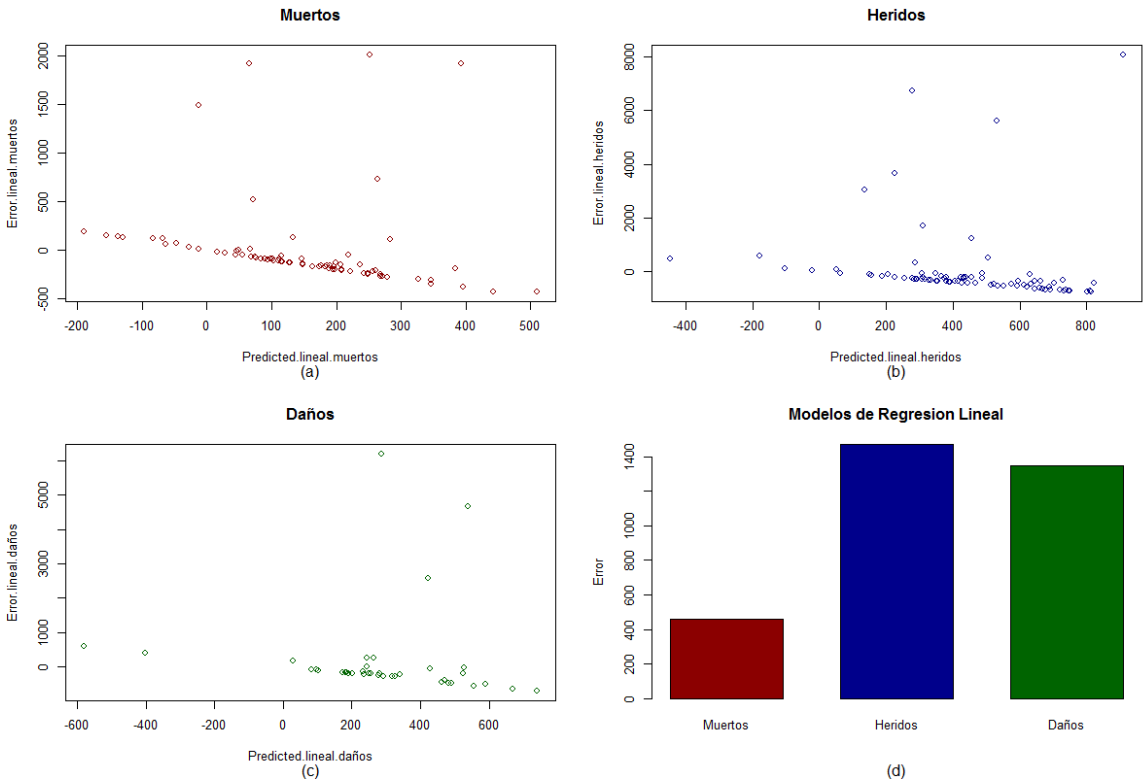


Figura 25. Comparación de errores de los mejores modelos de regresión lineal.

8.2. Resultados modelo RF

A continuación, se presentan los resultados de los RMSE obtenidos al predecir eventos con el modelo Random Forest.

Tabla 10.

Valores RMSE para los modelos ajustados con cuatro variables imputadas.

Variable respuesta	Random Forest	Random Forest Box-Cox	Random Forest Johnson
Muertos	**458,66**	478,36	461,29
Heridos	**1527,44**	1533,37	1532,92
Daños	**1252,11**	1393,35	1396,78

Nota. Incluye los resultados por cada variable respuesta y cada transformación por medio del modelo Random Forest con la imputación de cuatro variables predictoras.

Se ve claramente como en los tres casos (cada una de las variables respuesta) el modelo que muestra el mayor rendimiento es el que no tiene trasformación, mostrando una ventaja considerable sobre los demás.

En el desarrollo de los modelos para la base de datos imputada con siete variables, en cuanto a las variables muertos y heridos el mejor resultado se da con la trasformación de Johnson. Sin embargo, para la variable daños se presenta sin trasformación. Aún así, se tiene que los menores errores en predicción predominan en la base de datos con imputación de cuatro variables como se muestra en la tabla 11.

Tabla 11.

Valores RMSE para los modelos ajustados con siete variables imputadas.

Variables respuesta	**Random Forest**	**Random Forest Box-Cox**	**Random Forest Johnson**
Muertos	1434,50	1293	**1292,65**
Heridos	1851,72	1462,91	**1458,04**
Daños	**1891,66**	1949,50	1913.25

Nota. Incluye los resultados por cada variable respuesta y cada transformación por medio del modelo Random Forest con la imputación de siete variables predictoras.

