CONSTRUCCIÓN DE UN REPOSITORIO LIMPIO DE DATOS PARA LA DETECCIÓN DE PATRONES DE EVENTOS ERUPTIVOS DEL VOLCÁN GALERAS CON TÉCNICAS DE MINERÍA DE DATOS

LISBETH VIVIANA ROSERO LEGARDA
YEHIMY ALEXANDRA CABRERA CABRERA

UNIVERSIDAD DE NARIÑO
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS
SAN JUAN DE PASTO
2014

CONSTRUCCIÓN DE UN REPOSITORIO LIMPIO DE DATOS PARA LA DETECCIÓN DE PATRONES DE EVENTOS ERUPTIVOS DEL VOLCÁN GALERAS CON TÉCNICAS DE MINERÍA DE DATOS

LISBETH VIVIANA ROSERO LEGARDA
YEHIMY ALEXANDRA CABRERA CABRERA

Trabajo de grado presentado como requisito para optar al título de Ingeniero de Sistemas

Director del Proyecto
RICARDO TIMARÁN PEREIRA, Ph. D.

UNIVERSIDAD DE NARIÑO
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS
SAN JUAN DE PASTO
2014

NOTA DE RESPONSABILIDAD

”Las ideas y las conclusiones aportadas en el presente trabajo son responsabilidad exclusiva de sus autores”

Artículo 1, acuerdo No. 324 de octubre 11 de 1966, emanado por el Honorable Consejo Directivo de la Universidad de Nariño.

“La Universidad de Nariño no se hace responsable de las opiniones o resultados obtenidos en el presente trabajo y para su publicación priman las normas sobre el derecho de autor”.

Artículo 13, Acuerdo N. 005 de 2010 emanado del Honorable Consejo Académico.

Nota de aceptación:

_____
_____
_____
_____
_____
_____

_____

Firma del presidente del jurado

_____

Firma del jurado

_____

Firma del jurado

San Juan de Pasto, Mayo de 2014

DEDICATORIA

A Dios Por haberme permitido llegar hasta este punto y haberme dado salud para lograr mis objetivos, además de su infinita bondad y amor.

A mis Padres Mary y Silvio
Por haberme apoyado en todo momento, por sus consejos, sus valores, por la motivación constante que me ha permitido ser una persona de bien, pero más que nada, por su amor. Por los ejemplos de perseverancia y constancia que los caracterizan y que me han infundado siempre para salir adelante.

A mis familiares A mi hermana Aleyda por ser el ejemplo de una hermana mayor y de la cual aprendí aciertos y de momentos difíciles; a mi hermano Fabián por ser quien me inspira para también darle un buen ejemplo y me motiva a nunca decaer; a Wilmer Segovia por todo el apoyo brindado y a todos aquellos que participaron directa o indirectamente en la consecución de este logro.

Yehimmy Alexandra Cabrera Cabrera

DEDICATORIA

A Dios, por bendecirme, por darme las fuerzas necesarias para seguir adelante, por regalarme la compañía de mi familia.
A mi Mamita, por sus enseñanzas, por su ejemplo, por su valentía, Dedicación, esfuerzo y por la maravillosa Madre que fue, por cuidar de mi desde pequeña, y sé que desde el cielo lo sigue haciendo en cada paso que doy

A mi padre y mi Madre por su apoyo, perseverancia y valores inculcados que hicieron que todo esto posible.

A mis hermanos quienes me inspiran en todo momento.

Lisbeth Viviana Rosero Legarda

AGRADECIMIENTOS

A Dios por siempre estar ahí en todo el camino recorrido y su fortaleza para llegar hasta el final y alcanzar nuestro objetivo.

A nuestras familias por su apoyo, comprensión y amor.

A nuestro director de proyecto de grado al **Doctor Ricardo Timarán Pereira,** por su gran apoyo y motivación para la culminación de nuestros estudios profesionales y para la elaboración de este trabajo.

Al ingeniero del grupo de investigación **GRIAS**. En especial al profe **Andrés Oswaldo Calderón**, por su colaboración, apoyo, paciencia y tiempo dedicado en el desarrollo del proyecto.

Al Observatorio Vulcanológico y Sismológico de Pasto, A los ingenieros Diego Gómez y John Meneses por su tiempo, paciencia, colaboración y atención

A nuestros profesores por su tiempo compartido y por impulsar el desarrollo de nuestra formación profesional.

A nuestros amigos y compañeros, con quienes vivimos tantos momentos durante nuestra trayectoria en la Universidad hasta este punto, apoyándonos mutuamente en nuestra formación profesional.

A todas aquellas personas que contribuyeron en la realización de este proyecto de investigación.

RESUMEN

En este documento se presentan los resultados del proyecto de investigación cuyo objetivo fue la construcción de un repositorio limpio de datos para descubrir patrones de eventos eruptivos del volcán Galeras a partir del historial de emisiones y erupciones que posee el Observatorio Vulcanológico y Sismológico de Pasto (**OVSP**), situado en el departamento de Nariño en el Municipio de San Juan de Pasto (Colombia), utilizando técnicas de Minería de Datos. Utilizando la metodología CRISP-DM, se construyó un repositorio de datos con la información de los eventos dados en el periodo desde 1989 hasta el 2013. Se descubrieron reglas que conllevan a un resumen de posibles eventos eruptivos que se encuentran clasificados como erupciones y emisiones de gases o ceniza, utilizando las tareas de minería de datos como son clasificación, asociación y agrupación. El conocimiento generado permitirá tomar decisiones rápidamente para afrontar una posible catástrofe producto de una erupción del Volcán Galeras.

Palabras claves: Minería de datos, CRISP-DM, Volcán Galeras, Clasificación, Asociación y Agrupamiento, OVSP.

ABSTRACT

This document presents the results of a research project whose aim was to build a clean data repository to find out patterns of eruptive events from Galeras volcano based on the record of emissions and eruptions owned by the Volcanological and Seismological Observatory of Pasto, located in the Nariño department in the municipality of San Juan de Pasto (Colombia), by using Data Mining techniques. By using the CRISP-DM methodology, a data repository was built with the information on the events that took place in the period from 1989 to 2013. Rules that lead to a summary of possible eruptive events that are classified as eruptions and ash or gas emissions were discovered using the data mining tasks such as classification, association and clustering were discovered. The knowledge generated will allow to make decisions quickly to deal with a potential disaster product of an eruption of Galeras Volcano.

Keywords: Data Mining, CRISP- DM, Galeras Volcano, Classification, Clustering and Association

CONTENIDO

Pág.

LISTA DE TABLAS

Pág.

LISTA DE FIGURAS

Pág.

LISTA DE ANEXOS

Pág.

GLOSARIO

Algoritmo. Conjunto de operaciones y procedimientos que deben seguirse para resolver un problema.

Área de Conocimiento. Agrupación que se hace de los programas académicos, teniendo en cuenta cierta afinidad en los contenidos, en los campos específicos del conocimiento, en los campos de acción de la educación superior cuyos propósitos de formación conduzcan a la investigación o al desempeño de ocupaciones, profesiones y disciplinas.

CRISP-DM. Por sus siglas en ingles “Cross Industry Estándar Process for Data Mining”. Metodología usada para la aplicación de minería de datos.

Data Warehouse. Colección de datos orientados al tema, integrados y no volátiles para la toma de decisiones.

DCBD. Siglas de Descubrimiento de Conocimiento en Bases de Datos.

KDD. Por sus siglas en inglés “Knowledge Discovery from Data Bases” corresponde al descubrimiento de conocimiento en bases de datos.

SGBD. Sistema Gestor de Bases de Datos, es una colección de programas cuyo objetivo es servir de interfaz entre la base de datos, el usuario y las aplicaciones. Se compone de un lenguaje de definición de datos, de un lenguaje de manipulación de datos y de un lenguaje de consulta.

BALÍSTICO (FRAGMENTO): fragmento de roca expulsado violentamente por una erupción volcánica y que sigue una trayectoria balística, en forma de elipse.

BASALTO: roca de origen volcánico de color gris oscuro, que contiene menos de 53% de sílice. En estado fundido presenta una baja viscosidad, que implica una erupción generalmente no explosiva que produce flujos de lava.

BLAST: explosión volcánica de un cuerpo de magma cercano a la superficie. Este fenómeno puede deberse a un deslizamiento de una parte de un edificio volcánico. Un "blast" es una mezcla caliente de baja densidad de fragmentos de roca, ceniza y gases que se mueven a altas velocidades a través de la superficie terrestre.

CAÍDA DE CENIZA: fenómeno por el cual la ceniza (u otros materiales piroclásticos) cae por acción de la gravedad desde una columna eruptiva. La distribución de ceniza está en función de la dirección de los vientos.

CALDERA: gran depresión de origen volcánico, generalmente de forma circular o elíptica, de varios kilómetros hasta varias decenas de kilómetros de diámetro, formada por grandes erupciones volcánicas. La depresión (o anfiteatro) formada por el deslizamiento de un flanco de un volcán o colapso sectorial se denomina caldera de avalancha.

COLUMNA ERUPTIVA: El material expulsado por una erupción volcánica puede ascender verticalmente sobre el cráter, formando una nube de erupción o columna eruptiva.

CORTEZA: parte más externa y rígida de la tierra. Generalmente está constituida de rocas de composición balística (océanos) o de rocas más si silíceas (continentes).

CRÁTER: depresión de forma aproximadamente circular, de menos de 2 kilómetros de diámetro, con paredes muy empinadas, generalmente ubicada en la cima de un volcán y formada por la explosión o el colapso asociado/a una erupción volcánica.

DEFORMACIÓN: uno de los parámetros, que junto con la sismicidad y el control geoquímico permiten monitorear el estado de un volcán. El control de la deformación consiste en realizar medidas periódicas de la posición de puntos fijos y ver sus posibles variaciones en el tiempo. Estas medidas pueden ser realizadas por medio de inclinómetros.

DESASTRE: alteración interna en las personas, el medio ambiente que las rodea o sus bienes, generalmente por causas naturales, tecnológicas o por el hombre y que ocasiona un incremento en las demandas de atención medica de emergencias, excediendo su capacidad de respuesta. Los desastres son la materialización de unas condiciones de riesgo existentes,

DOMO: Abultamiento en forma de cúpula formada por la acumulación de lava viscosa, caracterizada por presentar flancos muy pendientes. Generalmente están formados por lavas de composición andesitíva, dacítica o riolítica y pueden alcanzar alturas de cientos de metros.

ERUPCIÓN: eyección de material volcánico como lava, piroclastos y gases volcánicos en la superficie de la tierra, que puede suceder desde una fisura o desde un cráter o boca eruptiva.

FLUJO DE LAVA: Derrame o corriente de roca fundida, originado en un cráter o en fracturas de los flancos de un volcán, por erupciones generalmente no explosivas. Los flujos de lava descienden por los flancos del volcán restringidos únicamente a las quebradas y pueden viajar ladera abajo hasta por varias decenas de kilómetros, desplazándose generalmente a bajas velocidades del

orden de decenas y raramente de centenares de metros por hora, para lavas de tipo andesitas a dacitas.

FLUJOS DE LODO Y ESCOMBROS (LAHARES): mezcla de materiales volcánicos, removilizados por el agua proveniente de la fusión de un casquete glaciar, de un largo craterico o de fuertes lluvias. Estos flujos se mueven ladera abajo, impulsados por la fuerza de la gravedad, a grandes velocidades (hasta 85 km/h), siguiendo los drenajes existentes; sin embargo pueden sobrepasar pequeñas barreras topográficas con relativa facilidad.

FLUJO PIROCLÁSTICO: mezcla caliente (300- 800 ºC) de gases, cenizas y fragmentos de roca, que descienden por los flancos del volcán, desplazándose a grandes velocidades (75- 150 km/h). Ocurren generalmente en erupciones grandes y explosivas o por el colapso del frente de un domo o un flujo de lava. Constituyen uno de los fenómenos más peligrosos asociados con las erupciones volcánicas.

FUMAROLA: orificio, fractura, grieta o fisura en el cráter o en los flancos de un volcán por donde emanan gases volcánicos y vapor de agua generalmente a altas temperaturas. La mayor parte de los gases emitidos son vapor de agua, sin embargo se encuentran otros gases como C02, CO, S02, H2S, CH4, CHI, etc.

INCLINÓMETROS ELECTRÓNICO: instrumento científico que permite detectar las variaciones en las pendientes del terreno.

INTENSIDAD: escala subjetiva que mide los efectos de un sismo sobre las personas, las edificaciones y la naturaleza. Para su medición se utiliza generalmente la escala Mercalli modificada.

LAPILLI: fragmento de roca de tamaño comprendido entre 2 y 64 mm emitido durante una erupción volcánica.

LLUVIA ÁCIDA: ciertos gases magmáticos (S02, Cl, entre otros) emitidos por un volcán en erupción, al entrar en contacto con el agua atmosférica forman ácidos fuertemente corrosivos que caen a la superficie en forma de lluvia.

MAGMA: roca fundida que contiene una fase liquida, gases disueltos, cristales de minerales y eventualmente burbujas de gas. Los magmas se forman a grandes profundidades en el manto o en la corteza terrestre. Cuando el magma ha perdido sus gases y alcanza la superficie se denomina lava. Si el magma se enfría al interior de la corteza terrestre forma la roca intrusiva.

MAGNITUD: valor que estima la energía liberada por un sismo. Se utiliza generalmente la escala de Richter.

MONITOREO: sistema que permite la observación, medición y evaluación continúa del progreso de un proceso o fenómeno a la vista, para tomar medidas correctivas. El monitoreo puede ser sismológico, vulcanológico, hidrometeorologico, radiológico, etc.

NUBE DE CENIZA: masa de gases y ceniza, generada por una explosión volcánica o derivada de un flujo piroclástico.

PIEDRA PÓMEZ: roca volcánica de color claro, llena de cavidades que la hacen muy poco densa (frecuentemente pueden flotar). Generalmente tienen una composición dacitica o riolitica. Las cavidades se forman por la expansión de los gases volcánicos durante la salida a la superficie.

PIROCLASTOS: fragmentos de roca volcánica fracturada, emitidos durante una emisión volcánica explosiva. Incluyen piedras pómez, ceniza y otros fragmentos de roca.

PLACAS TECTÓNICAS: grandes fragmentos que constituyen el envolvente externo del a Tierra. Son fragmentos de la corteza terrestre delimitada por zonas sísmicas y que llega a tener grandes dimensiones (varios miles de kilómetros cuadrados) y espesor de 30 a 40 kilómetros. Estas placas se encuentran flotando sobre una capa más dúctil y plástica del manto terrestre y se desplazan lentamente a una velocidad promedio de varios cm/año.

RIESGO VOLCÁNICO: representa los efectos dañinos de un peligro o amenaza volcánica. En términos probabilísticos constituye la probabilidad de pérdida de vidas humanas, destrucción de propiedad o pérdida de la productividad en un área afectada por un fenómeno volcánico.

SÍLICE: molécula formada por una átomo de silicio y dos átomos de oxigeno (Si02), que constituye la base de la estructura cristalina de la mayor parte de minerales. Es el más importante factor que controla la viscosidad de los magmas. Entre más alto sea el contenido del sílice, más alta es la viscosidad.

SISMO: sacudón del suelo producido por el movimiento abrupto y violento de una masa de roca a lo largo de una falla o fractura de la corteza terrestre. Los volcanes activos presentan una gran variedad de eventos sísmicos. Sismos de largo periodo (LP), asociados al movimiento de fluidos magmáticos bajo presión en los conductos volcánicos. Sismos volcano-tectonicos (VT), asociados a la fracturación de rocas bajo un volcán. Sismos híbridos, mezcla de varios tipos de señales sísmicas.

SISMÓGRAFO: instrumento científico de alta precisión que detecta, amplia y graba las vibraciones (ondas sísmicas) producidas por los sismos.

SISMOGRAMA: registro en papel (analógico) o en las computadoras (digital) de los eventos sísmicos.

VEI: el índice de Explosividad Volcánica (Volcanic Explosivity Index). Es una escala ampliamente utilizada para describir el tamaño de las erupciones volcánicas basadas entre otros factores, en el volumen de materia emitido. La escala VEI varía entre 0 y 8. Una erupción con VEI de 0 denota una erupción no explosiva sin importar el volumen de producto emitidos. Las erupciones con un VEI de 5 o más son consideradas “muy grandes” y ocurren raramente alrededor del planeta (en promedio una erupción cada década).

VISCOSIDAD: medida de resistencia de un material a fluir en respuesta a un esfuerzo. Entre más alto sea el contenido de sílice, más alta es la viscosidad.

VOLCÁN: orificio en la superficie de la tierra a través del cual el magma sale a la superficie. Con el mismo nombre se denomina a la montaña resultado de la acumulación de material volcánico.

VOLCÁN ACTIVO: es un volcán que ha tenido actividad durante los últimos 10.000 años y que presenta signos de actividad como sismos, fumarolas, emisiones de ceniza, o que está en erupción.

VOLCANOLOGÍA O VULCANOLOGÍA: rama de la geología y de la geofísica que estudia los volcanes y los fenómenos asociados.

VULCANIANA (ERUPCIÓN): tipo de erupción volcánica caracterizada por la ocurrencia de eventos explosivos de corta duración que emite material en la atmosfera hasta altitudes del orden de 20 km. Generalmente este tipo de actividad está asociada a la interacción entre el agua subterránea y el magma (erupción freatomagmatica)

VULNERABILIDAD: es la susceptibilidad o la predisposición intrínseca de un elemento o de un sistema de ser afectado gravemente. Es el factor interno del riesgo, debido a que esta situación depende de la actividad humana. La vulnerabilidad no es general, sino que debe entenderse en función de cada tipo de amenaza. Las condiciones o tipos de vulnerabilidad son los agentes que favorecen o facilitan las manifestaciones del desastre ante la presencia de los fenómenos.

INTRODUCCIÓN

La minería de datos nos permite contar con un nuevo y poderoso conjunto de herramientas para el análisis e interpretación de los datos. El objetivo final es aprovechar aún más los datos generados, amortizando el esfuerzo en tiempo y dinero que conlleva el registro de los mismos. Queda claro que la aplicación de las técnicas de minería de datos sólo se restringe a la disponibilidad de la información, la imaginación de los especialistas y el trabajo interdisciplinario en equipo. Es parte del proceso de Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases-KDD) que se define como la extracción no trivial, desde los datos, de información implícita, previamente desconocida y potencialmente útil.

La Minería de Datos (Data Mining-DM) se trata de la exploración y análisis, por medios automáticos o semiautomáticos, de grandes cantidades de datos con el fin de descubrir reglas y patrones significativos.
El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para un uso posterior.

En el Observatorio Vulcanológico y Sismológico de Pasto (OVSP), la minería de datos tiene diversos usos. Uno de ellos, que es el más importante, es la detección de las posibles erupciones o emisiones, con el fin de contribuir a procesos de gestión de riesgo en la zona de influencia del volcán Galeras.

En este documento se presenta el proyecto que tiene como objetivo construir un repositorio limpio de datos del OVSP, sin embargo se aplicaron superficialmente las técnicas de minería de datos, como clasificación, asociación y agrupación con el fin de obtener patrones de eventos eruptivos del volcán Galeras.

DESCRIPCIÓN DEL PROBLEMA

El OVSP es un instituto de investigación cuya principal actividad es la vigilancia del comportamiento del volcán Galeras. Para realizar este proceso se cuenta con equipos que registran los cambios de deformación, que se pueden presentar en el interior del volcán y del entorno. Estos registros son transmitidos al OVSP. Donde son almacenados en una base de datos “general” bajo el SGBD PostgreSQL. A través del sistema experto **SAIIG- SISTEMA PARA LA INTEGRACIÓN ANÁLISIS E INTERPRETACIÓN DE LA INFORMACIÓN GENERADA POR EL PROCESO DE VIGILANCIA Y MONITOREO DEL VOLCÁN GALERAS**, estos datos son procesados con el fin de generar modelos de comportamiento. Con el

análisis de la información obtenida con estos modelos, se establecen escenarios de evolución de la actividad al interior y en el entorno del Volcán Galeras.

La definición e interpretación de los procesos que se realizan en el interior y en el entorno del volcán Galeras no es una tarea simple ya que implica el análisis de varios parámetros. El objetivo es realizar el análisis de los factores a partir de varias disciplinas del conocimiento, por lo tanto es de vital importancia realizar minería de datos, a los datos obtenidos sobre el comportamiento del Volcán para complementar los patrones por otros medios y con ello tener más herramientas para el aporte al análisis del comportamiento volcánico.

OBJETIVOS

Objetivo General. Construir un repositorio limpio de datos para la detección de patrones de eventos eruptivos del volcán Galeras con técnicas de minería de datos que permita obtener información confiable y completa del comportamiento del volcán Galeras, para aplicar técnicas de minería de datos, cuyos resultados permitan predecir las posibles erupciones del volcán, para tomar medidas preventivas con tiempo.

Objetivos específicos:

- Apropiar el conocimiento sobre Comportamiento de Volcán Galeras y minería de datos por parte de los estudiantes investigadores.
- Categorización de las fuentes de datos internas y externas en el OVSP.
- Seleccionar los datos que incidan en las posibles erupciones.
- Diseñar y construir un repositorio con información de los eventos eruptivos del volcán Galeras.
- Aplicar técnicas de preprocesamiento y transformación de datos al repositorio con el fin de obtener datos correctos, consistentes y categorizados en un repositorio limpio y transformado.

JUSTIFICACIÓN

La capacidad de adquirir conocimiento en una determinada disciplina requiere de entrenamiento y experiencia; sin embargo, cuando se han definido relaciones que permiten reconocer un escenario, éstas pueden formalizarse y almacenarse para ser utilizadas posteriormente en una situación similar, como la que se presenta en el estudio de fenómenos volcánicos en donde la información es interpretada por diferentes especialistas los cuales necesitan conocer el escenario global en donde se desenvuelve el fenómeno.

En el momento de realizar la interpretación de la información obtenida por el proceso de vigilancia volcánica, esta implica varias áreas, disciplinas o parámetros geofísicos y geológicos (sismológicos, deformaciones, geoquímicas de gases y

suelos, potenciales, seguimientos de actividades superficiales), razón por la cual es necesario integrar la información obteniendo así una visión global del fenómeno volcánico y cómo los diferentes procesos interactúan entre sí.

Con la realización de este proyecto se utilizará una herramienta de minería de datos para realizar el respectivo análisis e interpretación del comportamiento del volcán Galeras, con lo cual, se mejora la forma de interpretar el comportamiento volcánico y por lo tanto el conocimiento del fenómeno, permitiendo al personal del OVSP una mejor respuesta a las autoridades y comunidades de la zona de influencia del Volcán Galeras

ALCANCE Y DELIMITACIÓN

El proceso de minería de datos a desarrollar se encargará de la utilización de las relaciones que usa el personal experto para interpretar la información que se tiene sobre el comportamiento del volcán Galeras para determinar posibles escenarios eruptivos. Para conseguirlo será necesario realizar primero la integración de la información derivada del proceso de vigilancia volcánica, la cual se almacenará en una base de datos íntegra para analizarla posteriormente y con ello detectar patrones de eventos eruptivos del volcán.

El repositorio quedará listo para su implementación en el OVSP, sirviendo como herramienta para agilizar el análisis e interpretación de la información, teniendo en cuenta algunas de sus variables, consideradas como las más importantes

ORGANIZACIÓN DEL DOCUMENTO

Este documento se encuentra organizado de la siguiente manera: en el capítulo 1 se aborda el contexto del volcán Galeras , en el capítulo 2 se presenta la información relacionada con el proceso de descubrimiento de conocimiento de bases de datos y la metodología para proyectos de minería de datos CRISP-DM, en el capítulo 3 se presenta el proceso de construcción de un repositorio limpio de datos para la detección de patrones de eventos eruptivos del volcán Galeras con técnicas de minería de datos siguiendo la metodología CRISP-DM, en el capítulo 4 las conclusiones y en el capítulo 5 las recomendaciones.

1. ERUPCIONES DEL VOLCÁN GALERAS

Los indígenas Quillacingas dieron el nombre de "Urcunina" (Montaña de Fuego) al volcán de Pasto, hoy conocido "volcán Galeras"; llamado así, por los primeros conquistadores españoles, por su semejanza con las Galeras o barcos que con sus velas navegaban en aquel entonces por el Mediterráneo. [21].

El volcán Galeras se localiza en el Departamento de Nariño, aproximadamente a 9 km al occidente de la Ciudad de San Juan de Pasto, capital de este departamento, en las coordenadas 1° 13' 43,8" de latitud norte y 77° 21' 33,0" de longitud al oeste y con una altura de 4276 msnm. De acuerdo con el Catálogo de Volcanes Activos del Mundo (CAVW) de la Asociación Internacional de Vulcanología y Química del Interior de la Tierra (IAVCEI), su código es el 1501-08. [21]. El Galeras es considerado en Colombia, como uno de los volcanes más activos, ya que presenta una alta tasa de períodos de actividad en comparación con los lapsos de tiempo en los que permanece en reposo.

Algunos rasgos Fisiográficos del Volcán son:

- Elevación 4276 metros sobre el nivel del mar
- Tipo de volcán Estratovolcán – calderico
- Diámetro de la base del Edificio volcánico 20 km
- Diámetro del cráter principal 320 m de diámetro y 80 m de profundidad. Posee otros cráteres aledaños más pequeños (cráteres secundarios) y varios campos fumarólicos.
- Altura de cono activo 150 metros sobre la cima de Galeras y 120 metros de diámetro.
- El actual cono activo tiene una edad estimada de 5.000 años.

El volcán Galeras que conocemos hoy en día es el centro eruptivo más reciente y actualmente activo del denominado Complejo Volcánico Galeras (CVG), el cual posee una forma cónica con su edificio destruido. En su evolución se identifican siete estados que del más antiguo al más reciente se han denominado como: Cariaco, Pamba, Coba Negra, La Guaca, Genoy, Urcunina y el actual Galeras. [21]. (Ver Figura 1).

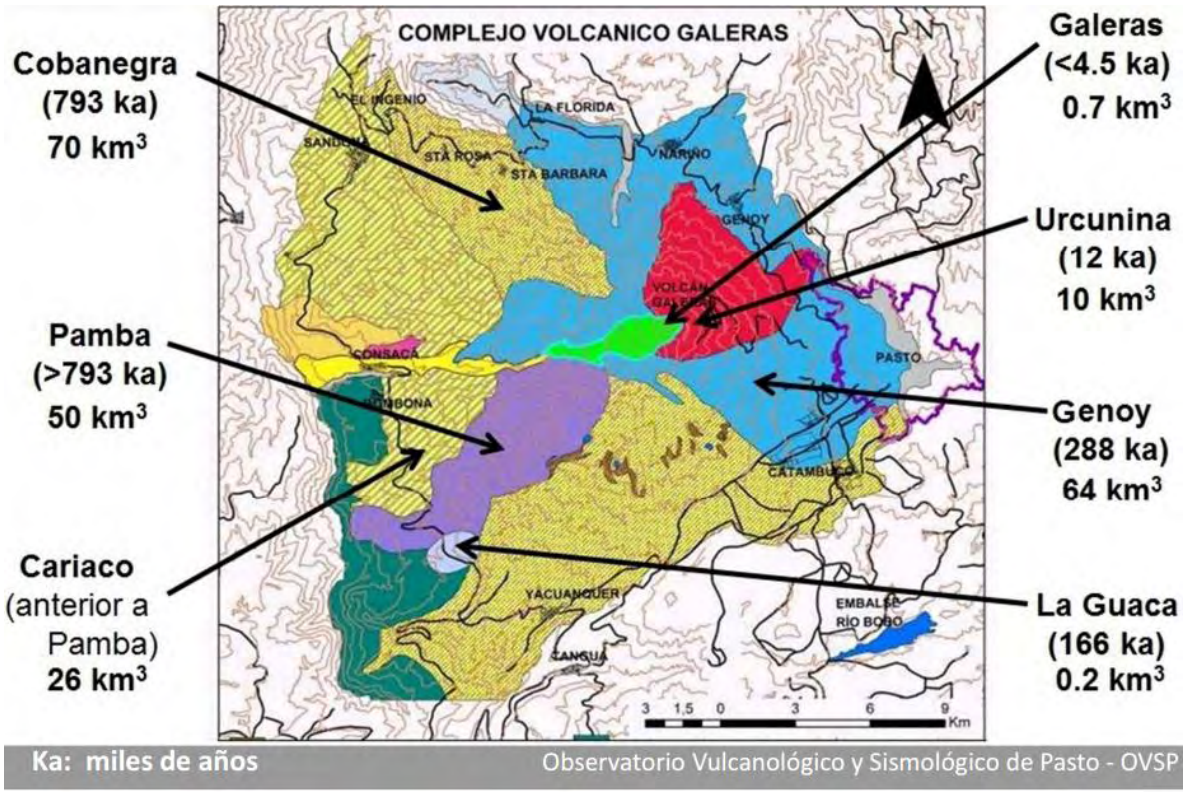

Figura 1. EVOLUCIÓN GEOLÓGICA DE DEPÓSITOS DEL CVG. [21]

1.1 EVENTO ERUPTIVO

1.1.1 Eventos antes de una erupcion volcánica. Las erupciones volcánicas son uno de los fenómenos más impredecibles de la naturaleza. Sin embargo, casi todas ellas tienen algo en común. No importa el tipo de volcán o su localización. En la mayoría de los casos, la erupción es precedida por temblores que se producen minutos, días o semanas antes de que el volcán se despierte, las erupciones magmáticas, por ejemplo, implican la salida del magma a la superficie y este movimiento ascendente normalmente genera terremotos detectables, puede deformar la superficie de la tierra y causar cambios en la temperatura y composición química del suelo y aguas de manantiales [1].

La gente que vive cerca de los volcanes puede detectar algunos de estos hechos antes de una erupción. Tanto la frecuencia como la intensidad de los terremotos tienden a aumentar antes de que dé comienzo. También pueden estar precedidas por actividad fumarólica y por áreas nuevas o más grandes de suelo caliente.

Sin embargo, la mayoría de los cambios precursores son demasiado sutiles para notarlos, y es esencial utilizar los medios más efectivos en la vigilancia del volcán, que incluyen una amplia gama de técnicas geofísicas, geodésicas y geoquímicas. Se utilizan sismógrafos para detectar y localizar temblores relacionados con la ascensión del magma [2]. Otras técnicas incluyen la medición de los cambios en el flujo de calor en el volcán. Cambios en la composición o abundancia relativa de gases fumarólicos también pueden preceder erupciones volcánicas y se pueden detectar a través de frecuentes análisis de gases. [1]

Estas técnicas pueden ser útiles a la hora de detectar signos de advertencia de una erupción inminente. Sin embargo, el éxito global de un sistema de vigilancia depende de la detección e interpretación de los signos precursores con suficiente antelación como para advertir y evacuar a la gente de las áreas amenazadas e iniciar otras medidas para mitigar los efectos de la erupción. Aunque estos sistemas de vigilancia pueden ser útiles para indicar un aumento en la probabilidad de actividad volcánica y su localización, no indican el tipo o escala de una erupción inminente, ni el Índice de Explosividad Volcánica que alcanzará.[1]

El ascenso del magma provoca terremotos. El lento ascenso produce un aumento progresivo de la intensidad y la frecuencia de estos temblores. Este proceso puede comenzar meses antes de la erupción. [2] El ascenso del magma también provoca ruidos subterráneos en el área en el que se encuentra el volcán. Estos ruidos pueden ser sordos o secos, como algo que se rompe. Si hay escarcha o hielo se produce la ruptura, hasta que se derrite por efecto del calor. Se produce un aumento de la temperatura en las aguas cercanas al volcán y en el suelo, que puede provocar la muerte del manto vegetal. Los animales pueden comportarse de manera anómala, al percibir vibraciones que nosotros no podemos captar. Es posible apreciar fumarolas, y si estas ya existían, se produce un aumento en el flujo de emisión y pueden percibirse incluso cambios en el color. Se producen deformaciones en el edificio volcánico. Se pueden medir y determinar sus pulsaciones mediante un medidor de inclinación. Cuando la erupción es inminente, hay pequeñas explosiones y emisiones de ceniza, que van aumentando de intensidad y frecuencia a medida que se acerca el momento en el que finalmente, el volcán entra en erupción. [1]

Eventos durante la explosión:

La salida del magma a la superficie se produce en tres formas: líquido (lavas), gases y proyección de fragmentos sólidos (piroclastos, de piros fuego y clasto fragmento) [3]. La cantidad de gas presente en el magma es el condicionante para que la erupción sea tranquila o explosiva, y de que predomine la emisión de lavas o de piroclastos. Recordemos que una explosión es el resultado de la expansión brusca del gas; un material explosivo corresponde a una reacción química que produce en muy poco tiempo una gran cantidad de gas [3].

