Evaluación de vulnerabilidad sísmica de la malla vial en Bucaramanga aplicando modelos de aprendizaje automático

Aldemar López Velandia

Fabián Leonardo Sandoval Bernal

Trabajo de Grado para Optar al Título de Ingeniero Industrial

Director:

Daniel Orlando Martínez Quezada

M.Sc en Ingeniería Industrial

Codirector:

Gustavo Chio Cho

Ph.D en Ingeniería de Caminos, Canales y Puertos

Universidad Industrial de Santander

Facultad de Ingeniería Fisicomecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2020

Agradecimientos

Dar gracias A Dios por permitirnos vivir esta linda experiencia a lo largo de estos años, agradecer al profesor Daniel Martínez Quesada, director de nuestro proyecto por apoyarnos, guiarnos, animarnos y compartir de su conocimiento y experiencia durante el desarrollo de nuestro proyecto, de igual forma agradecer a nuestro codirector el profesor Gustavo Chio Cho quien a través de su conocimiento nos brindó un aporte significativo en la ejecución del proyecto. También agradecer al grupo de investigación OPALO por acogernos en su grupo de investigadores permitiendo enfrentar con ellos nuevos retos de investigación.

Queremos Agradecer a la prestigiosa Universidad Industrial de Santander por abrirnos sus puertas y acogernos en esta gran familia universitaria años atrás, y que hoy en día nos forma para enfrentar un mundo laboral con valiosas herramientas competitivas y así poder asumir los nuevos y diferentes retos que nos de la vida, así mismo dar gracias a cada uno de los decentes de la institución por compartirnos sus conocimientos y experiencias vividas en el trascurso de todo este tiempo compartido.

Finalmente agradecer a las diferentes personas que de una u otra forma contribuyeron para el desarrollo y ejecución de nuestro proyecto.

Dedicatoria

Quiero dedicar primeramente este trabajo a Dios por permitirme culminar este bello proceso de formación, A mis padres Pedro José López Calderón y María Sofía Velandia junto con mis hermanos, quienes han sido mi gran apoyo y motor de impulso en todos estos años, sin ellos no podría ser la persona que soy hoy en día, su confianza cariño y esfuerzo son los que me motivan cada día a mejorar y seguir viviendo nuevas experiencias que contribuyen a mi desarrollo personal y profesional. Son ellos la razón especial de dedicar este trabajo. Agradecer a mi equipo de trabajo, los profesores Gustavo Chio Cho y Daniel Martínez y mi compañero de trabajo por su gran dedicación y esfuerzo. Finalmente quiero agradecer a amigos y personas allegadas que de una u otra forma conocí en el transcurso de mi carrera con quienes aprendí y viví grandes experiencias llenas de aprendizaje y enseñanzas para mi vida personal.

Aldemar López Velandia

Dedico este trabajo primeramente a Dios por darme salud y vida para poder cumplir mis sueños

A mi amada madre Luz del Carmen Bernal y hermana Claudia Sandoval quienes siempre ha sido mi gran fuente de motivación e inspiración para poder superarme cada día y alcanzar cada uno de mis propósitos.

Al profesor Jairo Sánchez Malaver quien siempre ha sido una inspiración para mí, con sus consejos e historias me han ayudado a formar como profesional y como persona.

Por último, dedico este trabajo a mis compañeros y amigos quienes estuvieron ahí para apoyarme y aconsejarme durante este largo y duro proceso, aquellos que hicieron que este sueño de ser ingeniero industrial se hiciera realidad.

Fabián Leonardo Sandoval Bernal

Tabla de contenido

Lista de tablas

Lista de figuras

Lista de Apéndices

(Los apéndices están adjuntos en el CD y puede visualizarlos en base de datos de la biblioteca UIS)

Apéndice A. Base de datos.

Apéndice B. Muestra aleatoria de 100 datos para evaluación.

Apéndice C. Modelos de aprendizaje automático en RStudio.

Apéndice D. Índice de vulnerabilidad de la malla vial de Bucaramanga (.xlsx, qgz).

Apéndice E. Artículo investigativo.

Resumen

TÍTULO: Evaluación de vulnerabilidad sísmica de la malla vial en Bucaramanga aplicando modelos de aprendizaje automático*

AUTOR: López Velandia Aldemar

Sandoval Bernal Fabián Leonardo**

PALABRAS CLAVE: Aprendizaje automático, vulnerabilidad, evaluación sísmica, red de carreteras.

DESCRIPCIÓN: La actividad sísmica es un fenómeno natural que hasta el día de hoy no se conoce el momento exacto en el que va a ocurrir ni la intensidad con la que se presentara ya que aún no se cuenta con las herramientas suficientes para predecir un fenómeno como estos, es por esto que ante una eventualidad como estas en la mayoría de casos no se cuenta con la preparación necesaria para atender dicha emergencia y aunque las entidades nacionales, se han tomado la tarea de mitigar los efectos que ocasionan, estos casi nunca son suficientes y se centran más en el post-desastre y no en el pre-desastre, olvidando la importancia de realizar una evaluación de vulnerabilidad sísmica de elementos importantes para la atención de desastres como lo son las edificaciones, redes viales, líneas de vida, entre otras. Pocos estudios se han enfocado en la evaluación de vulnerabilidad de las redes viales olvidando que esta línea de investigación de gran importancia en la intervención rápida y esfuerzos de recuperación.

Debido a lo anterior se planteó el diseño de modelos de aprendizaje automático para evaluación de la malla vial de Bucaramanga que ayude a la identificación de la vulnerabilidad en tramos viales ante una eventualidad sísmica. Estos modelos fueron desarrollados a partir de la experiencia y conocimiento de un experto. Para tal objetivo se comparó el algoritmo de aprendizaje supervisado K-vecinos próximos (K-NN) con un algoritmo ensamblado (BOOSTING). Donde se seleccionó el modelo que presentó mejor error en un conjunto de prueba, en este caso el método de Ensamblado (BO0STING) fue el que obtuvo el mejor error cuadrático medio (RMSE).

---

* Trabajo de grado

** Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director: Daniel Orlando Martínez, M.Sc en Ingeniería Industrial

Abstract

TÍTLE: Seismic Vulnerability Assessment of The Road Network, Applying Machine Learning Models: Case study Bucaramanga, Colombia*

AUTHOR (S): López Velandia Aldemar

Sandoval Bernal Fabián Leonardo**

KEYWORDS: Machine learning, vulnerability, assessment seismic, road network

DESCRIPTION: Seismic activity is a natural phenomenon which does not know the exact moment when it will happen or the intensity that it will present. There are not yet enough tools to predict a phenomenon like this. That is the reason why an event like this generally does not count with a necessary preparation to deal with this emergency. Although national entities have taken the requirements to mitigate the effects it can cause, there is a main focus on post-disaster rather than pre-disaster activities, underestimating the importance of a seismic vulnerability evaluation of important elements for disaster assistance such as buildings, road networks, lifelines, among others. Few studies have been focused on the vulnerability assessment of road networks ignoring that this line of research has a great importance in order to get a fast intervention and recovery results.

Furthermore, machine learning models have been fit for the evaluation of Bucaramanga highway network in order to help to identification of vulnerability in road sections before a seismic event. These models were developed from an expert experience. That's why, the k-Near Neighbors (K-NN) supervised learning algorithm was compared with an assembled methods (BOOSTING) and the model that presents the best error in a test set was selected. In this case, the method that obtained the best root mean squared (RMSE) was the method ensemble methods (BOOSTING).

* Bachelor thesis

** Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director: Daniel Orlando Martínez, M.Sc en Ingeniería Industrial

Introducción

Colombia se encuentra ubicada en una de las zonas sísmicas más activas de la Tierra, debido a que en la región convergen las placas tectónicas de Nazca y del Caribe contra la placa suramericana. Por consecuencia de esta ubicación, a través de la historia en Colombia se han presentado diferentes tipos de fenómenos naturales y ejemplos de ello están: el terremoto ocurrido en el año de 1979 en Tumaco con 7,9 grados de magnitud que dio lugar a un tsunami; el terremoto del eje cafetero de 6.2 de magnitud; la erupción del nevado del Ruiz en Armero, Tolima, en 1985; y uno más reciente fue el desastre natural de Mocoa, Putumayo, en 2017.

Para analizar más a fondo esta problemática se debe empezar afirmando que hoy, en pleno siglo XXI, la ciencia aún no cuenta con las herramientas suficientes para predecir con exactitud el día, la hora y el lugar en el que va a ocurrir un sismo y la magnitud con la que se pueda presentar, es por esto que ante una eventualidad como estas en la mayoría de casos no se cuenta con la preparación necesaria para atender dicha emergencia y aunque las entidades nacionales, departamentales y municipales se han tomado la tarea de mitigar los efectos que ocasionan, estos casi nunca son suficientes y se centran más en el post-desastre y no en el pre-desastre, olvidando la importancia de realizar una evaluación de vulnerabilidad sísmica de elementos importantes para la atención de desastres como lo son las edificaciones, redes viales, líneas de vida, entre otras.

Entre tanto, Bucaramanga es una ciudad que está ubicada al nororiente del país sobre la cordillera oriental, y debido a la alta sismicidad en la región no está exenta de ser afectada por un sismo de gran magnitud. Según el Servicio Geológico Colombiano, cerca del 50 % de los sismos en Colombia tienen como epicentro a Santander, razón por la cual el área metropolitana de

Bucaramanga es vulnerable a estos fenómenos(Campos García, Díaz Giraldo, Rubiano Vargas, & Costa Posada, 2012).

Teniendo en cuenta que Bucaramanga permanece en un alto riesgo sísmico la presente investigación construyó un modelo de regresión para la malla vial de Bucaramanga que permite la identificación de la vulnerabilidad en tramos viales ante una eventualidad sísmica. Este modelo se desarrolló bajo la experiencia y conocimiento de un experto en evaluación de vulnerabilidad sísmica y con la ayuda de inteligencia artificial y aprendizaje estadístico, se replicó el conocimiento del experto a partir del análisis de muestras.

Para el desarrollo de este proyecto se evaluaron dos modelos de regresión: K-Neighbors (K-NN), Ensemble Methods (BOOSTING) con el fin de estimar la vulnerabilidad de la malla vial de Bucaramanga replicando el análisis de un experto, con la finalidad de no incurrir en grandes costos económicos, pero que sea de gran ayuda para poder indagar el estado de las vías y permita al gobierno municipal o entidades correspondientes poder desarrollar estrategias para la prevención de desastres futuros.

Cumplimiento de objetivos

Tabla 1.
Cumplimiento de objetivos.

Objetivos Específicos	Apartado Relacionado
1. Realizar una revisión literaria sobre modelos usados en la evaluación de vulnerabilidad sísmica de sistemas viales y modelos ensamblados de clasificación supervisada.	2.0 - 3.0
2. Consolidar una base de datos del sistema vial de Bucaramanga que permita identificar niveles de vulnerabilidad vial ante una actividad sísmica.	Apéndice A
3. Seleccionar una muestra de la malla vial de Bucaramanga la cual pueda ser evaluada por un experto.	Apéndice B
4. Evaluar los diferentes modelos aprendizaje automático para un estudio de vulnerabilidad para seleccionar el modelo de clasificación supervisada que mejor describan la evaluación de vulnerabilidad de la malla vial en Bucaramanga.	5.0
5. Elaborar un artículo académico de carácter publicable basado en los resultados de la investigación.	Apéndice E

1. Generalidades de la Investigación

1.1. Planteamiento del problema

Según el Servicio Geológico Colombiano, cerca del 50 % de los sismos en Colombia tienen como epicentro a Santander, razón por la cual el área metropolitana de Bucaramanga se encuentra dentro de un ambiente sismo tectónico de reconocida actividad histórica, conocido como el nido sísmico ubicado en la mesa de los santos a 50 kilómetros de Bucaramanga todo esto debido a su posición geográfica, originando que esta ciudad sea una de las más vulnerables del territorio colombiano, presentándose movimientos telúrico de baja intensidad ya que en su mayoría presentan alta profundidad, no obstante, esto no es ninguna garantía que se presente un movimiento telúrico con altos niveles en la escala Richter y de poca profundidad, por ende, es necesario generar alertas tempranas para la mitigación del riesgo en el departamento.

Dado que en los últimos años se han realizado diferentes estudios aportando a esta temática de la gestión de desastre cabe resaltar que la mayoría de estos estudios sean enfocado a diferentes aspectos como la vulnerabilidad de edificaciones o planes de gestión de recursos y ayudas humanitarias y muy pocos han realizado estudios enfocados en la vulnerabilidad de las carreteras olvidando que esta línea de investigación también tiene gran importancia siendo vital en la intervención rápida y esfuerzos de recuperación. Terremotos recientes como los de Italia (2016, 2009), Nepal (2015), Japón (2011), Haití (2010), China (2008) o Indonesia (2004). Demostró que, independientemente de los preparativos para un terremoto considerable, los daños significativos y las disfunciones pueden afectar las áreas urbanas, su población y la infraestructura. Las redes de carreteras generalmente se pueden considerar como la infraestructura de transporte más

importante, ya que interconectan físicamente a personas de todo el mundo, en una manera directa, asequible y flexible. Como lo demuestran las experiencias de terremotos anteriores, las redes de carreteras pueden verse afectadas por dos grupos principales de factores: el daño directo (el colapso de estructuras tales como puentes y túneles, efectos de licuefacción, etc.) y daños indirectos (bloqueo debido a escombros de edificios, debido a la congestión del tráfico, etc.). La capacidad de una red para funcionar y adaptarse a nuevas condiciones repentinas es una característica importante que debe lograrse para evitar mayores pérdidas sociales y económicas (Fleischhauer 2008).

A partir de la experiencia de terremotos pasados revela que las carreteras son elementos vulnerables y su daño podría traducirse en demoras en las actividades de recuperación. La interrupción del tráfico puede afectar fuertemente las operaciones de emergencia y rescate inmediatamente después del terremoto, así como el esfuerzo de reconstrucción y otras actividades en el período siguiente (Argyroudis & Gehl, 2015).

Debido a que al momento de querer realizar una evaluación de vulnerabilidad en la malla vial, requiere una gran inversión de dinero lo que genera altos costos para su desarrollo, el presente trabajo es para proponer un modelo de aprendizaje automático teniendo en cuenta la participación y el criterio de un experto, una técnica que por medio del aprendizaje automático sea capaz de predecir un comportamiento similar de acuerdo a la información proporcionada, el cual facilite la toma de decisiones en la prevención para aquellas vías que presenten un riesgo significativo.

Para el desarrollo de este trabajo se debe considerar las aplicaciones del aprendizaje automático para el desarrollo de un sistema experto el cuál es el resultado de la colaboración de un experto humano especialista en el tema de estudio, para esto se utilizarán algoritmos de aprendizaje

automático, como lo son: método de ensamble (Boosting) y K- Vecinos próximos (K-NN) , los cuales nos permitirán realizar una evaluación de vulnerabilidad sísmica para la malla vial de Bucaramanga, siendo este un importante instrumento el cual puede servir de apoyo a las diferentes entidades encargadas para el desarrollo de estrategias para la preparación de desastres, con el fin de poder contribuir con la reducción de riesgos sobre estos eventos de tal magnitud.

1.2. Objetivos

1.1.1. Objetivo general.

Diseñar un modelo de aprendizaje automático para evaluar la vulnerabilidad sísmica de la malla vial en la ciudad de Bucaramanga a partir de datos proporcionados por un experto.

1.1.2. Objetivos específicos

- Realizar una revisión literaria sobre modelos usados en la evaluación de vulnerabilidad sísmica de sistemas viales y modelos ensamblados de clasificación supervisada.

- Consolidar una base de datos del sistema vial de Bucaramanga que permita identificar niveles de vulnerabilidad vial ante una actividad sísmica.

- Seleccionar una muestra de la malla vial de Bucaramanga la cual pueda ser evaluada por un experto.

- Evaluar los diferentes modelos de aprendizaje automático para un estudio de vulnerabilidad para seleccionar el modelo que mejor describa la evaluación de vulnerabilidad de la malla vial en Bucaramanga.

- Elaborar un artículo académico de carácter publicable basado en los resultados de la investigación.

2. Revisión de la literaria

Los terremotos son fenómenos naturales intrínsecamente impredecibles que pueden ocurrir en cualquier momento, pueden causar daños a diversas instalaciones de un área urbana (edificios, carreteras, puentes, líneas de vida, entre otros.). Pueden perturbar los servicios urbanos y las redes causan la pérdida de vidas humanas en un tiempo muy corto. Aunque, es prácticamente imposible predecir la hora exacta de una catástrofe, es posible estimar la probabilidad de ocurrencia de un terremoto. Por lo tanto, es de crucial importancia estar preparado si el área urbana es propensa a un riesgo sísmico significativo. (Benamar, Naili, & Amellal, 2018).

Para el estudio y evaluación de una amenaza sísmica de un área se basa en el estudio de sismología y en el análisis de las características sismológicas, sismo-genéticas y geológicas del sitio. Los estudios históricos están orientados a la definición de las principales características geofísicas. Características tales como: (epicentro, magnitud, aceleración del terreno, etc.). Con el objetivo de establecer medidas de prevención y modernización antisísmica, se han introducido metodologías para la evaluación del riesgo sísmico en edificios, esto con el fin de definir el daño esperado después de un terremoto en un área determinada e identificar todos los elementos en alto riesgo. Pero en la mayoría de las evaluaciones de riesgos, un componente importante a menudo se ha pasado por alto, o al menos, no se ha considerado de gran importancia el sistema de transporte, debido a que, si la red de transporte sigue siendo eficiente después de un evento ocurrido, la asistencia llegará rápida y fácilmente a las áreas afectadas. Como es el caso de terremotos ocurridos en California (1989 y 1994) y japoneses (Kobe, 1995) han demostrado trágicamente el papel

esencial que tienen las redes de transporte en una situación de emergencia después de un evento sísmico. (Andrea, Afiso, & Ondorelli, 2005).

Los artículos consultados que se presentaran a continuación fueron desarrollados para la etapa en donde se utiliza un modelo para poder predecir el grado de vulnerabilidad y poder mitigar o responder a diferentes consecuencias que se puedan presentar mediante un desastre sísmico.

2.1. Gestión de desastres

De acuerdo con Chandra Balijepalli y Olivia Oppong (2014), observan que después de un evento como estos no todo los enlaces en redes de carreteras se afecta de igual forma y que algunos de estos enlaces pueden llegar a ser más crítico que otros, por lo que la mayoría de los índices existentes diseñados para medir la vulnerabilidad ofrecen una buena medida a la accesibilidad de toda la red, pero estas rara vez consideran el alcance de la capacidad en los puntos críticos en redes de carreteras urbanas, por esta razón se atrevieron a proponer un nuevo estudio de vulnerabilidad sísmica en redes de carreteras utilizando técnicas de modelado de redes de tráfico combinadas con la utilización de Sistemas de Información Geográfica (SIG), para el cual considera la capacidad de servicio de los enlaces de carreteras. Finalmente utilizan estos resultados para proponer y describir un plan de desvío de tráfico en caso de que ocurra un desastre sísmico veintitrés.