Téngase en cuenta que los mejores modelos se dieron sin aplicar transformación alguna sobre sus variables respuesta. Con esto se da a entender, por ejemplo, que al predecir con un modelo Random Forest puede haber 459 muertos por encima o por debajo de la predicción obtenida.

Los modelos mostrados en la tabla 12 serán tenidos en cuenta para la comparación con los resultados de los demás modelos Regresión Lineal Múltiple y Maquinas de Soporte Vectorial.

Tabla 12.

Mejores resultados del modelo Random Forest.

Mejores resultados de RF			
Variable respuesta	Muertos	Heridos	Daños
Transformación	Sin transformar	Sin transformar	Sin transformar
RMSE	458,66	1527,44	1252.11

Nota. Incluye los valores de RMSE para los mejores resultados con el modelo de Random Forest.

En la figura 26a, 26b y 26c se observa el comportamiento del error con respecto al aumento de los valores predichos para muertos, heridos y daños respectivamente, se tiene una ligera proporcionalidad inversa, es decir, mientras el dato predicho va en aumento, el error disminuye. En la figura 26d se observa el error de los modelos predictores, siendo muertos la variable con el menor error de predicción.

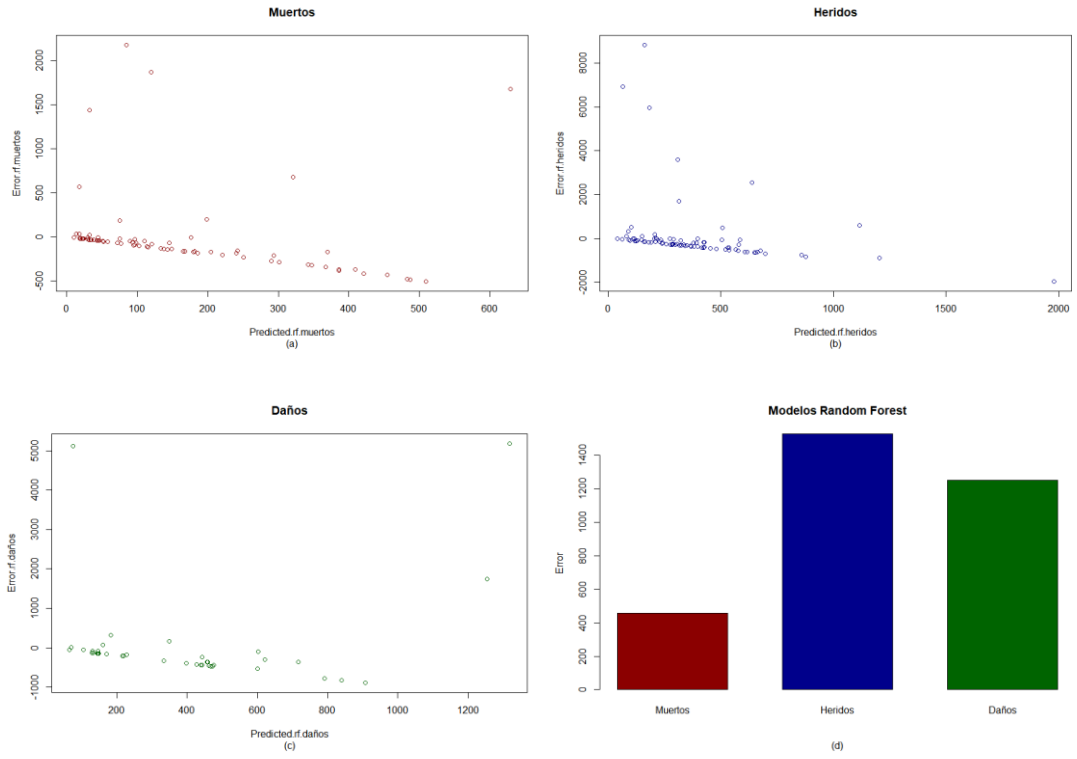


Figura 26. Comparación de errores de los mejores modelos de Random Forest.

8.3. Resultados modelo SVR

Como se mencionó en el capítulo anterior, la evaluación de los modelos con parámetros ajustados tuvo un mejor desempeño, basados en el RMSE, que el mismo modelo sin el ajuste, lo cual se refleja en la tabla 13.