Los fenómenos que ocurren en un volcán son bien conocidos desde hace mucho tiempo; sin embargo, para valorarlos en su aspecto directamente relacionado con el riesgo volcánico, es útil repasar las grandes catástrofes de las que tenemos noticias. Se observa, en líneas generales, que las pérdidas en vidas humanas han ocurrido por efectos indirectos (tsunamis, lahares, pérdida de cosechas, etc.) o por una mala gestión de la crisis, pues un volcán no pasa inmediatamente del más absoluto reposo a la más violenta actividad; todas las grandes erupciones vienen precedidas de actividad menor y con la suficiente antelación para tomar las medidas de evacuación de las poblaciones próximas.

La mayor parte de los eventos volcánicos sólo afectan a las proximidades del volcán, como la caída de bombas y las nubes de gases tóxicos, o bien presentan una movilidad baja, como las lavas. Incluso los grandes efectos del volcanismo explosivo están limitados a un entorno de pocos kilómetros, excepto la caída de cenizas arrastradas por el viento a grandes distancias. Otras catástrofes asociadas a los volcanes, como pueden ser los lahares o los deslizamientos de ladera pueden ocurrir sin erupción o terremoto, disparados simplemente por unas lluvias anormales que inestabilizan los materiales volcánicos. [3]

El estudio de la peligrosidad volcánica exige dividir cada uno de los episodios volcánicos en elementos muy sencillos que se evalúan independientemente. (Ver Tabla 1)

Factores de peligro	Tipo de daño
Proyección de bombas y escorias	Daños por impacto. Incendio
Caída de piroclastos	Recubrimiento por cenizas. Colapso de estructuras. Daños a la agricultura. Daños a instalaciones industriales
Dispersión de cenizas	Problemas en tráfico aéreo. Falta de visibilidad
Lavas y domos	Daños a estructuras. Incendios. Recubrimiento por lavas
Coladas y Oleadas Piroclásticas (Nubes ardientes)	Daños a estructuras. Incendios. Recubrimiento por cenizas
Lahares	Daños a estructuras. Arrastres de materiales. Recubrimiento por barros
Colapso total o parcial del edificio volcánico	Daños a estructuras. Recubrimiento por derrubios. Avalanchas. Tsunami inducido
Deslizamiento de laderas	Arrastres de materiales. Recubrimiento por derrubios. Daños a estructuras
Gases	Envenenamiento. Contaminación aire y agua
Onda de choque	Rotura de cristales y paneles
Terremotos y temblores volcánicos	Colapso del edificio volcánico. Deslizamiento de masas. Daños a estructuras
Deformación del terreno	Fallas. Daños a estructuras
Variaciones en el sistema geotérmico de acuíferos	Cambios en la temperatura y calidad del agua
Inyección de aerosoles en la estratosfera	Impacto en el clima. Efectos a largo plazo y/o a distancia

TABLA 1. "FACTORES DE PELIGRO VS DAÑO" [33].

Flujos de lava:

Las lavas son rocas de composición homogénea emitidas en forma líquida durante una erupción volcánica. Las propiedades físicas de la lava (especialmente la viscosidad), la variación de temperatura durante su recorrido, el volumen de

material emitido y las características del terreno por el que discurre, influyen sobre la morfología final que adquieren. Las lavas muy fluidas se extienden cubriendo grandes extensiones con un pequeño espesor. Las lavas viscosas poseen mayor altura, pero recorren distancias menores y el caso extremo son las lavas muy viscosas que se quedan sobre el propio centro de emisión, formando un domo. Es importante decir que las lavas se mueven lentamente, salvo casos muy excepcionales, y lejos de los centros de emisión se mueven a unos pocos metros por hora. Por ello, es muy difícil que causen pérdidas de vidas humanas.

La altura mínima que debe poseer una lava para que pueda moverse se conoce como altura crítica y depende de la cizalla umbral, es decir la cizalla (rozamiento) mínimo que debe aplicarse para que el fluido pueda moverse. La altura crítica va desde unos pocos centímetros hasta varias decenas de metros; las lavas de la erupción de Timanfaya (Lanzarote, Islas Canarias) poseen alturas críticas, moviéndose en el plano horizontal, entre 1.5 y 3 m. En el volcán Teide (Tenerife, Islas Canarias) podemos encontrar lavas con más de 20 m. de altura crítica. A medida que la colada se enfría, va aumentando su cizalla umbral y con ello la altura crítica, por eso, a grandes distancias del centro de emisión la colada tiene mayor espesor. En la anatomía de una lava podemos distinguir inicialmente la superficie en contacto con la atmósfera, cuyo aspecto depende del régimen de movimiento de la colada, después observamos el cuerpo de la colada, de aspecto masivo, ya que se enfría lentamente. En la base, encontramos una capa de escorias, formada por el enfriamiento rápido de la lava en contacto con el suelo, más los materiales que ha ido arrastrando y las alteraciones que haya producido por las elevadas temperaturas sobre el propio suelo. El aspecto superficial de una lava es muy espectacular, pero meramente anecdótico; ello es debido a la cizalla que el movimiento del interior de la colada ejerce sobre la superficie cuando ésta empieza a solidificarse. Si la cizalla es pequeña, simplemente provoca una leve ondulación en la superficie, que se conoce con el nombre hawaiano de lavas pahoe-hoe, que significa superficie por donde se puede caminar con los pies descalzos. Cuando la cizalla es lo suficientemente grande, rompe la capa superficial ya parcialmente solidificada, que después el movimiento irá triturando y redondeando. Las lavas al enfriarse, experimentan una contracción que produce sistemas de fracturas y disyunciones, siendo los principales tipos las disyunciones columnar y lenticular. Otro aspecto que se presenta es la disyunción esferoidal (en bolas de descamación), producidas por la meteorización e infiltración de la humedad a través de grietas ya existentes [3].

Gases:

Los gases, contenidos en el magma, se emiten a elevada temperatura y ascienden en forma de una columna convectiva, hasta llegar a la altura en la que columna y atmósfera tienen la misma temperatura, cesando entonces el ascenso. Esta columna tiene capacidad para arrastrar gran cantidad de piroclastos y materiales sólidos arrancados del conducto. Como ya se ha indicado anteriormente el gas es

el causante del mayor o menor grado de explosividad de la erupción. Además de la salida violenta por el cráter durante la erupción, el gas puede escapar por pequeñas fracturas del edificio volcánico y zonas próximas, dando lugar a fumarolas. También puede salir disuelto en el agua de los acuíferos existentes en el área, originando aguas termales y medicinales. Finalmente, algunos gases como el dióxido de carbono (CO_2) pueden escapar por difusión a través del suelo, incluso en áreas muy alejadas del volcán

Los gases procedentes del magma circulan por el sistema de fracturas, interaccionando con los distintos acuíferos y saliendo a la superficie en forma de fumarolas o de fuentes termales. El SO_2 y el CO_2 se consideran los componentes más significativos de la presencia de magma. Para obtener información completa sobre la composición del gas volcánico, la única forma consiste en realizar un muestreo directo de las fumarolas, analizándose posteriormente en el laboratorio mediante las técnicas químicas habituales. Esto se debe, fundamentalmente, a que los gases se disipan rápidamente y son fácilmente contaminables, además de salir a elevada temperatura y ser corrosivos, imposibilitando con ello la instalación de sensores de forma permanente [3].

Flujos y caida de piroclastos:

Los fragmentos sólidos o piroclastos expulsados durante una erupción volcánica proceden de la fragmentación del magma producida por la expansión violenta de las burbujas del gas que contiene. Los piroclastos abarcan una gran variedad de tamaños, recibiendo distintos nombres según sus dimensiones:

- Bloques - mayor de 64 mm
- Lapilli - Entre 64 mm y 2 mm
- Ceniza - Menor de 2 mm

Estos materiales fragmentarios son arrastrados violentamente por el gas hasta la boca de emisión. Los más grandes son proyectados balísticamente, incluso a grandes distancias, mientras que los más pequeños se incorporan a la columna. Una parte de estos materiales se acumula alrededor del centro emisor formando un cono de escoria.

Algunos fragmentos de magma del tamaño lapilli a bloque son expulsados en forma líquida, enfriándose parcialmente durante su trayectoria de caída, adoptando formas redondeadas o fusiformes que reciben el nombre de bombas. Las escorias se forman por la soldadura de varios fragmentos que al caer no están totalmente fríos. Las pumitas son materiales fragmentarios llenos de pequeñas cavidades producidas por la expansión de las burbujas de gas, generalmente de color claro y densidad inferior al agua.

En otros casos, la columna no posee suficiente fuerza ascensional para elevar todo el material incorporado, produciendo el colapso de la misma; este material cae sobre el volcán, descendiendo rápidamente por las laderas y formando densos flujos que se mueven a gran velocidad, temperaturas elevadas, con gran capacidad de transporte y pueden recorrer hasta 100 km de distancia. Este fenómeno se conoce como colada piroclástica y es uno de los más violentos que pueden ocurrir en una erupción. También existe otro tipo de flujos, producidos cuando la cantidad de gas es muy superior a la cantidad de ceniza, llamadas oleadas piroclásticas y su movimiento presenta un carácter turbulento.

Los flujos piroclásticos, característicos del volcanismo explosivo, descritos anteriormente, son los procesos más violentos que pueden ocurrir en un volcán. Una gran masa de gases y cenizas, a temperaturas superiores a 700ºC se mueven con una velocidad de 540 Km/h y pueden recorrer distancias de hasta 100 Km. La alta velocidad de estos flujos se explica porque se mueven sobre un colchón formado por el propio gas. Del flujo se escapan gases y cenizas muy finas, que forman una nube acompañante. Al avanzar el flujo, transporta junto con la ceniza, fragmentos de rocas, arrancados en el momento de la explosión o de las paredes del conducto y fragmentos de pómez aplastados por la presión. El flujo se detiene al perder el gas y si la temperatura es todavía lo suficientemente alta, las cenizas se sueldan. Los depósitos procedentes de las coladas piroclásticas se conocen como ignimbritas. Los piroclastos incorporados a la columna de gas, pueden ser arrastrados por el viento y caer en forma de lluvia de cenizas a grandes distancias.

Las oleadas piroclásticas, al ser menos densas, forman depósitos de poca entidad de carácter turbulento y con estructuras de estratificación cruzada, duna y antiduna. Estos flujos se adaptan en su desplazamiento a la topografía preexistente en el terreno, pero con capacidad suficiente para remontar algunos obstáculos. Es importante reconocer los depósitos de los materiales volcánicos en relación con los procesos que los originan. [3]

Lahares

Consisten en una avalancha de materiales volcánicos no consolidados, especialmente cenizas que se han acumulado sobre el cono, y que son movilizados por agua. El conjunto se mueve ladera abajo, canalizándose por los barrancos y cargándose de rocas, troncos, etc., pudiendo recorrer grandes distancias con gran poder destructivo. El agua necesaria para iniciar el proceso puede proceder de lluvias intensas o de la fusión parcial del hielo presente en la cima del volcán (Nevado de Ruiz, Colombia, 1985). Los lahares suelen desencadenarse después de la erupción cuando se combina el máximo de material no consolidado con la presencia de agua y en las grandes erupciones siguen generándose varios años después de finalizada la erupción. [3].

Colapso

Un fenómeno muy peligroso es el colapso del edificio volcánico, formado por la acumulación de los materiales de sucesivas erupciones sin cohesión entre ellos. La superposición de materiales duros y blandos da lugar a una estructura que, en algunos casos, puede resultar inestable y producir el colapso de una parte del edifico; las capas de materiales blandos y el agua pueden facilitar el movimiento del conjunto. También, la intrusión de un gran volumen de magma en el edificio volcánico puede desestabilizarlo y producir su colapso, como ocurrió en el volcán St. Helens (USA) en 1980. [3]

Calderas

El término caldera es de carácter morfológico y se aplica a relieves en forma de caldero. Actualmente en volcanología se utiliza para caracterizar las estructuras de colapso, formadas después de la salida rápida de un gran volumen de magma que vacía total o parcialmente la cámara magmática, provocando el hundimiento de la estructura que hay encima. Este colapso reactiva el dinamismo volcánico, generando fases de alta explosividad. El resultado final es una depresión, generalmente de dimensiones kilométricas, con paredes verticales formadas principalmente por los materiales emitidos en esa etapa. En el cráter de algunos volcanes se forma un lago de lava que, al vaciarse por disminución de la presión del magma o derrame lávico, da origen a estructuras de tipo caldera. El volcán Masaya en Nicaragua puede servir de ejemplo de este proceso. Los mares, producidos en explosiones freáticas presentan también el aspecto de una pequeña caldera [3]

Terremotos

La actividad sísmica presente en un volcán activo es difícil de clasificar y depende de cada escuela. En general, esta actividad incluso en periodos de reposo, puede ser muy intensa, con una gran cantidad de eventos de poca magnitud (menores de 2 en la escala de Richter) que suelen presentarse en grupos o enjambres, además de los sismos tectónicos que ocurren en la zona. El aumento de la actividad del volcán lleva asociado un incremento de la sismicidad. Estos eventos sísmicos son de pequeña magnitud debido a la escasa energía disponible que puede liberarse como energía sísmica. La fase gaseosa genera leves movimientos sísmicos que son superficiales y sólo pueden ser registrados por estaciones muy próximas. Las explosiones que acompañan a las erupciones también producen un tipo de evento sísmico muy característico, aunque de poca energía. El estudio de las explosiones se realiza combinando un sismómetro con un micrófono, de forma que se pueda separar la onda que llega por el terreno, de la onda sonora que viaja por el aire. [3]

Despues de una erupcion volcánica:

Efectos inducidos de las erupciones Las erupciones producen efectos indirectos que también repercuten sobre la vegetación: Formación de tsunamis asociados a la formación de calderas o a grandes deslizamientos, cambios de diferente duración en el clima, en el suelo, en la cantidad de CO2 y SO4H2 en la Atmósfera esto causa daños, fundamentalmente en las cosechas, por falta disminución de la luz solar, descensos en las temperaturas que provocan heladas tempranas y tardías, lluvias abundantes y nevadas fuera de temporada, y en consecuencia, disminución del crecimiento de algunas especies y falta de maduración. [4]

Emisiones de gas y lluvia ácida

Durante las erupciones volcánicas se emiten a la Atmósfera ingentes cantidades de gases contenidos en el magma. Estos gases forman parte de las columnas eruptivas, son el elemento imprescindible en la formación y desplazamiento de los flujos piroclásticos, y están contenidos en las lavas que se desplazan sobre la superficie, escapándose de ellas de manera más o menos violenta a lo largo de su recorrido. Los que se inyectan a partir de potentes columnas en violentas erupciones, incrementan el contenido global en CO2 y en compuestos de azufre. Estos últimos llevan a la formación de ácido sulfúrico. La presencia de aerosoles en la Atmósfera puede provocar una disminución de la radiación solar que llega a la superficie, necesaria para la vida de las plantas. También pueden dar lugar a descensos de hasta medio grado en la temperatura durante años. Los efectos de la lluvia ácida provocan efectos nocivos en el crecimiento y normal desarrollo de la vegetación a largo plazo. En regiones volcánicas activas y en las que, aunque no se hayan producido erupciones a lo largo de miles o cientos de miles de años, pero exista una emanación difusa y continuada de gases, incrementos puntuales en la emisión de los mismos, pueden llegar a afectar de forma negativa al normal desarrollo de la vegetación y los cultivos en los espacios próximos al lugar de salida. [4]

Cambios en las condiciones edáficas:

El territorio afectado directamente por la deposición del material emitido en una erupción volcánica sufre unos cambios que repercuten en el proceso de regeneración de la vegetación afectada. En el suelo se llevan a cabo unas transformaciones drásticas y duraderas.

De hecho el suelo existente antes de la erupción es recubierto por el nuevo material, en ocasiones con potencias de decenas de metros. Este material queda sometido a los procesos de meteorización, que será más o menos activa en función de las condiciones climáticas, topográficas y de la naturaleza de los depósitos, pero que posiblemente tarde varios siglos en lograr el desarrollo de las condiciones edáficas que permitan la colonización de esos depósitos por nuevas

especies vegetales u otras similares a las que allí existían antes de que la erupción tuviera lugar. Las coladas de lava son las que oponen más dificultades a los procesos de restauración vegetal, si bien hay que distinguir entre las que presentan una superficie lisa, con meteorización uniforme y lenta, y las que tienen una superficie escoriácea, cuyas oquedades pueden estar recubiertas de piroclastos o material arrastrado por el agua o el viento sobre el que puede producirse un rápido enraizamiento de las plantas. En este sentido es destacable la repoblación de algunos sectores de las coladas del Paricutín, aunque también puede ocurrir que la topografía local favorezca un arrastre de material fino susceptible de una pronta meteorización lo que retrasaría la colonización. En otras ocasiones, y aunque hayan transcurrido varios siglos desde el emplazamiento de los flujos lávicos, las condiciones ambientales solo han permitido la aparición de líquenes como en algunas lavas de la erupción de 1730-36 en la isla de Lanzarote. En otras ocasiones, en las diaclasas y los bordes de las coladas, puede formarse el suelo necesario para permitir el desarrollo de la vegetación.

Sobre los depósitos de ceniza, de flujos piroclásticos poco consolidados, de avalanchas y lahares, las condiciones de generación del suelo pueden ser localmente favorables a la regeneración vegetal relativamente rápida. Aparición de especies en menos de una década se ha observado sobre los depósitos de fango de la erupción del Saint Helens. [4]

La recuperación de la vegetación natural:

Después de una erupción, y en un periodo de tiempo generalmente largo, se lleva a cabo una recuperación de la vegetación. Esta puede producirse a partir de las especies que existían antes del evento, o con especies nuevas procedente de áreas alejadas del espacio afectado. En este último caso puede producirse una sustitución, al menos parcial, de las especies tradicionales por otras Esto lleva a que después de un tiempo unas especies son sustituidas por otras. Las plantas, en el primer año de rebrote, es posible que tengan una menor altura, hasta un 75% menor, así como una floración más pobre. [4]

1.2 VOLCÁN GALERAS

El actual cono activo tiene una edad estimada en cerca de 5000 años, la edad mínima en la formación del volcán se estima en 1,1 millon de años. En línea horizontal dista 8 Km. De la ciudad de San Juan de Pasto. [48]. En el estudio geológico de las amenazas y riesgos naturales se han identificado seis episodios eruptivos importantes registrados en los años: 4500, 4000, 2900, 2300 y 1100 años antes del presente y la erupción de 1866. Durante los últimos 500 años, la mayoría de las erupciones se han catalogado como vulcanianas, con columnas inferidas de baja altura (menores a 10 Km.), que han producido emisiones de gases y cenizas, pequeños flujos de lava y erupciones explosivas con la

generación de flujos piroclásticos, cuyos depósitos han alcanzado distancias de hasta 9,5 km desde el cráter. [5].

La morfología estructural actual del edificio volcánico, se ha formado a partir de roca fundida y gases calientes. La composición del magma es andesítico y sale a una temperatura que oscila entre los 900°C y los 1.000°C; es de viscosidad alta, fluye lentamente y presenta dificultades para la salida de los gases, convirtiendo la erupción en explosiva. Esto da lugar a la formación de piroclastos (lapilli, tobas volcánicas, piedra pómez y ceniza volcánica, entre otros). [48].

En su zona de influencia, la población es cercana a los 500.000 habitantes, en tanto a las personas que viven hoy en zona de amenaza alta en los municipios de Pasto, Nariño y La Florida, ascienden a 7.935. El asentamiento y crecimiento de poblaciones en zona de alta peligrosidad, ha conllevado a que se incremente la vulnerabilidad de esas poblaciones y consecuentemente con mayores niveles de riesgo, especialmente por los antecedentes del Volcán Galeras de generación de flujos piroclástico, considerados como la mayor amenaza.

Por su continua actividad, el Volcán Galeras es uno de los volcanes mejor documentado de Colombia, permite formarse una idea del tipo de magnitud de las erupciones que han ocurrido.

1.2.1 Actividad eruptiva del volcán galeras. El proceso de inicio del ciclo de actividad del Volcán es en junio de 1988, después de un relativo reposo, se asoció con una fase de limpieza y abertura de conductos volcánicos, el cual se caracterizó por el incremento en la actividad sísmica y manifestaciones de actividad superficial, desde un cráter secundario denominado El Pinta localizado en el sector oriental del cono, con emisiones de ceniza y gases volcánicos.

Inicialmente entre el 4 y 9 de mayo de 1989 se presentaron erupciones freáticas desde el cráter secundario El Pinta. En este mismo año se observó incandescencia en diferentes sectores del cono activo destacándose: abril en el cráter El Pinta, con temperatura estimada en superficie de 600°C; el 5 de septiembre en un campo fumarológico al SWW del cono activo denominado Las Chavas, con temperatura de 300°C; el 29 de noviembre así como el 2 de agosto y en septiembre de 1990, se observó incandescencia en la pared occidental interna del cráter secundario Las Potrillas.

Posteriormente, hasta octubre de 1991, se presentaron grandes cambios morfológicos en el cono al tiempo que ocurrió una intrusión magmática. El inicio de 1991 se caracterizó por la actividad explosiva, emisiones de ceniza e incandescencia desde el cráter principal. Entre julio y noviembre, se dio el emplazamiento y extrusión de un domo de lava andesítico con un proceso de deformación del edificio volcánico, el cual fue registrado por los inclino metros

electrónicos. Cráter (900 m al E) y Peladitos (1,6 Km al SE). Adicionalmente se observó el incremento en el registro de sismicidad asociada al movimiento de fluidos al interior del sistema volcánico. El 9 de octubre de 1991 se observó por primera vez el domo de lava, con una altura de 50 m respeto a la base del cráter volcánico, con diámetro entre 80 a 100 m y un volumen estimado de 400.00 m^3. El proceso paulatino de desgasificación y solidificación del domo en la base del cráter principal, obstruyo la libre interacción entre el interior y exterior del volcán, ocasionando la acumulación de presión, así como procesos de enfriamiento y cristalización,

Entre diciembre de 1991 y julio de 1992, la actividad en superficie mostró una clara disminución terminando con la destrucción del domo el 16 de julio. El importante resaltar que el 11 de julio de 1992, la actividad sísmica característica había mostrado una notable disminución y es cuando se registra, por primera vez en Galeras una señal sísmica de forma inusual hasta ese entonces, asociada también con movimientos de fluidos. A partir de esa fecha comienza el registro de ese tipo especial de sismos denominado "tornillos", que se presentaron hasta el día 16 de julio, unas horas antes del evento eruptivo. Además el 15 de julio se registró un enjambre de señales sísmicas asociadas con fracturamientos y movimientos de fluido a niveles muy superficiales y de muy pequeña magnitud. [5]

- El 16 de Julio de 1992 siendo las 4:40 pm se registra una erupción volcánica, destruyó aproximadamente el 90% de la expresión superficial del domo. La onda choque fue sentida en varias poblaciones alrededor del volcán. Material emitido: bloques andesítico y ceniza. Volumen emitido: 277.300 m^3. Altura de columna 6 Km, dispersándose en dirección NNW [22].

- Enero 14 de 1993, a la 1:41 pm se registra un Tornillo antes de la erupción de 4 horas. La erupción destruyo el material restante del domo, emitió ceniza y material incandescente. La columna alcanzo una altura entre 2 y 3 Km y se dispersó en dirección SSW-NNE. No fue posible estimar el volumen exacto de material emitido, pero se considera un valor mínimo de 30.000 m^3.

- Marzo 13 de 1993, a las 10:39 pm, se registran 74 eventos Tornillo en el periodo previo a la erupción. La columna de gases alcanzo una altura de 8 Km. El volumen emitido se estimó en 835.000 m^3. Fue la erupción que produjo los mayores cambios morfológicos en el cono activo, como la formación de una fisura radial de 20 m de longitud, que se denominó Novedad y de otros cráteres campos fumarólicos a los que se les llamo Marte, Florencia y La Joya, con diámetros entre 20 y 50 m hacia el sector sur del cono activo. Genero también colapsos de material en el sector occidental del cono y en la pared de un cráter ubicado al norte del cono activo denominado El Paisita, mostrando actividad superficial. La presión en un campo fumarológico localizado al Sur y conocido como Las Deformes y una grieta en el costado W del cono llamada Besolima disminuyo.

- Abril 4 de 1993 a las 4:03 pm, El evento eruptivo no fue precedido por Tornillos, La columna tuvo una altura de 6 Km. La ceniza se distribuyó hacia SW, hasta una distancia de 32 Km. El volumen emitido se calcula en 180.000m^3.

- Abril 13 de 1993, a las 03:21 am, previamente se registraron 6 eventos tipo Tornillo. La columna de emisión tuvo una altura de 6 Km dispersando la ceniza hacia el NW. El volumen emitido fue estimado en 217.000 m^3.

- Junio 7 de 1993 entre las 03:42 am y las 9:37 pm se registraron 103 sismos Tornillos antes de la erupción. El evento eruptivo tuvo dos fases: la primera asociación de una explosión y la segunda fase, 18 horas después, a emisión de gases y partículas sólidas principalmente dinos, con una altura de columna 7 Km en dirección NNE. La columna tuvo una altura de cientos de metros. Se estimó el volumen emitido en 1'255.00m^3, que es el mayor ciclo registrado para actividad de Galeras

- Marzo 21 de 2000 a las 4:28 pm, se presenta el episodio eruptivo y el proceso de relajación que tuvieron una duración aproximada de 3 horas. Esta actividad se asoció con cambios morfológicos en el sector fumarológico de Las Chavas

- Abril 5 y abril 22 del 2000 se registran eventos eruptivos cortos con duraciones de 25 y 40 minutos.

- Mayo 18 de 2000, se registra el evento menos energético de este año, compuesto por dos fases que combinaron el registro de señal de una señal sísmica asociada a tránsito de material fluido. La primera fase tuvo una duración de 200 segundos y las segunda de 45 minutos.

- Junio 7 de 2002 a las 2:08 pm, se presenta la emisión de ceniza y material no juvenil que marco un nuevo proceso de actividad por el cráter El Pinta, inactivo desde 1992. La fase previa al evento tuvo sismicidad de baja energía, principalmente tipo hibrido correspondiente al tránsito de fluidos y fracturamiento de material solido; se midieron temperatura de gases en diferentes sectores del cono, con valores entre 88 y 344°C.

- Julio 8 de 2005, durante aproximadamente 3 horas se registraron alrededor de 100 sismos asociados a fracturamiento de material rígido (VT) con magnitudes de duración menores de 3 grados, profundidades entre 1,5 y 3 Km. Epicentralmente localizados a 1 km a W del cráter. Este proceso se relacionó con intrusión de magma en niveles relativamente superficiales.

- El 17 de enero de 2008 a las 08:06 p.m. ocurrió una erupción de carácter explosivo, presentando una columna compuesta de vapor y ceniza de aproximadamente 8 km de altura, con emisión de piroclastos y una onda de

choque que fue percibida en gran parte de la ciudad de Pasto y las poblaciones aledañas.

- El 24 de abril de 2009 a las 7:30 p.m., se registró una erupción de carácter explosivo, arrojando piroclastos alrededor del domo del volcán. Se produjo onda de choque, columna de humo y ceniza.

- El 6 de junio de 2009 a las 07:18 a.m., se registró una erupción la cual estuvo acompañada por ondas acústicas, sin que se generaran efectos vibratorios notables en las poblaciones localizadas en la zona de influencia del volcán.

- El 7 de junio de 2009 a las 06:38 a.m. una nueva erupción de carácter explosivo, la segunda en menos de 24 horas. [21].

- El 20 de noviembre de 2009 a las 8:37 a.m., se registró una nueva erupción del volcán Galeras de carácter explosivo. La incandescencia asociada con la erupción fue observada desde varios sectores de la zona de influencia del volcán durante algunos minutos. Se estimó una altura de la columna eruptiva cercana a los 10 km, con dispersión hacia los Municipios de Nariño, La Florida y Chachagüí.

- El 2 de enero de 2010 a las 7:43 p.m., se registra un evento eruptivo de carácter explosivo, acompañado de onda de choque, con una columna de 10km de ceniza y un manto de piroclastos cayendo alrededor del cráter. La incandescencia permaneció en las faldas del edificio volcánico hasta varias horas después de la explosión.

- El 25 de agosto de 2010 se inició un evento eruptivo a las 4:00 a.m. presentando un bajo nivel de explosividad y en medio de las nubes se observó una columna de erupción ancha y anomalías térmicas detectadas con la cámara infrarroja desde el OVSP. No se detectó la presencia de rocas emitidas a manera de proyectiles balísticos. Por la presión de esta erupción se abrió un nuevo cráter quedando con cuatro [21].

Se resume la actividad eruptiva histórica, con el índice de explosividad volcánica correspondiente. (Ver Tabla 2).

FECHA (aa o aa/mm/dd)	DESCRIPCION	MAG	VEI
1535	Erupción explosiva del cráter central	4	3
1547	Actividad fumarólica	1	
1559-1560	Ciclo eruptivo: lavas? bombas?	3	
1574	Fumarolas, explosiones	3	
1580/12/07	Erupción explosiva del cráter central	5	4
1616/06/04	Erupción explosiva del cráter central, lava?, represamiento	3	3
1641-1643	Explosión, bombas?, lavas?	4	4
1687	Erupción	3	2
1696	Erupción	3	3
1727	Erupción	3	3
1754-1756	Erupción	3	2
1796/11-1801	Erupción explosiva del cráter central, lava		2
1823/06/17	Explosión	3	2
1823/06/24	Erupción	4	2
1828/10/24/-1834	Erupción explosiva del cráter central	2-3	3
1836	Erupción explosiva del cráter central		2
11865/0/02	Explosión	3	3
1866	Flujos Lava	3	3
1866-1869	Explosión	3	
1869/03/27	Explosión, bombas	3	
1869/07/09	Erupción	3	3
1891	Erupción explosiva del cráter central, lava?,		2
1923	Erupción explosiva del cráter central		2
1924/12/14-18	Fumarolas, ceniza, sonidos, Erupción, lava y domo	3	3
1925/05/25	Explosiones	3	
1925/07/01	Explosión y bombas	3	
1925/08/04	Explosión, bombas y ceniza	4	4
1925/11/21	Explosión, bombas, lavas?, flujos de lodo	4	4
1925/12/31	Explosión, bombas	3	
1926/03/21	Explosión, ceniza	3-4	3-4
1926/09/17	Explosión, ceniza	3	3
1927/05/01	Erupción explosiva del cráter central	1-2	
1930/04/17	Fumarola, ceniza, ruido, explosiones	2	
1932/10/10	Erupción explosiva del cráter central		2
1936/02/09	Explosión	3	2
1936/08/27	Explosión, flujo piroclástico, bombas	3	3
1989/05/05	Erupción freática principalmente desde un cráter secundario		2
1992/07/16	Erupción explosiva del cráter central		2
1993/01/14	Erupción explosiva del cráter central		2
1993/03/23	Erupción explosiva del cráter central		2
1993/04/04	Erupción explosiva del cráter central		1
1993/04/13	Erupción explosiva del cráter central		1
1993/06/07	Erupción explosiva del cráter central		2
2004/07/16	Emisiones de ceniza desde un cráter secundario		1
2004/07/21	Emisiones de ceniza desde un cráter secundario		1
2004/07/24-08/04	Emisiones de ceniza desde un cráter secundario		1
2004/08/11-12	Erupciones explosivas del cráter central		2
2004/10-11	Emisiones de ceniza desde el cráter central		1
2004/11/21	Erupción explosiva del cráter central		2
2005/11/24	Erupción desde el cráter central		1
2005/12/23-27	Emisiones de ceniza desde el cráter central		1

MAG: factor establecido para cuantificar la magnitud de la erupción
VEI: Volcanic Explosivity Index (Indice de Explosividad Volcánica)

Tabla 2. Resumen de la actividad eruptiva del volcán Galeras. [5].