Christos Polykretis, Maria Ferentinou y Christos Chalkias (2015), quienes tuvieron por objetivo desarrollar un estudio, de comparar el rendimiento de un método estadístico convencional como el índice de susceptibilidad de deslizamiento (LSI) y un método basado en redes neuronales artificiales (ANN, por sus siglas en inglés). La relación entre deslizamientos de tierra y diversos factores condicionantes que contribuyen a su ocurrencia se investigó a través del análisis basado en un Sistema de Información Geográfica. Se realizó un inventario de deslizamientos utilizando

fotos aéreas, imágenes satelitales y estudios de campo. Se consideraron ocho factores condicionantes, que incluyen la cobertura del suelo, la geología, la elevación, la pendiente, el aspecto, la distancia a la red de carreteras, la distancia a la red de drenaje, la distancia a los elementos estructurales. Después, los mapas de LS se produjeron utilizando LSI y ANN, y luego se compararon y validaron en consecuencia. Los resultados arrojados por esta extrapolación demuestran que ambos modelos se pueden usar para mitigar los peligros relacionados con los deslizamientos de tierra y para ayudar en un uso generalizado.

Enrico Quagliarini, Gabriele Bernardini, Silvia Santarelli y Michele Lucesoli (2018), quienes por medio de un método holístico preliminar y rápido para el riesgo sísmico obtienen la evaluación y estimación del nivel de daño de posibles vías de evacuación. Para ello deben tener en cuenta en primer lugar, los datos sobre seguridad y factores de influencia, es decir, uso y exposición de la trayectoria, características geométricas, características físico-estructurales, vulnerabilidad intrínseca que se relaciona con los elementos que componen la calle en sí, los elementos infraestructurales relacionados como pavimentos de calles, fundaciones, terraplenes, líneas de vida y los elementos interferentes como estructuras subterráneas. La vulnerabilidad extrínseca que se refiere a los elementos que no pertenecen directamente a la ruta en sí, pero pueden comprometerla o bloquearla (es decir, edificios que pueden colapsar y bloquear las calles debido a la formación de escombro). Entonces, según datos del mundo real, se propone una correlación sobre los niveles de riesgo-daño en el camino con el propósito adicional de evaluar las capacidades del método para describir escenarios post-terremoto.

Pierre Gehl, Francesco Cavalieri y Paolo Franchin (2018), estos autores propusieron un modelo de red bayesiana (BN) aproximado de las simulaciones de Monte-Carlo para un sistema de infraestructura expuesto a peligros sísmicos. Las cuales permitirán predecir métricas complejas de

rendimiento del sistema contribuyendo a una respuesta rápida posterior al terremoto. Para ello se selecciona un conjunto reducido de componentes de infraestructura, cuya importancia se clasifica a través de un algoritmo de bosque aleatorio, para predecir el rendimiento del sistema. Este enfoque fue aplicado a una red de carreteras francesa, donde solo se mantienen de 5 a 10 componentes de los 58 para estimar la distribución de las métricas de rendimiento del sistema que se basan en el flujo de tráfico. Los estudios de sensibilidad sobre el número de componentes seleccionados, el número de ejecuciones de simulación fuera de línea y la desratización de las variables revelan que la reducción de la BN aplicada a este ejemplo específico genera estimaciones confiables.

Yaohui Liu, Zhiqiang Li, Benyong Wei, Xiaoli Li y Bo Fu (2019), presentaron un enfoque integrado para una evaluación de vulnerabilidad macro sísmica compuesta por métodos de extracción de datos de tecnología y ciencia SIG, para ello establecieron un esquema de clasificación de vulnerabilidad EMS-98 y dos modelos de aprendizaje automático, máquina de vectores de soporte y reglas de asociación. Los mapas de riesgo sísmico se construyeron a través de datos que consistían en daños directos a edificios y víctimas humanas. Los resultados indicaron que los dos métodos de extracción de datos podrían lograr precisiones y estabilidades deseables al estimar la vulnerabilidad sísmica.

2.2. Aplicaciones del aprendizaje automático

El aprendizaje automático es una disciplina científica del ámbito de la inteligencia artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos. La máquina que realmente aprende es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros. Y automáticamente implica que

estos sistemas se mejoran de forma autónoma con el tiempo sin intervención humana (Tinaquero, 2018).

Según su historia, situar el principio de la inteligencia artificial es bastante complicado, sin embargo, parece haber cierto consenso en que Warren McCulloch y Walter Pitts dieron el pistoletazo de salida a esta joven ciencia y en 1943 gracias a sus trabajos en los que propusieron el primer modelo de red neuronal artificial, un modelo simple en el cual lograron demostrar que era capaz de aprender y resolver funciones lógicas. (Serrano, 2012).

En 1950, Alan Turing crea la prueba de Turing, en un artículo llamado Computing Machinery and Intelligence, la prueba busca determinar si una máquina es pensante o no, tras esto el defendía la idea que por medio de computación el pensamiento humano podía ser imitado o emulado. Ese mismo año Claude Shannon detalla un juego de ajedrez como proceso de búsqueda en su artículo Programmin a Computer for Playing Chess.

En 1956 tras la Conferencia de Dartmouth, se da vida al término Inteligencia artificial y por primera vez recibe un significado: hacer que una máquina se comporte como lo haría un ser humano, de tal manera que se la podría llamar inteligente. Ese mismo año se llevó acabo la primera demostración de un programa de inteligencia artificial, Allen Newell, J.C. Shaw y Herbert Simon lo hicieron con el The Logic Theorist.

A finales del Siglo XX, se desarrollan una serie de propuestas y aplicaciones de inteligencia artificial, entre los que se pueden destacar: en 1961, el primer programa de integración simbólica, escrito en LISP, para la solución de problemas de cálculo de nivel colegial llamado SAINT, creado por James Slagle.(Serna A, Acevedo M, & Serna M, 2017).

En 1964 En el MIT, Danny Bobrow, en su disertación, demuestra que equipos pueden entender lenguaje natural lo suficientemente bien como para resolver correctamente problemas verbales de álgebra. Ese mismo año, Bert Raphael, en su disertación, demuestra con el programa SIR el poder de la representación lógica de conocimientos para sistemas basados en preguntas y respuestas.

En 1965, John Alan Robinson inventó el método de resolución que permitió a programas trabajar de forma eficaz usando la lógica formal como lenguaje de representación. Paralelo a esto, Joseph Weizenbaum, en el MIT construyó a ELIZA, un programa interactivo que dialogaba en inglés sobre cualquier tema.

En 1966, Ross Quillian en su disertación demostró la utilidad de las redes semánticas en aplicaciones de IA. En este mismo año, el Automatic Language Processing Advisory Committee publica un informe negativo sobre traducción automática de máquina que detendrá por muchos años, hasta fines de la década de los años 80, la investigación en Procesamiento de Lenguaje.

En 1967, Edward Feigenbaum, Joshua Lederberg, Bruce Buchanan, Georgia Sutherland publican en Stanford el programa Dendral, capaz de interpretar la espectrometría de masas en compuestos químicos orgánicos. Este fue el primer programa basado en conocimiento para razonamiento científico. También ese año, Joel Moses en su trabajo de doctorado en el MIT, con el programa Macsyma, demostró el poder de razonamiento simbólico para los problemas de integración. Fue el primer programa exitoso, basado en conocimientos, para matemáticas.

Richard Greenblatt en el MIT construyó un programa basado en conocimiento para jugar ajedrez, MacHack. Y durante los años siguientes hubo grandes avances y mejoramientos para los trabajos ya existentes además de nuevas IA como PROLOG, ARCH, SHRDLU y MYCIN; a mediados de los 80 apareció una serie de aplicaciones basadas en redes neuronales artificiales,

entrenadas por el algoritmo de Backpropagation , entre otros grandes avances, dando comienzo a una robótica más avanzada y estable.(Serna A et al., 2017).

2.3. Aplicaciones de K-vecinos próximos (K-NN)

Sadra Karimzadeh ,Masashi Matsuoka, Jianming Kuang y Linlin Ge (2019), quienes hicieron uso de algoritmos de aprendizaje automatico como K-vecinos próximos (KNN), Maquina de Vectores de Sorporte (SVM), Bayes (NB) y Bosques Aleatorios (RDF). Con el fin de predecir los patrones de réplica de del terremoto de Kermanshah en Irán. Para este estudio utilizaron mapas de la distribución de deslizamiento y el cambio de estrés de Coulomb asociado con el terremoto de Kermanshah, junto con mapas de fallas de la región, además se recogieron registros de réplicas de este terremoto desde el primer segundo después del evento ocurrido. El cambio de tensión en la falla de la fuente (deducida de las imágenes de radar de apertura sintética) y las orientaciones de las fallas activas vecinas aportan para la dicha predicción. El setenta por ciento de las réplicas se utilizaron para el entrenamiento basado en una lógica binaria ("sí" o "no") para predecir las ubicaciones de todas las réplicas. Los resultados de las características operativas del receptor del mismo conjunto de datos indican que los métodos de ML superan a los mapas de Coulomb de rutina con respecto a la predicción espacial de los patrones de réplica, especialmente cuando los detalles de fallas activas vecinas están disponibles. El método KNN y los clasificadores RDF mostraron mejores rendimientos que el algoritmo NB. Las actuaciones del KNN y los algoritmos RDF para la clasificación en diferentes umbrales fueron mejores, lo que significó una mayor verdad se podría lograr una tasa positiva y una tasa de falsos positivos más baja.

Alireza Motevalli, Seyed Amir Naghibi, Hossein Hashemi, Ronny Berndtsson, Biswajeet Pradhan y Vahid Gholami (2019), quienes utilizaron una metodología que puede determinar

inversamente el tipo y la ubicación de la fuente principal de contaminantes de nitrato. Basada en dos técnicas de minería de datos de última generación, el árbol de regresión potenciado (BRT) y el vecino k-más cercano (KNN). Estas técnicas se utilizan para producir un mapa de vulnerabilidad a la contaminación por nitratos. La metodología puede mitigar los efectos del juicio subjetivo sobre la determinación de la importancia de diferentes fuentes y mecanismos para el transporte de nitrato. Los mecanismos investigados son factores hidrogeológicos, hidrológicos, antropogénicos, topográficos y de acondicionamiento del suelo. Por lo tanto, la metodología propuesta se utiliza para separar los procesos naturales y los efectos antropogénicos sobre la contaminación por nitratos. Para calcular los mapas de vulnerabilidad del agua subterránea, se seleccionó una concentración de nitrato de agua subterránea de 40 mg / L (sugerida por la OMS con un margen de riesgo del 20%) como un umbral general para identificar áreas contaminadas que resultaron en 96 pozos contaminados. Las ubicaciones no contaminadas se seleccionaron de los datos del pozo con una concentración de nitrato inferior a 15 mg / L (96 no contaminadas). Los modelos fueron entrenados en 70% de datos contaminados y 70% de sitios no contaminados. Los datos restantes, Se utilizaron 30% de sitios contaminados y 30% de sitios no contaminados para validar los resultados de la simulación. Los resultados mostraron que el BRT produjo resultados con un rendimiento más alto que el algoritmo KNN. Los resultados de la clasificación final basados en el modelo BRT mostraron la mayor importancia de la conductividad hidráulica, la densidad del río, el suelo, el porcentaje de pendiente, la recarga neta y la distancia desde las aldeas, en orden, en relación con otros factores.

Albert Comelli, Alessandro Stefano, Giorgio Russo, Samuel Bignardi, Maria Gabriella Sabini, Giovanni Petrucci, Massimo Ippolito y Anthony Yezzi (2019), estos autores propusieron un algoritmo para la delineación del tumor en la tomografía por emisión de positrones (PET). La

segmentación se logra mediante un algoritmo de contorno activo local, integrado y optimizado con el método de clasificación de k-vecino más cercano (KNN), que aprovecha la estrategia de validación cruzada estratificada de k-fold. El enfoque propuesto se evalúa considerando la delineación de cánceres localizados en diferentes distritos corporales (es decir, cerebro, cabeza y cuello y pulmón), y considerando diferentes trazadores radiactivos de PET. Los datos se procesan previamente para expresarse en términos de valor de absorción estandarizado. El algoritmo utiliza una región inicial seleccionada por el operador que contiene la lesión e identifica automáticamente una región óptima de interés independiente del operador alrededor del tumor. Sucesivamente, se utiliza un algoritmo de segmentación de contorno activo local de marcha corte por corte. La novedad clave del enfoque propuesto consiste en una forma novedosa de la energía que se minimizará durante la segmentación, que se mejora incorporando la información proporcionada por un clasificador KNN. El proceso de delineación y su finalización son completamente automáticos, de modo que la intervención del usuario se reduce al mínimo. Debido al alto nivel de automatización, la lesión segmentada final es independiente de la variación entre operadores en la entrada inicial del usuario, lo que hace que todo el proceso sea robusto y el resultado sea completamente repetible. Para evaluar el rendimiento bajo diferentes escenarios de relación de contraste, primero se evaluó el método propuesto en cinco conjuntos de datos fantasmas donde se valuó la aplicabilidad del método en el entorno de radioterapia mediante la investigación de cincuenta casos clínicos y dos radio-trazadores PET diferentes. Los experimentos fantasmas demostraron la efectividad del método en objetivos bien conocidos de bordes afilados. El método de segmentación propuesto mostró una sensibilidad superior al 90% para los experimentos relacionados con esferas con un diámetro superior a 17 mm. Además, a investigación muestra que

el método propuesto puede aplicarse en entornos clínicos y produce segmentaciones precisas e independientes del operador, logrando una buena precisión en condiciones realistas.

Xianyuan Dong y Mingjie Lu (2019), propusieron un algoritmo óptimo que puede ayudar a los gerentes de tráfico a tomar decisiones más precisas a partir del conocimiento previo de casos de accidentes de tráfico. El algoritmo basado en *k* -vecino más cercano determina el valor de peso de cada característica de caso de accidente basado en el índice de entropía de información, establece una base de recuperación de caso de accidente de tráfico utilizando el algoritmo de agrupación de dos pasos y propone un modelo de similitud global de casos de accidente de tráfico. Luego, se presenta un nuevo índice de evaluación integral llamado grado de correspondencia. Y luego, se desarrolla un sistema prototipo para realizar experimentos de recuperación de casos para verificar el rendimiento del algoritmo propuesto para la recuperación de casos de accidentes de tráfico. El resultado de los experimentos demuestra claramente la efectividad de este algoritmo de recuperación de casos para la gestión de accidentes de tráfico en tiempo real.

2.4. Aplicaciones de los métodos ensamblados (Boosting)

Hidetake Hirayama, Ram C. Sharma, Mizuki Tomita y Keitarou Hara (2018). El objetivo principal de su investigación fue evaluar la efectividad de dos sistemas de clasificación múltiple (MCS) el cual combina resultados de diferentes clasificadores supervisados como un medio para reducir pixeles aislados cuando se utilizan imágenes satelitales de alta resolución espacial. Esta investigación fue realizada en Tohoku, donde el gran terremoto del 2011 y el posterior tsunami han cambiado en gran medida el uso de la tierra regional y la cobertura de la tierra. Se prepararon múltiples características de entrada (cinco canales espectrales RapidEye, cinco índices espectrales, modelo de elevación digital y pendiente) de modelos de aprendizaje automático. Se recopilaron

datos de la verdad sobre el terreno que pertenecen a seis clases principales de cobertura del suelo (bosque, arbusto / pradera, tierra de cultivo, área urbana, cuerpo de agua, suelo desnudo). Se utilizaron 6 clasificadores de aprendizaje automático Extra Gradient boosting (XGBoost), Bosques aleatorios (RF), Maquina de soporte vectorial (SVM), K-vecinos próximos (KNN), Bagging y Red neuronal (NNET), estos algoritmos se probaron individualmente para determinar la precisión de la clasificación y los resultados se compararon con los de las combinaciones MCS. El XGBoost proporcionó la mayor precisión general (0.987), sin embargo, presentó un gran número de píxeles aislados con respecto a los demás modelos. Los autores esperan que este sistema resulte efectivo para monitorear cambios a gran escala en la cubierta terrestre y contribuya a conservar los paisajes y los servicios de los ecosistemas en la región afectada por el terremoto y el tsunami.

Lihui Chen y Pin Wang (2018), aplican el algoritmo de refuerzo adaptativo (AdaBoost) para investigar los factores más significativos para dos grupos de edad (grupos de conductores mayores y jóvenes) basados en datos de accidentes Del mundo real en California. Los factores de accidente incluyen género, tipo de carretera, condición del pavimento, clima, hora del día, comportamiento del vehículo, etc., así como sus subfactores correspondientes. Para el desarrollo del modelo primero entrenan a algunos estudiantes débiles para encontrar importancia y luego combinamos linealmente a esos estudiantes débiles en un estudiante unificado más fuerte. El método propuesto tiene varias ventajas: (1) capacidad para manejar datos no balanceados, (2) ningún requisito sobre el supuesto de distribución de datos y (3) ser robusto para diferentes conjuntos de datos.

Khawaja M, Asim Adnan Idris, Talat Iqbal y Francisco Martínez Álvarez (2018), proponen un sistema de predicción de terremotos combinando indicadores sísmicos junto con el método de conjunto basado en programación genética (GP) y AdaBoost (GP-AdaBoost). Los indicadores sísmicos se calculan a través de una metodología novedosa en la cual, los indicadores se calculan

para obtener la máxima información sobre el estado sísmico de la región. Los indicadores sísmicos calculados se utilizan con el algoritmo GP-AdaBoost para desarrollar un sistema de predicción de terremotos (EP-GPBoost). La configuración se ha organizado para proporcionar predicciones de terremotos de magnitud 5.0 y superiores, quince días antes del terremoto. Las regiones de Hindukush, Chile y el sur de California son consideradas para experimentación. El EP-GPBoost ha producido una notable mejora en predicción debido a la colaboración de fuertes capacidades de búsqueda y refuerzo de GP y AdaBoost, respectivamente. El sistema de predicción de terremotos muestra resultados mejorados en términos de precisión y coeficiente de correlación Matthews para las tres regiones consideradas en comparación con los resultados ya establecidos.

3. Marco de referencias

3.1. Marco de antecedentes

Álvaro Caballero (2007), en su trabajo de tesis de maestría titulado “Determinación de la vulnerabilidad sísmica por medio del método del índice de vulnerabilidad en las estructuras ubicadas en el centro histórico de la ciudad de Sincelejo, utilizando la tecnología del sistema de información geográfica”; quien emplea el método del índice de vulnerabilidad para poder determinar el daño esperado para diferentes aceleraciones sísmicas, utilizando como herramienta principal, la tecnología de Sistemas de Información Geográfica SIG, este acompañado por un estudio de zonificación geotécnica. Para el estudio del índice de vulnerabilidad identifica los parámetros más importantes que controlan el daño en edificaciones causadas por terremotos, este

método califica diversos aspectos de los edificios para distinguir la diferencia en un mismo tipo de construcción o tipología, material o año de construcción y las escalas de intensidad. Además, tales aspectos como estudios de configuración de planta y elevación, el tipo de calidad de materiales utilizados, la posición y segmentación del edificio, estados de conservación de las estructuras entre otros. Los aspectos ya mencionados son evaluados y clasificados otorgándoles una cifra en cantidad o porcentajes bajo el juicio y experiencias de expertos, permitiendo obtener de este modo un índice de vulnerabilidad sísmica.