Se observan algunos valores NaN en las tablas 13 y 14, estos se deben a la transformación Box-Cox, ya que el modelo predice un valor negativo cuando se tiene la variable repuesta trasformada y al momento de realizar la transformación inversa, esta obedece a la forma $\sqrt{Y}$, por lo tanto, el resultado no sería un valor real, lo cual se refleja entonces como un NaN.

Tabla 13.

Comparación RMSE (4 variables imputadas).

		SVR Sin transformar		**SVR Box-Cox**		**SVR Johnson**	
		CA	SA	CA	SA	CA	SA
MUERTOS	lineal	676.596	685.499	694.750	695.284	695.228	695.248
	polinomial	677.536	882.056	694.744	695.041	694.408	694.892
	radial	664.627	686.209	692.906	695.300	694.115	695.167
	sigmoide	678.273	1,416,371.000	695.624	NaN	695.654	1.10E+20
HERIDOS	lineal	924.190	938.619	962.088	962.097	962.331	962.702
	polinomial	938.584	946.897	971.789	14,960.990	963.174	2,403.153
	radial	939.233	939.549	963.671	964.668	962.927	964.489
	sigmoide	2,352.207	3,165.415	971.384	NaN	974.186	4,219.990

		SVR Sin transformar		SVR Box-Cox		SVR Johnson	
		CA	SA	CA	CA	SA	CA
DAÑOS	lineal	1,168.336	1,171.180	1,236.234	1,236.306	1,226.300	1,226.321
	polinomial	1,171.057	1,229.773	2,378.134	2,987.926	1,192.129	2,560.195
	radial	1,168.508	1,168.743	1,184.520	1,187.432	1,186.794	1,187.169
	sigmoide	1,134.556	1,290.70	1,189.36	2,742.545	1,195.630	1,206.177

Nota: Las siglas CA y SA, quieren decir: con ajuste y sin ajuste respectivamente, el ajuste hace referencia a la búsqueda de los parámetros óptimos para el modelo.

Tabla 14.

Comparación RMSE (7 variables imputadas).

		SVR Sin transformar		SVR Box-Cox		SVR Johnson	
		CA	SA	CA	SA	CA	SA
MUERTOS	lineal	854.705	854.707	866.586	866.659	865.982	866.768
	polinomial	856.107	931.102	887.480	NaN	867.712	8.11E+15
	radial	856.296	856.343	863.554	863.843	863.744	864.391
	sigmoide	969.729	12,607.150	867.879	NaN	868.007	1.09E+56
HERIDOS	lineal	1,611.728	1,611.852	1,656.839	1,658.906	1,651.754	1,658.584
	polinomial	1,611.489	1,894.937	1,657.776	4.46E+15	1,656.194	76,187.040
	radial	1,617.804	1,624.454	1,654.531	1,654.705	1,653.252	1,654.518
	sigmoide	3,911.130	29,488.570	1,661.753	NaN	1,659.856	1.02E+50
DAÑOS	lineal	2,047.303	2,060.507	2,001.383	2,003.453	1,956.620	2,010.958
	polinomial	1,774.885	1,967.777	1,557.844	1,826.842	1,740.891	1,773.390
	radial	1,856.500	1,875.308	1,938.215	1,938.932	1,933.423	1,933.769
	sigmoide	2,269.317	4,028.717	2,114.365	NaN	2,116.676	6,402.609

Nota: Las siglas CA y SA, quieren decir: con ajuste y sin ajuste respectivamente, el ajuste hace referencia a la búsqueda de los parámetros óptimos para el modelo.

Teniendo en cuenta lo anterior se hizo la selección de los errores más pequeños para cada una de las variables respuesta, así:

Tabla 15.

Mejor RMSE por variable respuesta según transformación 1.

Variable respuesta	**SVR Sin transformar**	**SVR Box-Cox**	**SVR Johnson**
Muertos	664.63	692.91	694.12
Variable respuesta	SVR Sin transformar	SVR Box-Cox	SVR Johnson
Heridos	924.19	962.09	962.33
Daños	1,134.56	1,184.52	1,186.79

Nota: Para la base de datos con 4 variables imputadas

Tabla 16.

Mejor RMSE por variable respuesta según transformación 2.

Variable respuesta	**SVR Sin transformar**	**SVR Box-Cox**	**SVR Johnson**
Muertos	854.71	863.55	863.74
Heridos	1,611.49	1,654.53	1,651.75
Daños	1,774.89	1,557.84	1,740.89

Nota: Para la base de datos con 7 variables imputadas

Finalmente se tiene,

Tabla 17.