1.2.2 Vigilancia de volcán galeras. El monitoreo o vigilancia volcánica de Galeras se realiza a través de la utilización de métodos geofísicos, geodésicos, geoquímicos y geológicos, apoyados por tecnologías de electrónica, comunicaciones e informática. A continuación se da una explicación de estos métodos y el tipo de información que se obtiene mediante su uso.

Área de Deformación volcánica (Geodesia):

Para medir los cambios horizontales y verticales que ocurren por las deformaciones en la superficie del edificio volcánico debido a su actividad interna, se emplean métodos de inclinometría seca, inclinometría electrónica, líneas cortas de nivelación, medidas electrónicas de distancias y sistemas de posicionamiento global (GPS), las mediciones se realizan en sobre el volcán.

Para esto utilizan varias tecnologías como son: los inclinómetros electrónicos, Mediciones Electrónicas de distancia **EDM** y Sistema global de navegación satelital **GNSS** o GPS permanentes. [5]

En el método de inclinometría seca se puede captar o medir variaciones verticales con rangos de precisión entre 2-3 microradianes mediante el empleo de niveles geodésicos de alta precisión un microradián es el ángulo formado por el desplazamiento vertical de 1 mm en el extremo de una línea de 1.0 km de largo. Las mediciones son hechas sobre un triángulo equilátero de 40 metros de lado. Las líneas cortas de nivelación permiten encontrar movimientos verticales sobre puntos fijos preestablecidos. La medición electrónica de distancias E.D.M. tiene como propósito encontrar cambios horizontales a lo largo de líneas base permanentes, empleando equipos de sistema láser o infrarrojo.

EDM es una estación total o un taquímetro, se ubican en diferentes partes, las bases están ubicadas sobre rocas de muy altas proporciones para mayor precisión, influye la temperatura y la presión, se tienen en cuenta las características del equipo, y del prisma con el que se toma la muestra, si disminuye se lo considera como una inflación y si aumenta es una deflación, la tasa de muestreo es cada 15 días. (Ver Figura 3.)

Figura 2. EDM. [5].

Los inclinómetros electrónicos permiten medir cambios de nivel muy pequeños en un punto sobre el volcán. Es un nivel en 2 dimensiones, los datos que llegan desde un inclinómetro que se encuentra en área para su interpretación y procesamiento de datos se tiene en cuenta dos constantes que son la constante del fabricante y la otra es la constante de la tarjeta de digitalización de datos, también se tiene en cuenta la temperatura porque influye mucho en los datos aplicando una formula. En Galeras están ubicados 8 inclinómetros electrónicos y la tasa de muestreo es de cada 10 minutos. [5]. (Ver Figura 2).

Figura 3. Inclinómetro Electrónico. [5].

GNSS Sistema Global de Navegación Satelital detecta mayor cantidad de señales de constelaciones hace referencia al nivel de penetración, se los instala sobre rocas de muy altas proporciones. (Ver Figura 4).

Figura 4. GNSS. [5].

El GPS ayuda a obtener los vectores de desplazamiento vertical en zonas de alta deformación cercanas a los cráteres o zonas fuente de deformación.

Área Geofísica:

La sismología es la rama de la geofísica que se encarga del estudio de los eventos sísmicos, naturales o inducidos. Para el caso de los sismos generados por la actividad volcánica, se utilizan redes con equipos portátiles y telemétricos localizadas sobre el volcán y en sus cercanías. Los instrumentos utilizados para captar y registrar las señales sísmicas son conocidos como sismógrafos. Registradores sismográficos analógicos y digitales (computador).

Las señales sismológicas de origen volcánico se deben al movimiento de fluidos en el sistema volcánico (desde la "cámara magmática, hasta el cráter). Normalmente se clasifican como:

Sismos volcanotectónicos (VT): Origen asociado a fracturas que ocurren como respuesta a cambios de esfuerzos en las áreas activas por movimiento de fluidos. Su frecuencia generalmente es > 5 Hz [50].

Sismos de largo período (LPS): Se atribuyen a la resonancia en grietas, cavidades y conductos, debido a cambios de presión en los fluidos que existen en los volcanes [45]. En este tipo de eventos predominan las bajas frecuencias.

Tremor volcánico (Tremor): Se caracteriza por el registro de formas de onda de manera persistente o sostenida en el tiempo. El tremor refleja una vibración continua del suelo o pequeños sismos muy frecuentes cuyas ondas se traslapan. Si la señal es monotónica (de frecuencia constante) se denomina tremor armónico; de otra parte si la señal varía significativamente en frecuencia o amplitud, se llama tremor espasmódico [44].

Sismos tipo híbridos (HYB): son una combinación entre un LPS y un volcanotectónico.

Sismos tornillo (TOR): son un tipo particular de eventos sísmicos de periodo largo registrado en volcanes andesíticos [43].

De los parámetros que se monitorean en los volcanes, la sismología es el más importante, debido a que nos da más información que los otros parámetros que se monitorean; para monitorear se utilizan unos sismómetros, los cuales son transductores de velocidad; toman la velocidad con la que se toma el suelo y la transforman en señal de voltaje.

Existen muchos tipos de sismógrafos, pero el que se maneja en el OVSP para el Volcán Galeras son:

- Sismómetros de corto periodo, estos pueden captar el movimiento del suelo con frecuencias desde aproximadamente 1hz hacia arriba.
- Sismómetros de banda ancha estos pueden captar el movimiento del suelo con frecuencias por debajo de 1hz, nos permiten capturar mayor información

La primera información que nos da la Sismología es la concurrencia sísmica; que es el número de sismos por día, por mes, etc., de un tipo de sismicidad. La concurrencia sísmica da una visión de que tan activa esta la fuente y que fuente está activa.

Otro aspecto fundamental que se monitorea es la energía sísmica, la cual se calcula para los diferentes tipos de sismos, esto para calcular cuanta energía se está liberando por sismo o por día o mes.

Desde el año 1989 hasta el año 2004 para medir la magnitud de un sismo se utilizaba los sismógrafos análogos (Ver Figura 5)
Desde el año 2004 hasta el momento se utilizan sismógrafos electrónicos (Ver Figura 6).

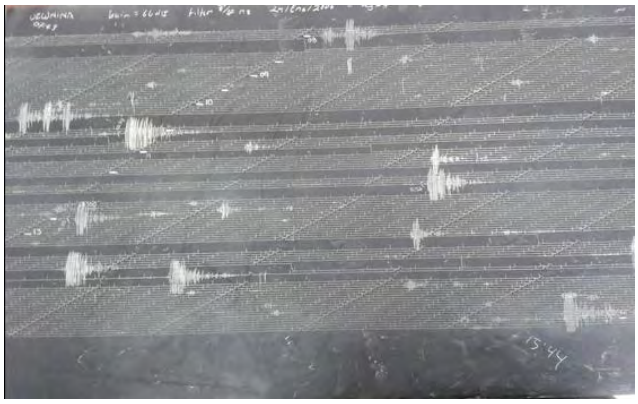

Figura 5. Lectura de sismógrafo Análogo [41]

Figura 6. Lectura de sismógrafo Digital. [41].

Área de geoquímica:

La geoquímica se encarga del estudio de la composición química de La Tierra y del comportamiento de los elementos en ella, tanto en materiales sólidos, como en líquidos y gaseosos. Para el estudio geoquímico en vigilancia de volcanes se recurre a muestreos sistemáticos de las diferentes emanaciones gaseosas y líquidas características de la actividad, tales como aguas termales y gases volcánicos, incluidos radón y dióxido de carbono (CO_2) contenidos en suelos. Tales muestras son sometidas a análisis para conocer las condiciones en su origen y en su camino a la superficie.

El COSPEC es un espectrómetro que mide la emisión de S02 de los volcanes.

Área de Geología:

La aplicación de la geología en el conocimiento e investigación de los volcanes se hace a través de estudios de las rocas y depósitos originados en eventos eruptivos pasados y actuales, lo cual permite conocer la historia eruptiva del volcán, las fechas de ocurrencia de las erupciones, cómo fueron dichas erupciones y el posible comportamiento en eventos futuros. Así, se hace posible la elaboración de los mapas de amenaza volcánica, herramienta indispensable en la planificación y en prevención de los desastres ocasionados por las erupciones volcánicas que afectan las poblaciones y el ambiente.

Un complemento necesario de la geología son las investigaciones históricas sobre erupciones observadas y registradas por cronistas o insertas en leyendas.

Debido a que la masa y la tasa de erupción varían en diferentes ordenes de magnitud, una escala logarítmica es necesaria para categorizar los tamaños de las erupciones volcánicas. [51].

En esta área se mide la llamada Índice de Explosividad Volcánica (VEI). El VEI (Ver Tabla 3) usa una escala integradora desde un valor de 0 a 8 para describir el volumen y la altura de la pluma de una erupción dada. El índice está basado en la información de la magnitud (volumen erupcionado) y la intensidad (altura de la columna eruptiva). Por ejemplo, una erupción de VEI=4 esta definida por tener un volumen total de 0.1-1 km3 de tefra y una altura de columna entre 10-25 km.

El índice de explosividad volcánica puede ser aplicado para erupciones recientes como antiguas, por otro lado esta escala no es usada para erupciones de lava, las cuales primeramente son no-explosivas y por lo tanto reciben una clasificación de 0/1. La supuesta asunción del índice de explosividad volcánica es que la magnitud y la intensidad de las erupciones están relacionadas de alguna manera, de modo que un solo número puede describir los diferentes aspectos del tamaño de una erupción, subsecuentes trabajos han mostrado que no existe una simple relación entre la magnitud y la intensidad de varias erupciones; por lo tanto son necesarias dos escalas diferentes para describir la magnitud e intensidad. [51].

TABLE 1. The Volcanic Explosivity Index[a]

VEI index	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
General description	Nonexplosive	Small	Moderate	Moderate–large	Large	Very large			
Qualitative description	Gentle	Effusive	← Explosive →		← Cataclysmic, paroxysmal →				
Maximum erupted volume of tephra (m^3)	10^4	10^6	10^7	10^8	10^9	10^{10}	10^{11}	10^{12}	10^{13}
Eruption cloud column height (km)	<0.1	0.1–1	1–5	3–15	10–25	>25			

[a] Adapted from Newhall and Self, 1982.

Tabla 3 Índice de Explosividad Volcánica. [49].

Por políticas de seguridad se cataloga como Emisiones y Erupciones, para el Volcán Galeras se maneja el índice de VEI es:

- Si el VEI es de 0 - 1 es una Emisión.
- Si el VEI es de 2 - 8 es una Erupción. [42].

2. PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS

2.1 DEFINICIÓN E INTRODUCCIÓN A KDD

La Extracción de conocimiento está principalmente relacionada con el proceso de descubrimiento conocido como Knowledge Discovery in Databases (KDD), conocido en español como el Proceso de Descubrimiento de Conocimiento en Bases de Datos (DCBD), es uno de los procesos más utilizados y la minería de datos, es una etapa de análisis de KDD, la cual consta de una serie de etapas o fases que se deben desarrollar para obtener conocimiento de alta calidad a partir de los datos.

El proceso KDD es iterativo por naturaleza, y depende de la interacción para la toma de decisiones, de manera dinámica [23].

Las investigaciones en DCBD, se centraron inicialmente en definir nuevas operaciones de descubrimiento de patrones y desarrollar algoritmos para éstas [24], [25], [26]. Investigaciones posteriores [27], [28], [29], [30], [31], [32], se han enfocado en el problema de integrar DCBD con Sistemas Gestores de Bases de Datos (SGBD), haciendo de ésta un área activa de investigación [6].

Como consecuencia del explosivo crecimiento en el volumen de los datos de las bases de datos, que superan los métodos tradicionales de análisis basados en hojas de cálculo y consultas ad-hoc [7] [8] [9], se ha generado una urgente necesidad de contar con nuevas técnicas y herramientas que puedan, inteligente y automáticamente, transformar los datos en información útil: en conocimiento [10]. Estas técnicas y herramientas son el objeto de estudio del proceso KDD.

La Extracción de conocimiento está principalmente relacionada con el proceso de descubrimiento conocido como KDD, que se refiere al proceso no-trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en un repositorio de información. No es un proceso automático, es un proceso iterativo que exhaustivamente explora volúmenes muy grandes de datos para determinar relaciones. Es un proceso que extrae información de calidad que puede usarse para extraer conclusiones basadas en relaciones o modelos dentro de los datos [10], por lo tanto se debe velar que el conocimiento extraído sea:

- **Valido.** Los patrones deben seguir siendo precisos para datos nuevos (con cierto grado de certidumbre), y no solo para aquellos que han sido usados en su obtención [10].

- **Novedoso.** Que aporte algo desconocido tanto para el sistema y preferiblemente para el usuario [10].

- **Potencialmente útil.** La información debe conducir acciones que reporten algún tipo de beneficio para el usuario [10].

- **Comprensible.** La extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión y validación y uso en la toma de decisiones. De hecho, una información incomprensible no proporciona conocimiento [10].

El proceso KDD se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a los patrones encontrados o que se van a encontrar posteriormente. Así el valor real de los datos reside en la información que se puede extraer de ellos, información que ayude a tomar decisiones o mejorar la comprensión de los fenómenos que nos rodean. Este proceso involucra métodos de minería de datos (algoritmos) para extraer (identificar) lo que se considera como conocimiento de acuerdo a la especificación ciertos parámetros usando una base de datos junto con el pre-procesamiento y post-procesamiento [11].

El descubrimiento de conocimiento en base de datos en los últimos años ha ganado preponderancia, se viene desarrollando y utilizando ampliamente ya como una disciplina con un cuerpo teórico y muy estructurado. Uno de sus componentes más importantes es la minería de datos que integra técnicas de análisis de datos y extracción de modelos.

2.2 COMPONENTES DEL PROCESO KDD

El proceso KDD se compone de: (Ver Figura 7)

- **Conocimiento del dominio y preferencias del usuario.** El diccionario de datos, información adicional de las estructuras de datos, restricciones entre campos, metas o preferencias del usuario, campos relevantes, listas de clases, jerarquías de generalización, modelos causales o funciones [12].

- **Control del descubrimiento.** Toma el conocimiento del dominio, lo interpreta y decide qué hacer [12].

- **Interfaces.** Entre la base de datos y el usuario [12].

- **Foco de atención.** Específica a qué tablas, campos y registros acceder. Se debe especificar mecanismos de selección aleatoria de registros tomando muestras estadísticamente significativas, puede usar predicados para seleccionar un subconjunto de registros que comparten cierta característica. Entre las técnicas

para enfocar la atención se encuentran la agregación, la partición de datos, la proyección. [12].

- **Extracción de patrones**: Donde patrón se refiere a cualquier relación entre los elementos de la base de datos. Pueden incluir medidas de incertidumbre. Aquí se aplican una gran cantidad de algoritmos de aprendizaje y estadísticos. [12].

- **Evaluación:** Un patrón es interesante en la medida que sea confiable, novedoso y útil respecto al conocimiento y los objetivos del usuario. La evaluación normalmente se le deja a los algoritmos de extracción de patrones que generalmente están basados en significado estadístico (sin embargo, no es ni debe ser el único criterio). [12].

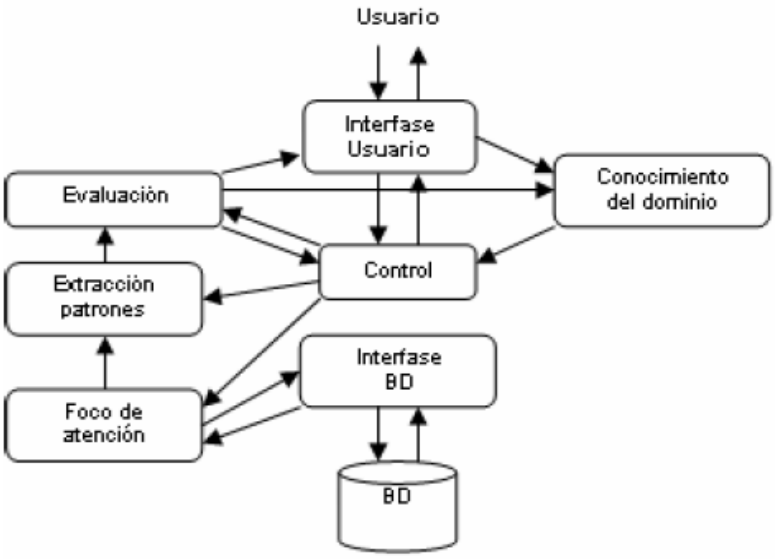

Figura 7. Componentes del proceso KDD. [34].

2.3 ETAPAS DEL PROCESO KDD

Las etapas del procesamiento de KDD se dividen en 5 etapas como muestra la figura 6, las cuales son: (Ver Figura 8)

- Selección de datos: En esta etapa se determinan las fuentes de datos y el tipo de información a utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos desde la o las fuentes de datos. [13]

- Preprocesamiento. Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera

de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación. [13]

- Transformación. Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente. [13].

- Data Mining. Es la fase de modelamiento propiamente tal, en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u "ocultos" en los datos. [13]

- Interpretación y Evaluación. Se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos. [13]

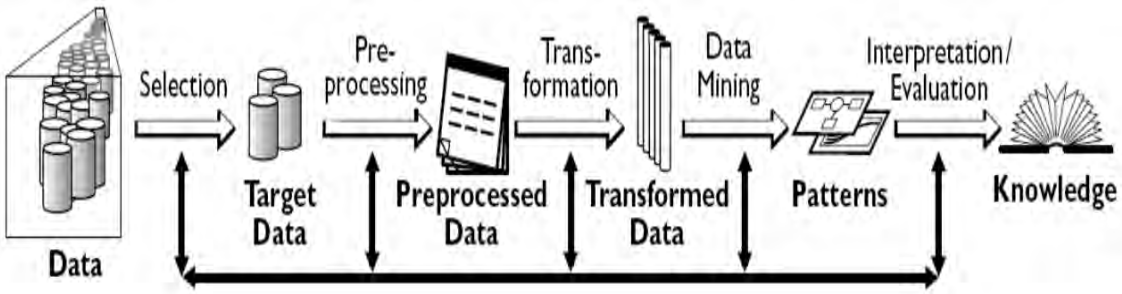

Figura 8. Fases del proceso KDD. [35].

Etapa de selección:

Es de suma importancia conocer el objetivo del negocio para el que se está encontrando una solución, con el fin de obtener conocimiento de alta calidad y cumplir todas las etapas de forma, satisfactoria y con mayor razón en esta fase donde de la selección de los datos radica en gran medida el éxito del proceso de minería de datos.

En la etapa de selección una vez se han escogido mediante un estudio minucioso las fuentes externas e internas de datos que van a satisfacer las necesidades de las siguientes etapas, se procede a seleccionar y preparar el subconjunto de datos que se van a minar, que constituyen lo que se conoce como "vista minable", debido a que esas fuentes se componen de una gran cantidad de datos que son irrelevantes o afectan de alguna medida el proceso de minería de datos. Esta etapa puede ser sencilla de completar dependiendo de la cantidad de datos a

analizar, si la cantidad de datos es pequeña no se presentaran muchos inconvenientes de lo contrario se convierte en una tarea compleja que necesita de una gran cantidad tiempo, de metodologías y tecnologías recientes como por ejemplo los “almacenes de datos” (data warehouses) [10].

Un Data Warehouse es una colección de datos orientados a un dominio, integrados, variable en el tiempo y no volátiles, con el fin de ayudar en la toma de decisiones de alto nivel [10], permitiendo aplicar eficientemente herramientas para resumir, describir y analizar los datos [13]. (Ver Figura 9)

Etapa de preprocesamiento/limpieza:

Una vez se tiene la “vista minable”, se procede a realizar el análisis de calidad de los datos, debido a que estos al proceder de diversas fuentes y al ser integrados pueden estar inconsistentes, de tal manera que se pueden encontrar datos “ruidosos” (noisy data), desconocidos (missing y empty), nulos, duplicados, [6] entre otros que afectan el análisis y la extracción de conocimiento de forma satisfactoria.

La integración de datos es un proceso que se realiza durante la recopilación de datos, como también por medio de un almacén de datos durante el proceso de carga mediante el sistema de ETL (Extraction, Transformation, Load). Esta integración puede ser susceptible a una disparidad de formatos, nombres, rangos entre otros [10].

Figura 9. Componentes de un Data Warehouse. [13]

Los datos ruidosos son valores que están significativamente fuera del rango de valores esperado. Se deben principalmente, a errores humanos, a cambios en el

sistema, a información no disponible a tiempo y a fuentes heterogéneas de datos. Los datos desconocidos empty son aquellos a los cuales no les corresponde un valor en el mundo real y los missing son aquellos que tienen un valor que no fue capturado. Los datos nulos son datos desconocidos que son permitidos por los sistemas gestores de bases de datos relacionales (SGBDR) [6].

Esta etapa tiene como objetivo eliminar o corregir la mayor cantidad de datos inconsistentes e irrelevantes, durante o después de la integración de datos, detectar y tratar valores faltantes y nulos es la tarea más importante a realizar en este punto del proceso KDD.

Valores faltantes:

Los valores faltantes pueden ser originados en primer lugar porque no existen en la realidad, porque los usuarios los omiten por desconocimiento o no quieren brindar ese tipo de información, también se generan en la integración de fuentes de datos donde generalmente se los combina mas no se hace una intersección con ellos [10].
Para tratarlos se puede:

- **Ignorar.** Se los puede dejar pasar como lo hacen algunos algoritmos muy robustos.

- **Eliminar.** Cuando se procede a quitar todo el atributo por la gran cantidad de nulos.

- **Filtrar la fila.** Sesgando los datos.

- **Reemplazar el valor.** Manualmente si la cantidad es pequeña o de forma automática por medio de un valor obtenido mediante funciones estadísticas, así como también por medio de algoritmos que permiten predecir un valor a partir de otro.

- **Segmentar.** Segmentar las tuplas por los valores que tengan disponibles [10].

- **Modificar la política de datos**. Esperar hasta que los datos faltantes estén disponibles [10].

Valores erróneos:

Detectar este tipo de valores se lo puede hacer por medio del formato de origen y contenido de un campo, cuando se analizan atributos numéricos se procede a buscar valores anómalos atípicos o extremos (outliers) también conocidos como valores aislados, exteriores o periféricos. Corregir este tipo de valores es

importante para la obtención de patrones de forma satisfactoria, de no hacerlo solo generaran ruido en la etapa de minería de datos [10].

Para tratarlos se puede:

- **Ignorar.** Se los puede dejar pasar como lo hacen algunos algoritmos muy robustos

- **Eliminar.** Cuando se procede a quitar todo el atributo, es preferible reemplazar la columna por otra discreta donde se especifica si el valor era normal u erróneo. Los anómalos se pueden reemplazar por no anómalo, anómalo superior o anómalo inferior [10].

- **Filtrar la fila.** Sesgando los datos.

- **Reemplazar el valor.** Por 'nulo', por máximos, mínimos o por medias. En casos especiales utilizando un algoritmo que permita predecir valores a partir de los datos ya existentes.

- **Discretizar.** Transformar un valor continuo en uno discreto (ej. Muy alto, alto, medio, bajo muy bajo), de esta manera los anómalos pueden quedar como muy alto o como muy bajo [10].

Etapa de transformación/reducción:

Para hacer más fácil el proceso de minería de datos se necesita disminuir el número efectivo de variables y la forma de representar los datos, para ello se utilizan métodos de reducción de dimensiones o de transformación [6].

Existen métodos de reducción de dimensiones que simplifican una tabla de una base de datos de forma horizontal o vertical. La reducción horizontal implica la eliminación de tuplas idénticas como producto de la sustitución del valor de un atributo por otro de alto nivel, en una jerarquía definida de valores categóricos o por la discretización de valores continuos. La reducción vertical implica la eliminación de atributos que son insignificantes o redundantes con respecto al problema, como la eliminación de llaves, la eliminación de columnas que dependen funcionalmente. Entre las técnicas de reducción más conocidas se tiene: agregaciones, compresión de datos, histogramas, segmentación, discretización basada en entropía y muestreo [6].

La transformación de atributos agrupa operaciones que transforman atributos en otros, que generan nuevos atributos y que cambian el tipo mediante numerización discretización aunque también se lo puede hacer por medio de rangos.

Discretización:

Se entiende como discretización, cuantización también conocida como “binning” al proceso de transformar valores numéricos en atributos discretos o nominales, de tal manera que se los puede trabajar como atributos categóricos con un menor número de valores [10].

Numerización:

La numeración es el proceso inverso a la discretización, no es tan utilizada como la discretización pero es útil cuando los métodos de minería de datos exigen trabajar con datos de tipo nominal por ejemplo en los métodos de modelización estadística (regresión logística o multinomial) [10].

Normalización de rangos:

Muchas técnicas de minería de datos requieren que se normalicen todos los atributos al mismo rango. La normalización más común es la normalización lineal uniforme y se normaliza a una escala genérica entre cero y uno utilizando la siguiente formula [10]:

$$v' = \frac{v - min}{max - min}$$

El resultado de esta normalización garantiza que el cociente entre valores se mantenga, y solo es necesario conocer el valor máximo y mínimo del atributo.

Etapa de minería de datos:

Esta etapa es una de las más importantes del proceso KDD ya que es donde se aplican los algoritmos de análisis de datos a los datos que se seleccionaron, limpiaron y transformaron en etapas anteriores con el fin de obtener patrones que puedan servir al usuario para la toma de decisiones. En síntesis es la etapa donde se produce conocimiento de alta calidad para el estudio que se haya abordado.

Se deben tener en cuenta 3 puntos importantes si se quiere realizar correctamente el proceso de minería de datos, en primer lugar se debe determinar el tipo de tarea de minería más apropiado, una vez se ha hecho esta elección se procede a elegir el tipo de modelo a utilizar y finalmente se escoge el algoritmo de minería que resuelva la tarea y obtenga el tipo de modelo que se está buscando [10].

Dependiendo del algoritmo seleccionado se obtendrá un formato diferente para la salida.

Algunas de las tareas de minería de datos son:

- **Clasificación.** Cada registro de la base de datos pertenece a una clase. En cada caso existe un conjunto de atributos donde uno de ellos es el atributo clase, aquí el resto de los atributos relevantes a la clase se utilizan para predecir la clase [10].

- **Agrupamiento.** Conocida como “clustering” o segmentación, identifica grupos naturales a través de los datos. Se agrupan los datos basándose en el principio de maximizar la similitud entre los elementos de un grupo minimizando la similitud entre los distintos grupos [10].

- **Regresión.** Es similar a la clasificación la diferencia radica en que se busca patrones para determinar un valor numérico [10].

- **Correlaciones.** Tarea descriptiva que se utiliza para examinar el grado de similitud entre dos valores [10].

- **Reglas de asociación.** Su objetivo es identificar relaciones entre atributos categóricos. Por lo general se formulan de la siguiente manera: “Si el atributo X toma el valor a entonces el atributo Y toma el valor b”. No implican una relación causa-efecto [10].

Etapa de interpretación / evaluación:

Esta etapa comprende la interpretación y evaluación de los resultados obtenidos por medio de patrones en la etapa de minería de datos. Partiendo de que un patrón descubierto debe ser preciso, comprensible e interesante [10], el usuario debe analizar si los resultados son los esperados, de lo contrario debe volver a aplicar los algoritmos según otros criterios o cambiar el algoritmo de minería de datos.

Esta etapa del proceso KDD incluye: la visualización de los patrones extraídos, la remoción de los patrones redundantes o irrelevantes y la traducción de los patrones útiles en términos que sean entendibles para el usuario [6].

Los resultados obtenidos pueden ser documentados para ser reportados a los usuarios interesados como también se pueden integrar como procedimientos almacenados en un SGBD o en sistemas expertos entre otras formas.

2.4 METODOLOGIA CRISP-DM

Para la implementación de una tecnología dentro de un negocio se requiere de una metodología y en el caso de proyectos de minería de datos se destaca

CRISP-DM (Cross Industry Standard Process for Data Mining), que es uno de los modelos principalmente utilizados en los ambientes académico e industrial. CRISP-DM es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Minería de Datos. (Ver Figura 10).

Los orígenes de CRISP-DM, se remontan hacia el año 1999 cuando un importante consorcio de empresas europeas tales como NCR (Dinamarca), AG (Alemania), SPSS (Inglaterra), OHRA (Holanda), Teradata, y Daimer-Chrysler, proponen a partir de diferentes versiones de KDD, el desarrollo de una guía de referencia de libre distribución denominada CRISP-DM [14].

El estándar incluye un modelo y una guía, estructurados en seis fases, La sucesión de fases no es necesariamente rígida. Cada fase es estructurada en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, donde finalmente se describen las acciones que deben ser desarrolladas para situaciones específicas, pero en ningún momento se propone como realizarlas [14]. Algunas de estas fases son bidireccionales, lo que significa que algunas fases permitirán revisar parcial o totalmente las fases anteriores [15]. (Ver Figura 11).

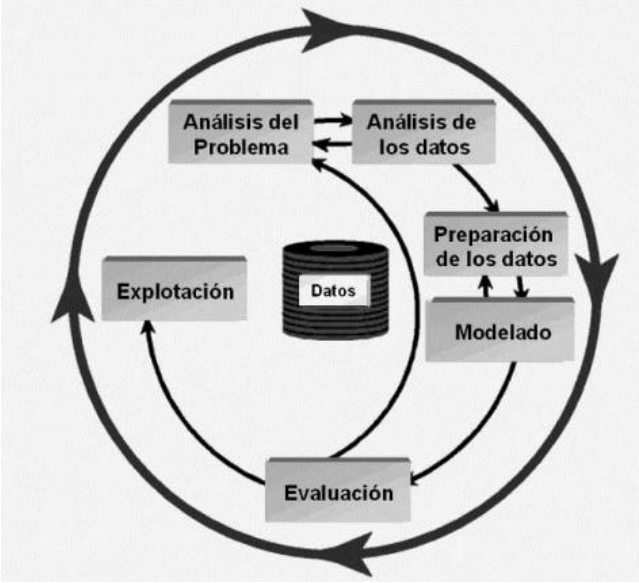

Figura 10. Modelo de proceso CRISP-DM. [36].

Las fases en las que se divide CRISP DM son:

- Entendimiento del negocio (Análisis del problema)
- Entendimiento de los datos
- Preparación de los datos
- Modelado
- Evaluación
- Implementación

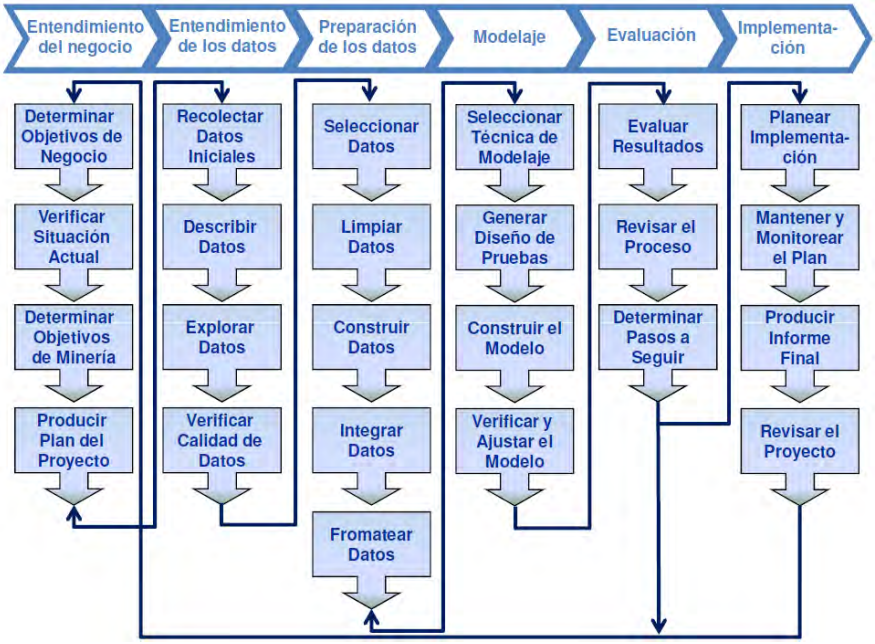

Figura 11. Ciclo de vida de un proyecto. Metodología CRISP-DM. [37].