Herleym Rodríguez (2011), en su trabajo "Análisis y evaluación de riesgos sísmico en líneas vitales. Caso de estudio Bogotá D.C." esta investigación se plantea un marco metodológico para el estudio del riesgo sísmico en líneas vitales de cualquier tipología: redes de acueducto, alcantarillado, eléctricas, de gas e hidrocarburos, de telecomunicaciones, tanques, puentes y vías. Para su aplicación se diseña e implementa una base de datos geográfica, y se desarrolla bajo metodología del Proceso Unificado de Racional RUP (Por sus siglas en inglés), el software "Riesgo Sísmico en Líneas Vitales - RSLV" utilizando Java con ArcGis Engine y ArcObjects. Con la información geográfica en la geodata-base, y el software RSLV, como resultado se presenta un caso de estudio del riesgo sísmico para las líneas vitales de la ciudad de Bogotá especialmente sobre las redes de acueducto y alcantarillado. La amenaza sísmica insumo para este caso de estudio corresponde a los escenarios con periodos de retorno de 50, 100, 200, 475 y 1000 años con información de aceleración, velocidad y desplazamiento pico del terreno, PGA, PGV y PGD respectivamente, para todas las fuentes sismogénicas integradas (fallas cercanas, intermedias y lejanas).

Daniel Martínez (2017), en su trabajo de investigación para optar por el título de magister de Ingeniería Industrial, de la Universidad Industrial de Santander en el año 2017 titulado "Diseño de

un sistema de apoyo a la toma de decisiones - DSS para la gestión de las etapas pre- desastres de sismos en Bucaramanga, basado en técnicas de aprendizaje automático(Machine learning)" en cuya investigación propuso un diseño DSS DM (aplicación de DSS a la gestión de desastres) el cual permite una toma de decisiones en actividades relacionadas a la preparación y mitigación de daños en la ciudad de Bucaramanga, generando escenarios de desastres basados en datos históricos e información georreferenciada, utilizando modelos de aprendizaje automático. Los cuales fueron tomados como parámetros de modelos de optimización que permitan entregar a un tomador de decisiones un conjunto de soluciones para la planificación de actividades logística inmersas en la gestión de desastres.

Yulima Cifuentes y Jessica Jerez (2018), quienes realizan su trabajo de grado para optar por el título de Ingeniería Industrial, de la Universidad Industrial de Santander en el año 2018 titulado "Evaluación de vulnerabilidad sísmica actual de albergues temporales en Bucaramanga aplicando algoritmos de clasificación supervisada" esta investigación tiene como objetivo, evaluar un índice de vulnerabilidad sísmica de albergues temporales en la ciudad de Bucaramanga aplicando dos modelos de clasificación supervisada, con base del criterio dado por un experto. Esto con el fin de validar refugios para una respectiva reubicación de las personas afectadas ante un desastre natural como lo es un sismo, además poder anticipar medidas de prevención y reducción de impactos. Para tal objetivo los dos métodos utilizados de clasificación supervisada son las máquinas de soporte vectorial y árboles de decisión, estos modelos permitieron predecir un nivel de vulnerabilidad sísmica de albergues temporales.

3.2. Marco teórico

3.2.1. Aprendizaje automático (Machine Learning). El Aprendizaje Automático es una rama de la Inteligencia Artificial que se encarga del diseño y desarrollo de algoritmos que permiten a una computadora mejorar un comportamiento automáticamente a través de la experiencia. Es por ello por lo que se le da el nombre de Aprendizaje Automático o Aprendizaje de Máquinas, ya que una máquina es capaz de aprender por sí sola por medio de datos empíricos. De forma más concreta, un algoritmo de AA es capaz de extraer características y patrones comunes de un conjunto de datos (ejemplos o datos de entrenamiento), y aplicarlas a nuevos conjuntos de datos (datos de pruebas). Esta disciplina se utiliza en diferentes campos, tales como diagnósticos médicos, análisis de mercados, robótica, banca, reconocimiento del habla y lenguaje escrito, detección de patrones en imágenes, clasificación de secuencias de ADN, o marketing (Nombela Escobar, 2011).

Otros autores definen este concepto de la siguiente manera:

- El machine learning no es otra cosa que hacer que las máquinas ejecuten acciones sin que tengas que programar el acto, es decir, que a partir de lo que esté sucediendo en el ambiente, la máquina o computadora pueda tomar la decisión de hacer tal o cual cosa (Bormann & Brauchitsch, 2017).

- Enseñar a un computador a aprender conceptos usando datos, sin ser explícitamente programado para ello (Contreras B, 2016).

- Cualquier cambio en un sistema que le permita desempeñar la misma tarea de manera más eficiente la próxima vez (Simón, 1983)

- Se dice que un programa de ordenador aprende por medio de la experiencia E con respecto a alguna clase de tareas T y medida de rendimiento P, si su desempeño en tareas en T, medida por P, mejora con la experiencia E. (Tom Mitchel, 1998).

Ejemplo: Jugando al ajedrez.

E = La experiencia de jugar muchos juegos de ajedrez.

T = La tarea de jugar ajedrez.

P = La probabilidad que el programa gane el próximo partido.

- Campo de estudio que da a los ordenadores la habilidad de aprender sin la necesidad de ser explícitamente programados (Arthur Samuel, 1959).

Existen diferentes tipos de algoritmos clasificados en base a la salida de estos. La distinción más significativa sería entre el aprendizaje supervisado y el aprendizaje no supervisado. Aunque en la literatura se pueden encontrar otros tipos de algoritmos, cómo el aprendizaje Semi-supervisado y por refuerzo, los cuales desglosan diferentes métodos entre los cuales esta, método de regresión, método de clasificación y método de agrupación, pero para el alcance de este proyecto solo se efectuará modelos del aprendizaje supervisado.

3.2.1.1. Tipos de aprendizaje automático.

3.2.1.1.1. Aprendizaje supervisado. En este tipo de aprendizaje se deduce una función a partir de un conjunto de datos de entrenamiento que ya están etiquetados, de ahí que reciba el nombre de supervisado. En estos algoritmos se proporciona un conjunto de datos de entrenamiento en forma de un vector. El algoritmo, tras analizar los vectores de entrada infiere una función conocida

como clasificador o función de regresión dependiendo de si la etiqueta es discreta o continua. La función inferida debe ser capaz de predecir la clase para cualquier vector de entrada (Nombela Escobar, 2011). A continuación, se muestra el funcionamiento del aprendizaje supervisado en mayor detalle en la figura 1.

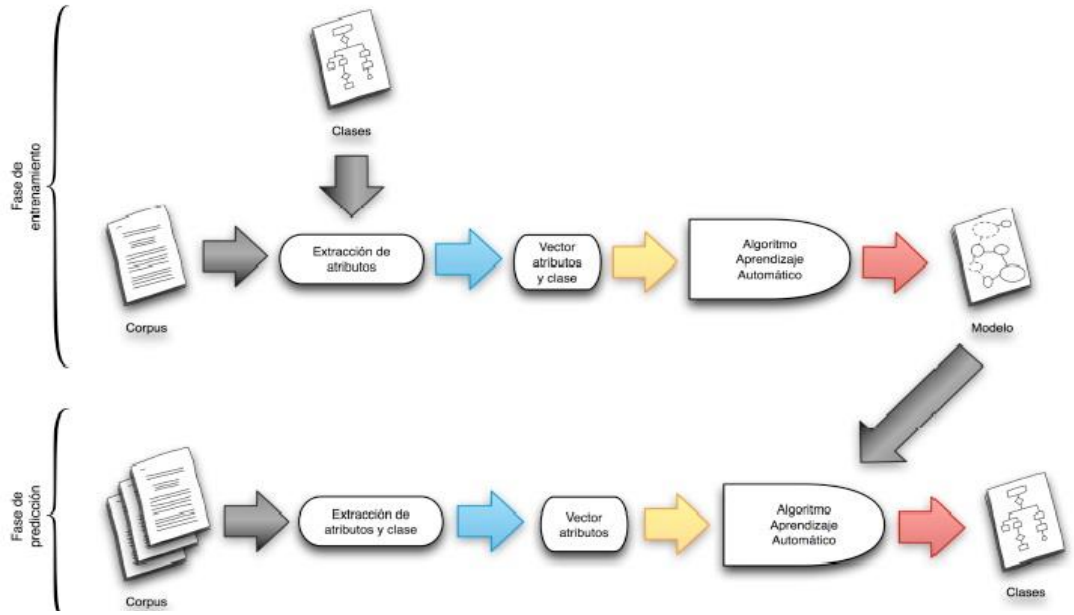


Figura 1. Aprendizaje supervisado.

En primer lugar, se lleva a cabo la fase de entrenamiento. En esta fase a partir de un corpus se extraen los atributos relevantes para el clasificador. Estos atributos se habrán clasificado previamente, por lo que a priori ya se conoce su clase. De la extracción de atributos sale un conjunto de vectores que formarán la entrada al algoritmo, el cual a partir de los datos y sus clases generará un modelo que generalice las características extraídas del corpus. En la siguiente fase, la de predicción, se realizará el mismo proceso de extraer atributos para generar un vector, pero ya no conoceremos la clase de estos ejemplos. El algoritmo, como ya habrá sido capaz de aprender un modelo, podrá aplicarlo a los datos y de esta manera predecir las clases de estos ejemplos.

3.2.1.1.2. Aprendizaje no supervisado. En este tipo de aprendizaje se intenta determinar la estructura de los datos, es decir, se intentan agrupar grandes cantidades de datos en función de las características que comparten. Recibe ese nombre debido a que no se conoce nada a priori. Es una manera de encontrar jerarquía y orden en un conjunto de datos sin estructura. Estas agrupaciones se pueden ver como un conjunto de elementos similares en algunos sentidos, pero diferentes de los elementos que pertenecen a otros grupos.

Uno de los principales métodos del aprendizaje no supervisado es el clustering. Conceptualmente se suele representar viendo qué elementos forman grupos en un eje (x, y), e introduciéndolos en un círculo(Nombela Escobar, 2011).

En el aprendizaje no supervisado para un conjunto de datos de entrada, no conocemos de antemano los datos de salida. Se utiliza para obtener una agrupación coherente de los datos en función de las relaciones entre las variables definidas en la data (Contreras B, 2016).

3.2.1.1.3. Aprendizaje Semi-supervisado. La técnica que utiliza aprendizaje supervisado y no supervisado se llama aprendizaje Semi-supervisado. El principal desafío del aprendizaje Semi-supervisado trata sobre explorar la información que contienen los datos no etiquetados de manera eficiente y efectiva.

Como ya se ha visto obtener datos en el aprendizaje supervisado es generalmente mucho más costoso que obtener datos en el aprendizaje no supervisado. El aprendizaje Semi-supervisado permite aprovechar los datos no supervisados y obtener un modelo predictivo que puede funcionar mejor que el que sólo utiliza datos supervisados. Desde otro punto de vista, permite utilizar una menor cantidad de datos supervisados y obtener el mismo nivel de resultados, es decir, se reduce el esfuerzo en etiquetar, lo que disminuye los costos. Por otro lado, también se ha demostrado que

el aprendizaje Semi-supervisado no ayuda siempre, más bien, depende de si el modelo elegido para utilizar aprendizaje Semi-supervisado es el correcto.(Herrera & Figueroa, 2016).

El aprendizaje Semi-supervisado se puede dividir en dos clases principales, transductive learning e inductive learning:

- Transductive learning: El objetivo en esta clase de aprendizaje es predecir la etiqueta correcta en la fracción de datos no etiquetada (aprendizaje no supervisado) que es utilizada para el aprendizaje Semi-supervisado.
- Inductive learning: En esta clase se intenta obtener la función de aprendizaje que permita predecir la etiqueta correcta en datos nuevos no etiquetados (aprendizaje no supervisado) que no se han utilizado para el aprendizaje Semi-supervisado.

3.2.1.1.4. Aprendizaje por refuerzo (RL). El aprendizaje por refuerzo (reinforcement learning (RL)) es un área dentro del aprendizaje automático (machine learning) dedicada al desarrollo de algoritmos que permiten a un agente (sistema, robot, personaje de videojuego, etc.) aprender a realizar una tarea donde se tienen que tomar decisiones secuenciales para alcanzar un objetivo, maximizando un valor acumulado de recompensa. En RL, un agente puede ser una instancia de diferentes tipos, por ejemplo, un robot, un personaje virtual de un juego de video, o bien puede ser simplemente un sistema o algoritmo encargado de controlar una planta industrial. Existe un gran número de aplicaciones prácticas a este tipo de problemas, donde la principal ventaja es que no se requiere de un experto para encontrar la solución al problema, sino simplemente se debe formular el problema de manera adecuada, especificando las recompensas o penalizaciones para que el agente lo pueda aprender (Omar & Rodríguez, 2015).

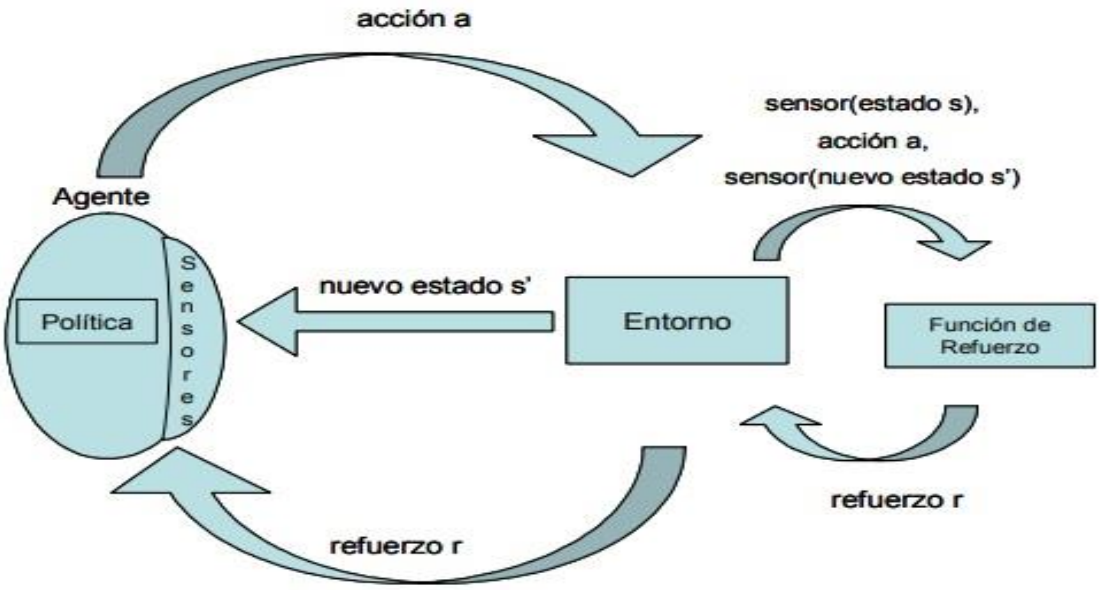


Figura 2. Aprendizaje por refuerzo. El agente interactúa con su ambiente ejecutando una acción, al hacerlo cambia de estado y recibe una recompensa.

3.2.1.2. Métodos de aprendizaje automático.

3.2.1.2.1. Modelos de regresión. Este método se utiliza para predecir el valor de un atributo continuo. Consiste en encontrar la mejor ecuación que atraviese de forma óptima un conjunto de puntos (n-dimensiones). Se utiliza cuando la precisión no es crítica y el número de variables es pequeño (Contreras B, 2016).

Ejemplo: Predecir el precio de una vivienda, dado su tamaño (Figura 3).

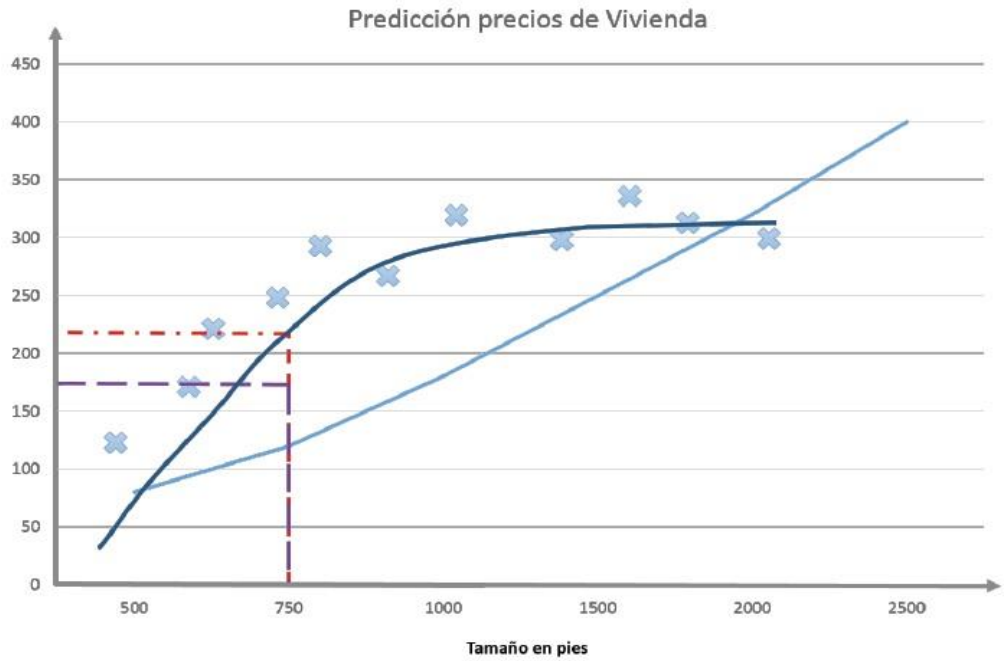


Figura 3. Predicción precios de una vivienda, dado su tamaño.

3.2.1.2.2. Modelos de clasificación.

Método utilizado para predecir un resultado de un atributo con valor discreto (a, b, c, …) dadas unas características $(X_0, X_1, X_2, X_3 \ldots X_n)$. El método simple de clasificación es el binario, donde se clasifica un registro de variables de entrada en 1 o 0. La clasificación múltiple es una extensión de la clasificación binaria.

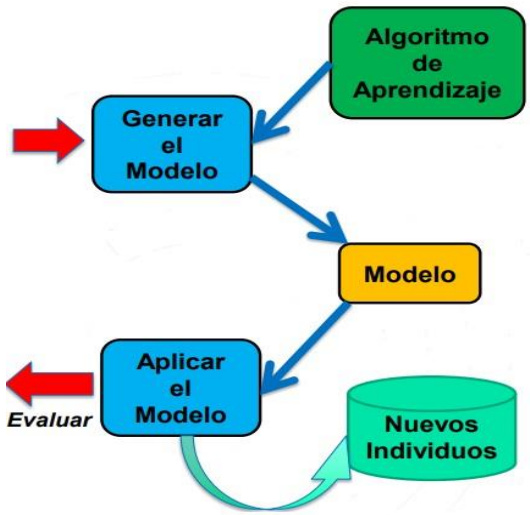


Figura 4. Modelo general de los métodos de clasificación.