Mejores resultados modelo SVR.

Mejores resultados de SVR			
Variable respuesta	Muertos	Heridos	Daños
Transformación	Sin transformar	Sin transformar	Sin transformar
RMSE	664.627	924.19	1134.556

Nota: Todos los valores corresponden a la base de datos con 4 variables imputadas

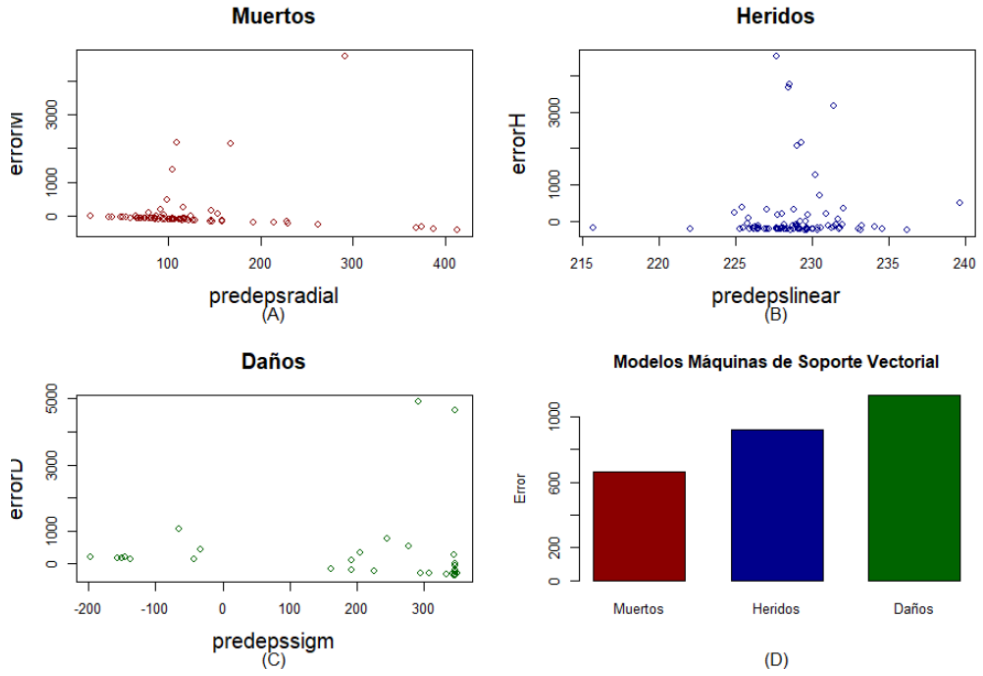


Figura 27. Comparación de errores de los mejores modelos de MSV.

En la figura 27a se observa una ligera tendencia del error a disminuir en cuanto aumenta el valor predicho en la variable muertos. Sin embargo, en las variables heridos y daños no tiene una tendencia clara en las figuras 27 b y 27c respectivamente, aunque cabe resaltar que los puntos se encuentran más dispersos en la variable daños, así mismo, se observa en la figura 27d que la variable que presenta más error en su predicción es daños y, por el contrario, muertos presenta el mejor resultado.

8.4. Comparación RMSE

Tabla 18.

Resultados finales para cada modelo de predicción.

	Modelo Lineal	RF	SVR
Muertos	460.6	**458.66**	664.627
Heridos	1472.79	1527.44	**924.19**
Daños	1348	1252.11	**1134.556**

Nota: Incluye los RMSE de los modelos que presentaron mejor comportamiento según cada variable respuesta.

En la tabla 18 se muestran los resultados de los errores mínimos de cada modelo según las variables respuesta, la cual indicaría que el modelo que presenta el mejor comportamiento de los tres es Maquinas de Soporte Vectorial en cuanto a las variables **heridos** y **daños**, para la variable muertos el que presenta el mínimo error es el modelo Random Forest.