Entendimiento del negocio:

La primera fase denominada fase de comprensión o entendimiento del negocio, es probablemente la más importante y reúne las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional. Para obtener el mejor provecho de Data Mining, es necesario entender de manera más completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados. En esta fase, es muy importante la capacidad de poder convertir el conocimiento adquirido del negocio, en un problema de Data Mining y en un plan preliminar cuya meta sea el alcanzar los objetivos del negocio [14].

Una descripción de cada una de las principales tareas que componen esta fase es la siguiente:

- **El establecimiento de los objetivos del negocio**. Básicamente se determina cuál es el problema que se desea resolver, por qué la necesidad de utilizar Data Mining y definir los criterios de éxito [14]. Todo lo que se refiere al contexto inicial, objetivos y los mencionados criterios de éxito.

- **Evaluación de la situación.** Se considera aspectos tales como: ¿cuál es el conocimiento previo disponible acerca del problema?, ¿se cuenta con la cantidad de datos requerida para resolver el problema?, ¿cuál es la relación coste beneficio de la aplicación de Data Mining?, etc. En esta fase se definen los requisitos del

problema, tanto en términos de negocio como en términos de Data Mining [14]. En resumen se realiza un inventario de recursos, requerimientos, supuestos y terminologías propias del negocio.

- **Establecimiento de los objetivos de la minería de datos.** Esta tarea tiene como objetivo representar los objetivos del negocio en términos de las metas del proyecto de DM [14].

- **Generación del plan del proyecto.** Descripción de los pasos a seguir y las técnicas a emplear en cada paso [14]. (plan, herramientas, equipo y técnicas).

Entendimiento de los datos:

Esta fase implica la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las próximas dos fases, son las que demandan el mayor esfuerzo y tiempo en un proyecto de Data Mining [14]. En general se requiere familiarizarse con los datos teniendo presente los objetivos del negocio [15].

- **Recopilación inicial de los datos.** Adecuación de datos para el futuro procesamiento. Se elaboran informes con una lista de los datos adquiridos, su localización, las técnicas utilizadas en su recolección, los problemas y soluciones inherentes a este proceso [14].

- **Descripción de los datos.** Este proceso involucra establecer volúmenes de datos (número de registros y campos por registro), su identificación, el significado de cada campo y la descripción del formato inicial [14].

- **Exploración de los datos.** El propósito es encontrar una estructura general para los datos. Esto involucra la aplicación de pruebas estadísticas básicas, que revelen propiedades en los datos recién adquiridos, se crean tablas de frecuencia y se construyen gráficos de distribución [14].

- **Verificación de calidad de datos.** Se determina la consistencia de los valores individuales de los campos, la cantidad y distribución de los valores nulos, y para encontrar valores fuera de rango, los cuales pueden constituirse en ruido para el proceso. Aquí se asegura la completitud y corrección de los datos [14].

Preparación de los datos:

Esta fase corresponde a la preparación de los datos para adaptarlos a las técnicas de Data Mining que se utilicen posteriormente, tales como técnicas de visualización de datos, búsqueda de relaciones entre variables u otras medidas

para exploración de los datos [14]. En la preparación de los datos se obtiene la vista minable o dataset [15].

La preparación de datos incluye las siguientes tareas generales de:

- **Selección de los datos.** En esta tarea, se selecciona un subconjunto de los datos adquiridos en la fase anterior, apoyándose en criterios previamente establecidos en las fases anteriores tales como calidad de los datos en cuanto a completitud y corrección de los datos y limitaciones en el volumen o en los tipos de datos que están relacionadas con las técnicas de Data Mining seleccionadas [14].

- **Limpieza de datos.** Esta tarea complementa a la anterior, y es una de las que más tiempo y esfuerzo consume, debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos a objeto de prepararlos para la fase de modelación. Algunas de las técnicas a utilizar para este propósito son, la normalización de los datos, desratización de campos numéricos, tratamiento de valores ausentes, reducción del volumen de datos, entre otras [14].

- **Construcción de datos.** Esta tarea incluye las operaciones de preparación de los datos tales como la generación de nuevos atributos a partir de atributos ya existentes, integración de nuevos registros o transformación de valores para atributos existentes [14].

- **Integración de datos.** Involucra la creación de nuevas estructuras, a partir de los datos seleccionados, por ejemplo, generación de nuevos campos a partir de otros existentes, creación de nuevos registros, fusión de tablas campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen [14].

- **Formateo de datos.** Está tarea consiste principalmente, en la realización de transformaciones sintácticas de los datos sin modificar su significado, esto, con la idea de permitir o facilitar el empleo de alguna técnica de Data Mining en particular, tales como la reordenación de los campos y/o registros de la tabla o el ajuste de los valores de los campos a las limitaciones de las herramientas de modelación (eliminar comas, tabuladores, caracteres especiales, máximos y mínimos para las cadenas de caracteres, etc.) [14].

Modelado:

En esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto de Data Mining específico. Las técnicas a utilizar en esta fase se eligen en función de ciertos criterios, en donde, la técnica debe ser apropiada según el problema, se debe disponer de datos adecuados así también como cumplir los requisitos del problema, disponer del tiempo adecuado para

obtener un modelo y tener conocimiento de la técnica [14]. En resumen en esta fase se aplica las técnicas de minería de datos al conjunto de datos (dataset) [15].

Se incluye las siguientes tareas generales de:

- **Selección de la técnica de modelado.** Consiste en la selección de la técnica de Data Mining más apropiada al tipo de problema a resolver [14].

- **Diseño de la evaluación.** Una vez construido un modelo, se debe generar un procedimiento destinado a probar la calidad y validez del mismo [14].

- **Construcción del modelo.** Después de seleccionada la técnica, se ejecuta sobre los datos previamente preparados para generar uno o más modelos [14].

- **Evaluación del modelo.** En esta tarea, los ingenieros de Data Mining interpretan los modelos de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos [14].

Evaluación:

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema [14]. De los modelos que se obtuvieron en las fases anteriores se determina si son útiles a las necesidades del negocio [15].

- **Evaluación de resultados.** Esta tarea involucra la evaluación del modelo en relación a los objetivos del negocio [14].

- **Revisar el proceso.** Se califica el proceso entero de Data Mining, a objeto de identificar elementos que pudieran ser mejorados [14].

- **Establecimiento de los siguientes pasos o acciones**. Según los resultados de la evaluación y la revisión de proceso, el equipo de proyecto decide cómo proceder. El equipo decide si hay que terminar este proyecto y tomar medidas sobre el desarrollo como iniciar más iteraciones, si es necesario o comenzar nuevos proyectos de minería de datos. Esta tarea incluye los análisis de recursos restantes y del presupuesto, que puede influir en las decisiones [16].

Implementación:

En esta fase, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio [14], Se requiere explotar la utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización [15].

- **Planificación de despliegue.** Resumir la estrategia de desarrollo, incluyendo los pasos necesarios y como realizarlos [16].

- **Planificación de la monitorización y del mantenimiento.** Resumir la estrategia de supervisión y mantenimiento incluyendo los pasos necesarios y como realizarlos [16].

- **Generación de informe final.** Esto incluye todo el desarrollo anterior, el resumen y la organización de los resultados [16].

- **Revisión del proyecto.** Resumir las experiencias importantes ganadas durante el proyecto. Evaluar lo que fue correcto y lo que se equivocó, lo que fue bien hecho y lo que necesita para ser mejorado [16].

3. CONSTRUCCIÓN DEL REPOSITORIO

Para la construcción del repositorio se abordaron cinco fases de la metodología CRISP-DM que son: el entendimiento del negocio, el entendimiento de los datos, la preparación de los datos, modelado y evaluación

3.1 FASE DE ENTENDIMIENTO DEL NEGOCIO

Objetivo del negocio:

Para esta fase es necesario tener pleno entendimiento del objetivo del negocio, en este caso la construcción de un repositorio limpio y transformado con los episodios eruptivos y emisiones del comportamiento del Volcán Galeras, utilizando el SGBD PostgreSQL, para su respectiva aplicación de técnicas de minería de datos, de tal manera que se puedan obtener patrones de eventos eruptivos.

El Volcán Galeras no pertenece a la categoría de los volcanes más destructivos, pero su importancia radica como se mencionó, en la pronta recurrencia de su actividad y además de que en su zona de influencia, se encuentran asentados siete municipios, el principal de ellos Pasto y un gran número de corregimientos y veredas, que en total albergan cerca de 500.000 habitantes. En varias ocasiones, tanto sus habitantes como la actividad económica se han visto afectados por las diversas manifestaciones del volcán. El problema fundamental con Galeras, se relaciona con el asentamiento de poblaciones en zonas de muy alta peligrosidad, especialmente por la probable afectación de flujos piroclásticos, lo cual incrementa notoriamente el nivel de vulnerabilidad y consecuentemente el riesgo, especialmente de aquellos habitantes que se ubican en zona de amenaza volcánica alta.

La definición e interpretación de los procesos que se realizan en el interior y en el entorno del volcán Galeras no es una tarea simple ya que implica el análisis de varios Parámetros. El objetivo es realizar el análisis de los factores a partir de varias disciplinas del conocimiento, por lo tanto es de vital importancia realizar minería de datos, a los datos obtenidos sobre el comportamiento del volcán para complementar los patrones por otros medios y con ello tener más herramientas para el aporte al análisis del comportamiento volcánico.

Evaluación de la situación actual:

El complejo volcánico Galeras, es vigilado y monitoreado por una sola entidad pública llamada Observatorio Vulcanológico y Sismológico de Pasto (OVSP), los

datos que facilitó esta entidad para el presente estudio están comprendidos entre el año 1989 hasta el año 2013, los cuales usan el SGBD PostgreSQL. Dichos datos son fundamentales para el proceso de la construcción del repositorio y por ende para el proceso de minería de datos.

Los datos están clasificados en diferentes áreas como son: Sismología Volcánica, Geoquímica, Climatológica, Deformación Volcánica, además de las bitácoras de emisión y erupción fundamentales para el desarrollo del proceso.

Una vez obtenida la base de datos del Volcán Galeras se analizaron los esquemas y sus respectivas tablas con el fin de seleccionar los datos más relevantes y construir el diccionario de datos con los mismos.

Esquemas:

A partir de la única fuente de datos, se analizaron los diferentes esquemas de base de datos para extraer los datos necesarios según las necesidades y los objetivos de la investigación. La base de datos está conformada por 34 tablas. (Ver Figura 12)

Base de datos OVSP
Public (34 tablas y 2 vistas)

Figura 12. Esquema de la base de datos [41].

Selección de tablas:

Después de un minucioso análisis y posterior selección de los esquemas, se efectuó la selección de las tablas que contienen la información necesaria para satisfacer los objetivos del proyecto. Estas relaciones se detallan en (Ver Tabla 4)

Diccionario de datos:

Para efectos de comprender mejor el funcionamiento y la dinámica de las relaciones existentes entre todas las tablas de las bases de datos, se hace necesario construir el diccionario de datos. De esta manera, para las tablas de la base de datos “OVSP”, el diccionario de datos se encuentra en (Ver Anexo1).

Se requirió de una constante comunicación con el personal del Observatorio Sismológico y Vulcanológico de Pasto, con el fin de solucionar las dudas que se presentaron conforme se avanzaba con el análisis de la base de datos. La colaboración por medio de reuniones y por correo electrónico facilitó avanzar en esta parte vital de la investigación.

Diagrama entidad relación-inicial:

A partir de los esquemas que contienen las tablas con los atributos relevantes, se elaboró el primer diagrama Entidad-Relación que se denominó Ovspbd. (Ver Figura 13).

Esquema public	
Relaciones	**Descripción**
Bitacoras_act_superf	Historial de los eventos que se han presentado en el volcán
Bitacoras_emision	Historial de emisiones del volcán
Bitacoras_erupcion	Historial de erupciones del volcán
Bitacora_observacion	Historial de las observaciones que se han realizado al volcán
Climatología	Datos del clima
Conteomanual	Datos de conteo de eventos presentados
Datosinclinomurad	Datos de la estación de inclinometría
Energiatotal	Datos de la energía total
Estacion_doas	Datos registrados por la estación sobre la distancia del cráter, altura de referencia, dirección de compas y cobertura de escaneo
Estacioninclinom	Datos registrados por la estación sobre la cobertura y altitud de la estación
Hyb	Datos sobre el movimiento de fluidos
Leclogashyb	Datos de la lectura análoga de hyb
LecanalogasLPS	Datos de la lectura análoga de LPS
Lecanalogastre	Datos de la lectura análoga de tre
Lecanalogasvt	Datos de la lectura análoga de vt
LecdigitalesLPS	Datos de la lectura digital de hyb
Lecdigitalestre	Datos de la lectura digital de tre
Lecdigitalesvt	Datos de la lectura digital de vt
Línea_radon	Datos sobre la línea de radón.
Localizaciones	Datos de la localización de los eventos
LPS	Datos del tipo de evento de larga duración registrado
Medida_radon	Datos de la cantidad de radón que se registro
Omi	Lectura del gas
Punto_radon	Datos del lugar donde hay presencia de radón
Radón_conc	Datos de la concentración de radón
Vige_radon	Datos del electro instalado para regular la cantidad de radón
So2	Descripción del so2 que se registro

Esquema public	
Relaciones	**Descripción**
So2_flux	Descripción del so2 flux que se registro
Tremor	Datos de los sismos tipo tremor que se registraron
Ventanatremor	Lectura de parámetros de un segmento de tremor
Volcán	Datos del volcán
Vt	Datos de los signos volcatectónicos

Tabla 4. Relaciones del esquema public– Ovsp [41].

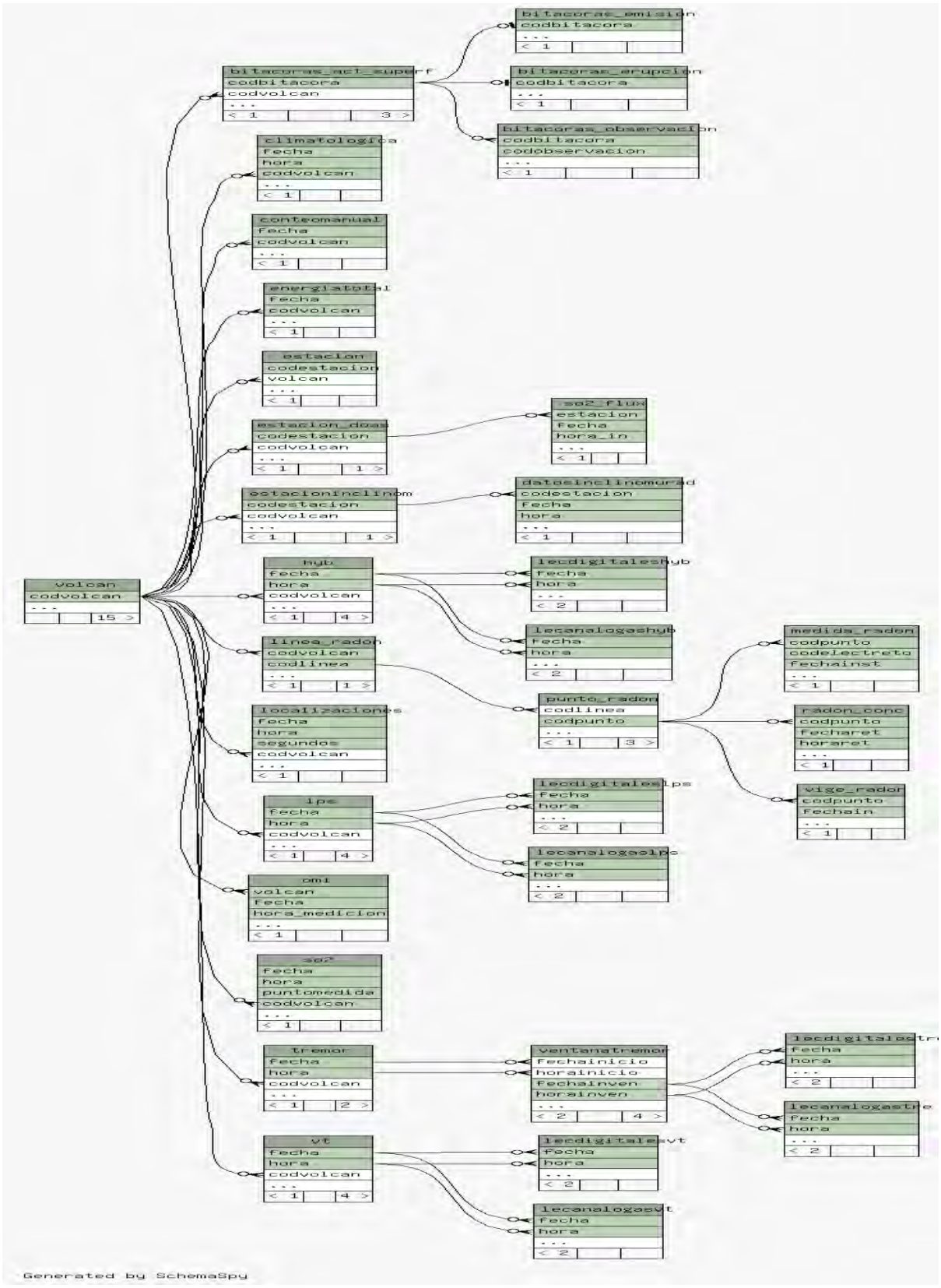

Figura 13. Diagrama Entidad – Relación. [41].

3.2 FASE DE ENTENDIMIENTO DE LOS DATOS

Para establecer las relaciones más evidentes que permitan definir las primeras hipótesis y lograr familiarizarse con los datos teniendo presente los objetivos del negocio y una vez creada la estructura con la mayor cantidad de atributos relevantes se realizó el primer análisis de calidad de datos (Ver Anexo2). En el cual se analizaron los datos y se observaron las inconsistencias que presentaba la información, por lo general son errores causados en la entrada del usuario; es aquí donde las técnicas de preprocesamiento de datos son útiles para mejorar la calidad de estos. Se presentaron inconsistencias en tablas que se relacionan con el área de sismología, debido a que una parte de la información se encuentra almacenada con atributos y medidas diferentes, dado a que se realizaron cambios en los equipos que toman estos datos.

Creación de tablas temporales:

A fin de desarrollar el procesamiento y la adecuación de datos se crearon tablas temporales, las cuales sirven para realizar la limpieza sin afectar las tablas originales. En éstas se llenaron los valores faltantes, se realizó corrección de datos manualmente, de las tablas del área sismológica (hyb, LPS, tre, vt) se realizó una conversión de unidades y con valores específicos para sacar la velocidad del movimiento del terreno y la unidad de medida es cm/segundo. Debido a que ciertos atributos aun no existían, fue necesario crear nuevos campos e insertar los valores correspondientes. Así mismo, fue necesario adicionar tablas que no se encontraban en las bases de datos originales. Todo esto con el fin de facilitar la construcción del repositorio final y la limpieza y transformación del mismo. (Ver Tabla 5).

TABLAS TEMPORALES PARA LIMPIEZA DE ATRIBUTOS			
* localizaciones * LPS * radon_conc * omi * so2	* bitacoras_act_superf * bitacoras_emision * bitacoras_erupcion * conteomanual	* tmp_vel_tre * tmp_vel_LPS * datosinclinomurad * hyb	* tmp_vel_vt * tmp_vel_hyb * vt * tremor

Tabla 5. Tablas temporales [41].

Para complementar y corregir los datos faltantes, fue necesario corroborar con el OVSP algunas fechas, nombres de estaciones, filtros y ganancias de los sismógrafos específicamente donde se encontraban datos absurdos. Como por ejemplo en la tabla conteo manual existía una fecha de "0009-05-01".

Limpieza de tablas temporales:

Al realizar una análisis de las tablas tmp_vel_tre, temp_vel_LPS, temp_vel_vt, temp_vel_hyb, la mayoría de los atributos contenidos en estas era ruidosa, es decir que existían valores que están significativamente fuera del rango de valores esperados o los valores eran incoherentes por errores humanos, cambios en el sistema y por fuentes heterogéneas de los mismos. Se analizaron detalladamente cada uno de ellos para determinar el tipo de estrategia a utilizar para el manejo de datos desconocidos o duplicados, se analizaron los campos mediante agrupación y consultas SQL y se realizaron los cambios respectivos mediante update, insert, delete. La descripción de la limpieza llevada a cabo en las tablas temporales. (Ver Tabla 6)

Tabla	Tipo de SQL	Descripción Limpieza
Bitacoras_emision	Update	Se cambió Color “NO DETARMINADO” por “NO DETERMINADO”
Bitacoras_emision	Update	Se cambió fenómenos sonido “no” por “NO”
Bitacoras_emision	Update	Se cambió en el atributo fenómenos_olor “fuerte olor a azufre”, “a azufre”, “azufre”, “Sulfuro de Hidrogeno” a “Azufre” “no” a “NO”
Bitacoras_emision	Update	Se cambió Color “NO DETARMINADO” por “NO DETERMINADO”
Hyb	Delete	Eliminación de sp, etotoc, etotor, drtotaloc, drtotalor, periodo, frecuencia, magnitudcoda, magnitudcodal, energiacodal,
Hyb	Update	Se cambió estación “ OLGA”, “ OLGA_” por “OLGA”, “ CUVZ”, “CUVZ_” por “CUVZ” “ ANGV”, “ ANGV_”, “ ANGV0” por “ANGV” , “ CR2R” “CR2R_” por “CR2R”
Hyb	Delete	Eliminación de observaciones toda la columna está vacía
Localizaciones	Delete	Eliminación de archivo, modelo, profitera, codlocalizacion
LPS	Update	Se cambió tipo “LPS” por “LPS”, “GLP*” , “GLP+”, “plp” por “GLP” “GTO*”, “GTO+” por “GTO”
LPS	Delete	Se elimina tipo “ERU”
LPS	Update	Se cambia el atributo “frecuancia” por “frecuencia”
LPS	Update	Se cambió estacion “ OLGA”, “ OLGA_” por “OLGA”, “ CUVZ”, “CUVZ_” por “CUVZ” “ ANGV”, “ ANGV_”, “ ANGV0” por “ANGV”

Tabla	Tipo de SQL	Descripción Limpieza
		, “ CR2R” “CR2R_” por “CR2R”
LPS	Delete	Se elimina el atributo etotalor, etotaloc, drtotaloc, drtotalor
LPS	Delete	Eliminación de observaciones toda la columna está vacía, magnitudcodal
LPS, hyb, bitacoras_act_superf, bitacoras_emision, bitacoras_erupcion, conteomanual, tremor, vt, SO2	Delete	Se elimina el atributo codVolcán debido a que toda la información pertenece al código 1 que es Galeras
Omi	Delete	Se eliminan los atributos Volcán debido a que toda la información pertenece al código 1 que es Galeras
Omi	Delete	Se eliminan los atributos fuente y observación
radon_conc	Delete	Se eliminan los atributos horaret, fecharet, codpunto
SO2	Delete	Se eliminan los atributos azimut, instrumento
SO2	Update	Se cambia Punto_medida “SANTA BÁRBARA” por “Santa Barbara” , “ALTO TINAJILLAS” por “Alto Tinajillas”, “PAS-CON” por “PASTO-CONSACA”, “CON-PAS” por “CONSACA- PASTO”
Tremor	Delete	Se eliminan los atributos ampmax, ampmin, permin, permax, etotoc, etotor, drtotaloc, drtotalor, observaciones, tipoevento,
Vt	Delete	Eliminación de sp, etotoc, etotor, drtotaloc, drtotalor, periodo, frecuencia, magnitudcoda, magnitudcodal, energiacodal,
Vt	Update	Se cambió estacion “ OLGA”, “ OLGA_” por “OLGA”, “ CUVZ”, “CUVZ_” por “CUVZ” “ ANGV”, “ ANGV_”, “ ANGV0” por “ANGV” , “ CR2R” “CR2R_” por “CR2R”
Vt	Delete	Eliminación de observaciones; toda la columna está vacía
Temp_vel_hyb	Insert	Inserta la velocidad del sismo tipo hibrido
Temp_vel_LPS	Insert	Inserta la velocidad del sismo tipo LPS
Temp_vel_tre	Insert	Inserta la velocidad del sismo tipo tremor
Temp_vel_vt	Insert	Inserta la velocidad del sismo tipo volcanotectónico

Tabla 6. Descripción limpieza tablas temporales. [41].

3.2.1 Construcción de la base de datos repositorio galeras. Se procedió a construir la base de datos *repositoriogaleras*, a partir de las tablas temporales que se limpiaron, con el fin de facilitar la construcción del repositorio final y por lo tanto para desarrollar satisfactoriamente la siguiente fase de la preparación de los datos.

La nueva base de datos se compone de diecisiete (17) tablas (Ver Tabla 7) y cuenta con integridad referencial, por lo que cada relación cuenta con sus respectivas llaves primarias (Ver Figura 14).

TABLAS	
* localizaciones	* bitacoras_act_superf
* tmp_vel_LPS	* bitacoras_emision
* radon_conc	* bitacoras_erupcion
* omi	* conteomanual
* so2	* tmp_vel_vt
* datosinclinomurad	* tmp_vel_tremor
* tmp_vel_hyb	* vt
* hyb	*tremor
*LPS	

Tabla 7. Tablas base de datos repositoriogaleras. [41].

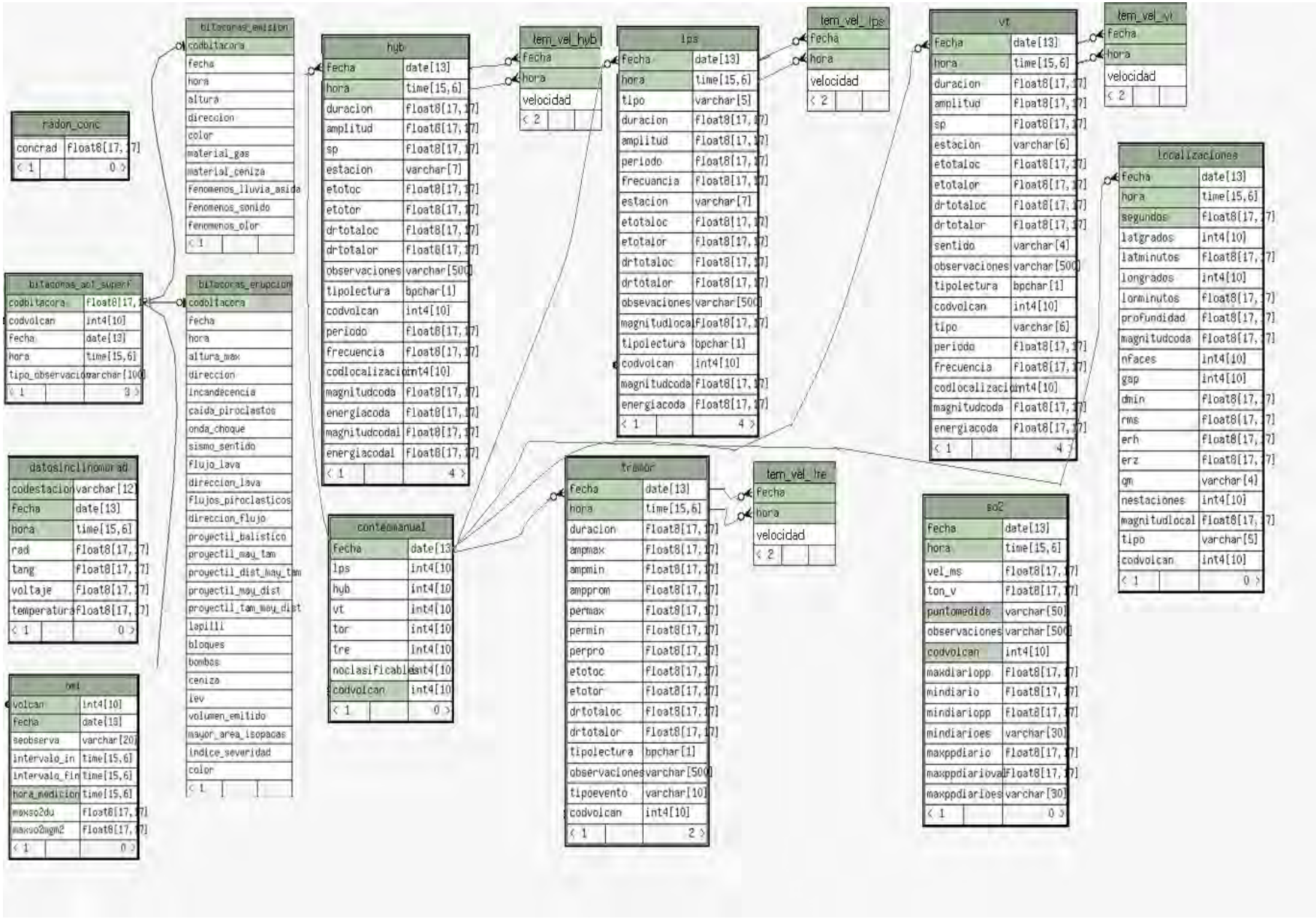

Figura 14. Diagrama entidad-relación base de datos repositoriogaleras. [41].

Descripción base de datos repositoriogaleras:

A continuación se presenta la descripción de las relaciones de la base de datos repositoriogaleras. (Ver Tabla 8).

Relaciones	Descripción
bitacoras_act_superf	Tabla que contiene las bitacoras de actividad superficial
bitacoras_emision	Bitácora de las emisiones de gas o ceniza reportadas al Observatorio
bitacoras_erupcion	Contiene los reportes de erupción
conteomanual	Conteo de número de eventos por día.
datosinclinomurad	Datos obtenidos con los inclinómetros electrónicos.
localizaciones	Contiene la localización de los eventos relacionados a fracturamiento de material cortical.
Hyb	Tabla que contiene los eventos tipo hibrido
LPS	Tabla que contiene los eventos relacionados a movimiento de fluido de fuente transitoria; con una duración de 5 a 10 segundos
so2	Datos de medidas de so2, máximos diarios, SCANDODAS, MOVILDOAS, COSPEC.
tremor	Tabla que contiene los eventos relacionados a movimiento de fluido de fuente persistente en el tiempo; con una duración de minutos, horas días-
radon_conc	Datos de concentración radón.
vt	Tabla que contiene los eventos relacionados a fracturas del edificio volcánico; tipo volcanotectónicos
Tmp_vel_tre	Tabla que contiene la velocidad relacionada al movimiento de fluido de fuente persistente en el tiempo.
Tmp_vel_LPS	Tabla que contiene la velocidad relacionada a movimiento de fluido con una duración de 5 a 10 segundos.
Tmp_vel_vt	Tabla que contiene la velocidad de eventos relacionados a fracturas del edificio volcánico; tipo volcanotectónicos
Tmp_vel_hyb	Tabla que contiene la velocidad de eventos tipo hibrido
Omi	

Tabla 8. Descripción de las tablas que conforman la base de datos repositoriogaleras. [41].

3.2 FASE DE PREPARACIÓN DE LOS DATOS

Vista minable o dataset, es aquí donde se realiza la preparación de los datos para adaptarlos a las técnicas de minería de datos que se utilicen posteriormente.
SELECCIÓN DE DATOS

Una vez construida la base de datos *repositoriogaleras* que servirá de fuente de alimentación del repositorio y que contendrá los atributos necesarios para la aplicación de técnicas de minería de datos. Se llevó a cabo la selección de los 59 atributos más relevantes de esta base de datos (Ver Tabla 9), a partir de los cuales se da inicio a las tareas de limpieza, construcción, integración y transformación de datos que culminarán con el repositorio final.