3.2.1.2.3. Modelos de agrupación (Clustering). Este método se utiliza cuando se necesita clasificar las instancias de datos, pero no se conocen previamente las categorías. Esta agrupación permite construir grupos (cluster) coherentes de instancias teniendo en cuenta las variables de la data. En palabras sencillas, permite encontrar qué se tiene en los datos.

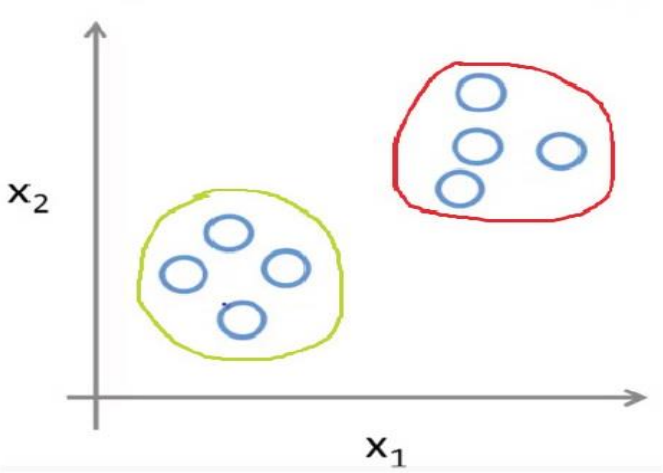


Figura 5. Aprendizaje no supervisado (unsupervised learning).

3.2.2. Sistema experto. Un sistema experto puede definirse como un sistema informático (hardware y software) que simula a un experto humano en un área de especialización dada. Como tal, un sistema experto debería ser capaz de procesar y memorizar información, aprender y razonar en situaciones deterministas e inciertas, comunicar con los hombres y/u otros sistemas expertos, tomar decisiones apropiadas, y explicar por qué se han tomado tales decisiones. Se puede pensar también en un sistema experto como un consultor que puede suministrar ayuda a (o en algunos casos sustituir completamente) el experto humano con un grado razonable de fiabilidad. Es decir, un sistema experto es generalmente el resultado de la colaboración de uno o varios expertos humanos especialistas en el tema de estudio, con los usuarios en mente. Los expertos humanos

suministran el conocimiento básico en el tema de interés, y los ingenieros del conocimiento trasladan este conocimiento a un lenguaje, que el sistema experto pueda entender. La colaboración del experto humano, los ingenieros del conocimiento y los usuarios es, quizás, el elemento más importante en el desarrollo de un sistema experto.

Poder contar con un sistema de experto por lo general es muy caro, pero el mantenimiento y el coste marginal de su uso repetido es relativamente bajo. Por otra parte, la ganancia en términos monetarios, tiempo, y precisión. Resultantes del uso de los sistemas expertos son muy altas, y la amortización es muy rápida. Hay varias razones para utilizar sistemas expertos. Las más importantes son:

- Con la ayuda de un sistema experto, personal con poca experiencia puede resolver problemas que requieren un conocimiento de experto. Esto es también importante en casos en los que hay pocos expertos humanos. Además, el número de personas con acceso al conocimiento aumenta con el uso de sistemas expertos.

- El conocimiento de varios expertos humanos puede combinarse, lo que da lugar a sistemas expertos más fiables, ya que se obtiene un sistema experto que combina la sabiduría colectiva de varios expertos humanos en lugar de la de uno solo.

- Los sistemas expertos pueden responder a preguntas y resolver problemas

- Son mucho más rápidos que un experto humano. Por ello, los sistemas son muy valiosos en casos en los que el tiempo de respuesta es crítico.

- En algunos casos, la complejidad del problema impide al experto humano resolverlo. En otros casos la solución de los expertos humanos no es fiable. Debido a la capacidad de los

ordenadores de procesar un elevadísimo número de operaciones complejas de forma rápida y aproximada, los sistemas expertos suministran respuestas rápidas y fiables en situaciones en las que los expertos humanos no pueden.

- Los sistemas expertos pueden ser utilizados para realizar operaciones monótonas, aburridas e inconfortables para los humanos.

Se pueden obtener enormes ahorros mediante el uso de sistemas expertos. El uso de los sistemas expertos se recomienda especialmente en las situaciones siguientes:

- Cuando el conocimiento es difícil de adquirir o se basa en reglas que solo pueden ser aprendidas de la experiencia.
- Cuando la mejora continua del conocimiento es esencial y/o cuando el problema está sujeto a reglas o códigos cambiantes.
- Cuando los expertos humanos son caros o difíciles de encontrar.
- Cuando el conocimiento de los usuarios sobre el tema es limitado (Castillo, Guti, & Castillo, n.d.).

3.2.3. Validación cruzada. En estadística o minería de datos, una tarea típica es aprender un modelo de los datos disponibles. Tal modelo puede ser un modelo de regresión o un clasificador. El problema con la evaluación de dicho modelo es que puede demostrar una capacidad de predicción adecuada en los datos de entrenamiento, pero puede fallar en predecir datos futuros no vistos. La validación cruzada es un procedimiento para estimar el rendimiento de generalización en este contexto, este método estadístico procura evaluar y comparar algoritmos de aprendizaje

dividiendo los datos en dos segmentos: uno usado para aprender o entrenar un modelo y el otro usado para validar el modelo. En la validación cruzada típica, los conjuntos de capacitación y validación deben cruzarse en rondas sucesivas de modo que cada punto de datos tenga la posibilidad de ser validado.(Payam, Lei, & Huan, 2008)

3.2.3.1. K-fold o validación cruzada de k interacciones: En la validación cruzada de K iteraciones o K-fold cross-validation los datos de muestra se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento (ver figura 6). El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. Este método es muy preciso puesto que evaluamos a partir de K combinaciones de datos de entrenamiento y de prueba. (Londoño, 2016).

K-fold cross-validation es un caso especial de validación cruzada donde se itera sobre un conjunto de datos k veces. En cada ronda, se divide el conjunto de datos en k partes: una parte se usa para la validación, y las partes k - 1 restantes se fusionan en un subconjunto de entrenamiento para la evaluación del modelo, como se muestra en la (ver figura 6), que ilustra el proceso de 5 puntos cruzados de validación. (Raschka, 2018)

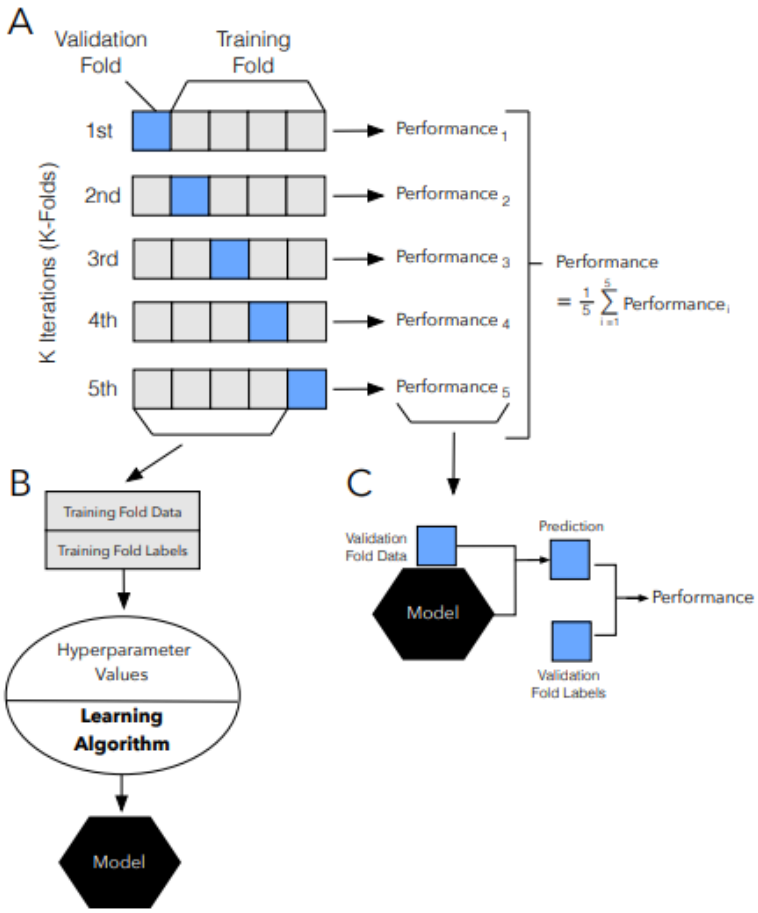


Figura 6. k-fold o validación cruzada para k interacciones.

3.2.3.2. Hold-out. El Hold-out validation o (validación de retención) evita la superposición entre los datos de entrenamiento y los datos de prueba, lo que proporciona una estimación más precisa del rendimiento de generalización del algoritmo. La desventaja es que este procedimiento no utiliza todos los datos disponibles y los resultados dependen en gran medida de la elección de la división de entrenamiento / prueba. Las instancias elegidas para su inclusión en el conjunto de pruebas pueden ser demasiado fáciles o demasiado difíciles de clasificar y esto puede sesgar los resultados. Para hacer frente a estos desafíos y utilizar los datos disponibles al máximo, se utiliza la validación cruzada k-fold (Payam et al., 2008).

El método Hold-out es indiscutiblemente la técnica de evaluación de modelos más simple; se puede resumir de la siguiente manera. Primero, tomamos un conjunto de datos etiquetado y lo dividimos en dos partes: una formación y un conjunto de pruebas. Luego, ajustamos un modelo a los datos de entrenamiento y predecimos las etiquetas del conjunto de pruebas. La fracción de predicciones correctas, que se puede calcular comparando las etiquetas predichas con las etiquetas de verdad básica del conjunto de prueba, constituye nuestra estimación de la precisión de predicción del modelo (Raschka, 2018)

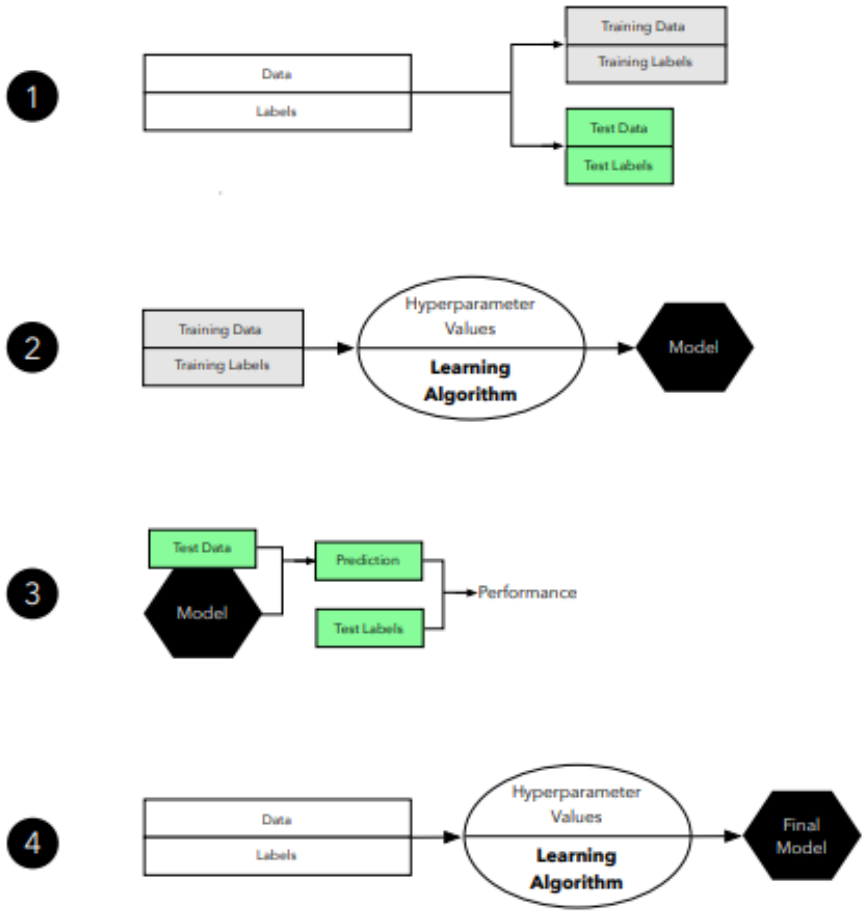


Figura 7. Illustration of k-fold cross-validation for model selection.

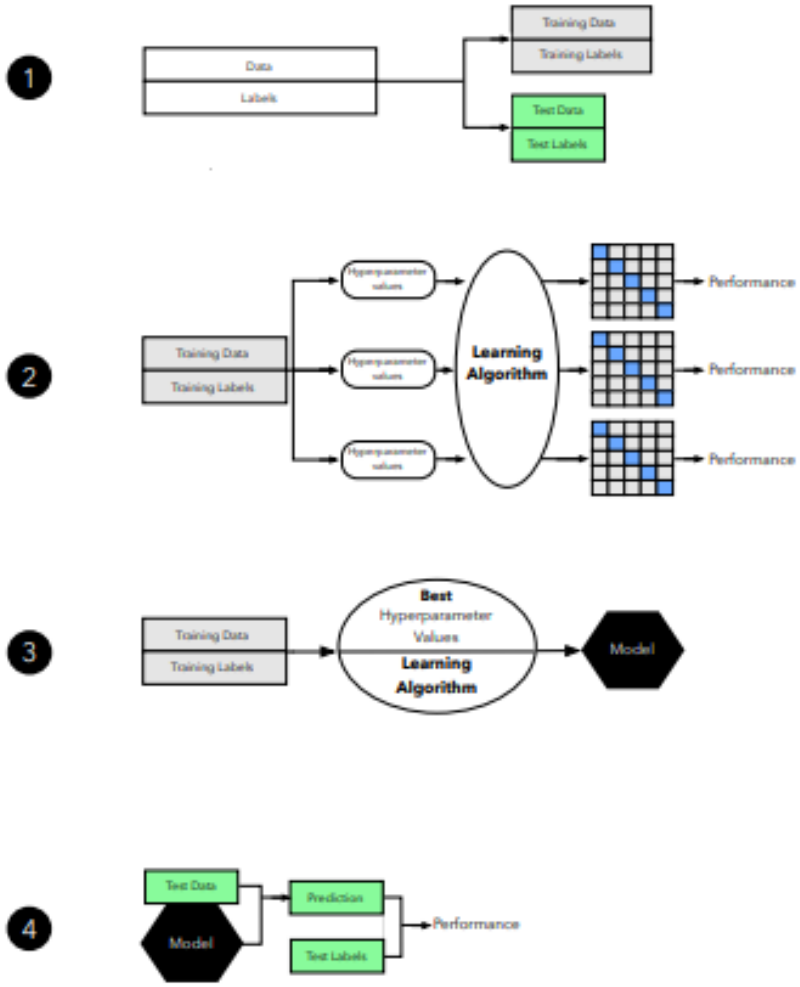


Figura 8. Visual summary of the holdout validation method.

3.2.4. Métodos ensamblados (Ensemble Methods). Los métodos ensamblados se basan en la construcción de un metamodelo que resulta de combinar, usando una determinada regla, un conjunto de modelos como se puede observar en la figura 9, de forma que la diversidad existente en los mismos permite obtener una solución consensuada y, en general, con más efectividad que un modelo único, aunque esto no siempre está garantizado. Intuitivamente, se trata de una técnica muy presente en el día a día para el ser humano.

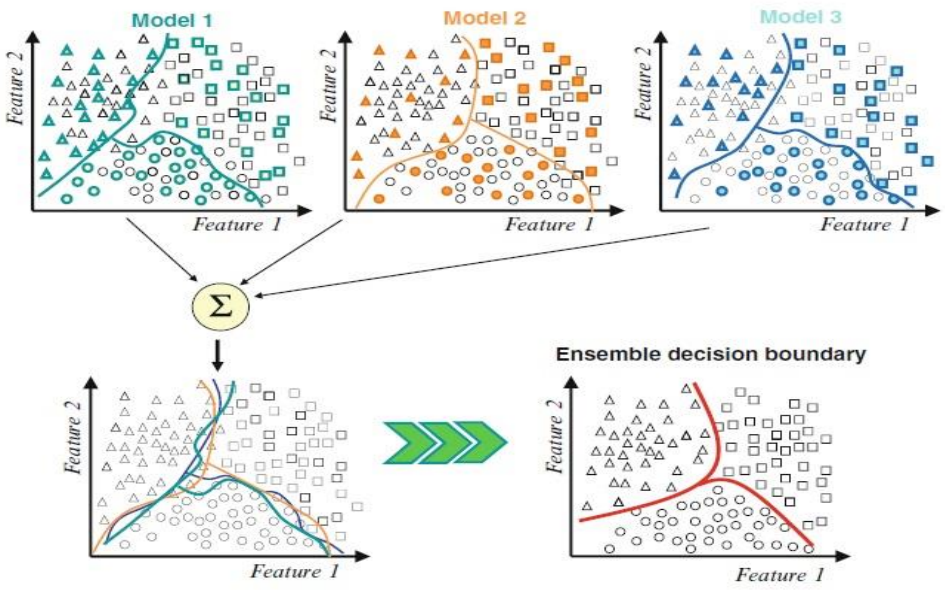


Figura 9. Reducción de la variabilidad mediante Métodos ensamblados. (Variability reduction using Ensemble methods).

Entre los métodos de aprendizaje supervisado, los métodos de ensamble suelen tener un destacado rendimiento en la práctica, sobre todo cuando existe diversidad entre los modelos que combinan. Esa diversidad puede lograrse generando muestras de entrenamiento con ciertas diferencias para cada uno de los modelos. Es decir, los modelos se entrenan con una distribución de los datos distinta de la original. Por ese motivo, los modelos ensamblados pueden ser especialmente apropiados en aquellos problemas en los que existen cambios en la distribución y que además dichos cambios se pueden caracterizar. La idea es entrenar los distintos modelos con distintos conjuntos de entrenamientos generados basándose en los cambios esperados en la distribución (Pablo, n.d.).

Actualmente en la literatura se encuentran distintos métodos de ensamblados, entre los cuales se encuentran, Baggin, Boosting (AdaBoost, XGBoost), Mezcla de expertos. Para el desarrollo de actual proyecto se utilizará el método Bossting más específicamente la variante XGBoost.