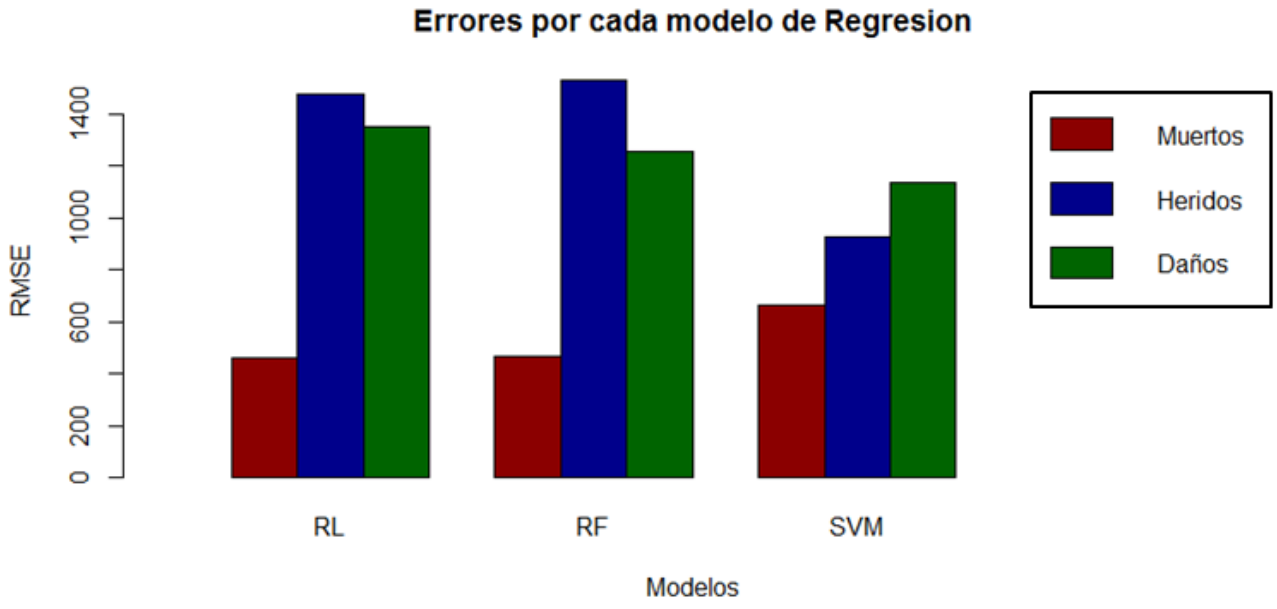


Figura 28. Comparación gráfica final de errores.

En la figura 28 se observa el comportamiento de los RMSE para cada una de las variables respuesta con respecto a cada uno de los modelos utilizados en el estudio, en la cual, se presenta que Random Forest es el modelo que mejor arroja resultados en cuanto a muertos, y Maquinas de Soporte Vectorial es el modelo mejor adaptado para heridos y daños.

9. Conclusiones

En este estudio se proponen 3 modelos de predicción – Regresión Lineal, Random Forest y Máquinas de Soporte Vectorial – para la predicción de los efectos de un sismo y/o terremoto como desastre natural, los cuales son, muertos, heridos y daños en millones de dólares. Este modelo se desarrolla con una amplia base de datos donde se recopilaron por parte de más de 5000 eventos a lo largo de la historia y datos recopilados del Banco Mundial de Desarrollo.

Se desarrollan los 3 modelos utilizando técnicas de aprendizaje automático para la predicción de muertos, heridos y daños, para el primero se destacó el modelo Random forest, ofreciendo una análisis más rápido y menos costoso, y para las dos siguientes variables, las máquinas de soporte vectorial tuvieron un mejor resultado, que aunque más costos computacionalmente que los demás modelos, este respondió mejor, teniendo en cuenta que son 4 kernels, cada uno con varios parámetros por ajustar y definir.

Se resalta la importancia de realizar un tratamiento de datos exhaustivo, con todas las técnicas que se tengan a disposición, ya que esto influencia de manera significativa en los resultados de la investigación. Aquí, la imputación de datos faltantes, por ejemplo, fue contraproducente en algunas de las variables y afecto todo el modelo; debido a los datos faltantes y/o la alta variabilidad que hubo entre ellos, por lo tanto, se debe recurrir a la apreciación y juicio del grupo de investigación para tomar una decisión sobre el tema.

Se sabe que los desastres naturales son difíciles de predecir, particularmente los sismos y/o terremotos, por el impredecible comportamiento de la tierra, por ende, la importancia de esta investigación radicó en lograr que la predicción de los efectos de estos eventos, en este caso de estudio, muertos, heridos o daños, fuesen lo más cercanos a la realidad con la ayuda de la inteligencia artificial y los modelos de predicción que en ésta se pueden implementar, apoyando de manera teórica a las organizaciones tanto gubernamentales como las que no lo son, a tomar decisiones en pro de mitigar los efectos de los desastres.

En la evaluación de los modelos se observa que estos no necesariamente deben tener una distribución normal en la variable respuesta, ya que, según los resultados obtenidos estos funcionaron mejor cuando no se realizó transformación alguna.