Atributo	Tipo	Tamaño	Descripción
Fecha	Date	13	Fecha de registro del evento
ton_v	Float8	17,7	Máximo valor de So2 por día (Toneladas/día)
Tipo	Character varying	5	Tipo de evento localizado.
Qm	Character varying	4	Calidad de la localización [A,B,C,D] A: epicenter excellent, focal depth good B: epicenter good, focal depth fair C: epicenter fair, focal depth poor D: epicenter poor, focal depth poor
Velocidad	Float8	17,7	Velocidad del sismo
Conrad	Float8	17,7	concentración de radón calculada
Maxso2du	Double precision		Máximo S02 (Unidades dobson)
Maxso2mgm2	Double precision		Máximo S02 (mg/m^2)
Codestacion	Carácter varying	7	Código de la estación
Voltaje	Float8	17,7	Valor de la componente voltaje (voltios), el valor 9999999 corresponde a NaN.
Temperatura	Float8	17,7	Valor de la componente temperatura (grados centígrados), el valor 9999999 corresponde a NaN.
Segundos	Float8	17,7	Segundos de ocurrencia del evento con decimal.
Latgrados	Integer	10	Latitud localización del evento (grados)
Latminutos	Float8	17,7	Latitud localización del evento (minutos)
Longrados	Integer	10	Longitud localización del evento

Atributo	Tipo	Tamaño	Descripción
			(grados).
Longminutos	Float8	17,7	Longitud localización del evento (minutos).
Profundidad	Float8	17,7	profundidad focal (km)
Rad	Float8	17,7	Valor de la componente radial (microradianes), el valor 9999999 corresponde a NaN.
Etotaloc	Float8	17,7	Energía de ondas de cuerpo de todo el evento (ergios).
Etotalor	Float8	17,7	Energía de ondas de superficie de todo el evento (ergios)
Etotoc	Float8	17,7	Energía de ondas de cuerpo de todo el evento (ergios).
Etotor	Float8	17,7	Energía de ondas de superficie de todo el evento (ergios)
Drtotaloc	Float8	17,7	Desplazamiento reducido de ondas de cuerpo de todo el evento (cm^2).
Drtotalor	Float8	17,7	Desplazamiento reducido onda de superficie de todo el evento (cm^2).
LPSoc	Float8	17,7	Energía total ondas de cuerpo de los eventos tipo LPS clasificados.
LPSor	Float8	17,7	Energía total ondas de superficie de los eventos tipo LPS clasificados.
Hyboc	Float8	17,7	Energía total ondas de cuerpo de los eventos tipo HYB clasificados.
Hybor	Float8	17,7	Energía total ondas de superficie de los eventos tipo HYB clasificados.
Vtoc	Float8	17,7	Energía total ondas de cuerpo de los eventos tipo VT clasificados.
Vtor	Float8	17,7	Energía total ondas de superficie de los eventos tipo VT clasificados.
Treoc	Float8	17,7	Energía total ondas de cuerpo de los eventos tipo TRE clasificados.
Treor	Float8	17,7	Energía total ondas de superficie de los eventos tipo TREMOR clasificados
Altura	Float8	17,7	Altura máxima de la emisión (metros sobre la cima).
Dirección	Character varying	-30	Dirección a la cual se desplaza la columna.
Color	Character varying	-30	Color de la emisión.

Atributo	Tipo	Tamaño	Descripción
material_gas	Character varying	-5	Si se observa salida de gas [S, N]
Cenizas	Character varying	-5	Si se observó salida de ceniza[S,N]
fenomenos_olor	Character varying	-70	Si se asocia algún tipo de olor a la emisión.
Emision	Character	-1	Si hay emisión [S, N].
Incandecencia	Character varying	(5),	Si se observa incandescencia [S,N]
altura_max	Double precision		Altura máxima de la columna
caida_piroclastos	Character varying	(5),	Si se observa caída de piroclastos [S, N].
onda_choque	Character varying	(5),	Si se observa onda de choque [S, N].
proyectil_balistico	Character varying	(5),	Si se registran proyectiles balísticos.
Erupcion	Character	-1	Si hay erupción [S, N].
sismo_sentido	Character	-1	Si se registra un sismo sentido relacionado con la erupción [S,N]
LPS	character varying	10	Numero de eventos LPS clasificables para la fecha dada.
Vt	character varying	10	Numero de eventos VT clasificables para la fecha dada.
Tre	character varying	10	Numero de eventos tipo TREMOR para la fecha dada.
Tor	character varying	10	Numero de eventos tipo TOR para la fecha dada.
Hyb	character varying	10	Numero de eventos HYB clasificables para la fecha dada.

Tabla 9. Atributos seleccionados de la base de datos repositoriogaleras. [41].

Limpieza de datos:

En la etapa de Limpieza se analiza la calidad de los datos, se aplican operaciones básicas como la remoción de datos ruidosos, se seleccionan estrategias para el manejo de datos desconocidos, datos nulos, datos duplicados y técnicas estadísticas para su reemplazo. En esta etapa, es de suma importancia la interacción con el usuario o analista.

Los datos ruidosos son valores que están significativamente fuera del rango de valores esperado. Se deben principalmente, a errores humanos, a cambios en el

sistema, a información no disponible a tiempo y a fuentes heterogéneas de datos. Los datos desconocidos son aquellos a los cuales no les corresponde un valor en el mundo real y son aquellos que tienen un valor que no fue capturado. Los datos nulos son datos desconocidos que son permitidos por los sistemas gestores de bases de datos relacionales (SGBDR). En el proceso de limpieza todos estos valores se ignoran, se reemplazan por un valor por omisión, por el valor más cercano o se usan métricas de tipo estadístico como la media, mínimo, máximo, para reemplazarlos.

De los datos existentes en la base de datos, se decidió seleccionar los registros de los diferentes atributos mencionados anteriormente y se decidió unificar registros por medio de atributos evaluados por día.

Se analizaron los datos contenidos en cada uno de los atributos, a fin de optimizar la calidad de los datos y que el análisis sirva para determinar cuáles atributos se conservarán para construir el repositorio final y cuales se eliminarán, todo esto se hace con el objeto de preparar los datos para la fase de aplicación de algoritmos de minería de datos. Los resultados del análisis de calidad de los datos (Ver Anexo3). Resumen de calidad de datos (Ver Tabla 10.).

Atributo	% Nulos
Fecha	0.0%
Maxso2du	76.24%
Maxso2mgm2	76.24%
Ton_v	78.12%
Codestacion	9.39%
Voltaje	9.34%
Temperatura	12.1%
Tipo	24.64%
Segundos	37.21%
Latgrados	37.21%
Latminutos	37.21%
Longrados	37.21%
Lonminutos	37.21%
Profundidad	37.21%
Qm	37.21%
Velocidad	0.0%
Rad	5.9%
Altura	93.54%
Dirección	93.54%
Color	93.54%
Material_gas	93.54%

Atributo	% Nulos
Cenizas	93.54%
Fenómenos_olor	99.94%
Emisión	92.79%
Altura_max	99.95%
Incandecencia	99.83%
caida_piroclastos	99.83%
Onda_choque	99.83%
Proyectil_balistico	99.83%
Erupción	99.79%
Sismo_sentido	98.43%
Otro	94.34%
Etotaloc	66.48%
Etotalor	66.48%
Etotoc	40.45%
Etotor	40.45%
Drtotaloc	24.65%
Drtotalor	24.65%
LPS	0.06%
Hyb	0.06%
Vt	0.06%
Tor	0.06%
Tre	0.06%
LPSoc	0.0%
LPSor	0.0%
Hyboc	0.0%
Hybor	0.0%
Vtoc	0.0%
Vtor	0.0%
Treoc	0.0%
Treor	0.0%
RLPSoc	0.0%
RLPSor	0.0%
Rhyboc	0.0%
Rhybor	0.0%
Rvtoc	0.0%
Rvtor	0.0%
Rtreoc	0.0%
Rtreor	0.0%

Tabla 10. Análisis de calidad de datos de la tabla de 59 atributos. [41].

Construcción de datos:

Se continúa con las operaciones de preparación de los datos, se realiza la generación de nuevos atributos a partir de atributos ya existentes en el repositorio, se integran registros y se transforman los valores para algunos atributos.

En la construcción de nuevos atributos tales como comportamientoton_v, comportamientorad comportamientovel y clase. Para comportamientorad se evaluó el atributo rad donde considerando el anterior valor de cada tupla se lo clasificó como “subió” “estable”, y “bajó”. Para comportamientovel se evaluó el atributo velocidad donde considerando el anterior valor de cada tupla se lo clasificó como “subió” “estable”, y “bajó”.
Para comportamientoton_v se evaluó el atributo tonv donde considerando el anterior valor de cada tupla se lo clasificó como “subió” “estable”, y “bajó”. Para clase_sismos, se evaluó los atributos latgra, latmin, longmin, longgrados, profundidad, hyb, LPS, tre, tor, vt y tipo, con los cuales se clasificó en “S” y “N”, lo que quiere decir es que si se registró algún evento sismológico entonces en clase_sismos se clasifica como si “S”, de lo contrario se clasifica como no “N”. Se describe los atributos adicionados según las necesidades de la investigación. (Ver Tabla 11).

Atributos Añadidos	Descripción
Comportamientoton_v	Considerando el atributo ton_v se lo clasifica en “subió” “estable”, y “bajó”.
Comportamientovel	Considerando el atributo velocidad se lo clasifica en “subió” “estable”, y “bajó”.
Comportamientorad	Considerando el atributo rad se lo clasifica en “subió” “estable”, y “bajó”.

Tabla 11. Atributos a añadir. [41].

Integración de datos:

Una vez adicionados los nuevos atributos se procede a realizar nuevamente otro análisis de calidad de datos de la nueva tabla que se denominó **eventos001**. (Ver Tabla 12).

Atributo	% Nulos
Fecha	0.0%
Maxso2du	76.24%
Maxso2mgm2	76.24%

Atributo	% Nulos
Ton_v	78.12%
Codestacion	9.39%
Voltaje	9.34%
Temperatura	12.1%
Tipo	24.64%
Segundos	37.21%
Latgrados	37.21%
Latminutos	37.21%
Longrados	37.21%
Lonminutos	37.21%
Profundidad	37.21%
Qm	37.21%
Velocidad	0.0%
Rad	5.9%
Altura	93.54%
Dirección	93.54%
Color	93.54%
Material_gas	93.54%
Cenizas	93.54%
Fenómenos_olor	99.94%
Emisión	92.79%
Altura_max	99.95%
Incandecencia	99.83%
caida_piroclastos	99.83%
Onda_choque	99.83%
Proyectil_balistico	99.83%
Erupción	99.79%
Sismo_sentido	98.43%
Otro	94.34%
Etotaloc	66.48%
Etotalor	66.48%
Etotoc	40.45%
Etotor	40.45%
Drtotaloc	24.65%
Drtotalor	24.65%

Atributo	% Nulos
LPS	0.06%
Hyb	0.06%
Vt	0.06%
Tor	0.06%
Tre	0.06%
LPSoc	0.0%
LPSor	0.0%
Hyboc	0.0%
Hybor	0.0%
Vtoc	0.0%
Vtor	0.0%
Treoc	0.0%
Treor	0.0%
RLPSoc	0.0%
RLPSor	0.0%
Rhyboc	0.0%
Rhybor	0.0%
Rvtoc	0.0%
Rvtor	0.0%
Rtreoc	0.0%
Rtreoc	0.0%
Comportamientorad	5.91%
Coportamientoton_v	78.13%
Comportamientovel	0.0%
Clase	0.0%

Tabla 12. Análisis de calidad de datos de la tabla eventos001. [41].

Hecho el análisis, se procede a eliminar campos que no son relevantes y que contienen información ruidosa y que arrojan porcentajes significativamente altos de datos nulos, también se hace la modificación de algunos atributos y se reemplazan por otros los cuales se generan mediante atributos ya existentes. Los campos que se eliminaron y correlacionaron se muestran (Ver Tabla 13).

Se remplaza el atributo erupción por clase_erupcion, donde se evaluó los atributos propios de erupción como son la incandescencia, altura_maxima, caida_priroclasticos y onda_choque, con los cuales se clasificó en “S” y “N”, lo que quiere decir es que si se registró una onda de choque y/o una caída de piroclastos

y/o una incandescencia entonces en clase_erupcion se clasifica como si “S”, de lo contrario se clasifica como no “N”.

Se remplaza el atributo emisión por clase_emision, donde se evaluó los atributos propios de emisión como son la altura, material_gas, cenizas y fenómenos_olor, con los cuales se clasificó en “S” y “N”, lo que quiere decir es que si se registró una altura de emisión de más de 2000 metros y/o se observa salida de gas y/o se observa salida de ceniza y/o se asocia algún tipo de olor a la emisión entonces en clase_emision se clasifica como si “S”, de lo contrario se clasifica como no “N”.

Atributo	Razón eliminación
Codestacion	Poca relevancia para el proceso de minería
Voltaje	
Temperatura	
Drtotalor	
Drtotaloc	
Etotaloc	
Etotoc	
Etotor	
Etotalor	
Sismo_sentido	Alto porcentaje de nulos o datos ruidosos
Otro	
Hyboc	Correlacionado con “hyb”
Hybor	
Rhyboc	
Rhybor	
LPSoc	Correlacionado con “LPS”
LPSor	
RLPSoc	
RLPSor	
Vtoc	Correlacionado con “vt”
Vtoc	
Rvtoc	
Rvtor	
Treoc	Correlacionado con “tre”
Treor	
Rtreor	
Rtreoc	
Latgrados	Correlacionado con “qm”
Latminutos	
Longminutos	

Longgrados	
Segundos	
Profundidad	

Tabla 13. Atributos eliminados. [41].

Como resultado del análisis anterior y teniendo en cuenta los nuevos atributos generados, se descartaron 71 registros que no presentaban coherencia en su información (no coincidían con las fechas de emisión o erupción) o que contenían una gran cantidad de campos nulos**.**

Se crearon cuatro (4) tablas de eventos eruptivos del Volcán Galeras; ellas se denominan **TERU17A18**, **TEMI648A19, TERUT8729A18 y TEMIT8729A19**

La estructura de la tabla **TERU17A18** se presenta con más detalle (Ver Tabla 14), en esta tabla se encuentran solamente las erupciones del volcán Galeras hasta el año 2013, con los atributos propios de erupciones. La estructura de la tabla **TEMI648A19** se presenta con más detalle (Ver Tabla 15), En esta tabla se encuentran solamente las emisiones del volcán Galeras desde el año 1989 hasta el año 2013, con los atributos propios de emisiones. La estructura de la tabla **TERUT8729A18** se presenta con más detalle (Ver Tabla 16), en esta tabla se encuentran todos los datos significativos con los atributos propios de erupciones. La estructura de la tabla **TEMIT8729A19** se presenta con más detalle (Ver Tabla 17), en esta tabla se encuentran todos los datos significativos con los atributos propios de emisiones.

Atributo	Tipo	Tamaño	Descripción
Fecha	Date	13	Fecha de registro del evento
comportamientoton_v	Character varying	17,7	Subió o bajó según el anterior
Tipo	Character varying	5	Tipo de evento localizado.
Qm	Character varying	(4)	Calidad de la localización [A,B,C,D] A: epicenter excelent, focal depth good B: epicenter good, focal depth fair C: epicenter fair, focal depth poor D: epicenter poor, focal depth poor
comportamientovel	Float8	17,7	Subió o bajó según la Velocidad del sismo anterior
Comportamientorad	Float8	17,7	Subió o bajó según el radón del sismo anterior
Incandecencia	Character varying	(5),	Si se observa incandescencia [S,N]

Atributo	Tipo	Tamaño	Descripción
altura_max	Double precision,		Altura máxima de la columna
caida_piroclastos	Character varying	(5),	Si se observa caída de piroclastos [S, N].
onda_choque	Character varying	(5),	Si se observa onda de choque [S, N].
proyectil_balistico	Character varying	(5),	Si se registran proyectiles balísticos.
LPS	character varying	10	Numero de eventos LPS clasificables para la fecha dada.
Vt	character varying	10	Numero de eventos VT clasificables para la fecha dada.
Tre	character varying	10	Numero de eventos tipo TREMOR para la fecha dada.
Tor	character varying	10	Numero de eventos tipo TOR para la fecha dada.
Hyb	character varying	10	Numero de eventos HYB clasificables para la fecha dada.
Clase_sismos	character varying	10	Se relacionó algún sismo [S,N]
Clase_erupcion	character varying	10	Si hay erupción [S,N]

Tabla 14. Estructura de la tabla TERU17A18. [41].

Atributo	Tipo	Tamaño	Descripción
Fecha	Date	13	Fecha de registro del evento
comportamientoton_v	Character varying	17,7	Subió o bajó según el anterior
Tipo	Character varying	5	Tipo de evento localizado.
Qm	Character varying	(4)	Calidad de la localización [A,B,C,D] A: epicenter excelent, focal depth good B: epicenter good, focal depth fair C: epicenter fair, focal depth poor D: epicenter poor, focal depth poor
comportamientovel	Float8	17,7	Subió o bajó según la Velocidad del sismo anterior
Comportamientorad	Float8	17,7	Subió o bajó según el radón del sismo anterior
Altura	Float8	17,7	Altura máxima de la emisión (metros sobre la cima).
Dirección	Character varying	(30)	Dirección a la cual se desplaza la columna.

Color	Character varying	(30)	Color de la emisión.
material_gas	Character varying	(5)	Si se observa salida de gas [S, N]
Cenizas	Character varying	(5)	Si se observó salida de ceniza[S,N]
fenomenos_olor	Character varying	(70)	Si se asocia algún tipo de olor a la emisión.
Clase_emision	Character	(1)	Si hay emisión [S, N].
LPS	character varying	10	Numero de eventos LPS clasificables para la fecha dada.
Vt	character varying	10	Numero de eventos VT clasificables para la fecha dada.
Tre	character varying	10	Numero de eventos tipo TREMOR para la fecha dada.
Tor	character varying	10	Numero de eventos tipo TOR para la fecha dada.
Hyb	character varying	10	Numero de eventos HYB clasificables para la fecha dada.
Clase_sismos	character varying	10	Se relacionó algún sismo [S,N]

Tabla 15. Estructura de la tabla TEMI648A19 [41].

Atributo	Tipo	Tamaño	Descripción
Fecha	Date	13	Fecha de registro del evento
comportamientoton_v	Character varying	17,7	Subió o bajó según el anterior
Tipo	Character varying	5	Tipo de evento localizado.
Qm	Character varying	(4)	Calidad de la localización [A,B,C,D] A: epicenter excelent, focal depth good B: epicenter good, focal depth fair

			C: epicenter fair, focal depth poor D: epicenter poor, focal depth poor
comportamientovel	Float8	17,7	Subió o bajó según la Velocidad del sismo anterior
Comportamientorad	Float8	17,7	Subió o bajó según el radón del sismo anterior
Incandecencia	Character varying	(5),	Si se observa incandescencia [S,N]
altura_max	Double precision,		Altura máxima de la columna
caida_piroclastos	Character varying	(5),	Si se observa caída de piroclastos [S, N].
onda_choque	Character varying	(5),	Si se observa onda de choque [S, N].
proyectil_balistico	Character varying	(5),	Si se registran proyectiles balísticos.
LPS	character varying	10	Numero de eventos LPS clasificables para la fecha dada.
Vt	character varying	10	Numero de eventos VT clasificables para la fecha dada.
Tre	character varying	10	Numero de eventos tipo TREMOR para la fecha dada.
Tor	character varying	10	Numero de eventos tipo TOR para la fecha dada.
Hyb	character varying	10	Numero de eventos HYB clasificables para la fecha dada.
Clase_sismos	character varying	10	Se relacionó algún sismo [S,N]
Clase_erupcion	character varying	10	Si hay erupción [S,N]

Tabla 16. Estructura de la tabla TERUT8729A18. [41].

Atributo	Tipo	Tamaño	Descripción
Fecha	Date	13	Fecha de registro del evento
comportamientoton_v	Character varying	17,7	Subió o bajó según el anterior
Tipo	Character varying	5	Tipo de evento localizado.

Atributo	Tipo	Tamaño	Descripción
Qm	Character varying	(4)	Calidad de la localización [A,B,C,D] A: epicenter excelent, focal depth good B: epicenter good, focal depth fair C: epicenter fair, focal depth poor D: epicenter poor, focal depth poor
comportamientovel	Float8	17,7	Subió o bajó según la Velocidad del sismo anterior
Comportamientorad	Float8	17,7	Subió o bajó según el radón del sismo anterior
Altura	Float8	17,7	Altura máxima de la emisión (metros sobre la cima).
Dirección	Character varying	(30)	Dirección a la cual se desplaza la columna.
Color	Character varying	(30)	Color de la emisión.
material_gas	Character varying	(5)	Si se observa salida de gas [S, N]
Cenizas	Character varying	(5)	Si se observó salida de ceniza[S,N]
fenomenos_olor	Character varying	(70)	Si se asocia algún tipo de olor a la emisión.
Clase_emision	Character	(1)	Si hay emisión [S, N].
LPS	character varying	10	Numero de eventos LPS clasificables para la fecha dada.
Vt	character varying	10	Numero de eventos VT clasificables para la fecha dada.
Tre	character varying	10	Numero de eventos tipo TREMOR para la fecha dada.
Tor	character varying	10	Numero de eventos tipo TOR para la fecha dada.
Hyb	character varying	10	Numero de eventos HYB clasificables para la fecha dada.
Clase_sismos	character varying	10	Se relacionó algún sismo [S,N]

Tabla 17. Estructura de la tabla TEMIT8729A19. [41].

Formateo de datos:

Implica la realización de transformaciones de los datos, para permitir o facilitar la aplicación de técnicas de Minería de Datos, se elabora los ajustes necesarios de valores de los campos según los límites de las herramientas a usar para la aplicación de algoritmos de minería.

Se discretizarón o categorizaron todos los valores de los atributos de las tablas **TERU17A18 y TERUT8729A18** que contenían valores continuos u homogéneos. (Ver Tabla 18), (Ver Tabla 19), (Ver Tabla 20), (Ver Tabla 21), (Ver Tabla 22), (Ver Tabla 23).

LPS	
Discretización	**# de eventos LPS**
S	Mayor o igual a 1
N	Si es 0

Tabla 18. Discretización atributo LPS. [41].

Vt	
Discretización	**# de eventos vt**
S	Mayor o igual a 1
N	Si es 0

Tabla 19. Discretización atributo vt. [41].

Tre	
Discretización	**# de eventos tre**
S	Mayor o igual a 1
N	Si es 0

Tabla 20. Discretización atributo tre. [41].

Tor	
Discretización	**# de eventos tor**
S	Mayor o igual a 1
N	Si es 0

Tabla 21. Discretización atributo tor [41].

Hyb

Discretización	# de eventos hyb
S	Mayor o igual a 1
N	Si es 0

Tabla 22. Discretización atributo hyb [41].

clase_sismos	
Discretización	**# de eventos LPS, tre, hyb, tor, vt**
S	Mayor o igual a 1
N	Si es 0

Tabla 23. Discretización atributo clase_sismos. [41].

Se discretizarón o categorizaron todos los valores de los atributos de las tablas **TEMI648A19 y TEMIT8729A19** que contenían valores continuos u homogéneos. (Ver Tabla 24), (Ver Tabla 25), (Ver Tabla 26), (Ver Tabla 27), (Ver Tabla 28), (Ver Tabla 29).

LPS	
Discretización	**# de eventos LPS**
S	Mayor o igual a 1
N	Si es 0

Tabla 24. Discretización atributo LPS. [41].

Vt	
Discretización	**# de eventos vt**
S	Mayor o igual a 1
N	Si es 0

Tabla 25. Discretización atributo vt. [41].

Tre	
Discretización	**# de eventos tre**
S	Mayor o igual a 1
N	Si es 0

Tabla 26. Discretización atributo tre. [41].

Tor	
Discretización	**# de eventos tor**

S	Mayor o igual a 1
N	Si es 0

Tabla 27. Discretización atributo tor [41].

Hyb	
Discretización	**# de eventos hyb**
S	Mayor o igual a 1
N	Si es 0

Tabla 28. Discretización atributo hyb. [41].

clase_sismos	
Discretización	**# de eventos LPS, tre, hyb, tor, vt**
S	Mayor o igual a 1
N	Si es 0

Tabla 29. Discretización atributo clase_sismos. [41].

Repositorio final:

Finalmente se cuenta con el repositorio de datos limpio del volcán Galeras, listo para aplicar técnicas de minería de datos y obtener los patrones de eventos eruptivos del volcán Galeras. La base de datos resultante está conformada por una tabla llamada **TEMI648A19,** (Ver Tabla 30), donde se presenta el historial de las emisiones del volcán Galeras unidas con atributos que influyen para una emisión de gases o ceniza. Presenta un total de 648 registros y 19 atributos. Y otra tabla llamada **TERU17A18,** (Ver Tabla 31) donde se presenta el historial de erupciones del volcán Galeras unidas con los mismos atributos anteriormente mencionados, Presenta un total de 17 registros y 18 atributos.

Atributo	Tipo	Tamaño	Descripción
Fecha	Date	13	Fecha de registro del evento

comportamientoton_v	Character varying	10	Subió, estable o bajó la medida de so2 según el anterior
Tipo	Character varying	5	Tipo de evento localizado.
Qm	Character varying	(4)	Calidad de la localización [A,B,C,D] A: epicenter excelent, focal depth good B: epicenter good, focal depth fair C: epicenter fair, focal depth poor D: epicenter poor, focal depth poor
comportamientovel	Character varying	10	Subió, estable o bajó según la Velocidad del sismo anterior
Comportamientorad	Character varying	10	Subió, estable o bajó según el comportamiento del radón anterior
Altura	Character varying	10	Altura máxima de la emisión (metros sobre la cima).
Dirección	Character varying	(30)	Dirección a la cual se desplaza la columna.
Color	Character varying	(30)	Se observó color de la emisión [S, N].
material_gas	Character varying	(5)	Si se observa salida de gas [S, N]
Cenizas	Character varying	(5)	Si se observó salida de ceniza[S,N]
fenomenos_olor	Character varying	(70)	Si se asocia algún tipo de olor a la emisión [S, N].
Clase_emision	Character	(1)	Si hay emisión [S, N].
LPS	character varying	10	Se observó eventos LPS clasificables para la fecha dada. [S,N]
Vt	character varying	10	Se observó eventos VT clasificables para la fecha dada. [S,N]
Tre	character varying	10	Se observó eventos tipo TREMOR para la fecha dada. [S,N]
Tor	character varying	10	Se observó eventos tipo TOR para la fecha dada. [S,N]
Hyb	character varying	10	Se observó eventos HYB para la fecha dada [S,N]
Clase_sismos	character varying	10	Se relacionó algún sismo [S,N]

Tabla 30. Estructura del repositorio final (TEMI648A19). [41].

Atributo	Tipo	Tamaño	Descripción
Fecha	Date	13	Fecha de registro del evento

comportamientoton_v	Character varying	17,7	Subió, estable o bajó la medida de so2 según el anterior
Tipo	Character varying	5	Tipo de evento localizado.
Qm	Character varying	(4)	Calidad de la localización [A,B,C,D] A: epicenter excelent, focal depth good B: epicenter good, focal depth fair C: epicenter fair, focal depth poor D: epicenter poor, focal depth poor
comportamientovel	Character varying	10	Subió, estable o bajó según la Velocidad del sismo anterior
Comportamientorad	Character varying	10	Subió, estable o bajó según el comportamiento del radón anterior
Incandecencia	Character varying	(5),	Si se observa incandescencia [S,N]
altura_max	Double precision,		Altura máxima de la columna
caida_piroclastos	Character varying	(5),	Si se observa caída de piroclastos [S, N].
onda_choque	Character varying	(5),	Si se observa onda de choque [S, N].
proyectil_balistico	Character varying	(5),	Si se registran proyectiles balísticos.
LPS	character varying	10	Se observó eventos LPS clasificables para la fecha dada. [S,N]
Vt	character varying	10	Se observó eventos VT clasificables para la fecha dada. [S,N]
Tre	character varying	10	Se observó eventos tipo TREMOR para la fecha dada. [S,N]
Tor	character varying	10	Se observó eventos tipo TOR para la fecha dada. [S,N]
Hyb	character varying	10	Se observó eventos HYB para la fecha dada [S,N]
Clase_sismos	character varying	10	Se relacionó algún sismo [S,N]
Clase_erupcion	character varying	10	Si hay erupción [S,N]

Tabla 31. Estructura del repositorio final (TERU17A18). [41].

Análisis final de calidad de datos:

Para mayor información (Ver Anexo16)

Tabla TERU17A18 (Ver Tabla 32).

Atributo	% Nulos
Fecha	0.0%
Coportamientoton_v	0.0%
Tipo	5.56%
Qm	11.11%
comportamientovel	0.0%
Comportamientorad	0.0%
Altura_max	77.78%
incandescencia	0.0%
caida_piroclastos	0.0%
Onda_choque	0.0%
Proyectil_balistico	0.0%
Clase_erupcion	0.0%
LPS	0.0%
Vt	0.0%
Hyb	0.0%
Tre	0.0%
Tor	0.0%
Clase_sismo	0.0%

Tabla 32. Análisis de Calidad de datos a la tabla Tabla TERU17A18. [41].

Atributo	% Nulos
Fecha	0.0%
Coportamientoton_v	0.0%
Tipo	15.28%
Qm	16.20%
comportamientovel	0.0%
Comportamientorad	0.0%
Altura	12.96%
Dirección	0.0%
Color	0.0%
Material_gas	0.0%
Cenizas	0.0%
Fenómenos_olor	0.0%

Atributo	% Nulos
Clase_emision	0.0%
LPS	0.0%
Vt	0.0%
Hyb	0.0%
Tre	0.0%
Tor	0.0%
Clase_sismo	0.0%

Tabla 33. Análisis de Calidad de datos a la tabla Tabla TEMI648A19. [41].

3.3 FASE DE MODELADO

En esta fase se seleccionan las técnicas de modelado más apropiadas para el proyecto de Minería de Datos.

Las tareas de minería de datos escogidas para el proceso de descubrimiento de patrones de eventos eruptivos, se seleccionaron las tareas de minería de datos Asociación y Agrupamiento, Sin embargo se aplicó la tarea de Clasificación, pero no se seleccionó debido a que el resultado no era muy significativo.

Se escogió la herramienta Weka (Waikato Environment for Knowledge Analysis) para realizar estas tareas. Weka fue desarrollada en la Universidad de Waikato (Nueva Zelanda) bajo licencia GPL. Esta herramienta permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario. Weka es una de las suites más utilizadas en el área de descubrimiento de conocimiento en los últimos años.

Tarea de clasificación:

La clasificación es el proceso por medio del cual se encuentra propiedades comunes entre un conjunto de objetos de una base de datos y se los cataloga en diferentes clases, de acuerdo al modelo de clasificación. [46]. Tomando como clase los valores de los atributos clase_emision y clase_erupcion de las tablas **TEMIT8729A19 y TERUT8729A18** correspondientemente se realizaron las pruebas.

Parámetros de Evaluación del Modelo de Clasificación:

Para la tarea de clasificación se utilizó la técnica de árboles de decisión. El modelo de clasificación basado en árboles de decisión, es probablemente el más utilizado y popular por su simplicidad y facilidad para entender [17]. La clasificación con árboles de decisión considera clases disjuntas, de forma que el árbol conducirá a

una y solo una hoja, asignando una única clase a la predicción. Para esta tarea, se escogió como clase el atributo clase_emision y clase_erupcion que determina si hay erupción, emisión o no.