Bossting: es un algoritmo que aprovecha el resultado conjunto de otros algoritmos, para mejorar la capacidad que tendrían estos últimos de forma individual. Con el fin de mejorar la capacidad predictiva. (Jesús Garcia, 2018)

XGBoost: El método de XGBoost o (Extra-Gradiente Boosting) es una implementación de código abierto popular y eficiente del algoritmo de árboles aumentados de gradientes. La potenciación de gradientes es un algoritmo de aprendizaje supervisado que intenta predecir de forma apropiada una variable de destino mediante la combinación de un conjunto de estimaciones a partir de un conjunto de modelos más simples y débiles. XGBoost ha funcionado bastante bien en las competiciones de aprendizaje automático gracias a que utiliza con eficacia una amplia gama de tipos de datos, relaciones y distribuciones, y al gran número de hiperparámetros que pueden modificarse y afinarse para mejorar los ajustes. Esta flexibilidad hace que XGBoost sea una opción sólida para los problemas en la regresión y clasificación (binaria y multiclase). (Chen & Guestrin, 2016)

A través del Boosting, el XGBoost ejecuta K iteraciones donde en cada una se añade un nuevo árbol de decisión que intentará corregir el error cometido en las anteriores. En cada iteración del algoritmo XGBoost se añade un nuevo árbol de decisión, que se desarrolla con el objetivo de minimizar el error acumulado por los árboles anteriores. El objetivo del XGBoost es minimizar el error de predicción optimizando la siguiente función:

$$obj(\theta) = \sum_{I}^{N} l(y_i, \mathrm{y}_i) + \sum_{K=1}^{K} \Omega(f_k)$$

Donde el primer término es una función que representa el error cometido entre las predicciones estimadas por XGBoost y los valores reales. Esta función por defecto suele ser el error cuadrático

medio (MSE) en el caso de regresión o la precisión (ACC) para clasificación. Este primer término es personalizable por el usuario en función del objetivo del modelo concreto(Brasas Estéves, 2019).

El segundo término de la fórmula es una función que penaliza la complejidad del modelo para reducir el sobreajuste del modelo, ya que es un problema recurrente al trabajar con XGBoost. Este sobreajuste se debe a la relación de dependencia estadística entre los árboles construidos. Además, el tiempo de construcción del modelo aumenta al aumentar el número de iteraciones.

Los árboles de XGBoost siguen una estructura que ejecuta en cada nodo interno la bifurcación de una variable de la forma $variable < umbral$. Clasifica de esta forma las variables de entrada repetidamente hasta el nodo hoja correspondiente La variable comprobada en cada nodo y sus umbrales se llevan a cabo de forma que minimice lo máximo posible la función objetivo. Una vez en cada nodo hoja se obtiene un valor para la variable en cuestión, el resultado final es el promedio ponderado en función de los pesos de cada árbol del Ensemble como se puede ver en la (figura 10) (Brasas Estéves, 2019).

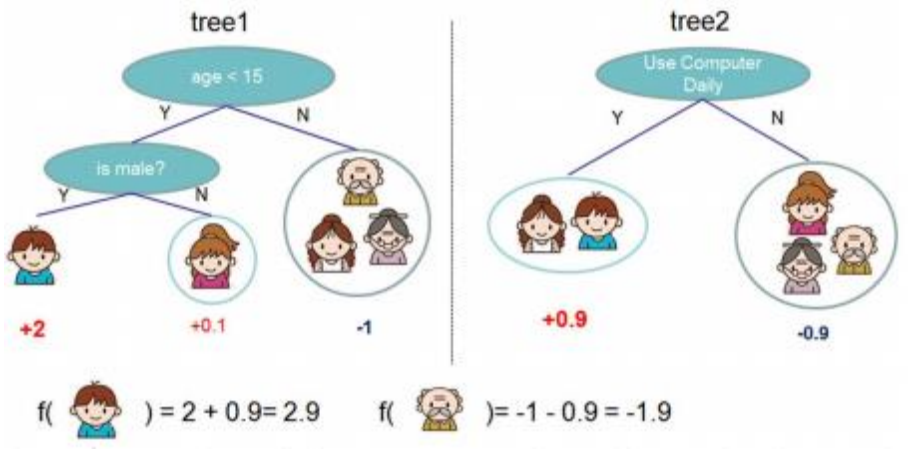


Figura 10. Ejemplo sencillo de XGBoost, Comprueba variables en cada nodo interno hasta llegar al nodo hoja obteniendo así un valor para las variables de entrada.

El XGBoost al igual que el AdaBoost pertenecen al método de ensamble Boosting esto quiere decir que funcionan de igual forma, donde eligen clasificadores fuertes y clasificadores débiles sin embargo AdaBoost se utiliza principalmente para problemas de clasificación y el método de XGBoost es un método flexible ya que permite solucionar problemas tanto de clasificación como problemas de regresión(Brasas Estéves, 2019).

AdaBoost*:* El término de AdaBoost es una contracción de "Adaptive Boosting" el cual fue creado por Freund y Schapiere, es un diseño mejorado del Boosting. Este algoritmo lo que busca es crear un clasificador fuerte, a partir de una combinación de clasificadores débiles o simples. Para esto AdaBoost propone entrenar una serie de clasificadores débiles de manera iterativa, de modo que cada clasificador se enfoque en los datos que fueron erróneamente clasificados, obteniendo como resultado un clasificador global el cual minimizará el error de clasificación (ver figura 10).Para esto cada clasificador simple se aprende dándole un peso diferente a cada una de la iteraciones que se van realizando consecutivamente, a todos los datos se les asigna inicialmente el mismo peso $\alpha = (1/m)$, este peso es el que se va actualizando en cada una de las iteraciones según los ejemplos que estén mal clasificados, por ejemplo a los modelos predichos incorrectamente por el paso anterior se les asignara un peso mayor y los que fueron predichos como correctos tendrán un peso menor, de esta forma lo que se busca es minimizar el error esperado y en enfocarse en poder clasificar en el siguiente paso correctamente los datos que presentan un mayor peso (Schapire, 1996). Es por esta razón que, a medida, en que avanza las interacciones los aprendices que son difíciles de predecir comiencen a recibir una influencia mayor así cada aprendiz subsiguiente débil se vea obligado a concentrase en los ejemplos que los anteriores no observan, de modo que al final se realiza una suma de todos los clasificadores que fueron previamente generados y se obtiene una hipótesis cuya predicción es más acertada.

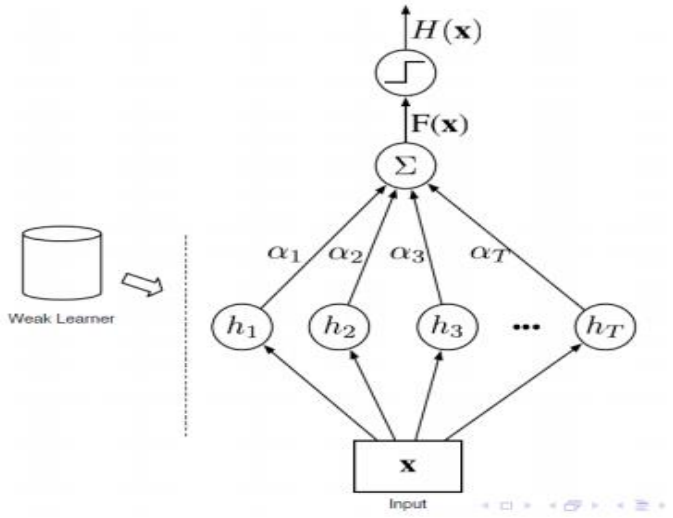


Figura 11. Diagrama de flujo del funcionamiento de AdaBoost.

Para que un clasificador sea efectivo y preciso en sus predicciones, se requiere que cumpla las siguientes condiciones. (1) debe haber sido entrenado con suficientes datos; (2) debe tener un error de entrenamiento bajo; y (3) debe ser simple. (N. Vapnik, 2013).

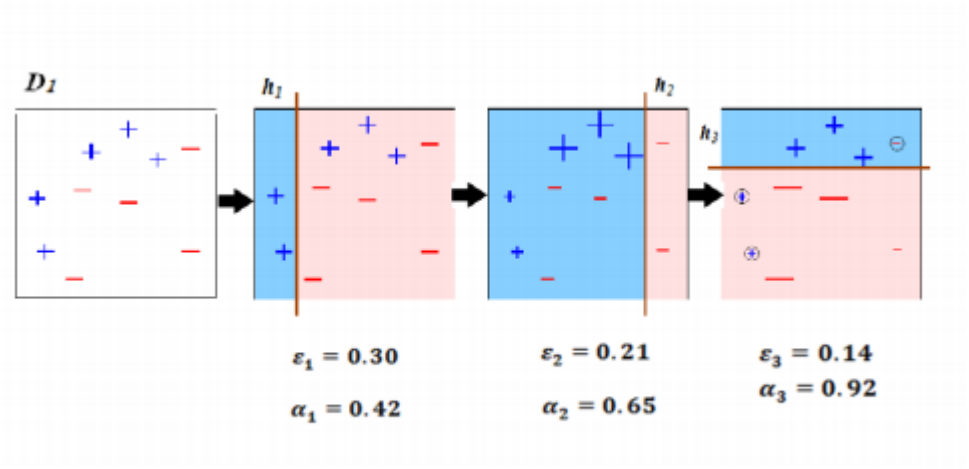


Figura 12. Ejemplo de procedimiento de AdaBoost.

En la figura 12 se observa que se tiene un conjunto de Datos D_1, dónde se puede observar que no es posible encontrar una línea que separe perfectamente los datos positivos de los negativos,

en el caso de AdaBoost y XGBoost se va a suponer que se tiene clasificadores simples, en su primer intento h_1, el clasificador realiza un corte donde se puede evidenciar que hay tres datos positivos que no los toma en cuenta, generando que en su siguiente intento h_2 genera un incremento en el peso de estos tres datos positivos que se clasificaron con error. Por lo que en la siguiente interacción hará que este clasificador presente un mayor enfoque, dándole más importancia a estos puntos haciéndolo quedar condicionado, debido a esto, traza nuevamente otra línea de corte como se puede visualizar en la fig. 12 logrando clasificar a los datos positivos que tenían más peso, pero también incluye a los que tenían un peso menor, igualmente se puede decir que en el segundo intento al clasificar los datos se toman tres negativos quedando estos nuevamente mal clasificados, debido a esto, al realizar un nuevo entrenamiento se incrementará el peso a estos errores y se les disminuye a aquellos que ya fueron seleccionados correctamente para hacer una tercera interacción. Al realizar el tercer clasificador de igual manera este Clasificador se enfocará nuevamente en los ejemplos que contengan un mayor peso y generará un nuevo corte que permitirá resolver el problema.

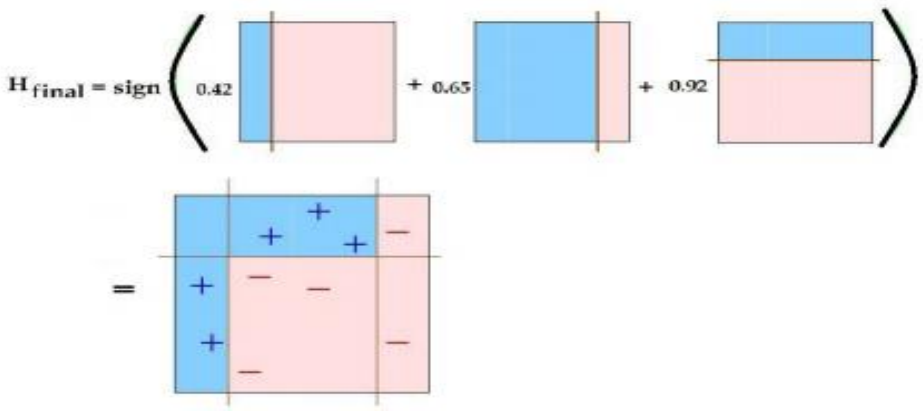


Figura 13. Clasificador Final.

Como se observa en la figura 13 finalmente todos los clasificadores se van a combinar sumando todos los clasificadores creados por cada solución y esta teóricamente debería ser mejor que cada clasificador por separado (Schapire, 1996).

3.2.5. K-vecinos próximos (K-NN). La clasificación KNN fue desarrollado a partir de la necesidad de realizar análisis discriminante cuando son desconocidos o difíciles de determinar estimaciones paramétricas fiables de densidades de probabilidad.

El algoritmo KNN es un método para clasificar objetos según los ejemplos de entrenamiento más cercanos en el espacio de características. KNN es un tipo de aprendizaje basado en instancias, o de aprendizaje lento, donde la función sólo es aproximada a nivel local y todos los cálculos se aplazan hasta la clasificación. El KNN es la fundamental y más simple técnica de clasificación cuando hay poco o ningún conocimiento previo acerca de la distribución de los datos.

Esta regla simplemente retiene todo el conjunto de entrenamiento durante el aprendizaje y asigna a cada consulta una clase representada por el sello mayoría de sus k-vecinos más cercanos en el conjunto de entrenamiento. La regla de vecino más cercano (NN) es la forma más simple de KNN cuando K = 1. En este método cada muestra debe ser clasificada de manera similar a sus muestras circundantes. Por lo tanto, si se desconoce la clasificación de una muestra, entonces podría ser predicha teniendo en cuenta la clasificación de sus muestras de vecinos más cercanos. Dada una muestra desconocida y un conjunto de entrenamiento, todas las distancias entre la muestra desconocida y todas las muestras en el conjunto de entrenamiento se pueden calcular. La distancia con el valor más pequeño corresponde a la muestra en el conjunto de entrenamiento más cercano a la muestra desconocida. Por lo tanto, la muestra desconocida puede clasificarse en base a la clasificación de este vecino más cercano. La Figura 14 muestra la regla de decisión KNN para

K = 1 y K = 4 para un conjunto de muestras dividida en 2 clases. En la Figura 14 (a), una muestra desconocida se clasifica mediante el uso de sólo una muestra conocida; en la Figura 14 (b) se utiliza más de una muestra conocida (Imandoust & Bolandraftar, 2013).

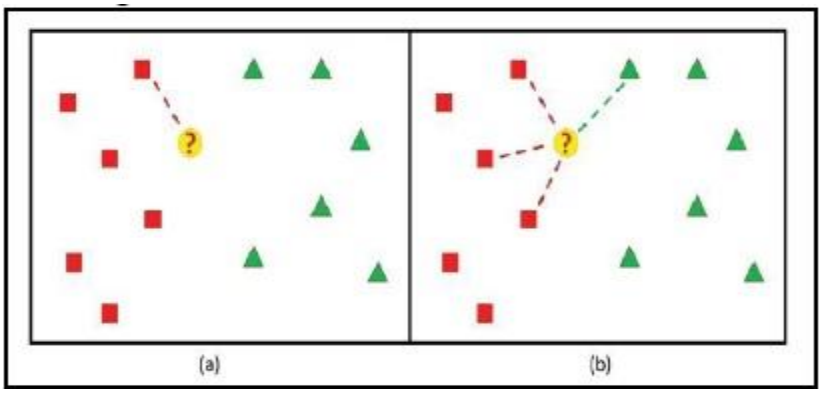


Figura 14. a) Regla de decisión para k=1, b) regla de decisión para k=4.

3.3. Vulnerabilidad

Los términos que a continuación se explicaran son tomados y adaptados según la terminología de la Estrategia Internacional Para la Reducción de Desastres de las Naciones Unidas (UNISDR), cuyo propósito es promover un entendimiento y la utilización en común de conceptos relativos a la reducción del riesgo de desastres, al igual que prestar asistencia a los esfuerzos dirigidos a la reducción del riesgo de desastres por parte de las autoridades, los expertos y el público en general.

Vulnerabilidad: Esta se define como "las características y las circunstancias de una comunidad, sistema o bien que los hace susceptibles a los efectos dañinos de una amenaza". Tal como lo indica la UNISDR, existen diversos aspectos de la vulnerabilidad que surgen de varios factores físicos, sociales, económicos y ambientales. Entre los ejemplos se incluyen el diseño inadecuado y la construcción deficiente de edificios, puentes y vías, la protección inadecuada de los bienes, la falta de información y de concientización pública, un reconocimiento oficial limitado

del riesgo y de las medidas de preparación y la desatención a una gestión ambiental sensata o prudente.

Las siguientes definiciones ejemplifican las principales nociones sobre el concepto de vulnerabilidad

- **Susceptibilidad:** Es el grado de fragilidad interna de un sujeto, objeto o sistema para enfrentar una amenaza y recibir un posible impacto debido a la ocurrencia de un evento adverso.

- **Exposición:** Es la condición de desventaja debido a la ubicación, posición o localización de un sujeto, objeto o sistema expuesto al riesgo.

- **Resiliencia:** Es la capacidad de un sistema, comunidad o sociedad expuestos a una amenaza para resistir, absorber, adaptarse y recuperarse de sus efectos de manera oportuna y eficaz, lo que incluye la preservación y la restauración de sus estructuras y funciones básicas.

VULNERABILIDAD = EXPOSICIÓN x SUSCEPTIBILIDAD / RESILIENCIA

- **Incapacidad:** falta de aptitud o capacidad para reaccionar a un evento riesgoso o también se refiere a la aptitud de una sociedad para actuar ante un peligro.

- **Análisis de riesgos:** Proceso de comprender la naturaleza del riesgo para determinar el nivel de riesgo, es la base para la evaluación de riesgos y las decisiones sobre las medidas de reducción del riesgo y preparación para la respuesta. Incluye la estimación del riesgo

- **Amenaza:** Peligro latente de que un evento físico de origen natural, o causado, o inducido por la acción humana de manera accidental, se presente con una severidad suficiente para causar pérdida de vidas, lesiones u otros impactos en la salud, así como también daños y pérdidas en los bienes, la infraestructura, los medios de sustento, la prestación de servicios y los recursos ambientales.

Colombia se encuentra ubicada en una de las zonas más sísmicas de la tierra, conocida como el anillo circumpacífico, el cual corresponde al borde del océano pacifico donde han ocurrido los últimos grandes sismos como en Chile, Japón y Ecuador. Debido a esta ubicación hace que la vulnerabilidad en todo el territorio sea alta y de acuerdo con estudios de amenaza sísmica realizados a nivel nacional por la Asociación Colombiana de Ingeniería Sísmica –AIS–, cerca del 40% de los colombianos se encuentra en zonas de amenaza sísmica alta y 47% de la población del país está ubicada en zonas de amenaza sísmica intermedia, es decir, el 87% de la población colombiana se encuentra bajo un nivel de riesgo sísmico considerable.

El gran problema que se presenta no es solo por el nivel de amenaza sísmica sino también por el grado de vulnerabilidad que pueden presentar no solo las edificaciones sino también la infraestructura como puentes, carreteras, presas, redes eléctricas, entre otras. Y el caso de la ciudad de Bucaramanga se investigó que la vulnerabilidad vial ante un evento sísmico es inminente y que es ningún caso se ha evaluado el impacto que tendría el colapso de la malla vial en la ciudad.

3.4. Tipo de suelos

3.4.1. Roca no consolidada. Las rocas no consolidadas o sedimentos no consolidados son materiales sueltos, que van desde la arcilla, limo, arenas y gravas, de origen fluvial, aluvial y glacial entre otros se caracterizan porque el agua subterránea fluye a través de los espacios entre

los granos. Los procesos geológicos asimismo pueden erosionar y metamorfosear los sedimentos no consolidados, los terremotos, por ejemplo, pueden licuar sedimentos no consolidados (pero no sedimentos consolidados). Los terrenos de sedimentos no consolidados son vulnerables ante eventos sísmicos, sin embargo, estos terrenos se hacen favorables o menos vulnerables cuando estos sedimentos se encuentran totalmente saturados en agua. (Benjumea, Teixidó, Martínez, & Valls, 2005).