Del experimento se concluye que tanto los modelos Random Forest y Máquinas de Soporte Vectorial tienen mejor comportamiento que una Regresión Lineal Múltiple, lo cual, se esperaba desde el inicio del experimento, por la robustez que estos dos modelos tienen, tanto estadísticamente como en el análisis de grandes bases de datos.

10. Recomendaciones

Las estadísticas que los países recopilan sobre diferentes temas, como sociales, culturales, económicos, de desarrollo, educación, economía, entre otros, son escasos y en muchos, deficientes, generando una alta incertidumbre del comportamiento de cada país. Esto precisamente afecta el avance de investigaciones que se fundamentan en estos datos, por lo cual se hace un llamado a los estados para que esta labor se realice con mayor rigor para aportar conocimiento a la población.

Este trabajo, apoya las demás investigaciones que se han llevado a cabo sobre el tema central que es logística humanitaria, aportando nuevas herramientas para las futuras investigaciones, donde el principal objetivo sea ayudar y motivar a países, ciudades y personas alrededor del mundo

continuar con esta línea de investigación y así aportar su granito de arena por las generaciones venideras.

Al observar los resultados de este trabajo, se recomienda seguir con la línea de investigación en la logística humanitaria realizando experimentos con diferentes modelos de regresión y otros posibles factores como variables predictoras, esto con el fin de buscar una mayor eficiencia en la predicción.

Referencias Bibliográficas

Aggarwal, Charu C. (2015). Data Mining: The Textbook. New York, USA. Springer.

Aguirre R. (2 de octubre de 2017). Mocoa: así ha sido la lenta recuperación tras la tragedia. Recuperado de http://www.elcolombiano.com

Ansfield, Valentine J. (2017). Earthquake magnitudes and intensities. Salem press encyclopedia of scienc.

Apte, A. (2010). Humanitarian logistics: A new field of research and action. Foundations and Trends® in Technology, Information and Operations Management, 3(1), 1-100.

Asim, K. M., Martínez-Álvarez, F., Basit, A., & Iqbal, T. (2017). Earthquake magnitude prediction in Hindukush region using machine learning techniques. Natural Hazards, 85(1), 471-486.

Banco Mundial. Recuperado de: https://datos.bancomundial.org/indicador/

Barnston, A. G. (1992). Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. Weather and Forecasting, 7(4), 699-709.

Barrera, A y Hernández, A. (2015). Un algoritmo evolutivo para el problema de distribución de recursos post-desastres sísmicos en la ciudad de Bucaramanga (Tesis de pregrado). Universidad Industria de Santander, Bucaramanga, Colombia. Recuperado de: http://tangara.uis.edu.co/

Barreto, M. y Niño, P. (2016). Un algoritmo memético para el problema de localización-ruteo con ventanas de tiempo para la atención de desastres sísmicos en la ciudad de Bucaramanga (Tesis de pregrado). Universidad Industrial de Santander, Bucaramanga, Colombia. Recuperado de: http://tangara.uis.edu.co/

Betancourt, G. A. (2005). Las máquinas de soporte vectorial (MSVs). Scientia et technica, 1(27).

Blanco Bernardeau, A., Alonso Sarría, F., y Gomariz Castillo, F. (2014). Elaboración de un mapa de carbono orgánico del suelo en la Región de Murcia.

Breiman, L. (2001). Random forests. Machine learning. Berkeley, USA.

Cambronero, C. G., & Moreno, I. G. (2006). Algoritmos de aprendizaje: knn & kmeans. Inteligencia en Redes de Comunicación, Universidad Carlos III de Madrid.

Campos, A., Nielsen, N., Díaz, C., Rubiano, D., Costa, C., Ramírez, F., & Dickson, E. (2012). Análisis de la gestión del riesgo de desastres en Colombia.

Chang, C. C., & Lin, C. J. (2011). LIBMSV: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, *2*(3), 27.

Chen, F., Yu, B., & Li, B. (2017). A practical trial of landslide detection from single-temporal Landsat8 images using contour-based proposals and random forest: a case study of national Nepal. Landslides, 1-12.

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

Cozzolino, A. (2012). Humanitarian logistics: cross-sector cooperation in disaster relief management. Springer Science & Business Media.

David, M. (2017). Support Vector Machines: The Interface to libMSV in Package e1071. David. Meyer@ R-Project. org.