Para la poda del árbol se tuvo en cuenta el factor de confianza C (*confidence level*), que influye en el tamaño y capacidad de predicción del árbol construido. El valor por defecto de este factor es del 25% y conforme va bajando este valor, se permiten más operaciones de poda y por lo tanto llegar a árboles cada vez más pequeños [38]. Otro factor que se utilizó para variar el tamaño del árbol fue a través del parámetro M que especifica el número mínimo de instancias o registros por nodo del árbol. [39].

Para evaluar la calidad del modelo, dividiendo el repositorio de datos en dos conjuntos: entrenamiento y prueba, se escogió el método validación cruzada con n pliegues (*n-fold cross validation*). Este método consiste en dividir el conjunto de entrenamiento en *n* subconjuntos disjuntos de similar tamaño llamados pliegues (*folds*) de forma aleatoria. Posteriormente se realizan *n* iteraciones (igual al número de subconjuntos definido), donde en cada una se reserva un subconjunto diferente para el conjunto de prueba y los restantes *n-1* (uniendo todos los datos) para construir el modelo (entrenamiento). En cada iteración se calcula el error de muestra parcial del modelo. Por último se construye el modelo con todos los datos y se obtiene su error promediando los obtenidos anteriormente en cada una de las iteraciones. [40].

En este estudio se utilizó n=10 particiones, que es el valor que comúnmente se usa y que se ha probado que da buenos resultados. [40].

Con el fin de detectar los patrones más confiables de eventos eruptivos, utilizando árboles de decisión, para todos los casos registrados entre 1989 y 2013, se realizaron diferentes pruebas para obtener los mejores parámetros de poda y por ende obtener el árbol con el mayor número y porcentaje de instancias correctamente clasificadas. Para realizar estas pruebas se utilizó la librería *RWeka* del paquete estadístico *R* que contiene el algoritmo J48.

Descubrimiento de Patrones con Árboles de Decisión:

Con el fin de detectar patrones de eventos eruptivos del volcán Galeras utilizando arboles de decisión para los casos registrados desde 1989 y 2013, se realizaron diferentes pruebas para obtener los mejores parámetros y obtener el árbol con el mayor número y porcentaje de instancias correctamente clasificadas.

El modelo de **clasificación** para la tabla **TEMIT8729A19,** se realizó con las siguientes pruebas:

Prueba 1. El número de registros por nodo M fue variable con un incremento de 2 en 2 y el factor confianza C también. El factor M varió en el rango entre M=2 a M=30 y se varió el factor C en el rango entre C=0.1 a C=0.5. De acuerdo a los resultados obtenidos, el árbol construido con los parámetros M=2 y C=0.5 fue el mejor con un porcentaje de confianza de 93.89%, las cuales se puede observar con más detalle (Ver Anexo4), el mejor árbol podado se puede observar en (Ver Figura 15) y se presentan los parámetros de precisión (Ver Figura 16).

Prueba 2. El número de registros por nodo M fue variable con un incremento de 2 en 2 y el factor confianza C también. El factor M varió en el rango entre M=2 a M=20 y se varió el factor C en el rango entre C=0.1 a C=0.5. De acuerdo a los resultados obtenidos, el árbol construido con los parámetros M=2 y C=0.5 fue el mejor con un porcentaje de confianza de 93.89%, las cuales se puede observar con más detalle (Ver Anexo5). el mejor árbol se puede observar en (Ver Figura 17) y se presentan los parámetros de precisión (Ver Figura 18).

En la tabla **TERUT8729A18** se realizó las siguientes pruebas:

Prueba 1. Se mantuvo constante el número de registros por nodo M como también el factor confianza C. El factor M se mantuvo en 1 y el factor C en C=0.99, De acuerdo a los resultados obtenidos, el árbol construido con los parámetros M=1 y C=0.99 las cuales se puede observar con más detalle (Ver Anexo9), el mejor árbol se puede observar en (Ver Figura 19) y se presentan los parámetros de precisión (Ver Figura 20),

Prueba 2. El número de registros por nodo M varió y el factor confianza C se mantuvo constante. El factor M varió entre el rango de M=2 hasta M=20 y el factor C se mantuvo constante C= 0.99, De acuerdo a los resultados obtenidos, el árbol construido con los parámetros M=2 y C=0.99 fue el mejor, con un porcentaje de confianza de 99.81%, las cuales se puede observar con más detalle (Ver Anexo10). el mejor árbol se puede observar en (Ver Figura 21) y se presentan los parámetros de precisión (Ver Figura 22),

Prueba 3. El número de registros por nodo M se mantuvo constante como también el factor confianza C. El factor M se mantuvo constante M=1 y el factor C se mantuvo constante C= 0.99, De acuerdo a los resultados obtenidos, el árbol construido con los parámetros M=1 y C=0.99 fue el mejor, con un porcentaje de confianza de 99.83%, las cuales se puede observar con más detalle (Ver Anexo11). el mejor árbol se puede observar en (Ver Figura 23) y se presentan los parámetros de precisión (Ver Figura 24),

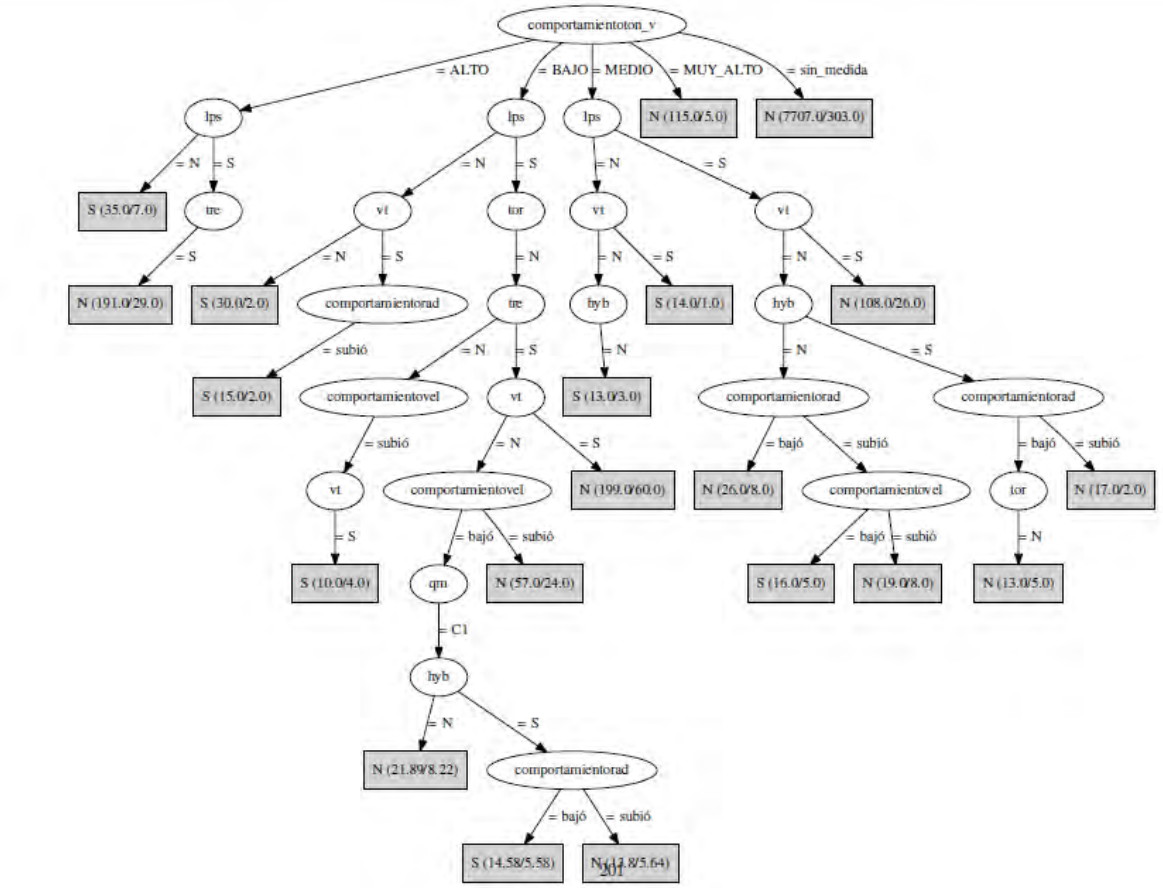

Figura 15. Mejor Árbol obtenido con algoritmo J48 (M = 2 y C = 0.5). Porcentaje de confianza=93.89% [41].

```
Correctly Classified Instances        8196               93.8939 %
Incorrectly Classified Instances       533                6.1061 %
Kappa statistic                          0.3411
Mean absolute error                      0.1045
Root mean squared error                  0.2284
Relative absolute error                 75.992  %
Root relative squared error             87.1299 %
Coverage of cases (0.95 level)          96.4715 %
Mean rel. region size (0.95 level)      55.1266 %
Total Number of Instances             8729
=== Confusion Matrix ===

   a    b   <-- classified as
8044   37 |    a = N
 496  152 |    b = S
```

Figura 16. Parámetros de precisión [41].

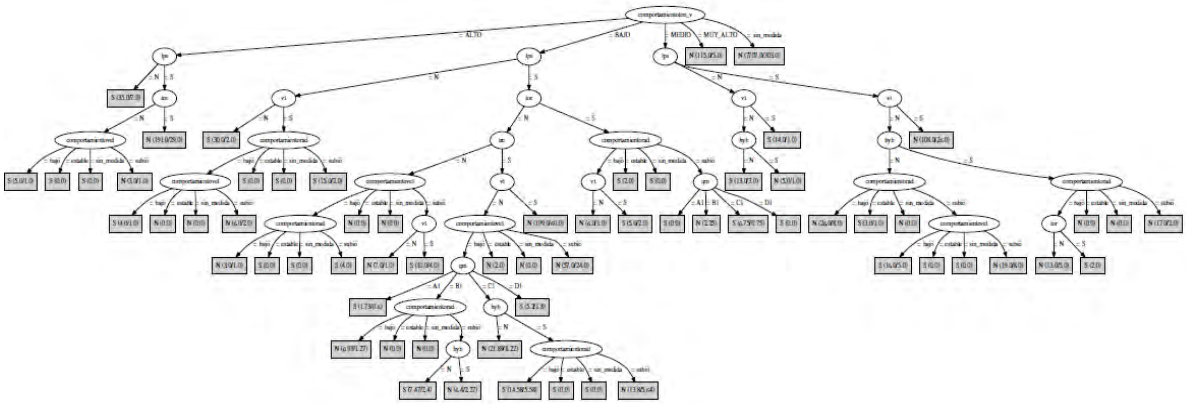

Figura 17. Árbol obtenido con algoritmo J48 (M = 2 y C = 0.5). Porcentaje de confianza=93.89% [41].

```
Correctly Classified Instances        8196               93.8939 %
Incorrectly Classified Instances       533                6.1061 %
Kappa statistic                          0.3411
Mean absolute error                      0.1045
Root mean squared error                  0.2284
Relative absolute error                 75.992  %
Root relative squared error             87.1299 %
Coverage of cases (0.95 level)          96.4715 %
Mean rel. region size (0.95 level)      55.1266 %
Total Number of Instances             8729
=== Confusion Matrix ===
   a    b   <-- classified as
8044   37 |    a = N
 496  152 |    b = S
```

Figura 18. Parámetros de precisión [41].

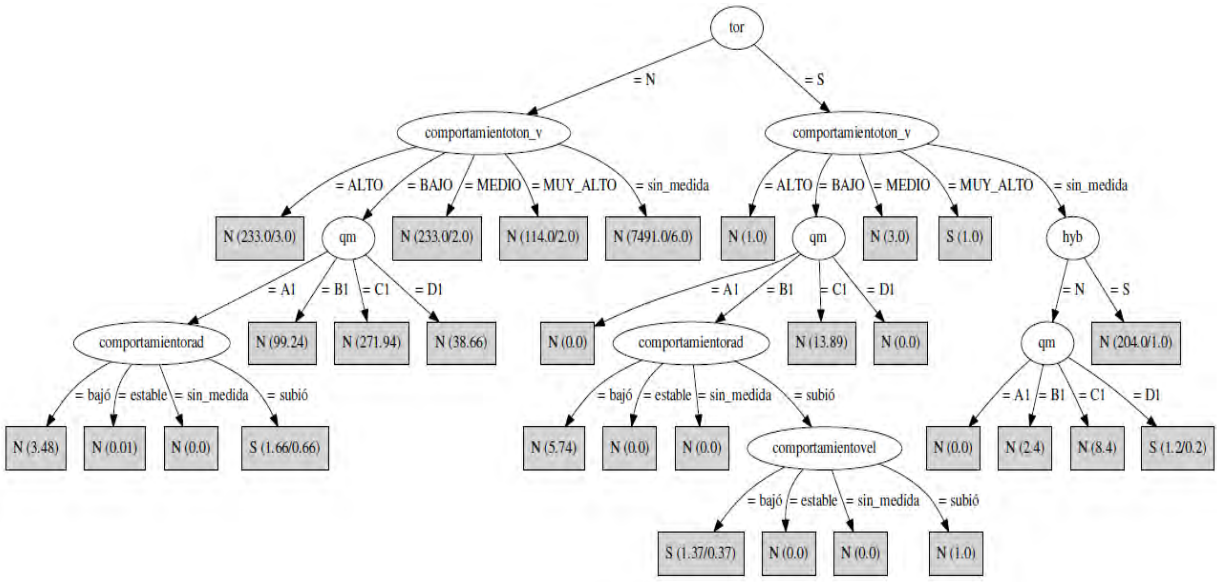

Figura 19. Árbol obtenido con algoritmo J48 (M = 1 y C = 0.99). Porcentaje de confianza=99.84% [41].

```
Correctly Classified Instances        8715               99.8396 %
Incorrectly Classified Instances        14                0.1604 %
Kappa statistic                          0.3632
Mean absolute error                      0.0034
Root mean squared error                  0.0404
Relative absolute error                 79.8125 %
Root relative squared error             89.0567 %
Coverage of cases (0.95 level)          99.8396 %
Mean rel. region size (0.95 level)      50.0344 %
Total Number of Instances             8729

=== Confusion Matrix ===

    a    b   <-- classified as
 8711    0 |    a = N
   14    4 |    b = S
```

Figura 20. Parámetros de precisión [41].

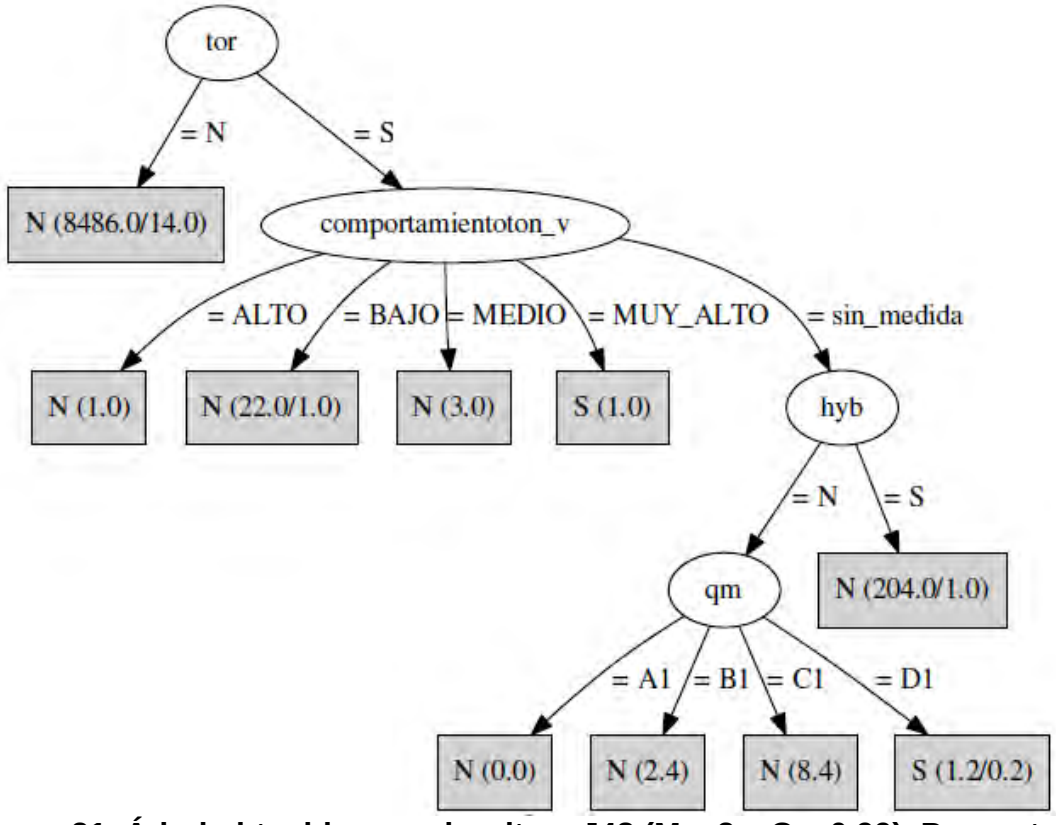

Figura 21. Árbol obtenido con algoritmo J48 (M = 2 y C = 0.99). Porcentaje de confianza=99.81% [41].

```
=== Summary ===

Correctly Classified Instances        8713               99.8167 %
Incorrectly Classified Instances        16                0.1833 %
Kappa statistic                          0.1997
Mean absolute error                      0.0037
Root mean squared error                  0.0428
Relative absolute error                 87.1852 %
Root relative squared error             94.2842 %
Coverage of cases (0.95 level)          99.8167 %
Mean rel. region size (0.95 level)      50.0172 %
Total Number of Instances             8729

=== Confusion Matrix ===

    a    b   <-- classified as
 8711    0 |    a = N
   16    2 |    b = S
```

Figura 22. Parámetros de precisión [41].

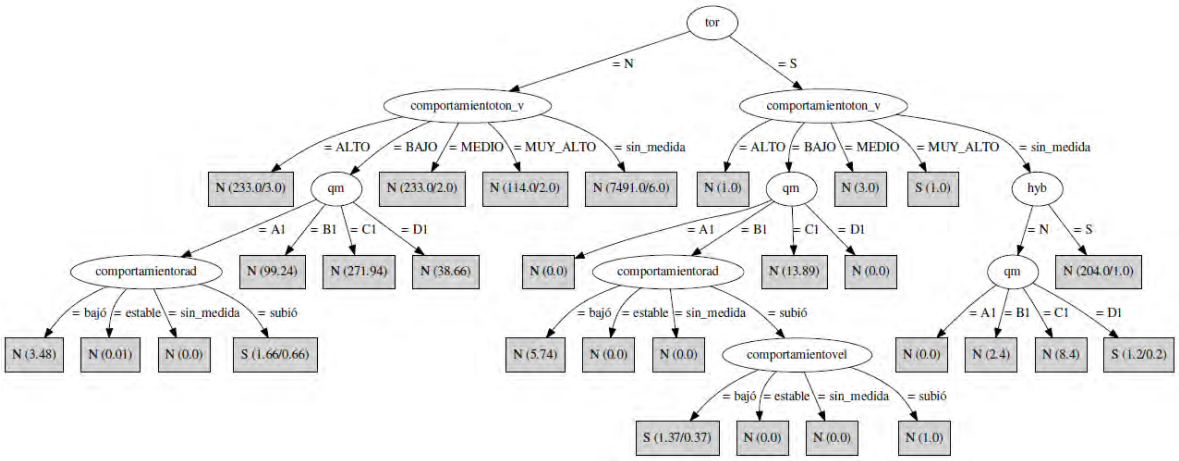

Figura 23. Árbol obtenido con algoritmo J48 (M = 1 y C = 0.99). Porcentaje de confianza=99.84% [41].

```
=== Summary ===

Correctly Classified Instances        8715               99.8396 %
Incorrectly Classified Instances        14                0.1604 %
Kappa statistic                          0.3632
Mean absolute error                      0.0034
Root mean squared error                  0.0404
Relative absolute error                 79.8125 %
Root relative squared error             89.0567 %
Coverage of cases (0.95 level)          99.8396 %
Mean rel. region size (0.95 level)      50.0344 %
Total Number of Instances             8729

=== Confusion Matrix ===

    a    b   <-- classified as
 8711    0 |    a = N
   14    4 |    b = S
```

Figura 24. Parámetros de precisión [41].

3.3.1 Tarea de asociación. La tarea de Asociación descubre patrones en forma de reglas, que muestran los hechos que ocurren frecuentemente juntos en un conjunto de datos determinado. En este problema, se da un conjunto de atributos y una colección de registros de una base de datos. La tarea es encontrar relaciones entre los atributos de esa base de datos para descubrir reglas de asociación que cumplan unas especificaciones mínimas dadas por el usuario, expresadas en forma de soporte y confianza [18] Tomando el conjunto de datos de las erupciones y emisiones que produjo el volcán Galeras, se extrajeron reglas que determinaron ciertas características.

Para la tarea de Asociación se utilizó el algoritmo *Apriori* [18], implementado en Weka en el paquete *WEKA.associations.Apriori*. Para evaluar las regla de asociación resultantes se utilizaron los parámetros soporte y confianza, dos

métricas que permiten conocer la calidad de la regla. El soporte o cobertura de una regla se define como el número de instancias en las que la regla se puede aplicar. La confianza o precisión mide el porcentaje de veces que la regla se cumple cuando se puede aplicar [19].

Descubrimiento de Reglas de Asociación:

Los parámetros de ejecución del algoritmo Apriori con el repositorio TEMI648A19, Se fijó como mínima confianza el 80% (0.8), un soporte mínimo inferior de 0.1 y un numero de reglas a generar de 100. Además se presentan en (Ver Figura 25) y las primeras 39 reglas generadas de 100, con una confianza del 100%, en (Ver Figura 26). Todos los resultados de esta prueba se muestran en (Ver Anexo6).

```
Apriori
=======

Minimum support: 0.3 (163 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 14

Generated sets of large itemsets:

Size of set of large itemsets L(1): 14

Size of set of large itemsets L(2): 46

Size of set of large itemsets L(3): 42

Size of set of large itemsets L(4): 13
```

Figura 25. Parámetros de ejecución del algoritmo Apriori con el repositorio TEMI648A19 [41].

```
Best rules found:

 1. tre=S 452 ==> clase_emision=S 452    conf:(1)
 2. tre=S tor=N 427 ==> clase_emision=S 427    conf:(1)
 3. lps=S 416 ==> clase_emision=S 416    conf:(1)
 4. lps=S tor=N 387 ==> clase_emision=S 387    conf:(1)
 5. lps=S tre=S 371 ==> clase_emision=S 371    conf:(1)
 6. qm=C1 366 ==> clase_emision=S 366    conf:(1)
 7. qm=C1 tor=N 349 ==> clase_emision=S 349    conf:(1)
 8. lps=S tre=S tor=N 346 ==> clase_emision=S 346    conf:(1)
 9. hyb=S 345 ==> clase_emision=S 345    conf:(1)
10. vt=S 329 ==> clase_emision=S 329    conf:(1)
11. tor=N hyb=S 317 ==> clase_emision=S 317    conf:(1)
12. vt=S tor=N 313 ==> clase_emision=S 313    conf:(1)
13. qm=C1 tre=S 298 ==> clase_emision=S 298    conf:(1)
14. lps=S hyb=S 285 ==> clase_emision=S 285    conf:(1)
15. tre=S hyb=S 285 ==> clase_emision=S 285    conf:(1)
16. qm=C1 tre=S tor=N 283 ==> clase_emision=S 283    conf:(1)
17. comportamientorad=subió 277 ==> clase_emision=S 277    conf:(1)
18. comportamientovel=bajó 274 ==> clase_emision=S 274    conf:(1)
19. qm=C1 lps=S 272 ==> clase_emision=S 272    conf:(1)
20. comportamientovel=subió 265 ==> clase_emision=S 265    conf:(1)
21. comportamientorad=subió tor=N 264 ==> clase_emision=S 264    conf:(1)
22. comportamientoton_v=sin_medida 261 ==> clase_emision=S 261    conf:(1)
23. vt=S tre=S 261 ==> clase_emision=S 261    conf:(1)
24. tre=S tor=N hyb=S 261 ==> clase_emision=S 261    conf:(1)
25. comportamientovel=bajó tor=N 257 ==> clase_emision=S 257    conf:(1)
26. lps=S tor=N hyb=S 257 ==> clase_emision=S 257    conf:(1)
27. qm=C1 lps=S tor=N 255 ==> clase_emision=S 255    conf:(1)
28. comportamientorad=bajó 254 ==> clase_emision=S 254    conf:(1)
29. comportamientovel=subió tor=N 253 ==> clase_emision=S 253    conf:(1)
30. comportamientoton_v=sin_medida tor=N 248 ==> clase_emision=S 248    conf:(1)
31. vt=S tre=S tor=N 247 ==> clase_emision=S 247    conf:(1)
32. lps=S tre=S hyb=S 246 ==> clase_emision=S 246    conf:(1)
33. lps=S vt=S 245 ==> clase_emision=S 245    conf:(1)
34. qm=C1 hyb=S 243 ==> clase_emision=S 243    conf:(1)
35. comportamientorad=bajó tor=N 240 ==> clase_emision=S 240    conf:(1)
36. qm=C1 lps=S tre=S 240 ==> clase_emision=S 240    conf:(1)
37. comportamientovel=bajó tre=S 239 ==> clase_emision=S 239    conf:(1)
38. comportamientorad=subió tre=S 229 ==> clase_emision=S 229    conf:(1)
39. lps=S vt=S tor=N 229 ==> clase_emision=S 229    conf:(1)
```

Figura 26. Mejores reglas generadas con Apriori con el conjunto de datos TEMI648A19 [41].

Los parámetros de ejecución del algoritmo Apriori con el repositorio **TERUT17A18**, Se fijó como mínima confianza el 80% (0.8), un soporte mínimo inferior de 0.1 y un numero de reglas a generar de 100. Además se presentan en (Ver Figura 27) y las primeras 32 reglas generadas de 100, con una confianza del 100%, en (Ver Figura 28). Todos los resultados de esta prueba se muestran en (Ver Anexo7).

```
Apriori
=======

Minimum support: 0.35 (5 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 13

Generated sets of large itemsets:

Size of set of large itemsets L(1): 11

Size of set of large itemsets L(2): 35

Size of set of large itemsets L(3): 52

Size of set of large itemsets L(4): 40

Size of set of large itemsets L(5): 15

Size of set of large itemsets L(6): 2
```

Figura 27. Parámetros de ejecución del algoritmo Apriori con el repositorio TERUT17A18 [41].

```
Best rules found:

 1. vt=S 14 ==> clase_erupcion=S 14    conf:(1)
 2. lps=S vt=S 14 ==> clase_erupcion=S 14    conf:(1)
 3. vt=S tre=S 14 ==> clase_erupcion=S 14    conf:(1)
 4. lps=S vt=S tre=S 14 ==> clase_erupcion=S 14    conf:(1)
 5. hyb=S 13 ==> clase_erupcion=S 13    conf:(1)
 6. lps=S hyb=S 13 ==> clase_erupcion=S 13    conf:(1)
 7. hyb=S tre=S 13 ==> clase_erupcion=S 13    conf:(1)
 8. lps=S hyb=S tre=S 13 ==> clase_erupcion=S 13    conf:(1)
 9. vt=S hyb=S 12 ==> clase_erupcion=S 12    conf:(1)
10. lps=S vt=S hyb=S 12 ==> clase_erupcion=S 12    conf:(1)
11. vt=S hyb=S tre=S 12 ==> clase_erupcion=S 12    conf:(1)
12. lps=S vt=S hyb=S tre=S 12 ==> clase_erupcion=S 12    conf:(1)
13. comportamientorad=subió 11 ==> clase_erupcion=S 11    conf:(1)
14. tor=N 11 ==> clase_erupcion=S 11    conf:(1)
15. comportamientorad=subió lps=S 11 ==> clase_erupcion=S 11    conf:(1)
16. comportamientorad=subió tre=S 11 ==> clase_erupcion=S 11    conf:(1)
17. lps=S tor=N 11 ==> clase_erupcion=S 11    conf:(1)
18. tre=S tor=N 11 ==> clase_erupcion=S 11    conf:(1)
19. comportamientorad=subió lps=S tre=S 11 ==> clase_erupcion=S 11    conf:(1)
20. lps=S tre=S tor=N 11 ==> clase_erupcion=S 11    conf:(1)
21. comportamientorad=subió vt=S 10 ==> clase_erupcion=S 10    conf:(1)
22. comportamientorad=subió hyb=S 10 ==> clase_erupcion=S 10    conf:(1)
23. vt=S tor=N 10 ==> clase_erupcion=S 10    conf:(1)
24. hyb=S tor=N 10 ==> clase_erupcion=S 10    conf:(1)
25. comportamientorad=subió lps=S vt=S 10 ==> clase_erupcion=S 10    conf:(1)
26. comportamientorad=subió lps=S hyb=S 10 ==> clase_erupcion=S 10    conf:(1)
27. comportamientorad=subió vt=S tre=S 10 ==> clase_erupcion=S 10    conf:(1)
28. comportamientorad=subió hyb=S tre=S 10 ==> clase_erupcion=S 10    conf:(1)
29. lps=S vt=S tor=N 10 ==> clase_erupcion=S 10    conf:(1)
30. lps=S hyb=S tor=N 10 ==> clase_erupcion=S 10    conf:(1)
31. vt=S tre=S tor=N 10 ==> clase_erupcion=S 10    conf:(1)
32. hyb=S tre=S tor=N 10 ==> clase_erupcion=S 10    conf:(1)
```

Figura 28. Mejores reglas generadas con Apriori con el conjunto de datos TERUT17A18 [41].

3.3.2 Tarea de agrupación. El *clustering* o segmentación es una de las tareas más frecuentes en la minería de datos. A diferencia de la tarea de clasificación, en el *clustering* no se conoce la clase a la cual pertenecen los datos, por esa razón se dice que es un proceso de aprendizaje no supervisado. En esta tarea se trata de encontrar grupos similares entre un conjunto de datos basado en el concepto de distancia [18].

El algoritmo más popular de los métodos de agrupamiento particionales es el *K-means [9]*. La idea de este algoritmo es tomar como entrada el parámetro *k* y particionar el conjunto de datos en *k* grupos o *clúster*, donde los casos o instancias pertenecientes al mismo grupo tengan características similares. Todo caso nuevo es comparado con los grupos formados y asociado a aquél que sea el más próximo, en los términos de una distancia que normalmente es la Euclidiana [47].

Para la tarea de agrupación se utilizó la técnica particional con el algoritmo *K-means* [20], implementado en Weka, como *SimpleKmeans*, en el cual se configura el número de grupos (*NumClusters*) a formar y la semilla (*seed*), que se utiliza en la generación de un número aleatorio, el cual es usado para hacer la asignación inicial de instancias a los grupos. Para evaluar los resultados del agrupamiento, se utilizó el propio conjunto de entrenamiento, (*Use tranining set)*, que indica que porcentaje de instancias se van a cada grupo.

Análisis de Datos con Agrupamiento o Clustering:

Con el fin de generar grupos similares entre los registros de datos **TEMI648A19 y TERU17A18** en los cuales se encuentran emisiones y erupciones del Volcán Galeras respectivamente, se configuró el parámetro K del algoritmo K-means en 2 y 4 con una semilla por defecto de 10.

Para encontrar grupos que relacionen los factores que influyen antes de una emisión del volcán Galeras se utilizó el repositorio **TEMI648A19** se configuró el número de clústers en 2, (K = 2), y el resultado se puede observar en (Ver Tabla 34) para mayor detalle se observa (Ver Anexo12).

Para encontrar grupos que relacionen los factores que influyen antes de una emisión del volcán Galeras se utilizó el repositorio **TEMI648A19** se configuró el número de clústers en 4, (K = 4), y el resultado se puede observar en (Ver Tabla 35) para mayor detalle se observa (Ver Anexo13).

Para encontrar grupos que relacionen los factores que influyen antes de una erupción del volcán Galeras se utilizó el repositorio **TERU17A18** se configuró el número de clústers en 2, (K = 2), y el resultado se puede observar en (Ver Tabla 36) para mayor detalle se observa (Ver Anexo14).
Para encontrar grupos que relacionen los factores que influyen antes de una emisión del volcán Galeras se utilizó el repositorio **TERU17A18** se configuró el

número de clústers en 4, (K = 4), y el resultado se puede observar en (Ver Tabla 37) para mayor detalle se observa (Ver Anexo15).