3.4.2. Limolitas con intercalaciones de arenitas, arcillolitas y calizas arenosas.

Limolitas: Shale o lutita se denominan así a las limolitas y arcillolitas mejor consolidadas (ver tabla 2). La marga es una lutita calcárea. Según el grado de consolidación diagenética, pueden clasificarse así:

- De bajo grado de consolidación. Arcillolita, lodolita y limolita.
- De mediano grado de consolidación. Shale arenoso, shale lodoso y limolita laminada.
- De alto grado de consolidación. Argilita, es una roca más resistente y menos deformable que las anteriores, sin significar ello que sea la más durable, pues las lutitas, pueden tener mucho o poco cementante, pero su durabilidad está supeditada a su naturaleza silícea, ferruginosa o calcárea.(Vallejo Velasquez, 2014)

Calizas (ver tabla 2): La caliza es una roca sedimentaria (generalmente de origen orgánico) carbonatada que contiene al menos un 50% de calcita ($CaCO3$), y que puede estar acompañada de dolomita, aragonito y siderita; de color blanco, gris, amarilla, rojiza, negra; y textura granular fina a gruesa, bandeada o compacta, a veces contiene fósiles. Minerales esenciales: calcita (más del 50%). Minerales accesorios: dolomita, cuarzo, goethita (limonita), materia orgánica. Las calizas

tienen poca dureza y en frío reportan efervescencia (desprendimiento burbujeante de CO2) bajo la acción de un ácido diluido. Contienen frecuentemente fósiles, por lo que son de gran importancia en estratigrafía, así como diversas aplicaciones industriales. Usos: el mayor consumo de caliza se efectúa en la fabricación de cementos; es materia prima de la industria química (grandes masas de caliza se utilizan anualmente como fundentes en la extracción de diversas menas metálicas). La caliza de grano fino se emplea en litografía y se denomina caliza 24 litográfica. Calizas de distintos tipos se emplean en construcción, tanto como piedra estructural, como para fachadas y recubrimientos sobre paredes de cemento, y como piedra de acabado para la ornamentación interior. También se usa en la producción de azúcar y en la industria del vidrio.(Suarez, 1987).

Tabla 2.
Características de las rocas arcillolita, caliza, limolita.

Roca	Componente	Características
Arcillolita	Partículas de arcilla	Más del 50% de arcilla
Caliza	Granos de Calcita	Más del 50% de calcita y menos de 25% de arcilla
Limolita	Partículas del tamaño de limos	Más del 50% de los granos menores de 0.06 mm y menos del 25% de arcillas.

Nota: Adaptada de (Deslizamientos: Análisis Geotécnico) Características de las rocas sedimentarias.

3.4.3. Capas rojas constituidas por arenitas, conglomerados y limolitas. Los conglomerados, las areniscas, las limolitas y las arcillolitas (ver tabla 3) son rocas detríticas que

se originan desde partículas que mantienen su integridad física durante el transporte; los carbonatos, las evaporitas, las ferruginosas y los fosfatos son de origen físico - químico, formadas por la precipitación de sustancias que se encuentran en disolución en un líquido de arrastre. Existe otro grupo de rocas sedimentarias llamadas biogénicas, porque en su formación intervienen seres vivos, tales como los carbonatos, los fosfatos y las silíceas, en este grupo se incluyen las que se forman por acumulación de organismos vivos como las calizas de arrecifes; también aquellas que después de su muerte son transportadas y acumuladas, caso de las diatomitas; se incluyen también las tobas calcáreas formadas por la precipitación de CaCO3 propiciada por la acción fotosintética de vegetales; finalmente las rocas orgánicas que son las formadas por acumulaciones de materia orgánica, caso del petróleo y el carbón.(Vallejo Velasquez, 2014).

Tabla 3.
Características de las rocas, conglomerados, limolitas, arenisca.

Roca	Componente	Características
Conglomerados	Partículas grandes redondeadas de roca y fragmentos de minerales	Más del 50% de los granos mayores de 2mm y menos de 25% de arcilla.
Limolita	Partículas del tamaño de limos	Más del 50% de los granos menores de 0.06 mm y menos del 25% de arcillas.
Arenisca	Partículas redondeadas menores que la roca	Más del 50% de los granos entre 2 y 0.06 mm y menos del 25% de arcilla.

Nota: Adaptada de (Deslizamientos: Análisis Geotécnico) Características de las rocas sedimentarias.

3.4.4. Cuarzo feldespático, cuarcita, granulitas y mármoles. Según el glosario técnico minero del ministerio de minas y energía de la república de Colombia (Ministerio de minas, 2003) definen algunos conceptos como se muestra a continuación:

Mármoles: Son aquellas rocas que después de un proceso de elaboración son aptas para ser utilizadas como materiales de construcción, elementos de ornamentación, arte funerario y para escultura, objetos artísticos y variados, y que conservan de manera íntegra su composición, textura y características fisicoquímicas originales.

Cuarcita: la cuarcita es el resultado de metamorfismo de la arenisca cuarzosa la cuales contiene granos de cuarzo de distintos tamaños unidos por un cemento silíceo; contiene también óxidos de hierro que le dan coloraciones muy variadas y, además, hojuela de mica y otros minerales.

Granulita: la granulita es un tipo de roca metamórfica de alto grado que está compuesta característicamente por agrupamientos de minerales anhidros y fueron formadas en condiciones extremas de temperatura y a gran profundidad, son particularmente susceptibles de metamorfismo retrogresivo sobre todo en presencia de fluidos residuales a temperaturas más bajas. En el metamorfismo progresivo hay transformaciones mineralógicas que suponen cada vez un mayor grado de metamorfismo. Las fases minerales son estables a alta presión y a alta temperatura.(Reyes Cortes, 2000)

Cuarzo feldespato: El cuarzo feldespato o también conocido como arcosa es una arenisca que además de contener cuarzo tiene feldespato en una cantidad del orden de 25%. Ambos minerales soportan la degradación mecánica durante el transporte, siendo el segundo más susceptible a la descomposición. Su aparición en proporciones mayores a las de un pequeño porcentaje, evidencian condiciones de aridez y de transporte corto o rápido(Duque Escobar, 2017).

4. Caracterización de la malla vial

4.1. Clasificación de la malla vial de Bucaramanga

Según el estudio realizado por Universidad Industrial de Santander la clasificación vial urbana (Industrial de Santander Universidad & Alcaldía de Bucaramanga, 2010) se realizó con base a la función, uso, localización y características geométricas que cumple dentro de la estructura urbana de la movilidad. Esta jerarquización se efectúo con el fin de planear los operativos de campo y poder definir los métodos a utilizar para el inventario. La siguiente es la clasificación de la malla vial de Bucaramanga:

- Vía primaria: Son las vías que por su alto nivel de tráfico son principales para la movilidad de la ciudad. Constituyen las vías que alimentan zonas urbanas y permiten conectarse con las vías intermunicipales y a su vez se identifican por su función de estructuración de actividades intraurbanas.

- Vía secundaria: Las vías secundarias son aquellas que se caracterizan por su función de polos de atracción de la actividad urbana y están orientadas a canalizar el tráfico lento, público y privado.

- Vía terciaria: La red terciaria corresponde al restante de los elementos viales que se integran y dan continuidad a la malla vial existente.

A continuación, (ver figura 15) se muestra como está distribuida la clasificación vial en Bucaramanga.

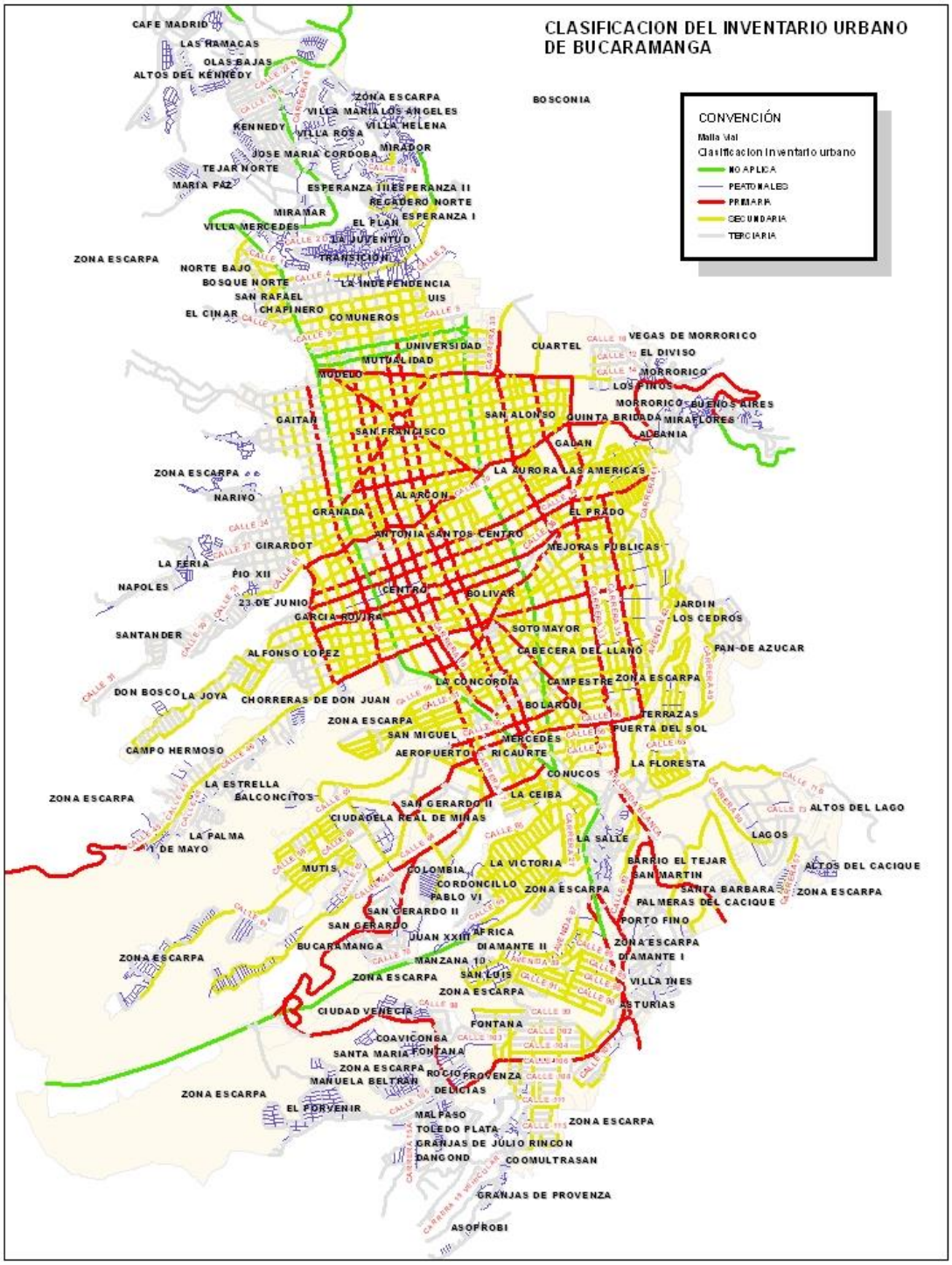


Figura 15. Clasificación vial en Bucaramanga.

4.2. Tipos de pavimento

Los pavimentos son estructuras constituidas por capas superpuestas, relativamente horizontales, conformadas por materiales apropiados y adecuadamente compactados, que resisten los esfuerzos de las cargas repetidas del tránsito durante el período para el cual fue diseñada. Los pavimentos deben cumplir los siguientes requisitos, con el fin de brindar comodidad y seguridad durante el tiempo de funcionamiento del mismo: Ser resistentes ante la acción de las cargas impuestas por el tránsito y los agentes de intemperismo, tener una textura superficial que brinde seguridad y comodidad al usuario, durable, capacidad de drenaje, silenciosa, económico. Los pavimentos según su superficie de rodadura se clasifican en: flexibles, rígidos, afirmados y articulados o en adoquines de concreto.(Hernández Cepeda, 2018)

Dentro del estudio realizado entre la Universidad Industrial de Santander y la Alcaldía de Bucaramanga, se menciona que la ciudad de Bucaramanga cuenta con estos 4 tipos de pavimento, los cuales según sus características se usan de acuerdo con el tipo y cantidad de flujo vehicular que circula sobre él. Estos son:

4.2.1. Pavimento rígido. Se compone de losas de concreto hidráulico que en algunas ocasiones presenta un armado de acero, una subbase y una capa llamada subrasante (ver figura 16). Esta tiene un costo inicial más elevado que el flexible, sin embargo, su periodo de vida varía entre 20 y 40 años; el mantenimiento que requiere es mínimo y solo se efectúa (comúnmente) en las juntas de las losas (Universidad Nacional Autónoma de México, n.d.). Los elementos y funciones de un pavimento rígido son:

- Subrasante: Es la capa de terreno de una carretera que soporta la estructura de pavimento y que se extiende hasta una profundidad que no afecte la carga de diseño que

corresponde al tránsito previsto. (ver figura 16). (Universidad Nacional Autónoma de México, n.d.)

- Subbase: Es la capa de la estructura de pavimento destinada fundamentalmente a soportar, transmitir y distribuir con uniformidad las cargas aplicadas a la superficie de rodadura de pavimento, de tal manera que la capa de subrasante la pueda soportar absorbiendo las variaciones inherentes a dicho suelo que puedan afectar a la subbase. La subbase debe controlar los cambios de volumen y elasticidad que serían dañinos para el pavimento. (ver figura 16). (Universidad Nacional Autónoma de México, n.d.)

- Losa de concreto hidráulico: Es la capa superior de la estructura de pavimento, construida con concreto hidráulico, por lo que, debido a su rigidez y alto módulo de elasticidad, basan su capacidad portante en la losa, más que en la capacidad de la subrasante, dado que no usan capa de base (ver figura 16).(Universidad Nacional Autónoma de México, n.d.)

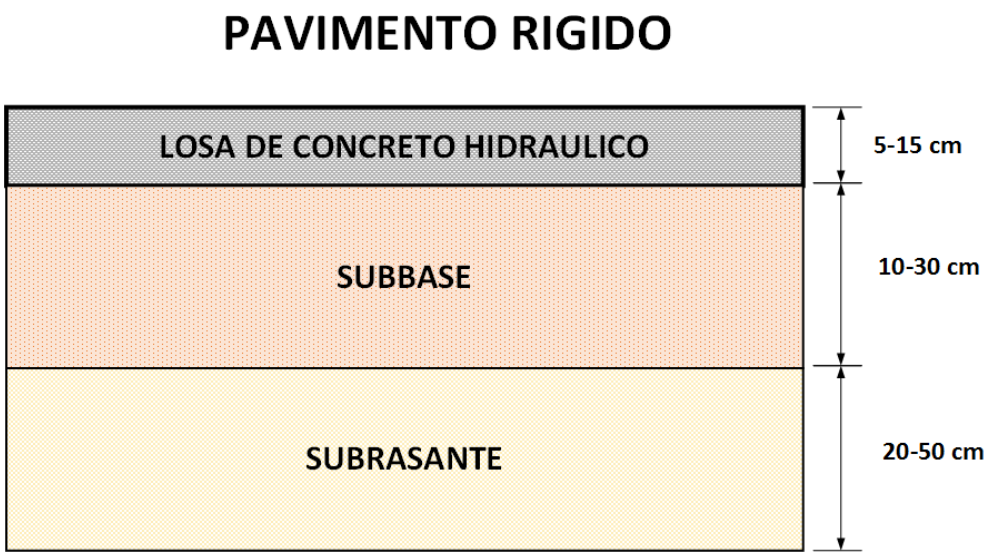


Figura 16. Capas de pavimento rígido.

4.2.2. Pavimento flexible. El pavimento flexible (ver figura 17) resulta ser el más económico por su construcción inicial, tiene un periodo de vida de entre 10 y 15 años, pero tienen la desventaja de requerir mantenimiento constante para cumplir con su vida útil. Este tipo de pavimento está compuesto principalmente de una carpeta asfáltica, construida sobre una capa de base y una capa de subbase las que usualmente son de material granular y por último estas capas descansan en una capa de suelo compactado, llamada subrasante ver figura. (Sánchez Rivera, 2006).

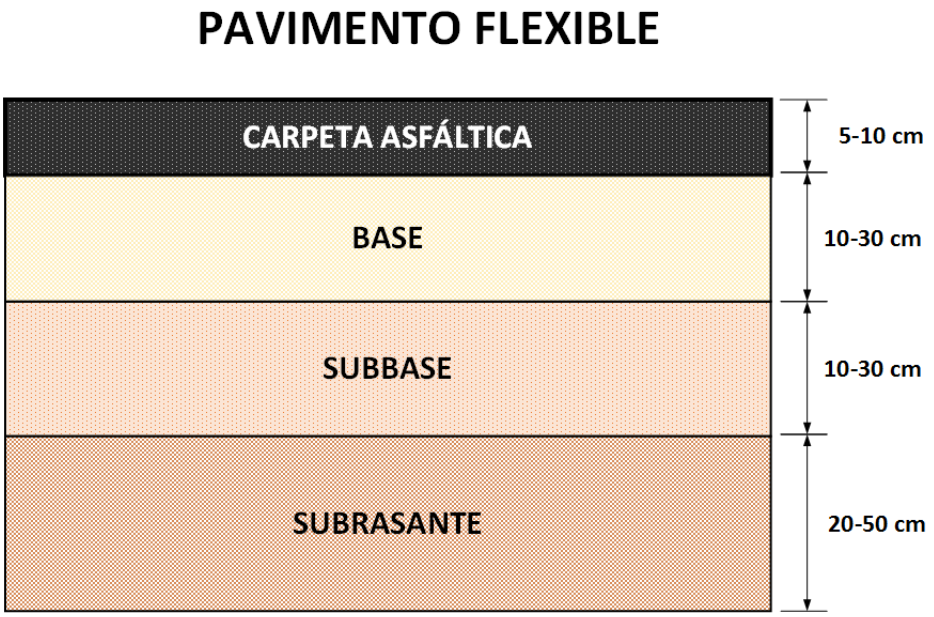


Figura 17. Constitución pavimento flexibles.

4.2.3. Pavimento adoquín. Los pavimentos articulados se definen como el conjunto de elementos prefabricados hechos de concreto, que son instalados sobre una superficie para brindar acabado, resistencia, durabilidad y vida útil. Dichos elementos son fabricados de manera mecánica garantizando un resultado homogéneo, que instalados sobre una base adecuada posibilitará tanto el tránsito de vehículos como de peatones (Hernández Cepeda, 2018).

Un pavimento articulado de concreto generalmente consiste en un suelo de subrasante, subbase granular (opcional), base granular, colchón de arena, el pavimento de bloques de concreto y el borde de confinamiento. El diseño y construcción de pavimento articulado varía con el clima, condiciones de disponibilidad de materiales, métodos de diseño, condiciones de suelo y cargas de tráfico. La colocación de los elementos puede ser manual o mecánica (Armijos Cuenca, 2011). La (figura 18) presenta los elementos que componen un pavimento articulado.