Del mundo para México: suman 440 toneladas de ayuda de 27 países. (27 de septiembre de 2017). Recuperado de http://www.eluniversal.com.mx

Espino, A. I. L., Mur, R. A., & de Miguel, M. A. S. (2004). Aprendizaje automático en conjuntos de clasificadores heterogéneos y modelado de agentes (Tesis Doctoral). Universidad Carlos III de Madrid, Departamento de Informática.

Expansion. Recuperado de: https://www.datosmacro.com/

Fatídico septiembre termina con 360 muertos por el sismo del 19. (30 de septiembre de 2017). Recuperado de http://www.proceso.com.mx

Federación internacional de sociedades de la cruz roja y de la media luna roja [IFRC]. (s.f.). Gestión de desastres. Recuperado de http://www.ifrc.org

González, L. (2003). Modelos de clasificación basados en máquinas de vectores de soporte. *Asociación científica europea de economía aplicada*.

Guerra de la Corte, A. (2016). Técnicas de selección de variables en minería estadística de datos. Universidad de Sevilla.

Gunn, S. R. (1998). Support vector machines for classification and regression. ISIS technical report, 14(1), 5-16.

Gutiérrez, J. D. (2007). Clasificación de Imágenes Usando Máquinas de Soporte Vectorial.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1999). *Análisis multivariante* (Vol. 491). Madrid: Prentice Hall.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning. 2001.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International journal of forecasting, 22(4), 679-688.

IBM Knowledge Center. (s.f). Valores perdidos. Recuperado de: https://www.ibm.com

Indexmundi. Recuperado de: https://www.indexmundi.com/

Jaiwei, H., & Kamber, M. (2006). Data mining: concepts and techniques. *ed: Morgan Kaufmann San Francisco*.

Kondratyev, K. Y., Krapivin, V. F., & Varostos, C. A. (2006). Natural disasters as interactive components of global-ecodynamics. Springer Science & Business Media.

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International Journal of Computer Science, 1*(2), 111-117.

Kunz, N., & Reiner, G. (2012). A meta-analysis of humanitarian logistics research. Journal of Humanitarian Logistics and Supply Chain Management, 2(2), 116-147.

Lagos, I. J., & Vargas, J. A. (2003). Sistema de familias de distribuciones de Johnson, una alternativa para el manejo de datos no normales en cartas de control. *Revista Colombiana de Estadística, 26*(1), 25-40.

Lee, J. H., Sameen, M. I., Pradhan, B., & Park, H. J. (2018). Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods. Geomorphology, 303, 284-298.

Lin, G. F., Chang, M. J., Huang, Y. C., & Ho, J. Y. (2017). Assessment of susceptibility to rainfall-induced landslides using improved self-organizing linear output map, support vector machine, and logistic regression. Engineering Geology, 224, 62-74.

Malaeb, Z. A. (1997). A SAS® code to correct for non-normality and non-constant variance in regression and ANOVA models using the Box–Cox method of power transformation. *Environmental Monitoring and Assessment, 47*(3), 255-273.

Martín Guareño, J. J. (2016). Support vector regression: propiedades y aplicaciones.

Mattioli, Glen S., Jansma, P. (2017). Earthquake prediction. Salem press encyclopedia of scienc.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). Machine learning: An artificial intelligence approach. Springer Science & Business Media.

Minitab Inc. (s.f.). ¿Qué son MAPE, MAD y MSD?. Recuperado de: https://support.minitab.com

Minitab Inc. (s.f.). Método de estimación de mínimos cuadrados y método de estimación de máxima verosimilitud. Recuperado de: https://support.minitab.com

Mitchell, T. M. (1997). Machine learning. 1997. Burr Ridge, IL: McGraw Hill.

Montgomery, D. C. (1996). Desing and analysis of experiments. *Jonh Wiley & Sons, New York, EUA*.

Montoya, L., & Masser, I. (2005). Management of natural hazard risk in Cartago, Costa Rica. Habitat International, 29(3), 493-509.

Nateghi, R., Guikema, S. D., & Quiring, S. M. (2014). Forecasting hurricane-induced power outage durations. Natural hazards, 74(3), 1795-1811.

National Centers for Environmental Information. Recuperado de: https://www.ngdc.noaa.gov

Ngwenya, N. K., & Naude, M. J. (2016). Supply chain management best practices: A case of humanitarian aid in southern Africa. Journal of Transport and Supply Chain Management, 10(1), 1-9.

Okulewicz, S. C. (2017). Earthquake hazards. Salem Press Encyclopedia Of Science.