	Full_Data	0	1
	(648)	(434)	(214)
comportamientoton_v	sin_medida	sin_medida	sin_medida
qm	C1	C1	C1
comportamientovel	bajó	bajó	subió
comportamientorad	subió	subió	bajó
lps	S	S	N
vt	S	S	S
tre	S	S	S
tor	N	N	N
hyb	S	S	S
clase_emision	S	S	S

Tabla 34. Clústers resultantes con K=2 con el repositorio TEMI648A19. [41].

	Full_Data	0	1	2	3
	(648)	(327)	(136)	(60)	(125)
comportamientoton_v	sin_medida	sin_medida	sin_medida	BAJO	BAJO
qm	C1	C1	C1	C1	C1
comportamientovel	bajó	bajó	subió	subió	bajó
comportamientorad	subió	subió	bajó	bajó	subió
lps	S	S	N	S	N
vt	S	S	S	S	N
tre	S	S	S	N	S
tor	N	N	N	N	N
hyb	S	S	S	S	N
clase_emision	S	S	S	S	S

Tabla 35. Clústers resultantes con K=4 con el repositorio TEMI648A19. [41].

	Full_Data	0	1
	(17)	(15)	(2)
comportamientoton_v	sin_medida	sin_medida	BAJO
qm	C1	C1	A1
comportamientovel	bajó	bajó	bajó
comportamientorad	subió	subió	subió
lps	S	S	S
vt	S	S	S
hyb	S	S	S
tre	S	S	S
tor	N	N	N
clase_erupcion	S	S	S

Tabla 36. Clústers resultantes con K=2 con el repositorio TERU17A18. [41].

	Full_Data	0	1	2	3
	(17)	(11)	(2)	(2)	(2)
comportamientoton_v	sin_medida	sin_medida	BAJO	MEDIO	ALTO
qm	C1	C1	A1	B1	B1
comportamientovel	bajó	bajó	bajó	bajó	bajó
comportamientorad	subió	subió	subió	subió	bajó
lps	S	S	S	S	S
vt	S	S	S	S	S
hyb	S	S	S	S	S
tre	S	S	S	S	S
tor	N	N	N	N	N
clase_erupcion	S	S	S	S	S

Tabla 37. Clústers resultantes con K=4 con el repositorio TERU17A18. [41].

3.4 FASE DE INTERPRETACIÓN Y EVALUACIÓN DE RESULTADOS

En esta fase, se mira si el modelo se ajusta a las necesidades establecidas en el proyecto. Se evaluarán los patrones que se descubran con el fin de determinar su validez, remover los patrones redundantes o irrelevantes y traducir los patrones útiles en términos que sean entendibles para el usuario, se interpretan los patrones descubiertos con el fin de consolidar el conocimiento descubierto e incorporarlo en otro sistema para posteriores acciones o para confrontarlo con conocimiento previamente descubierto.

3.4.1 Análisis de los resultados para los eventos de volcán galeras tipo emisión. En esta sección se realiza una evaluación e interpretación de los resultados obtenidos con los datos de las emisiones del volcán Galeras en el periodo comprendido entre 1989 hasta el año 2013 almacenados en los repositorios **TEMIT648A19 y TEMIT8729A19,** aplicando las tareas de minería de datos clasificación, asociación y agrupamiento.

Modelos de Clasificación:

Analizando los resultados de las dos pruebas de clasificación realizadas con el conjunto de datos **TEMIT8729A19**, en el cual se almacenan los datos sobre los factores de las emisiones del volcán Galeras que sucedieron en el periodo comprendido entre el año 1989 y el año 2013, donde se escogió el atributo *clase_emision* como clase, se puede observar que el árbol de decisión resultante de la prueba 1 es el mejor (ver Figura 15), con 8160 instancias correctamente clasificados, que corresponde a un porcentaje de precisión del 93.48%, y 569 instancias incorrectamente clasificadas, correspondiente a un porcentaje de error del 6,52%. Las instancias mejor clasificadas son aquellas cuando la clase_emision es N, que significa que no hay emisiones. El estadístico Kappa, que mide la

coincidencia de la predicción con la clase real de este modelo es de 0.2721, que se considera inaceptable (1.0 significa que ha habido coincidencia absoluta).

Por las anteriores consideraciones, podemos concluir que la tarea de clasificación **no** es apropiada para este caso.

Modelos de Asociación:

Se utilizó el conjunto de datos **TEMIT648A19** para obtener reglas de asociación que relacionen los factores antes de una emisión del volcán Galeras de 648 emisiones clasificadas en el OVSP en el periodo 1989 al 2013.

Las 100 reglas de asociación generadas (ver Anexo6) tienen una confianza del 100% y un mínimo soporte de 10%, lo que las convierte en reglas fuertes, por lo tanto interesantes, significativas y con una alta precisión.

Entre las reglas de asociación más representativas están:

- El 100% de las emisiones son con calidad de localización del sismo regular, hay registro de movimientos de fluidos tipo LPS de 5 a 10 segundos y hay movimientos de fluidos tipo tremor de larga duración. El 10% de las emisiones cumple con este patrón.
- El 100% de las emisiones registra movimientos de fluidos tipo LPS de 5 a 10 segundos, hay registro de una energía sísmica tipo tremor y hay un sismo tipo hibrido. El 10% de las emisiones cumple con este patrón.
- El 100% de las emisiones el comportamiento de la velocidad de sismo subió, hay un movimiento de fluidos tipo tremor de larga duración y no hay un sismo tipo tornillo en el mismo día. El 10% de las emisiones cumple con este patrón.
- El 100% de las emisiones el comportamiento del So2 subió de sismo subió, hay un movimiento de fluidos tipo LPS de 5 a 10 segundos y no hay un sismo tipo tornillo. El 10% de las emisiones cumple con este patrón.
- En el 100% de las emisiones hay un movimiento de fluidos tipo LPS de 5 a 10 segundos, hay un movimiento de fluidos tipo tremor de larga duración, no hay un sismo tipo tornillo y hay un sismo tipo hibrido. El 10% de las emisiones cumple con este patrón.
- En el 100% de las emisiones hay un movimiento de fluidos tipo LPS de 5 a 10 segundos de duración, hay una energía sísmica tipo Volcanotectónica y hay un movimiento de fluidos tipo tremor de larga duración. El 10% de las emisiones cumple con este patrón.
- En el 100% de las emisiones la calidad de localización del sismo es regular, hay un movimiento de fluidos tipo tremor de larga duración y hay un sismo tipo hibrido. El 10% de las emisiones cumple con este patrón.

- En el 100% de las emisiones la calidad de localización del sismo es regular, el comportamiento de la velocidad del sismo bajó y no hay un sismo tipo Tornillo en el mismo día. El 10% de las emisiones cumple con este patrón.
- En el 100% de las emisiones hay un movimiento de fluidos tipo LPS de duración de 5 a 10 segundos entonces, no hay un sismo tipo tornillo y hay un sismo tipo hibrido. El 10% de las emisiones cumple con este patrón.
- En el 100% de las emisiones el comportamiento del radón subió, hay un movimiento de fluidos tipo LPS de duración de 5 a 10 segundos y hay un movimiento de fluidos tipo tremor de larga duración. El 10% de las emisiones cumple con este patrón.

Modelos de Agrupación o Clustering:

Se utilizó el conjunto de datos **TEMIT648A19,** para aplicarle la técnica de clustering, con el fin de encontrar similitudes entre las emisiones del volcán Galeras en el periodo de 1898 y 2013, formando grupos similares que relacionen los factores que influyen en una emisión.

Como se puede observar en la (Ver Tabla 33), se formaron dos clústers (K = 2). En el clúster 0 se encuentran 434 emisiones y en el clúster 1 se encuentran 214 emisiones. Se agrupan así debido a los factores o atributos que influyen antes de una emisión del volcán Galeras, estos factores son: el comportamiento de la velocidad sísmica, comportamiento del radón y si existe un sismo tipo LPS el cual es un movimiento de fluidos con una duración de 5 a 10 segundos.

De acuerdo a las características o atributos que diferencian al clúster 0 del clúster 1, se pueden obtener los siguientes patrones:

Clúster 1: Porcentaje De Representatividad 67%

El 67% de las emisiones de volcán Galeras en el periodo comprendido del año 1989 hasta 2013, son aquellas cuya calidad de localización del sismo es regular tipo C1, el comportamiento de la velocidad del sismo bajó, el comportamiento del radón subió, existe un registro de movimientos de fluidos tipo LPS con una duración de 5 a 10 segundos, existen sismos Volcanotectónicos, existen movimientos de fluidos de larga duración catalogados como tremor, no existen eventos sísmicos tipo tornillo ocurridos en el mismo día de una emisión y existe registro de movimientos de fluidos y movimientos volcanotectónicos a la vez llamado Hibrido.

Clúster 2: Porcentaje De Representatividad 33%

El 33% de las emisiones de volcán Galeras en el periodo comprendido del año 1989 hasta 2013, son aquellas cuya calidad de localización del sismo es regular

tipo C1, el comportamiento de la velocidad del sismo subió, el comportamiento del radón bajó, no existe un registro de movimientos de fluidos tipo LPS con una duración de 5 a 10 segundos, existen sismos Volcanotectónicos, existen movimientos de fluidos de larga duración catalogados como tremor, no existen eventos sísmicos tipo tornillo ocurridos en el mismo día de una emisión y existe registro de movimientos de fluidos y movimientos volcanotectónicos a la vez llamado Hibrido.

Para encontrar grupos similares que relacionen los factores antes de una emisión como se puede observar en la (Ver Tabla 34), se formaron cuatro clústers (K = 4). En el clúster 0 se encuentran 327 emisiones, en el clúster 1 se encuentran 136 emisiones, en el clúster 2 se encuentran 60 emisiones y en el clúster 3 se encuentran 125 emisiones. Se agrupan así debido a los factores o atributos que influyen antes de una emisión del volcán Galeras, estos factores son: el comportamiento de la velocidad sísmica, comportamiento del radón, si existe un sismo tipo LPS el cual es un movimiento de fluidos con una duración de 5 a 10 segundos, si existe un sismo tipo volcanotectónico, si existe un sismo tipo tremor el cual es causado por un movimiento de fluidos y tiene una duración de minutos, días o semanas y si existe un sismo tipo hibrido el cual es causado por movimiento de fluidos y sismo tipo volcanotectónicos.

Teniendo en cuenta que el número de casos o soporte en cada clúster es superior al 10%, los patrones más representativos son:

Clúster 1: Porcentaje De Representatividad 51%

El 51% de las emisiones de volcán Galeras en el periodo comprendido del año 1989 hasta 2013, son aquellas cuya calidad de localización del sismo es regular tipo C1, el comportamiento de la velocidad del sismo bajó, el comportamiento del radón subió, existe un registro de movimientos de fluidos tipo LPS con una duración de 5 a 10 segundos, existen sismos Volcanotectónicos, existen movimientos de fluidos de larga duración catalogados como tremor, no existen eventos sísmicos tipo tornillo ocurridos en el mismo día de una emisión y existe registro de movimientos de fluidos y movimientos volcanotectónicos a la vez llamado Hibrido.

Clúster 2: Porcentaje De Representatividad 21%

El 21% de las emisiones de volcán Galeras en el periodo comprendido del año 1989 hasta 2013, son aquellas cuya calidad de localización del sismo es regular tipo C1, el comportamiento de la velocidad del sismo subió, el comportamiento del radón bajó, no existe un registro de movimientos de fluidos tipo LPS con una duración de 5 a 10 segundos, existen sismos Volcanotectónicos, existen movimientos de fluidos de larga duración catalogados como tremor, no existen eventos sísmicos tipo tornillo ocurridos en el mismo día de una emisión y existe

registro de movimientos de fluidos y movimientos volcanotectónicos a la vez llamado Hibrido.

Clúster 4: Porcentaje De Representatividad 19%

El 19% de las emisiones de volcán Galeras en el periodo comprendido del año 1989 hasta 2013, son aquellas cuyo comportamiento de So2 medido en toneladas bajó, la calidad de localización del sismo es regular tipo C1, el comportamiento de la velocidad del sismo bajó, el comportamiento del radón subió, no existe un registro de movimientos de fluidos tipo LPS con una duración de 5 a 10 segundos, no existen sismos Volcanotectónicos, existen movimientos de fluidos de larga duración catalogados como tremor, no existen eventos sísmicos tipo tornillo ocurridos en el mismo día de una emisión y no existe registro de movimientos de fluidos y movimientos volcanotectónicos a la vez llamado Hibrido.

3.4.2 Análisis de los resultados para los eventos de volcán galeras tipo erupción. En esta sección se realiza una evaluación e interpretación de los resultados obtenidos con los datos de las erupciones del volcán Galeras en el periodo comprendido entre 1989 hasta el año 2013 almacenados en los repositorios **TERU17A18 y TERUT8729A18,** aplicando las tareas de minería de datos clasificación, asociación y agrupamiento.

Modelos de Clasificación:

Analizando los resultados de las tres pruebas de clasificación realizadas con el conjunto de datos **TERUT8729A18**, en el cual se almacenan los datos sobre los factores de las erupciones del volcán Galeras que sucedieron en el periodo comprendido entre los años de 1989 al 2013, donde se escogió el atributo *clase_erupcion* como clase, se puede observar que el árbol de decisión resultante de la prueba 1 es el mejor (ver Figura 19), con 8715 instancias correctamente clasificados, que corresponde a un porcentaje de precisión del 99.84%, y 14 instancias incorrectamente clasificadas, correspondiente a un porcentaje de error del 0,16%. Las instancias mejor clasificadas son aquellas cuando la clase_erupcion es N, que significa que no hay erupciones. El estadístico Kappa, que mide la coincidencia de la predicción con la clase real de este modelo es de 0.3632, que se considera inaceptable (1.0 significa que ha habido coincidencia absoluta).

Por las anteriores consideraciones, podemos concluir que la tarea de clasificación **no** es apropiada para este caso

Modelos de Asociación:

Se utilizó el conjunto de datos **TERU17A18** para obtener reglas de asociación que relacionen los factores antes de una erupción del volcán Galeras de 17 eventos catalogados por el OVSP como erupciones en el periodo comprendido entre el año de 1989 al 2013.

Las 100 reglas de asociación generadas (Ver Anexo7) tienen una confianza del 100% y un mínimo soporte de 10%, lo que las convierte en reglas fuertes, por lo tanto interesantes, significativas y con una alta precisión.

Entre las reglas de asociación más representativas están:

En el 100% de las erupciones del volcán Galeras el comportamiento de la velocidad del sismo subió, existe un registro de movimientos de fluidos de 5 a 10 segundos llamado LPS, existen sismos tipo Volcanotectónicos, existen movimientos de fluidos de larga duración que pueden durar minutos, días o semanas llamados Tremores y no existen eventos sísmicos tipo tornillo en el mismo día de una erupción del volcán Galeras. El 10% de las erupciones cumple este patrón.

En el 100% de las erupciones del volcán Galeras el comportamiento de la velocidad del sismo subió, existe un registro de movimientos de fluidos tipo LPS el cual tiene una duración de 5 a 10 segundos, existen sismos provocados por movimientos de fluidos y volcanotectónicos a la vez llamados híbridos, existen movimientos de fluidos llamados tremores, los cuales pueden tener una duración de minutos, días o semanas y no existan eventos sísmicos tipo tornillo en el mismo día de una erupción. El 10% de las erupciones cumple este patrón.

En el 100% de las erupciones del volcán Galeras el comportamiento del radón subió, existe un registro de movimientos de fluidos tipo LPS el cual puede tener una duración de 5 a 10 segundos, existen sismos provocados por movimientos de fluidos y volcanotectónicos a la vez llamados híbridos. El 10% de las erupciones cumple este patrón.

En el 100% de las erupciones del volcán Galeras el comportamiento del radón subió, existe un registro de movimientos de fluidos tipo LPS, el cual tiene una duración de 5 a 10 segundos, existen sismos tipo volcanotectónicos. El 10% de las erupciones cumple este patrón.

En el 100% de las erupciones del volcán Galeras el comportamiento del radón subió, existe registro de movimientos de fluidos tipo tremor, el cual puede tener una duración de minutos, horas, días y semanas, existen sismos producidos por movimiento de fluidos y volcanotectónicos a la vez llamados híbridos. El 10% de las erupciones cumple este patrón.

En el 100% de las erupciones del volcán Galeras el comportamiento del radón subió, existe registro de movimientos de fluidos tipo tremor, el cual puede durar minutos, horas, días y semanas, existen sismos tipo volcanotectónico. El 10% de las erupciones cumple este patrón.

En el 100% de las erupciones del volcán Galeras el comportamiento del radón subió, existe registro de movimientos de fluidos tipo tremor, el cual puede durar minutos, horas, días y semanas, existe sismos tipo LPS el cual puede durar de 5 a 10 segundos. El 10% de las erupciones cumple este patrón.

Modelos de Agrupación o Clustering:

Se utilizó el conjunto de datos **TERU17A18,** para aplicarle la técnica de clustering, con el fin de encontrar similitudes entre las erupciones del volcán Galeras en el periodo comprendido entre el año 1898 y 2013, formando grupos similares que relacionen los factores que influyen en una erupción.

Como se puede observar en (Ver Tabla 35), se formaron dos clústers (K = 2). En el clúster 0 se encuentran 15 erupciones y en el clúster 1 se encuentran 2 erupciones. Se agrupan así debido a los factores o atributos que influyen antes de una erupción del volcán Galeras, estos factores son: el tipo de localización del sismo.

Teniendo en cuenta que el número de casos o soporte en cada clúster sea superior al 10% y de acuerdo a las características o atributos que diferencian al clúster 0 del clúster 1, se pueden obtener los siguientes patrones:

Clúster 1. Porcentaje de representatividad 88%

El 67% de las erupciones de volcán Galeras en el periodo comprendido del año 1989 hasta 2013, son aquellas cuya calidad de localización del sismo es regular, el comportamiento de la velocidad del sismo bajó, el comportamiento del radón subió, existen registros de movimientos de fluidos tipo LPS, los cuales tienen una duración de 5 a 10 segundos, existen sismos tipo Volcanotectónicos, existen movimientos de fluidos catalogados como sismos tipo tremor, los cuales tienen una duración de minutos, días o semanas, existen eventos sísmicos tipo Hibrido los cuales son sismos provocados por movimiento de fluidos y volcanotectónicos a las vez y no existen sismos tipo tornillo en el mismo día de una erupción.

Clúster 2. Porcentaje de representatividad 12%

El 12% de las erupciones del volcán Galeras en el periodo comprendido del año 1989 hasta 2013, son aquellas cuyo comportamiento del gas So2 medido en toneladas bajó, la calidad de localización del sismo es alto, el comportamiento de la velocidad del sismo bajó, el comportamiento del radón subió, existen registros de movimientos de fluidos tipo LPS, los cuales tienen una duración de 5 a 10 segundos, existen sismos tipo Volcanotectónicos, existen movimientos de fluidos catalogados como sismos tipo tremor, los cuales tienen una duración de minutos, días o semanas, existen eventos sísmicos tipo Hibrido los cuales son sismos provocados por movimiento de fluidos y volcanotectónicos a las vez y no existen sismos tipo tornillo en el mismo día de una erupción.

Para encontrar grupos similares que relacionen los factores antes de una erupción como se puede observar en (Ver Tabla 36), se formaron cuatro clústers (K = 4). En el clúster 0 se encuentran 11 erupciones, en el clúster 1 se encuentran 2 erupciones, en el clúster 2 se encuentran 2 erupciones y en el clúster 3 se encuentran 2 erupciones. Se agrupan así debido a los factores o atributos que influyen antes de una erupción del volcán Galeras, estos factores son: el comportamiento del gas So2, el comportamiento del radón, y la calidad de la localización del sismo.

Teniendo en cuenta que los clúster 1 y clúster 2 de K = 4, son iguales a los clúster 1 y 2 de K = 2 y que el número de casos o soporte en cada clúster sea superior al 10%, los patrones más representativos son:

Clúster 1. Porcentaje de representatividad 64%

El 64% de las erupciones de volcán Galeras en el periodo comprendido del año 1989 hasta 2013, son aquellas cuya calidad de localización del sismo es regular, el comportamiento de la velocidad del sismo bajó, el comportamiento del radón subió, existen registros de movimientos de fluidos tipo LPS, los cuales tienen una duración de 5 a 10 segundos, existen sismos tipo Volcanotectónicos, existen movimientos de fluidos catalogados como sismos tipo tremor, los cuales tienen una duración de minutos, días o semanas, existen eventos sísmicos tipo Hibrido los cuales son sismos provocados por movimiento de fluidos y volcanotectónicos a las vez y no existen sismos tipo tornillo en el mismo día de una erupción.

Clúster 2. Porcentaje de representatividad 12%

El 12% de las erupciones del volcán Galeras en el periodo comprendido del año 1989 hasta 2013, son aquellas cuyo comportamiento del gas So2 medido en toneladas bajó, la calidad de localización del sismo es alto, el comportamiento de la velocidad del sismo bajó, el comportamiento del radón subió, existen registros de movimientos de fluidos tipo LPS, los cuales tienen una duración de 5 a 10

segundos, existen sismos tipo Volcanotectónicos, existen movimientos de fluidos catalogados como sismos tipo tremor, los cuales tienen una duración de minutos, días o semanas, existen eventos sísmicos tipo Hibrido los cuales son sismos provocados por movimiento de fluidos y volcanotectónicos a las vez y no existen sismos tipo tornillo en el mismo día de una erupción.

Clúster 3. Porcentaje de representatividad 12%

El 12% de las erupciones del volcán Galeras en el periodo comprendido del año 1989 hasta 2013, son aquellas cuyo comportamiento del gas So2 medido en toneladas se mantuvo estable, la calidad de localización del sismo es buena, el comportamiento de la velocidad del sismo bajó, el comportamiento del radón subió, existen registros de movimientos de fluidos tipo LPS, los cuales tienen una duración de 5 a 10 segundos, existen sismos tipo Volcanotectónicos, existen movimientos de fluidos catalogados como sismos tipo tremor, los cuales tienen una duración de minutos, días o semanas, existen eventos sísmicos tipo Hibrido los cuales son sismos provocados por movimiento de fluidos y volcanotectónicos a las vez y no existen sismos tipo tornillo en el mismo día de una erupción.

Clúster 4. Porcentaje de representatividad 12%

El 12% de las erupciones del volcán Galeras en el periodo comprendido del año 1989 hasta 2013, son aquellas cuyo comportamiento del gas So2 medido en toneladas estuvo alto, la calidad de localización del sismo es buena, el comportamiento de la velocidad del sismo bajó, el comportamiento del radón bajó, existen registros de movimientos de fluidos tipo LPS, los cuales tienen una duración de 5 a 10 segundos, existen sismos tipo Volcanotectónicos, existen movimientos de fluidos catalogados como sismos tipo tremor, los cuales tienen una duración de minutos, días o semanas, existen eventos sísmicos tipo Hibrido los cuales son sismos provocados por movimiento de fluidos y volcanotectónicos a las vez y no existen sismos tipo tornillo en el mismo día de una erupción.

3.4.3 Discusión de los resultados. De acuerdo a los resultados obtenidos en las diferentes pruebas realizadas en la etapa de minería de datos con las tareas de asociación y agrupación, donde se analizaron 648 casos de emisiones del volcán Galeras, el patrón general es que en una emisión no se presentan sismos tipo tornillo, la calidad de localización de los sismos es regular, el comportamiento del So2 bajó, se presentan sismos tipo tremor y adicional a esto se presentan sismos tipo LPS los cuales pueden durar minutos, días o semanas.

Con Base en en la tarea de agrupación aplicada al repositorio de emisiones podemos concluir que la mejor partición es cuando el parámetro K es igual a 4 (K=4), debido a que contiene las mismas características que los dos patrones que

se obtuvieron cuando el parámetro K se configuró en 2 (K=2), además aparecen dos nuevos grupos con características que influyen en una emisión las cuales son:

- El 51% de las emisiones de volcán Galeras, se comportan de tal manera que la velocidad del sismo baja, el comportamiento del radón sube, se registran movimientos de fluidos catalogados como LPS y catalogados como tremor, se registran sismos catalogados como volcanotectónicos (Vt), y se registran movimientos de fluidos y movimientos volcanotectónicos a la vez catalogados como sismos Hibridos.

- El 21% de las emisiones de volcán Galeras, se comportan de tal manera que la velocidad del sismo sube, el comportamiento del radón baja, se registran movimientos de fluidos catalogados como tremor, se registran sismos catalogados como volcanotectónicos (Vt), y se registran movimientos de fluidos y movimientos volcanotectónicos a la vez catalogados como sismos Hibridos.

- El 9% de las emisiones de volcán Galeras, se comportan de tal manera que el comportamiento del gas SO2 baja, la velocidad del sismo sube, el comportamiento del radón baja, se registran movimientos de fluidos catalogados como LPS, se registran sismos catalogados como volcanotectónicos (Vt), y se registran movimientos de fluidos y movimientos volcanotectónicos a la vez catalogados como sismos Hibridos.

- El 19% de las emisiones de volcán Galeras, se comportan de tal manera que comportamiento del gas SO2 baja, la velocidad del sismo baja, el comportamiento del radón sube, se registran movimientos de fluidos catalogados como tremor.

Apoyado en los resultados obtenidos con las pruebas realizadas en la etapa de minería de datos con las tareas de asociación y agrupación, donde se analizaron 17 casos de erupciones del volcán Galeras, el patrón general es que en una erupción se presentan sismos tipo LPS y sismos tipo tremor con un 100% de probabilidad. En el caso de sismos tipo híbridos y volcanotectónicos se evidenció que están presentes en un 88,23%, así como también el comportamiento del radón subió en un 70% de los casos. Ademas, el comportamiento de la velocidad del sismo bajó.

En en la tarea de agrupación aplicada al repositorio de erupciones podemos concluir que la mejor partición es cuando el parámetro K es igual a 4 (K=4), debido a que contiene las mismas características que los dos patrones que se obtuvieron cuando el parámetro K se configuró en 2 (K=2), además aparecen dos nuevos grupos con características que influyen en una erupción las cuales son:

- El 64% de las erupciones de volcán Galeras, se comportan de tal manera que la velocidad del sismo baja, el comportamiento del radón sube, se registran movimientos de fluidos catalogados como LPS y catalogados como tremor, se

registran sismos catalogados como volcanotectónicos (Vt), y se registran movimientos de fluidos y movimientos volcanotectónicos a la vez catalogados como sismos Hibridos.

- El 12% de las erupciones de volcán Galeras, se comportan de tal manera que el comportamiento del gas SO2 baja, la velocidad del sismo baja, el comportamiento del radón sube, se registran movimientos de fluidos catalogados como LPS y catalogados como tremor, se registran sismos catalogados como volcanotectónicos (Vt), y se registran movimientos de fluidos y movimientos volcanotectónicos a la vez catalogados como sismos Hibridos.

- El 12% de las erupciones de volcán Galeras, se comportan de tal manera que el comportamiento del gas SO2 se mantiene estable, la velocidad del sismo baja, el comportamiento del radón sube, se registran movimientos de fluidos catalogados como LPS y catalogados como tremor, se registran sismos catalogados como volcanotectónicos (Vt), y se registran movimientos de fluidos y movimientos volcanotectónicos a la vez catalogados como sismos Hibridos.

- El 12% de las erupciones de volcán Galeras, se comportan de tal manera que el comportamiento del gas SO2 sube, la velocidad del sismo baja, el comportamiento del radón baja, se registran movimientos de fluidos catalogados como LPS y catalogados como tremor, se registran sismos catalogados como volcanotectónicos (Vt), y se registran movimientos de fluidos y movimientos volcanotectónicos a la vez catalogados como sismos Hibridos.

Basado en el conocimiento que el OVSP tiene antes de un evento eruptivo y los patrones no obtenidos por minería de datos, sino por análisis estadístico y gráfico, se tiene un parámetro que dice que cuando baja el S02 hay erupción hecho que se corrobora con la minería de datos.

3.4 **FASE DE IMPLEMENTACIÓN.**

En esta fase se trata de explotar la potencia de los modelos, integrarlos en los procesos de toma de desiciones de la organización y difundir informes sobre el conocimiento extraído.

La implementación de los patrones obtenidos mediante esta investigación dependerá del análisis y la evaluación por parte de los expertos del OVSP, una vez sea validada y aceptada por los mismos.

Así como también este conocimiento puede ser adicionado como base para dar apoyo en la toma de decisiones en cuanto a la actividad del complejo volcánico Galeras.

4. CONCLUSIONES

En este proyecto se presentó todo el proceso de construcción, limpieza y transformación del repositorio de eventos eruptivos de volcán Galeras. Siguiendo la metodología CRISP-DM, con el fin de obtener patrones eruptivos del volcán Galeras aplicando técnicas de minería de datos.

Se aplicaron tres fases de la metodología CRISP-DM: entendimiento del negocio, entendimiento de los datos y preparación de los datos. Como resultado de la primera fase se obtuvo el diagrama entidad-relación. Se implementó el diagrama entidad-relación en la base de datos *repositoriogaleras1* y finalmente como resultado se obtuvo el repositorio que contiene 4 tablas denominadas **TERU17A18, TEMI648A19, TERUT8729A18 y TEMIT8729A19** a partir de la base de datos *repositoriogaleras1*.

La fase de preparación de los datos que incluye la selección, limpieza, construcción, integración y transformación de datos fue la más costosa en esta investigación debido a la cantidad de datos faltantes e imprecisos que contenían la base de datos que se tomó como fuente para la construcción del repositorio final.

En la mayoría de los casos en donde se aplican técnicas de minería de datos aplica el modelo de clasificación, pero en este caso no fue optimo, debido a que las reglas que salieron no fueron significativas, el modelo de clasificación, no sirve en este caso

Esta investigación servirá de guía para futuras investigaciones que tengan que ver con la construcción de repositorios limpios de datos.

Esta investigación permitió aplicar profundizar los conocimientos adquiridos en la carrera de Ingeniería de Sistemas y principalmente en el área de base de datos y la inteligencia de negocios.

Con esta investigación, el Observatorio Vulcanológico y Sismológico de Pasto cuenta con un estudio que permite obtener información de calidad sobre factores en el área sismológica y geoquímica que inciden en una posible emisión o erupción del volcán Galeras. Los patrones descubiertos permitirán soportar la toma de decisiones eficaces a las directivas del OVSP y tomar las medidas necesarias que permitan disminuir los diferentes daños a la comunidad que pueden desencadenarse debido a la actividad del Volcán Galeras

5. RECOMENDACIONES

Continuar con este tipo de proyectos de investigación aplicada, en la modalidad de trabajos de grado que permiten acercar al estudiante a la solución de problemas reales en el área de minería de datos.

Implementar el conocimiento adquirido en esta investigación con el fin ayudar en la toma de decisiones en los posibles eventos eruptivos del volcán Galeras.

REFERENCIAS BIBLIOGRAFÍCAS

[1] Tilling RI, Lipman PW. Lessons in reducing volcanic risk. Nature 1993;364:277-80.

[2] Simkin T, Siebert L. Volcanoes of the world. 2nd. edition. Washington, D.C.: Smithsonian Institution; 1994.

[3] Portal Informativo Asociación Volcanes de Canarias White paper, disponible en: www.volcanesdecanarias.com, 05/12/2012.

[4] REES, J.D. (1979): Effects of the erption of Paricutin volcano on landforms, vegetation, and human occupancy. En: Volcanic activity and human ecology. Academic Press. New York, pp.249-292

[5] Instituto Colombiano de Geología y minería INGEOMINAS, White paper, disponible en: www.sgc.gov.co/getattachment/Pasto/Volcanes/Volcan-Galeras/Actividad-historica/Actividad_historica_galeras.pdf.aspx, 03/03/2013

[6] PostgresKDD Un Sistema para Descubrimiento de Conocimiento en Base de Datos Fuertemente Acoplado con el SGBD PostgreSQL Manual de Referencia, (2006).