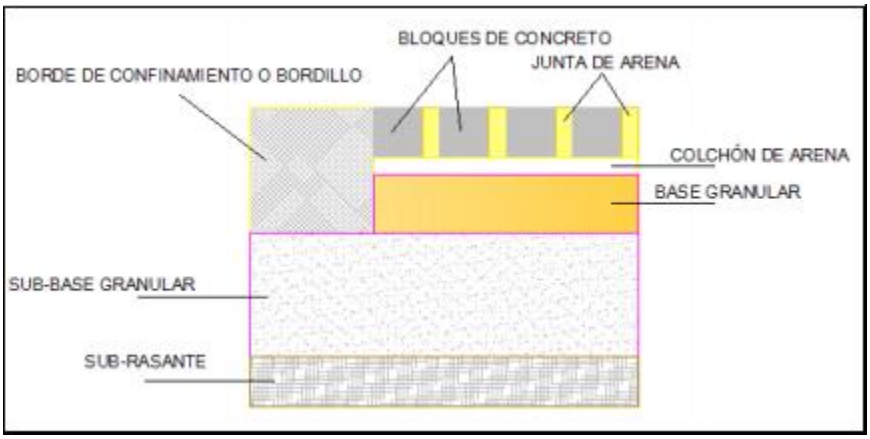


Figura 18. Componentes Tradicionales de un Pavimento Articulado de Concreto. Adaptado de Estudio del diseño estructural y constructivo de pavimentos articulados en base a bloques de asfalto (Armijos Cuenca, 2011).

- Subrasante: resultado posterior a la excavación y se constituye en la base del pavimento articulado. Esta debe encontrarse libre de materia orgánica y de ser necesario se compacta para brindar mayor estabilidad.(Hernández Cepeda, 2018).

- Base y Sub-base granular: estas capas que se encuentran entre la subrasante y el colchón de arena; dependiendo del tipo de suelo y especificaciones de diseño será necesario la

aplicación de una capa de subbase precediendo a la base. Su objetivo es aumentar la capacidad de soporte de la estructura del pavimento, por lo cual, estas capas están conformadas por materiales granulares, suelo estabilizado o con poca cantidad de concreto.(Hernández Cepeda, 2018).

- Colchón de arena: como su nombre lo indica para esta capa se implementa arena gruesa, limpia y sin residuos de materia orgánica; tiene como función asentar los adoquines y filtrar el agua que se presente en la estructura (Hernández Cepeda, 2018).

- Juntas de arena: como su nombre lo indica para esta capa se implementa arena gruesa, limpia y sin residuos de materia orgánica; tiene como función asentar los adoquines y filtrar el agua que se presente en la estructura (Hernández Cepeda, 2018).

- Adoquines: deben cumplir con las especificaciones del diseño y normativa, esto hace referencia a formas geométricas, color, textura, resistencia a la flexión y compresión.(Hernández Cepeda, 2018).

4.2.4. Pavimento afirmado. El pavimento afirmado es una capa compactada de material granular natural o procesado con gradación específica que soporta directamente las cargas y esfuerzos del tránsito (ver imagen 19). Debe poseer la cantidad apropiada de material fino cohesivo que permita mantener adheridas todas las partículas. Funciona como superficie de rodadura en gran parte de la red vial terciaria (Fernando et al., 2018).

Figura 19. Pavimento afirmado.

4.3. MDR (Modified Distress Rating)

El MDR (Modiffied disstres Rating) es un índice que nos muestra la calidad superficial de un pavimento, las variables de este índice son las fallas encontradas en cada tramo de vía estudiado (ver tabla 20):

Fisuras (longitudinal y transversal)	Grietas en bloque
Piel de cocodrilo	Ahuellamiento
Hundimiento	Ondulación
Baches	Descascaramiento
Cabezas duras	Desgaste superficial
Exudación	Parches
Grietas de esquina	

Este indicador del estado de pavimento MDR tiene valores entre 0 y 100. Donde 0 simboliza una vía completamente destruida y 100 una vía en perfecto estado (Santander, 2010).

Tabla 4.
Rangos MDR.

Estado superficial	Rango MDR
Muy buenos	100-80
Buenos	80-55
Regulares	55-30
Malos	0-30

Nota: tomado de (Industrial de Santander Universidad & Alcaldía de Bucaramanga, 2010).

4.4. Distancia a fallas

En geología, una falla es una fractura o zona de fracturas a lo largo de la cual ha ocurrido un desplazamiento relativo de los bloques paralelos a la fractura. Esencialmente, una falla es una discontinuidad que se forma debido a la fractura de grandes bloques de rocas en la Tierra cuando las fuerzas tectónicas superan la resistencia de las rocas. Cuando la actividad en una falla es repentina y brusca, se puede producir un gran terremoto, provocando incluso una ruptura en la superficie terrestre. Lo que genera y se evidencia en la superficie del terreno es una forma topográfica llamada escarpa de falla. (Instituto nacional de prevención sísmica – inpres- 1, 1980)

En este caso se les denomina distancia a fallas a la mínima distancia euclidiana que hay entre cada una de las fallas y cada tramo de vía de la malla vial de Bucaramanga; esta distancia se presenta en kilómetros y lo que indica es que tan cerca está el tramo de vía a cada falla para deducir así su vulnerabilidad.

Sin embargo, cada falla tiene asociada una serie de características presentadas a continuación:

Tabla 5.
Caracterización de las fallas de Bucaramanga.

FALLA	VELOCIDAD DE DESPLAZAMIENTO	LONGITUD	ZONAS AFECTADAS
Falla de Bucaramanga - Morrorico	10 mm/año	220 Km	Recorre el costado oriental del perímetro urbano correspondiente a los barrios de Pan de Azúcar y lagos del Cacique en dirección Norte-Sur.
Falla del Suárez - Chimitá	0.01 a 0.1 mm/año	170 Km	Sigue el trazo del Río de Oro, para luego chocar con la falla de Bucaramanga,
Falla Chitota – Río Suratá		Se extiende hacia el Noreste en el costado Norte de la ciudad, aproximadamente paralela al Río Suratá,	Escarpa Norte de la Meseta de Bucaramanga

Nota: Adaptado de unidad nacional para la gestión del riesgo de desastres- Colombia (2013) plan municipal del riesgo de desastres de Bucaramanga (página 10).

4.5. Caracterización de los tipos de suelos

La ciudad de Bucaramanga cuenta con diferentes tipos de suelo alrededor de su área, esto hace que algunas superficies sean más vulnerables que otras; gran parte del área de Bucaramanga está formada por roca no consolidada y en un porcentaje menor hay áreas conformadas por otro tipo de rocas o suelos.

Para la caracterización de los tipos de suelos en el área de Bucaramanga para la presente investigación se tuvo en cuenta información geológica del servicio geológico colombiano (sgc) en la cual se indagó en una serie de mapas en la cual presentaba la constitución geológica de Santander y más específicamente de Bucaramanga. Los datos encontrados en los mapas arrojaron cuatro tipos suelos geológicos de los cuales está constituido Bucaramanga:

- Roca no consolidada
- Limolitas con intercalaciones de arenitas, arcillolitas y calizas arenosas
- Capas rojas constituidas por arenitas, conglomerados y limolitas
- Cuarzo feldespático, cuarcita, granulitas y mármoles

Dicha información se pasó a formato Kml con el fin de que se pudiese modificar en el software RStudio, seguidamente se realizó una intersección con los datos de la malla vial con el fin de saber que tramos viales se encontraban en cada tipo de suelo. Y al final se pasó a variables binarias (0 y 1) donde 1 significaba que ese tramo vial se encontraba en ese tipo de suelo y 0 lo contrario.

4.6. Caracterización de la Vulnerabilidad.

Para la caracterización de la vulnerabilidad de la malla vía de Bucaramanga se tuvo en cuenta una cantidad de variables que a criterio del experto tenían mayor influencia sobre las vías tales como: tipo de pavimento, clase de inventario, MDR, distancia a fallas, tipo de geología presente; dichas variables fueron transformadas a variables binarias (0 y 1) con el fin de facilitarle al experto el análisis del cálculo.

Ante una evaluación de vulnerabilidad desarrollada por un experto es necesario saber que la percepción del riesgo es diferente en cada persona. Es por ello que los datos entregados fueron descritos a juicio del experto, quien realiza la etiqueta de los datos y estima los daños esperados en la malla vial de Bucaramanga ante un sismo de diseño.

5. Caso de estudio

Surge la necesidad de realizar un estudio de vulnerabilidad vial en la ciudad de Bucaramanga debido a que es una ciudad constantemente afectada por movimientos telúricos y además que carece de estudios de vulnerabilidad de la malla vial, ya que los organismos correspondientes no han identificado la importancia necesaria a la malla vial como para implementar estrategias y planes de acción en la prevención de desastres. Es por ello que este proyecto presenta una solución basada en técnicas de aprendizaje automático, que permita identificar e intervenir de forma correctiva aquellos tramos viales con el fin de disminuir o reducir los niveles de vulnerabilidad.

El caso de estudio se realiza para la malla vial del municipio de Bucaramanga, Empleando las técnicas de K-vecinos próximos (K-NN) y métodos ensamblados (“Boosting”). Para esta investigación se hizo la recolección de una base de información con aproximadamente 7044 datos los cuales contenían información de algunas características específicas sobre los tramos viales de la ciudad Bucaramanga: Clasificación de la malla vial, Tipo de pavimento, Daño superficial del pavimento (MDR), entre otros. Además, se añadió a esta a información la distancia (km) de las fallas geológicas a los tramos viales, las cuales se obtuvieron con el programa estadístico Rstudio.

Y la conformación del tipo de suelo sobre la que está construida la ciudad. De esta base de datos se extrajo una muestra de 100 datos, los cuales fueron evaluados y analizados por un experto quien a partir de estos determinó un índice de vulnerabilidad.

A partir de este estudio realizado por el experto se desea replicar su conocimiento por medio del aprendizaje automático y la modelación de dos algoritmos ya mencionados anteriormente. Para la ejecución de los modelos por medio del método de regresión, se hizo un entrenamiento en cada uno de los códigos utilizando el software Rstudio, para esto se obtuvo que considerar que todas las variables fueran de tipo cuantitativo por lo que las variables que tenían características cualitativas como la clasificación de las vías se tuvieron que pasar a una variable binaria asignado 1 si la vía era de clase primaria, secundaria o terciaria o de lo contrario se asignaba 0. Igualmente sucede con los tipos de pavimento y tipos de suelo, donde se aplica la misma metodología. Una vez transformadas las variables se procede realizar el entrenamiento para consecutivamente probar cada uno de los modelos con el conjunto de datos de prueba verificando la eficiencia en los resultados de la regresión con el RMSE (raíz del error cuadrático medio), con este valor se determinó cuál de los dos algoritmos tiene mayor precisión al momento de replicar los datos analizados por el experto

5.1. Conjunto de datos

Según lo planteado por el experto el nivel de vulnerabilidad existente es afectado por 12 variables, estas divididas en factores estructurales y factores no estructurales como se muestra en la tabla 6. Los factores estructurales son aquellos que están relacionados directamente con el elemento o con las condiciones físicas de la estructura y los factores no estructurales son aquellos que, aunque no

están relacionados directamente con la estructura pueden llegar a influir en su daño estructural al encontrarse en el mismo entorno.

Tabla 6.
Factores asociados a la vulnerabilidad vial.

TIPO	FACTOR
Factores estructurales	• Clase de inventario: primaria, secundaria, terciaria • Tipo de pavimento: flexible, rígido, adoquín, afirmado • MDR
Factores no estructurales	• Falla de Bucaramanga • Falla del rio Suratá • Falla Chitota- rio Suratá • Falla Chitota- rio Suratá 2 • Falla Suarez – Chimitá • Roca no consolidada • Limolitas con intercalaciones de arenitas, arcillolitas y calizas arenosas • Capas rojas constituidas por arenitas, conglomerados y limolitas • Cuarzo feldespático, cuarcita, granulitas, y mármoles

5.2. Selección de los parámetros

Se puede decir que el aprendizaje automático de un algoritmo tiene dos tipos de parámetros: uno de ellos es el conjunto de parámetros de modelo, datos que le pasamos para su entrenamiento y

sobre los que el algoritmo se ajusta. El otro conjunto de parámetro son los valores que podemos ajustar al mismo algoritmo para que este realice su aprendizaje a partir de estos, estos son conocidos como híper-parámetros, aquellos que no se aprenden dentro del mismo algoritmo en su entrenamiento (Zamorano Ruiz, 2018) la determinación de los parámetros óptimos empelando la búsqueda exhaustiva es la medida implementada que permite controlar el sobreajuste de los modelos.

El sobreajuste u "overfitting" se produce cuando un modelo obtiene muy buenos resultados con los datos de entrenamiento, pero su precisión es notablemente más baja con el conjunto de test. Esto se produce porque el modelo se ha adaptado a los valores del conjunto de entrenamiento y no es capaz de generalizar para los datos que no ha procesado (Lima Miranda, 2018).

Para la aplicación de los algoritmos se utilizó una base de datos que contiene información y características de la malla vial de la ciudad de Bucaramanga, de esta base de datos se seleccionó una muestra aleatoria de 100 datos los cuales fueron analizados por un experto quien asoció un índice de vulnerabilidad en una escala de 0 a 1. A partir de la muestra generada en la presente investigación se hizo una partición de entrenamiento que comprendía el 75% como datos de muestra y el 25% restante como datos de prueba, con los que se validaban los modelos.

Para la ejecución del proyecto se determinan dos tipos de modelo de aprendizaje automático para predecir el índice de vulnerabilidad a los tramos viales de la ciudad siendo estos: K-vecinos próximos (K-NN) y Método ensamblado (Boosting). La metodología para la selección de hiper-parámetros de los modelos anteriormente mencionados consistió en realizar una búsqueda exhaustiva que permitiera realizar diferentes iteraciones posibles entre los rangos definidos para cada uno de los modelos bajo una estrategia de validación cruzada. Las características

computacionales utilizadas para los algoritmos dependen del tipo de computador en el cual se ejecutó el código de programación.

Las características del ordenador empleado en la ejecución de los modelos:

Procesador: Intel(R) Core (TM) i3-4005U CPU @ 1.70 Hz 1Ram.70 GHz

Memoria RAM: 4,00 GB

Sistema operativo: 64 bits

En la tabla 7 se describen los pasos a seguir para la ejecución de los modelos

Tabla 7.
Proceso de ejecución de los modelos de aprendizaje automático.

Paso	Descripción
1	Importar las librerías Necesarias para la ejecución de los modelos.
2	Descargar la base de datos
3	Elegir el porcentaje de datos correspondiente al entrenamiento del modelo, de forma pseudo-aleatoria se selecciona el conjunto de datos de entrenamiento, el restante pasará a ser el conjunto de datos de prueba con los que se validaran los modelos de aprendizaje automático.
4	Convertir el conjunto de entrenamiento y prueba en matrices (solo para XGBoost)
5	Verificar que los datos seleccionados no contengan valores faltantes
6	Definir los rangos de valores a evaluar en cada parámetro para obtener los valores óptimos, técnica conocida como búsqueda exhaustiva.
7	Entrenar el modelo con los valores óptimos obtenidos en el paso anterior
8	Validar el modelo con el conjunto de datos de prueba
9	Determinar la eficiencia del modelo bajo el criterio del Error cuadrático (RMSE).

5.2.1. K-vecinos próximos (K-NN). Para el desarrollo de este algoritmo se hizo uso de la librería "caret" de Rstudio. Esta librería nos facilita la función K-NN y comprende la función Train-control, que requiere de un conjunto de parámetros los cuales nos permiten controlar el sobreajuste del modelo y así poder obtener el valor óptimo para K que corresponde al número de vecinos a tener en cuenta en la estimación de una nueva muestra, este parámetro es de gran importancia ya que depende de la calidad del modelo a obtener. El valor de k se da por el método de validación cruzada, explicado en la sección 3, consta de realizar diferentes iteraciones y repetir n veces este proceso de esta forma se obtiene el valor óptimo, este último se obtiene al hallar el menor RSME o el mayor valor para Rsquared. A continuación, se muestran algunos de los resultados obtenidos al probar distintos valores de K bajo el método validación cruzada con n=5. En la tabla 8 se muestran el RMSE y Rsquared, obtenidos para cada uno de los valores, se puede evidencia que a medida que K toma valores altos incrementa el error medio cuadrático.

Tabla 8.
Estimación RMSE y RSQUARED con diferentes valores de K en el modelo de K-vecinos próximos.

K	RMSE	RSQUARED
1	0.03019501	0.9148320
5	0.03808278	0.9032551
15	0.06492101	0.8490494
30	0.10240639	0.6737798
55	0.12597593	0.1719230
62	0.12663108	0.3096948
85	0.12663108	0.3096948
100	0.12663108	0.3096948

Para esta investigación estudio el valor que optimo es para K=1, debido a que representa el menor RMSE de 0,03019501 y un Rsquared de 0,9148320 el cual representa un porcentaje muy bueno donde se puede afirmar que los datos se ajusta a los datos.

Los datos de prueba y las estimaciones del modelo en este conjunto de prueba presentan una fuerte correlación, por lo que sus valores de RMSE y Rsquared fueron 0.0443 y 0.8675 respectivamente. Sin embargo, en la figura 20, se puede observar la existencia de un valor atípico el cual puede afectar en los resultaos de la predicción del modelo.

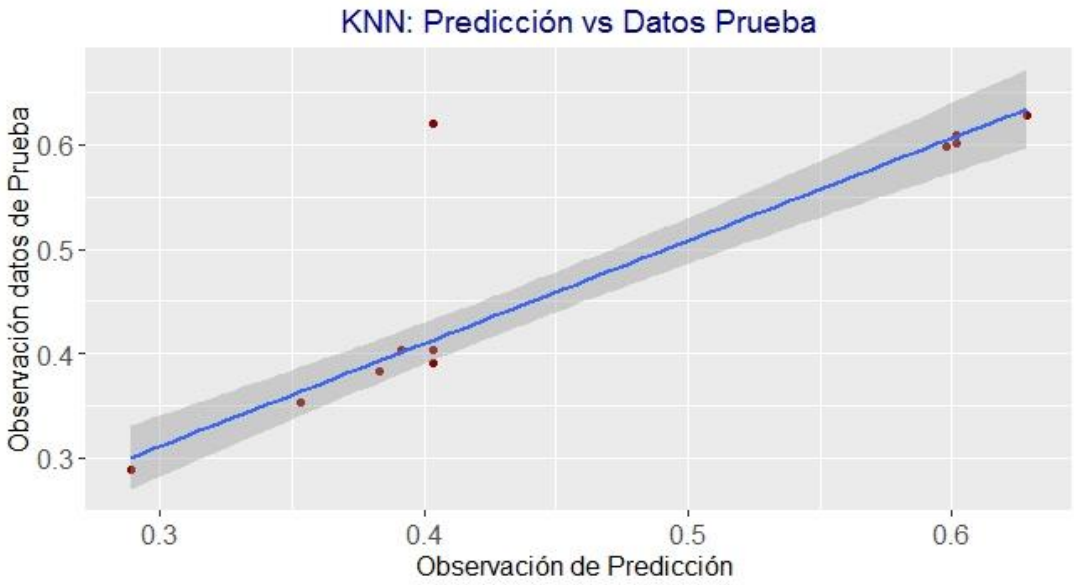


Figura 20. Comportamiento de la observación de datos de prueba contra la observación de la predicción en el modelo K-vecinos próximos.