Olsen, G. R., Carstensen, N., & Høyen, K. (2003). Humanitarian crises: what determines the level of emergency assistance? Media coverage, donor interests and the aid business. Disasters, 27(2), 109-126.

Pedroza, G., Prieto, A., & Goddard, J. (2007). Aplicación de las Maquinas de Soporte Vectorial a Reconocimiento de Hablantes. Universidad Autónoma Metropolitana, México.

Pérez Planells, L., Delegido, J., Rivera-Caicedo, J. P., & Verrelst, J. (2015). Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. *Revista Española de Teledetección, 2015, vol. 44, p. 55-65*.

Pérez, J. M. R. (2014). Máquinas de vectores soporte en entornos de supercomputación: aplicación a fusión nuclear (Doctoral dissertation, UNED).

Resendiz, J. A. (2006). Las Máquinas de Soporte Vectorial para identificación en Línea. (Título de Maestría). Instituto Politécnico Nacional.

Rodríguez, J. T., Vitoriano, B., Montero, J., & Kecman, V. (2011). A disaster-severity assessment DSS comparative analysis. OR spectrum, 33(3), 451-479.

Safeer, M., Anbuudayasankar, S. P., Balkumar, K., & Ganesh, K. (2014). Analyzing transportation and distribution in emergency humanitarian logistics. Procedia Engineering, 97, 2248-2258.

Sandın, J. M., Colomer, A. A., & Palacios, R. P. (2012). Técnicas de regresión para la estimación de la localización de la mirada.

Santana, A. (2015). El estado del arte de los modelos de optimización en la logística de atención a desastres (Tesis de pregrado). Universidad Industrial de Santander, Bucaramanga, Colombia. Recuperado de: http://tangara.uis.edu.co/

Schölkopf, B., Smola, A. J., Williamson, R. C., & Bartlett, P. L. (2000). New support vector algorithms. Neural computation, 12(5), 1207-1245.

Science for a Changing World. The Modified Mercalli Intensity Scale. Recuperado de: https://earthquake.usgs.gov/learn/topics/mercalli.php

Suman 366 los fallecidos en el terremoto del 19 de septiembre en México. (3 de octubre de 2017). Recuperado de http://www.elperiodico.com

Thomas, S., Pillai, G. N., & Pal, K. (2017). Prediction of peak ground acceleration using ϵ-SVR, ν-SVR and Ls-SVR algorithm. Geomatics, Natural Hazards and Risk, 8(2), 177-193.

University of California, Departament of Statistics (s.f.). Random Forest, Leo Breiman and Adele Cutler. Berkley, USA. Recuperado de: https://www.stat.berkeley.edu

Van Wassenhove, L. N. (2006). Blackett memorial lecture. Humanitarian aid logistics: Supply chain management in high gear 475–476.

Vapnik V. (1995). The nature of statistical learning theory. New York (NY): Springer-Verlag.

Vapnik, V. (1998). Statistical learning theory new york. NY: Wiley.

Villalibre Calderón, C. (2013). Concepto de urgencia, emergencia, catástrofe y desastre: revisión histórica y bibliográfica.

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2007). Probabilidad y estadística para ingeniería y ciencias. Pearson Educación.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., & Bai, X. (2015). Flood hazard risk assessment model based on random forest. Journal of Hydrology, 527, 1130-1141.

Wikipedia. Escala Sismológica de Mercalli. Recuperado de: https://es.wikipedia.org/wiki/Escala_sismol%C3%B3gica_de_Mercalli

Williams, G. (2011). Data mining with Rattle and R: The art of excavating data for knowledge discovery. Springer Science & Business Media.

World Confederation for Physical Therapy [WCPT]. (2016). What is disaster management? London, UK. Recuperado de: http://www.wcpt.org/disaster-management/what-is-disaster-management

Yang, Z. (1999). Estimating a transformation and its effect on Box-CoxT-ratio. *Test*, *8*(1), 167-190.

Yoon, D. K., & Jeong, S. (2016). Assessment of Community Vulnerability to Natural Disasters in Korea by Using GIS and Machine Learning Techniques. In Quantitative Regional Economic and Environmental Analysis for Sustainability in Korea (pp. 123-140). Springer, Singapore.

Zhang, K., Wu, X., Niu, R., Yang, K., & Zhao, L. (2017). The assessment of landslide susceptibility mapping using random forest and decision tree methods in the Three Gorges Reservoir area, China. Environmental Earth Sciences, 76(11), 405.