5.2.2. Método ensamblado (“Boosting”). Para el desarrollo del algoritmo Ensemble Methods más exactamente Boosting (XGBoost) se hizo el uso de un conjunto de librerías las cuales permitieron realizar la selección de los parámetros de una forma más sencilla.

Las librerías que se utilizan para el desarrollo del algoritmo son, XGBoost, Matrix, Caret entre otras, pero la que define exactamente el modelo del método ensamblado "Boosting" es la librería XGBoost esta librería facilita seleccionar los parámetros para el controlar el sobreajuste del modelo, entre los cuales esta Nround [60,100] cumple la función de controlar el número máximo de iteraciones asimismo debe ser ajustado por el método de validación cruzada n=5, Max_depth este nos especifica la profundidad máxima del árbol [10,15,20,25]; Colsample_bytree es la

porción de submuestra [0.1, 0.26, 0.42, 0.58, 0.74, 0.9] de columnas al construir cada árbol, eta[0.1, 0.26, 0.42, 0.58, 0.74, 0.9] cumple la función de controlar la velocidad de aprendizaje la cual varía dependiendo el valor; gamma [0,5,10,15] se encarga de la regularización del modelo es decir que evita en cierta parte el sobreajuste, entre más alto sea el valor mayor será el sobreajuste; Min_child_weight [0,5,10,15] corresponde al número mínimo de iteraciones necesarias en cada nodo; por último la función que cumple Subsample en el XGBoost es que muestrea un porcentaje de los datos de entrenamiento antes de implantar árboles, con el fin de evitar el sobreajuste en este caso se establece una variación de [0.1, 0.5, 0.9], los valores por parámetro se detallan en la Tabla 9.

Tabla 9.
Valores asignados al rango de parámetros para el método de ensamble Boosting (XGBoost).

Parámetros	Rango
Nrounds	60,100
Max_depth	10,15,20,25
Colsample_bytree	0.1, 0.26, 0.42, 0.58, 0.74, 0.9
eta	0.1, 0.26, 0.42, 0.58, 0.74, 0.9
gamma	0,5,10,15
Min_child_weight	0,5,10,15
Subsample	0.1, 0.5, 0.9

Establecido la variación de los parámetros se ejecuta el modelo y esta muestra las medidas de error por cada combinación establecida al variar cada uno de los parámetros bajo el método validación cruzada. En la tabla 10 se muestran los datos de RMSE y Rsquared obtenidos en la variación de cada uno de los parámetros, a partir de estos, se puede observar que los parámetros que obtuvieron el mejor RMSE: 0.0129 y Rsquared: 0.989, fueron: Eta=0.9, max_depth=10, gamma=0, Colsample_bytree= 0.9, Nround=60.

Tabla 10.
Resultados al ejecutar el modelo variando cada uno de los parámetros.

eta	**max_depth**	**gamma**	**colsample_bytree**	**Nrounds**	**RMSE**	**Rsquared**
0.1	10	0	0.1	60	**0.079**	**0.693**
0.1	10	0	0.1	100	**0.126**	**0.054**
0.1	10	0	0.3	60	**0.058**	**0.825**
0.1	10	0	0.3	100	**0.049**	**0.875**
0.1	10	0	0.5	60	**0.035**	**0.935**
0.1	10	0	0.5	100	**0.035**	**0.936**
0.1	10	0	0.7	100	**0.102**	**0.663**
0.1	10	0	0.7	60	**0.025**	**0.964**
0.1	10	0	0.9	60	**0.032**	**0.956**
0.1	10	0	0.9	100	**0.027**	**0.964**
0.9	10	0	0.9	60	**0.0129**	**0.989**

De la siguiente grafica (figura 21) se puede concluir que los datos de prueba en función de la predicción tienen fuerte correlación, es decir que los datos predichos por el modelo no se alejan de los datos de prueba, además se puede observar que los valores atípicos existentes no se encuentran muy alejados esto hace que no afecten la efectividad de la predicción.

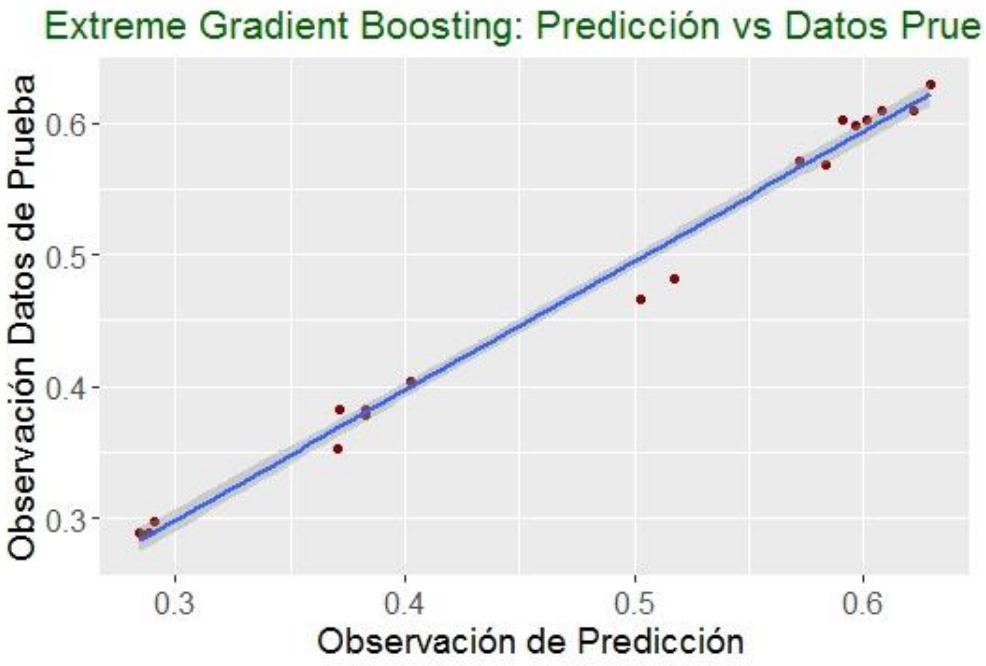


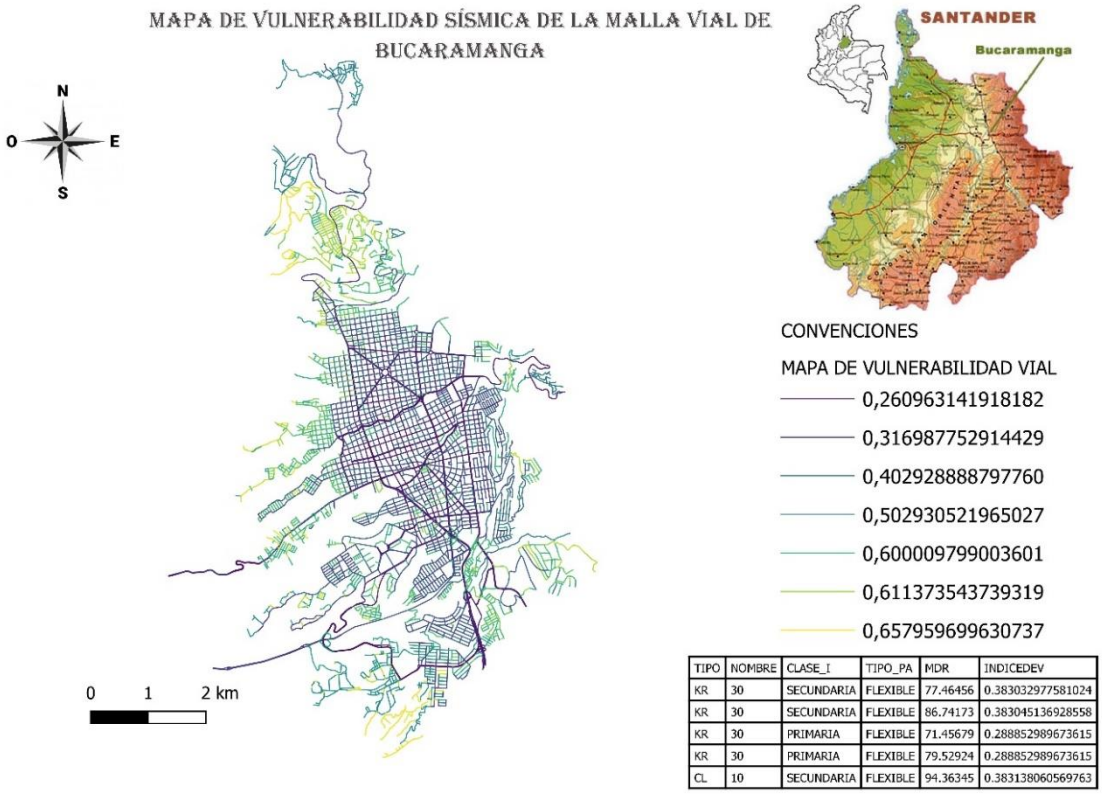
Figura 21. Comportamiento de la observación de datos de prueba contra la observación de la predicción en el modelo XGBoost.

Una vez se ejecutan los algoritmos y se validad los resultados de los valores de prueba en cada uno de los modelos propuestos en esta investigación, se concluye que el modelo que obtuvo una mayor precisión de replicar el conocimiento del experto es el Método ensamblado Boosting (XGBoost) ya que obtuvo la disminución más representativa del error medio cuadrático de 1.29%, como se muestra a continuación, tabla 11.

Tabla 11.
Resultado de los valores de prueba RMSE y Rsquared en cada uno de los modelos.

Modelo	RMSE	Rsquared
K-NN	**0.0443**	**0.8675**
XGBoost	**0.0129**	**0.989**

Una vez seleccionado el modelo óptimo, este es ejecutado para predecir el índice de vulnerabilidad de toda la malla vial de la ciudad de Bucaramanga y como resultado final se obtiene el siguiente mapa (ver figura 22), para una mejor visualización de los atributos ver (APÉNDICE D):



TIPO	NOMBRE	CLASE_I	TIPO_PA	MDR	INDICEDEV
KR	30	SECUNDARIA	FLEXIBLE	77.46456	0.383032977581024
KR	30	SECUNDARIA	FLEXIBLE	86.74173	0.383045136928558
KR	30	PRIMARIA	FLEXIBLE	71.45679	0.288852989673615
KR	30	PRIMARIA	FLEXIBLE	79.52924	0.288852989673615
CL	10	SECUNDARIA	FLEXIBLE	94.36345	0.383138060569763

Figura 22. Mapa de vulnerabilidad sísmica de la malla vial de Bucaramanga.

6. Conclusiones

En la presente investigación se desarrolló un modelo de aprendizaje automático supervisado que permitiera evaluar la vulnerabilidad sísmica de la malla vial de la ciudad de Bucaramanga a partir

de muestras etiquetadas por un experto, debido a que realizar estudios de vulnerabilidad por un experto pueden representar costos muy elevados incluso llevar mucho tiempo su análisis. Debido a lo anterior se decidió la realización de este proyecto con el fin de replicar el conocimiento de un experto en evaluaciones de vulnerabilidad sísmica a partir de conjuntos de datos pequeños.

Al indagar la documentación en la revisión de literatura se pudo constatar que la gran mayoría de estudios relacionados con vulnerabilidad sísmica se han enfocado generalmente en evaluaciones de vulnerabilidad edificaciones, dejando de lado la infraestructura vial. Razón por la cual se orienta este estudio a esta línea investigación que representa gran importancia en la intervención rápida o esfuerzos de recuperación ante un evento sísmico de gran magnitud.

A través de estudios realizados por diferentes entidades sobre la infraestructura vial de Bucaramanga se logra obtener información clara y precisa que al consolidarla presenta una serie de características que facilitan el desarrollo de este proyecto en la búsqueda de querer evaluar la vulnerabilidad sísmica de la malla vial.

Los modelos de aprendizaje automáticos propuesto en esta investigación, permite predecir el índice de vulnerabilidad para la malla vial de Bucaramanga logrando resultados competitivos. La métrica de error utilizada fue la raíz del error medio cuadrático (RMSE) en conjuntos de prueba, bajo este criterio se obtienen resultados de 4,43% y 1,29% para K-vecinos próximos (K-NN) y Método ensamblado (Boosting) respectivamente. En resumen, se concluye que el algoritmo que representa el mejor rendimiento para la evaluación de los datos y nos permite dar solución a esta investigación es el método de ensamblado Boosting (XGBoost), el cual presenta la mayor reducción del error (RMSE) y por tal motivo se aproxima más al criterio del experto.

7. Recomendaciones

Para futuros proyectos enfocados en la predicción de la vulnerabilidad vial, es importante que se indague más a fondo sobre los diferentes parámetros tales como: potencial de deslizamiento, intensidad sísmica, potencial de licuefacción entre otros; que pueden llegar a influir en la predicción del índice vulnerabilidad de la malla vial ante un evento sísmico debido a que se cree que los parámetros que se tuvieron en cuenta en esta investigación solo representan una pequeña parte necesaria para obtener una buena confiabilidad.

A partir de nuestra experiencia como autores, identificamos la importancia de consultar y obtener información a partir de la participación de diferentes expertos en el desarrollo de estos temas y así mismo proponer otros métodos de aprendizaje automáticos con el fin de hallar modelos que logren obtener una mayor precisión en la evaluación de vulnerabilidad sísmica.

Referencias bibliográficas

Andrea, A. N. D., Afiso, S. A. C., & Ondorelli, A. N. C. (2005). Pure and Applied Geophysics Methodological Considerations for the Evaluation of Seismic Risk on Road Network, 162, 767–782. https://doi.org/10.1007/s00024-004-2640-0

Argyroudis, S., & Gehl, P. (2015). Systemic Seismic Risk Assessment of Road Networks Considering Interactions with the Built Environment, 30, 524–540. https://doi.org/10.1111/mice.12136

Armijos Cuenca, V. F. (2011). Estudio del diseño estructural y constructivo de pavimentos articulados en base a bloques de asfalto. PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE.

Benamar, D. A., Naili, M., & Amellal, O. (2018). Author ' s Accepted Manuscript. International Journal of Disaster Risk Reduction. https://doi.org/10.1016/j.ijdrr.2018.06.014

Benjumea, B., Teixidó, T., Martínez, P., & Valls, P. (2005). Sísmica de alta resolución en sedimentos no consolidados : ejemplos en áreas de depósitos de origen fluvio-deltaico en Cataluña deltaic areas in Catalonia, 1–4.

Bormann, U., & Brauchitsch, B. Von. (2017). Artificial Intelligence and Robotics and Their Impact on the Workplace, (April).

Brasas Estéves, P. (2019). Análisis Estadístico de Datos Para el Mantenimiento Prendictivo de Maquinaria Industriale. Universidade Da Coruña.

Campos García, A., Díaz Giraldo, C., Rubiano Vargas, D., & Costa Posada, C. (2012). Análisis de la gestión del riesgo de desastres en Colombia. Banco Mundial, 1, 438.

Castillo, E., Guti, M., & Castillo, E. (n.d.). Sistemas Expertos y Modelos de Redes Probabilısticas.

Chen, T., & Guestrin, C. (2016). XGBoost : A Scalable Tree Boosting System. University of Washington.

Cifuentes Urrego, A. Y., & Fuentes Jerez, J. L. (2018). Evaluación de vulnerabilidad sísmica actual de albergues temporales en Bucaramanga aplicando algoritmos de clasificación supervisada.

Contreras B, F. A. (2016). Introducción a machine learning, metodos y tipos de machine learning, 1, 25.

Duque Escobar, G. (2017). Manual de geología para ingenieros, 25.

Fernando, L., Alzate, M., Lota, L. F., David, J., Gonzalez, B., Julio, C., & Laitón, T. (2018). Mejoramiento de vías terciarias - vías de tercer orden 15, 1.0.

Hernández Cepeda, Y. beatríz. (2018). Pavimentos de adoquines de concreto una solucion

ambiental en la construcción de infraestructura vial colombiana. UNIVERSIDAD MILITAR NUEVA GRANADA.

Herrera, I., & Figueroa, A. (2016). Aprendizaje Semi-Supervisado de Múltiples Vistas para Detectar Temporalidad de Preguntas, (August). https://doi.org/10.13140/RG.2.1.1403.1602

Imandoust, S. B., & Bolandraftar, M. (2013). Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background, 3(5), 605–610.

Industrial de Santander Universidad, & Alcaldía de Bucaramanga. Inventario de infraestructura de la malla vial (2010).

Instituto nacional de prevención sísmica – inpres- 1. (1980). Fallas Geológicas (p. 11).

Lima Miranda, E. (2018). Aplicación de algoritmos de aprendizaje automático para predecir la disfunción cognitiva en pacientes de esclerosis múltiple mediante de conectividad estructural. Universitat Oberta de Cataluña.

Londoño, A. F. (2016). Algoritmos de clasificación lineal para la identificación de zonas cerebrales. UNIVERSIDAD TECNOLÓGICA DE PEREIRA.

Ministerio de minas. (2003). GLOSARIO TÉCNICO MINERO, 168.

Nombela Escobar, B. (2011). Aplicación de técnicas de aprendizaje automático para la extracción de información en textos farmacológicos.Aplicación de técnicas de aprendizaje

automático para la extracción de información en textos farmacológicos.

Omar, E., & Rodríguez, G. (2015). Aprendizaje por refuerzo mediante transferencia de conocimiento cualitativo.

Pablo, P. (n.d.). Uso de Ensembles en Problemas con Cambios de Distribución.

Payam, R., Lei, T., & Huan, L. (2008). Cross-validation. ARIZONA STATE UNIVERSITY, 6.

Quezada Martínez, D. O. (2017). Diseño de un sistema de apoyo a la toma de decisiones - DSS para la gestión de las etapas pre- desastres de sismos en Bucaramanga, basado en técnicas de aprendizaje automático(Machine learning).

Raschka, S. (2018). Model Evaluation , Model Selection , and Algorithm Selection in Machine Learning arXiv : 1811 . 12808v2 [cs . LG] 3 Dec 2018, 49.

Reyes Cortes, M. (2000). Petrografia y petrologia metamorfica.

Sánchez Rivera, S. E. (2006). Ampliación y reconstrucción de la carretera federal México-Puebla de la ciudad de Cholula a Santa María Zacatepec. Universidad de las Americas Puebla.

Santander, U. I. de. (2010). Inventario de infraestructura vial, 1–178.

Schapire, E. (1996). Cap ´ Clasificadores D ´ ebiles - AdaBoost, 21–31.

Serna A, A., Acevedo M, E., & Serna M, E. (2017). Principles of Artificial Intelligence in Computer Science Principios de la Inteligencia Artificial en las Ciencias Computacionales, 354–361.

Serrano, A. G. (2012). Iinteligencia artificial. Fundamentos.

Suarez, J. (1987). Deslizamientos: Análisis geotécnico.

Tinaquero, D. M. (2018). Detector predictivo de conexiones fraudulentas.

Universidad Nacional Autónoma de México. (n.d.). Diseño de Pavimentos Rígidos, 14–47.

Vallejo Velasquez, J. C. (2014). Manual de geología: capítulo 9. Rocas sedimentarias. UNIVERSIDAD NACIONAL DE COLOMBIA SEDE MANIZALES.

Zamorano Ruiz, J. (2018). Comparativa y análisis de algoritmos.