[7] M. Chen, J. Han, and P. Yu, "Data Mining: An Overview from Database Perspective," in IEEE Transactions on Knowlegde and Data Engineering, (1996).

[8] T. Imielinski and H. Manila, "A database perspective on knowlegde discovery communications," Association for Computing Machinery, vol. 39, no. 11, Noviembre (1996).

[9] J. Han and M. Kamber, Data Mining Concepts and Techniques. San Francisco, Estados Unidos: Morgan Kaufmann, (2001).

[10] Hernández, J. y Ramírez, M. y Ramírez, C.: Introducción a la Minería de Datos, Editorial Pearson, Madrid, España, (2004).

[11] Universidad Nacional del Nordeste. Facultad de Ciencias Exactas, Naturales y Agrimensura. Trabajo de Adscripción: Minería de Datos. Sofía J. Vallejos. Materia: Diseño y Administración de Datos. Licenciatura en Sistemas de Información, Corrientes, Argentina. (2006).

[12] Universidad Nacional del Nordeste Facultad de Ciencias Exactas, Naturales y Agrimensura, Mgter. La Red Martínez, David Luis (2008)

[13] WebMining Consultores, Business Intelligence & Analytics, Data Mining, (2011)

[14]. José Alberto Gallardo Arancibia. ”Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM”, White paper, Disponible en la web: http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf, 20/09/2013.

[15]. Aníbal Goicochea**. “**CRISP-DM, Una metodología para proyectos de Minería de Datos”, [en línea]. Agosto 2009. White paper, Disponible en la web: http://anibalgoicochea.com/2009/08/11/crisp-dm-una-metodologia-para-proyectos-de-mineria-de-datos/, 16/06/2013.

[16]. Dataprix. “Metodología CRISP-DM para minería de datos”, White paper , Disponible en la web: http://www.dataprix.com/modelo_crisp-dm, 30/01/2013

[17]. K. Sattler and O. Dunemann, “SQL Database Primitives for Decision Tree Classifiers”. In: CIKM, Atlanta, Georgia, USA, 2001.

[18]. R. Agrawal and R. Srikant, R., “Fast Algorithms for Mining Association Rules”, en: VLDB Conference, Santiago de Chile, Chile, 1994.

[19]. J. Hernández, M.J. Ramirez and C. Ferri, “Introducción a la Minería de Datos”, Editorial Pearson Prentice Hall, ISBN 84-205-4091-9. Madrid, España, 2005.

[20]. J. Han and M. Kamber, “Data Mining Concepts and Techniques”, Ed. Morgan Kaufmann Publishers, San Francisco, USA, 2001.

[21]. Observatorio Vulcanológico y Sismológico de Pasto 2014.

[22]. CASTRO, M., VERA, L. y POSSO, H. (2006). Epidemiología del cáncer de cuello uterino: estado del arte. Revista Colombiana de Obstetricia y Ginecología, vol.57, No.3 Bogotá, Colombia.

[23] SK Gupta; Vasudha Bhatnagar; SK Wasan. A proposal for Data Mining Management System

[24]. AGRAWAL, R., IMIELINSKI, T. and SWAMI, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In: ACM SIGMOD. Washington DC, USA.

[25]. AGRAWAL, R. and SRIKANT, R.(1994). Fast Algorithms for Mining Association Rules. In: VLDB Conference, Santiago de Chile, Chile.

[26]. FAYYAD, U.; PIATETSKY-SHAPIRO, G. and SMYTH, P. (1996a). From Data Mining to Knowledge Discovery: An Overview. En: Advances in Knowlege Discovery and Data Mining. Menlo Park,

[27]. AGRAWAL, R., MEHTA, M., SAFER J. and SRIKANT, R. (1996).The Quest Data Mining System. In: 2º Conference KDD y Data Mining.Porland, Oregon, USA.

[28]. FAYYAD, U.; PIATETSKY-SHAPIRO, G. and SMYTH, P. (1996b). The KDD Process for Extracting Useful Knowledge from Volumes of Data. En: Communications of the ACM. Vol. 39, No 11 (nov). New York (NY, USA): ACM. p. 27-34. ISSN: 0001-0782

[29]. AGRAWAL, R., GHOSH S., IMIELINSKI, T., IYER, B. and SWAMI, A. (1992). An Interval Classifier for Database Mining Applications. In: Proceedings VLDB

[30]. HAN, J. and KAMBER, M. (2001). Data Mining Concepts and Techniques. Ed. Morgan Kaufmann Publishers, San Francisco, USA.

[31]. IMIELNSKI, T. and MANNILA, H. (1996). A Database Perspective on Knowledge Discovery. En: Communications of the ACM. Vol. 39, No.11 (nov): New York (NY, USA): ACM. p. 58-64. ISSN: 0001-0782

[32]. FAYYAD, U.; PIATETSKY-SHAPIRO, G. and SMYTH, P. (1996b). The KDD Process for Extracting Useful Knowledge from Volumes of Data. En: Communications of the ACM. Vol. 39, No 11 (nov). New York (NY, USA): ACM. p. 27-34. ISSN: 0001-0782

[33] “Factores de Peligro vs Daño” White paper, disponible en: www.volcanesdecanarias.com/index.php?lang=es, 04/04/2013

[34]. S. M. Weiss; Ñ. Indurkhya. Predective Data Mining. M. Kaufmann, Harcourt Intl., USA, 1998.

[35]. Maximiliano Silva “Minería de datos y descubrimiento de conocimiento” White paper, disponible en: http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/ Mineria_de_Datos_y_KDD.pdf, 18/08/2013.

[36]. Daedalus, “Técnicas de modelado predictivo de la contaminación en la ciudad sostenible”, White paper, disponible en: www.daedalus.es/blog/es/whitepaper-tecnicas-de-modelado-predictivo-de-la-contaminacion-en-la-ciudad-sostnible/, 05/05/2013.

[37]. Lucero Obando. “METODOLOGIA CRISP-DM (*Cross-Industry Standard Process for Data Mining*)”, [en línea]. Disponible en la Web: http://aprendest-022011.wikispaces.com/file/view/CRISP-DM.pdf.

[38]. GARCÍA, M. y ÁLVAREZ, A. (2010). Análisis de Datos en WEKA –Pruebas de Selectividad. Disponible en: *http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf* .

[39]. WITTEN, I. and FRANK, E. (2000). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Ed. Morgan Kaufmann Publishers, 365 p, ISBN: 1-55860-552-5. San Francisco, CA, USA.

[40]. HERNÁNDEZ, J., RAMIREZ, M.J y FERRI, C. (2005). Introducción a la Minería de Datos. Editorial Pearson Prentice Hall, ISBN 84-205-4091-9. Madrid, España.

[41]. Esta investigación.

[42]. Paola Narváez, Geóloga, Observatorio Vulcanológico y Sismológico de Pasto

[43]. H. Kumagai y B. Chouet. Acoustic properties of a crack containing magmatic or hydrothermal fluids, Journal of Geophysical Research, Vol 105, pages 25,493-25512, November 10, 2000

[44]. Sanders, C.O. and Ryall, F. (1983). Geometry of magma bodies beneath Long Valley, California determined from anomalous earthquake signals. Geophysical Research

[45]. houet, B., 1992. A seismic model for the source of long-period events and harmonic tremor. in: Gasparini et al., (eds.), Volcanic Seismology, IAVCEI Proceedings in Volcanology 3, p. 133-156.

[46]. J. Hernández, M. J. Ramírez, C. Ferri, Introducción a la Minería de Datos, Prentice Hall

[47]. SHARMA, N., BAJPAI, A. and LITORIYA, R. (2012). Comparison the various clustering algorithms of Weka tools. In: International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 5, ISSN: 2250-2459.

[48]. Historia de la actividad del volcán Galeras y percepción de los fenómenos telúricovolcánicos, Miniesterio de cultura. Bogota: El mal Pensante

[49]. Newhall C., Self S. (1982). The volcanic explosivity index (VEI): An estimate of explosive magnitude for historical volcanism. J. Geophys. Res. 87, 1231-1238

[50]. Stephens, C.D., B.A. Chouet, R.A. Page, J.C. Lahr, and J.A. Power, 1994. Seismological aspects of the 1989-1990 eruptions at Redoubt Volcano, Alaska: the SSAM perspective. Jour. Volcano. Geotherm. Res, v. 62, p. 153-182.

[51]. Pyle D. M. (2000). Sizes of Volcanic Eruption. Encyclopedia of Volcanoes. Academy Press. Part II, pp 263-269. San Diego California.

ANEXOS

Anexo 1. DICCIONARIO DE DATOS

TABLA bitacoras_act_superf

Nombre	Tipo	Definición
Codbitacora	double precisión	Código de la bitácora.
Codvolcan	Integer	Código del volcán al cual se relaciona la bitácora.
Fecha	Date	fecha de ocurrencia de la bitácora.
Hora	time without time zone	Hora de ocurrencia de la bitácora.
tipo_observacion	character varying(100)	tipo de actividad superficial [SISMO SENTIDO, ERUPCION, OTRO, EMISION].

TABLA bitacoras_emision

Nombre	Tipo	Definición
Codbitacora	double precisión	Código de la bitácora
Fecha	Date	Fecha de ocurrencia de la emisión.
Hora	time without time zone	Hora de ocurrencia de la emisión.
Altura	double precisión	Altura máxima de la emisión (metros sobre la cima).
Dirección	character varying(30)	Dirección a la cual se desplaza la columna de emisión.
Color	character varying(30)	Color de la columna de emisión.
material_gas	character varying(5)	Si se observa salida de gas [S, N]
material_ceniza	character varying(5)	Si se observa salida de ceniza[S, N]
fenomenos_lluvia_asida	character varying(5)	Si se asocia lluvia acida [S, N]
fenomenos_sonido	character varying(70)	Si se asocia sonido
fenomenos_olor	character varying(70)	Si se asocia algún tipo de olor a la emisión.

TABLA bitacoras_erupcion

Nombre	Tipo	Definición
Codbitacora	double precisión	código de la bitácora.
Fecha	Date	fecha de registro del evento.
Hora	time without time zone	Hora de registro del evento.
altura_max	double precisión	Altura máxima de la columna.
Dirección	character varying(30)	Dirección a la cual se desplaza la columna.
incandescencia	character	Si se observa incandescencia [S,N]

Nombre	Tipo	Definición
	varying(5)	
caida_piroclastos	character varying(5)	Si se observa caída de piroclastos [S, N].
onda_choque	character varying(5)	Si se registra onda de choque[S,N]
sismo_sentido	character varying(5)	Si se registra un sismo sentido relacionado con la erupción [S,N]
flujo_lava	character varying(5)	Si se registra un flujo de lava [S,N].
direccion_lava	character varying(30)	Dirección del flujo de lava.
flujos_piroclasticos	character varying(5)	Si se registra flujo piroclastico [S, N].
direccion_flujo	character varying(30)	Dirección del flujo.
proyectil_balistico	character varying(5)	Si se registran proyectiles balísticos.
proyectil_may_tam	double precisión	mayor tamaño de proyectil
proyectil_dist_may_tam	double precisión	Distancia a la cual se encuentra el proyectil de mayor tamaño.
proyectil_may_dist	double precisión	Mayor distancia de proyectil.
proyectil_tam_may_dist	double precisión	Tamaño del proyectil que se encuentra a mayor distancia.
Lapilli	character varying(5)	Si el material emitido contiene lapilli [S,N].
Bloques	character varying(5)	Si se emitieron bloques [S, N].
Bombas	character varying(5)	Si se emitieron bombas[S, N].
Ceniza	character varying(5)	Si se observó salida de ceniza[S,N]
Iev	double precisión	Índice de explosividad volcánica.
volumen_emitido	double precisión	Volumen de material emitido.
mayor_area_isopacas	double precisión	Mayor área de esópicas.
indice_severidad	double precisión	Índice de severidad.
Color	character varying(30)	Color de la emisión.

TABLA bitacoras_observacion

Nombre	Tipo	Definición
Codbitacora	double precisión	código de la bitácora
codobservacion	Integer	código de la observación
Fecha	Date	fecha del reporte
hora_in	time without time zone	Hora de inicio del reporte.
hora_fin	time without time zone	Hora de fin del reporte.
Reporta	character varying(100)	Entidad o persona que reporta.
Edita	character varying(100)	Persona quien edita.
Observación	character varying(7000)	reporte recibido.

TABLA climatológica

Nombre	tipo	Definición
fecha	date	Fecha de registro.
hora	time without time zone	Hora de registro.
wvkmh	double precision	velocidad del viento (km/h)
wdgs	double precision	Dirección del viento (grados Azimut -180°)
tmgc	double precision	temperatura.
hm	double precision	Humedad relativa.
pl	double precision	
pr	double precision	Presión.
plmmh	double precision	Datos de pluviómetro (mm/h)
wdaz	double precision	dirección de viento (grados Azimut)
codvolcan	integer	código del volcán

TABLA conteomanual

Nombre	tipo	Definición
Fecha	date	Fecha para conteo.
LPS	integer	Numero de eventos LPS clasificables para la fecha dada.
Hyb	integer	Numero de eventos HYB clasificables para la fecha dada.
Vt	integer	Numero de eventos VT clasificables para la fecha dada.
Tor	integer	Numero de eventos tipo TOR para la fecha dada.
Tre	integer	Numero de eventos tipo TREMOR para la fecha dada.
noclasificables	integer	Numero de eventos no clasificables para la fecha dada.
Codvolcan	integer	código del volcán.

TABLA datosinclinomurad

Nombre	tipo	Definición
codestacion	character varying(12)	Código de la estación de inclinometria.
fecha	date	Fecha del registro.
hora	time without time zone	Hora del registro.
rad	double precision	valor de la componente radial (microradianes), el valor 9999999 corresponde a NaN.
tang	double precision	valor de la componente tangencial (microradianes), el valor 9999999 corresponde a NaN.
voltaje	double precision	valor de la componente voltaje (voltios), el valor 9999999 corresponde a NaN.
temperatura	double precision	Valor de la componente temperatura (grados centígrados), el valor 9999999 corresponde a NaN.

TABLA energiatotal

Nombre	tipo	Definición
fecha	date	Fecha
LPSoc	double precision	Energía total ondas de cuerpo de los eventos tipo LPS clasificados.
LPSor	double precision	Energía total ondas de superficie de los eventos tipo LPS clasificados.
hyboc	double precision	Energía total ondas de cuerpo de los eventos tipo HYB clasificados.
hybor	double precision	Energía total ondas de superficie de los eventos tipo HYB clasificados.
vtoc	double precision	Energía total ondas de cuerpo de los eventos tipo VT clasificados.
vtor	double precision	Energía total ondas de superficie de los eventos tipo VT clasificados.
treoc	double precision	Energía total ondas de cuerpo de los eventos tipo TRE clasificados.
treor	double precision	Energía total ondas de superficie de los eventos tipo TREMOR clasificados.
codvolcan	integer	Código del volcán.

TABLA estacion

Nombre	tipo	Definición
codestacion	character varying(7)	Código de la estación.
nombre	character varying(50)	Nombre de la estación.
volcán	integer	código del volcán al cual pertenece la estación.
latitud	integer	Ubicación de la estación latitud (grados).
latitudm	double precision	Ubicación de la estación latitud (minutos).
longitud	integer	Ubicación de la estación longitud (grados).
longitudm	double precision	Ubicación de la estación longitud (minutos).
altitud	double precision	Altitud de la estación (metros).
observación	character varying(500)	Observación sobre la estación.

TABLA estacion_doas

Nombre	Tipo	Definición
Codestacion	character varying(30)	Código de la estación.
cod_espect	character varying(30)	Código del espectrómetro.
dist_crater	double precisión	Distancia de la estación al cráter (km).
altura_ref	double precisión	Altura de referencia de la estación.
dir_compas	Integer	Dirección de compas (Azimut - 180)
cobertura_escaneo	Integer	Cobertura de escaneo (Grados)
Codvolcan	Integer	Código del volcán al cual pertenece la estación.
Latitud	Integer	Ubicación de la estación latitud (Grados).
Latitudm	double precisión	Ubicación de la estación latitud (minutos).
Longitud	Integer	Ubicación de la estación longitud (Grados).
Longitudm	double precisión	Ubicación de la estación longitud (minutos)
Altitud	double precisión	Altitud de la estación (metros)

TABLA estacioninclinom

Nombre	tipo	Definición
nombre	character varying(30)	Nombre de la estación.
codestacion	character varying(12)	Código de la estación.
latitud	double precision	Ubicación de la estación latitud (grados).
latitudm	double precision	Ubicación de la estación latitud (minutos).
longitud	double precision	Ubicación de la estación longitud (grados).
longitudm	double precision	Ubicación de la estación longitud (minutos).
observación	character varying(500)	Observación de la estación.

Nombre	tipo	Definición
altitud	double precision	Altitud de la estación (metros).
codvolcan	integer	Código del volcán al cual pertenece la estación.

TABLA lecanalogashyb

Nombre	tipo	Definición
fecha	date	Fecha de registro del evento.
hora	time without time zone	Hora de registro del evento.
ganancia	integer	Factor de ganancia del registrador.
filtros	character varying(10)	Filtros del registrador.
estacion	character varying(10)	Estación en la cual se realiza la lectura.
distancia	double precision	Distancia a la cual se encuentra la estación en la cual se realiza la lectura.
sensibilidad	double precision	Sensibilidad de la estación en el registrador.

TABLA lecanalogasLPS

Nombre	tipo	Definición
Fecha	Date	Fecha de registro del evento.
Hora	time without time zone	Hora de registro del evento.
ganancia	Integer	Factor de ganancia del registrador.
Filtros	character varying(10)	Filtros del registrador.
estacion	character varying(10)	Estación en la cual se realiza la lectura.
distancia	double precision	Distancia a la cual se encuentra la estación en la cual se realiza la lectura.
sensibilidad	double precision	Sensibilidad de la estación en el registrador.

TABLA lecanalogastre

Nombre	tipo	Definición
Fecha	date	Fecha de registro del evento.
Hora	time without time zone	Hora de registro del evento.
ganancia	integer	Factor de ganancia del registrador.
Filtros	character varying(10)	Filtros del registrador.
estacion	character varying(10)	Estación en la cual se realiza la lectura.
distancia	double precision	Distancia a la cual se encuentra la estación en la cual se realiza la lectura.
sensibilidad	double precision	Sensibilidad de la estación en el registrador.

TABLA lecanalogasvt

Nombre	tipo	Definición
Fecha	date	Fecha de registro del evento.
Hora	time without time zone	Hora de registro del evento.
Ganancia	integer	Factor de ganancia del registrador.
Filtros	character varying(10)	Filtros del registrador.
Estación	character varying(10)	Estación en la cual se realiza la lectura.
Distancia	double precision	Distancia a la cual se encuentra la estación en la cual se realiza la lectura.
Sensibilidad	double precision	Sensibilidad de la estación en el registrador.

TABLA lecdigitaleshyb

Nombre	tipo	Definición
Fecha	Date	Fecha de registro del evento.
Hora	time without time zone	Hora de registro del evento.
Ganancia	integer	Ganancia de la tarjeta digitalizadora.
Amplitud	integer	Amplitud máxima leída.
nombrearchivo	character varying(15)	Nombre del archivo que contiene la traza del evento.
Programap	character varying(7)	Código del programa con el que realizo la lectura.
Horaamp	time without time zone	Hora de la máxima amplitud leída en el evento.

Nombre	tipo	Definición
Horacoda	time without time zone	Hora de terminación del evento.
Muestreo	double precision	Tasa de muestreo de la traza.
Horainven	time without time zone	Hora de inicio del archivo que contiene el evento.
Estación	character varying(7)	Estación en la cual se realizó la lectura del evento.

TABLA lecdigitalesLPS

Nombre	tipo	Definición
Fecha	date	Fecha de registro del evento.
Hora	time without time zone	Hora de registro del evento.
Ganancia	integer	Ganancia de la tarjeta digitalizadora.
Amplitud	integer	Amplitud máxima leída.
nombrearchivo	character varying(15)	Nombre del archivo que contiene la traza del evento.
Programap	character varying(7)	Código del programa con el que realizo la lectura.
Horaamp	time without time zone	Hora de la máxima amplitud leída en el evento.
Horacoda	time without time zone	Hora de terminación del evento.
Muestreo	double precision	Tasa de muestreo de la traza.
Horainven	time without time zone	Hora de inicio del archivo que contiene el evento.
Estación	character varying(7)	Estación en la cual se realizó la lectura del evento.

TABLA lecdigitalestre

Nombre	tipo	Definición
Fecha	date	Fecha de registro del evento.
Hora	time without time zone	Hora de registro del evento.
Ganancia	integer	Ganancia de la tarjeta digitalizadora.
Amplitud	integer	Amplitud máxima leída.
Nombrearchivo	character varying(15)	Nombre del archivo que contiene la traza del evento.
Programap	character varying(7)	Código del programa con el que realizo la lectura.

Nombre	tipo	Definición
Horaamp	time without time zone	Hora de la máxima amplitud leída en el evento.
Horacoda	time without time zone	Hora de terminación del evento.
Muestreo	double precision	Tasa de muestreo de la traza.
Horainven	time without time zone	Hora de inicio del archivo que contiene el evento.
Estación	character varying(7)	Estación en la cual se realizó la lectura del evento.

TABLA lecdigitalesvt

Nombre	tipo	Definición
Fecha	date	Fecha de registro del evento.
Hora	time without time zone	Hora de registro del evento.
Ganancia	integer	Ganancia de la tarjeta digitalizadora.
Amplitud	integer	Amplitud máxima leída.
Nombrearchivo	character varying(15)	Nombre del archivo que contiene la traza del evento.
Programap	character varying(7)	Código del programa con el que realizo la lectura.
Horaamp	time without time zone	Hora de la máxima amplitud leída en el evento.
Horacoda	time without time zone	Hora de terminación del evento.
Muestreo	double precision	Tasa de muestreo de la traza.
Horainven	time without time zone	Hora de inicio del archivo que contiene el evento.
Estación	character varying(7)	Estación en la cual se realizó la lectura del evento.

Tabla vt

Nombre	tipo	Definición
Fecha	date	fecha de registro del evento
Hora	time without time zone	hora de registro del evento
Duración	double precisión	duración del evento (segundos)
Amplitud	double precision	amplitud máxima del evento
Sp	double precision	Diferencia de tiempo entre el arribo de la onda

		s y la onda p
Estación	character varying(6)	Estación de lectura de parámetros del evento.
Etotaloc	double precision	Energía de ondas de cuerpo de todo el evento (ergios)
etotalor	double precision	energía de ondas superficiales de todo el evento (ergios)
Drtotaloc	double precision	Desplazamiento reducido de ondas de cuerpo de todo el evento (cm^2)
Drtotalor	double precision	Desplazamiento reducido de ondas superficiales todo el evento (cm^2)
Sentido	character varying(4)	Identifica si el evento fue sentido o no
observaciones	character varying(500)	observaciones sobre el evento
Tipolectura	character(1)	
Codvolcan	integer	código del volcán al cual pertenece el evento
Tipo	character varying(6)	tipo de evento [GVA, GVB]
Periodo	double precision	Periodo medido en la máxima amplitud
Frecuencia	double precision	frecuencia medida en la máxima amplitud
codlocalizacion	integer	código de la localización relacionada
magnitudcoda	double precision	magnitud de coda del evento
Energiacoda	double precision	energía calculada de acuerdo a la duración del evento

Tabla volcan

Nombre	**tipo**	**Definición**
codvolcan	integer	Código del volcán
nombre	character varying(40)	nombre del volcán
latitud	character varying(10)	coordenada volcán latitud en grados
longitud	character varying(10)	coordenada del volcán longitud en grados
elevación	double precision	elevación del volcán en metros sobre el nivel del mar

Tabla tremor

Nombre	tipo	Definición
Fecha	date	Fecha de registro del evento.
Hora	time without time zone	Hora de registro del evento.
Duración	double precision	Duración del evento (segundos).
Ampmax	double precision	Amplitud máxima del evento.
Ampmin	double precision	Amplitud mínima del evento.
Ampprom	double precision	Amplitud promedio del evento.
Permax	double precision	Periodo asociado a la amplitud máxima del evento.
Permin	double precision	Periodo asociado a la amplitud mínima del evento.
Perpro	double precision	Periodo asociado a la amplitud promedio del evento.
Etotoc	double precision	Energía de ondas de cuerpo de todo el evento (ergios).
Etotor	double precision	Energía de ondas de superficie de todo el evento (ergios)
Drtotaloc	double precision	Desplazamiento reducido de ondas de cuerpo de todo el evento (cm^2).
Drtotalor	double precision	Desplazamiento reducido onda de superficie de todo el evento (cm^2).
Tipolectura	character(1)	
Observaciones	character varying(500)	Observaciones del evento.
Tipoevento	character varying(10)	Tipo de evento
Codvolcan	integer	Código del volcán al cual se relaciona el evento.

Tabla so2

Nombre	tipo	Definición
Fecha	date	Fecha de la medida.
Hora	time without time zone	Hora de la medida.
Azimut	double precision	Dirección de la columna (grados Azimut).
vel_ms	double precision	Velocidad del viento con la cual se calculó la medida (m/s)
ton_v	double precision	Máximo valor de So2 por día (Toneladas/día)
Puntomedida	character varying(50)	Punto en el cual se toma la medida.

Instrumento	character varying(20)	Tipo de instrumento para la medida.
observaciones	character varying(500)	Observaciones sobre la medida.
Codvolcan	integer	Código del volcán.
Maxdiariopp	double precision	Porcentaje de pluma para el valor máximo diario.
Mindiario	double precision	Valor mínimo de SO2 en el día (Toneladas/día).
Mindiariopp	double precision	Porcentaje de la pluma de mínimo valor diario.
Mindiarioes	character varying(30)	Estación en la cual se toma el mínimo.
Maxppdiario	double precision	Máximo porcentaje de pluma diario.
maxppdiarioval	double precision	Valor de SO2 asociado al máximo porcentaje de pluma diario.
maxppdiarioes	character varying(30)	Estación asociada al máximo porcentaje de pluma.

Tabla radon_conc

Nombre	Tipo	Definición
codpunto	character varying(10)	Código del punto en el cual se tomó la medida.
fecharet	Date	Fecha de retiro del electrect.
horaret	time without time zone	Hora de retiro del electrect.
concrad	double precision	Concentración de radón calculada.

Tabla punto_radon

Nombre	Tipo	Definición
codlinea	character varying(30)	Código de la línea de radón.
codpunto	character varying(10)	Código del punto de radón.
punto	character varying(30)	Nombre del punto de radón.
observacion	character varying(1000)	Observación sobre el punto de radón.

Tabla omi

Nombre	Tipo	Definición
Volcan	Integer	Código del volcán al cual pertenece la medida.
Fecha	Date	Fecha de la medida.
Seobserva	character varying(20)	Registra si se observa o no emisión en la imagen.
intervalo_in	time without time zone	Hora inicio de intervalo.
intervalo_fin	time without time zone	Hora fin de intervalo.

hora_medicion	time without time zone	Hora de la medición.
maxso2du	double precision	Máximo S02 (Unidades dobson).
maxso2mgm2	double precision	Máximo so2 (mg/m^2)
Fuente	character varying(20)	Fuente de la cual se toma la imagen.
Observación	character varying(600)	Observaciones sobre la medida.

Tabla medida_radon

Nombre	Tipo	Definición
Codpunto	character varying(10)	Código del punto donde se toma la medida.
Codelectreto	character varying(15)	Código del electrect instalado.
Fechainst	Date	Fecha de instalación del electrect.
Horainst	time without time zone	Hora de instalación del electrect.
Fecharet	Date	Fecha para retirar el electrect.
Horaret	time without time zone	Hora para retirar el electrect.
Potencialin	double precisión	Potencial inicial del electrect.
Potencialfin	double precisión	Potencial final del electrect.
presion20psi	double precisión	Presión.
Observaciones	character varying(600)	Observaciones sobre la toma de la medida.
Diasexposicion	double precisión	Días que el electrect estuvo en el punto de medida.
Horasexposicion	double precisión	Horas a las que el electrect estuvo en el punto de medida.

Tabla LPS

Nombre	Tipo	Definición
Fecha	Date	fecha de registro del evento
Hora	time without time zone	Hora de registro del evento
Tipo	character varying(5)	tipo de evento [GLP(LPS), GTO(tornillo)]
Duración	double precision	Duración del evento (segundos)
Amplitud	double precision	máxima amplitud del evento
Periodo	double precision	Periodo de la onda de máxima amplitud del evento.
Frecuencia	double precision	Frecuencia de la onda de máxima amplitud del evento.
Estación	character varying(7)	Estación en la cual fue leído el evento.
Etotaloc	double precision	Energía de ondas de cuerpo de todo el evento (ergios).

Etotalor	double precision	Energía de ondas de superficie de todo el evento (ergios)
Drtotaloc	double precision	Desplazamiento reducido de ondas de cuerpo de todo el evento (cm^2).
Drtotalor	double precision	Desplazamiento reducido onda de superficie de todo el evento (cm^2).
Observaciones	character varying(500)	Observaciones sobre el evento.
Magnitudlocal	double precision	
Tipolectura	character(1)	
Codvolcan	Integer	Código del volcán al cual se relaciona el evento.
Magnitudcoda	double precision	Magnitud de coda del evento.
Energiacoda	double precision	Energía calculada teniendo en cuenta la amplitud y duración del evento.

Tabla localizaciones

Nombre	**Tipo**	**Definición**
Fecha	Date	Fecha de ocurrencia del evento.
Hora	time without time zone	Hora de ocurrencia del evento.
Segundos	double precision	Segundos de ocurrencia del evento con decimal.
Latgrados	Integer	Latitud localización del evento (grados)
Latminutos	double precision	Latitud localización del evento (minutos)
Longrados	Integer	Longitud localización del evento (grados).
Lonminutos	double precision	Longitud localización del evento (minutos).
Profundidad	double precision	profundidad focal (km)
magnitudcoda	double precision	Magnitud de coda (duración del evento).
Nfaces	Integer	Numero de faces que se tuvieron en cuenta en la localización.
Gap	Integer	Mayor separación azimutal entre las estaciones (grados).
dmin	double precision	Distancia epicentral a la estación más cercana (km).
rms	double precision	Error medio cuadrático de los residuales de tiempo.
erh	double precision	Error standard del epicentro (km).
erz	double precision	Error standard de la profundidad focal (km).
qm	character varying(4)	Calidad de la localización [A,B,C,D] A: epicenter excelent, focal depth good B:

		epicenter good, focal depth fair C: epicenter fair, focal depth poor D: epicenter poor, focal depth poor
nestaciones	Integer	Número de estaciones que se tuvieron en cuenta en la localización.
archivo	character varying(14)	Nombre del archivo que contiene la traza del evento.
modelo	character varying(3)	Modelo de localizaciones utilizado.
magnitudlocal	double precision	Magnitud local del evento.
profitera	double precision	Profundidad de iteración inicial.
codlocalizacion	double precision	Código de la localización.
tipo	character varying(5)	Tipo de evento localizado.

Tabla línea_radon

Nombre	Tipo	Definición
codvolcan	Integer	Código del volcán al cual pertenece la estación.
codlinea	character varying(30)	Código de la línea de radón.
observación	character varying(1000)	Observaciones de línea de radón.

Tabla hyb

Nombre	Tipo	Definición
Fecha	Date	Fecha de registro del evento.
Hora	time without time zone	Hora de registro del evento.
duracion	double precision	Duración del evento (segundos).
amplitud	double precision	Amplitud máxima del evento.
Sp	double precision	Diferencia de tiempo entre el arribo de ondas S y ondas P.
estacion	character varying(7)	Estación de lectura de parámetros del evento.
Etotoc	double precision	Energía de ondas de cuerpo de todo el evento (ergios).
Etotor	double precision	Energía de ondas de superficie de todo el evento (ergios)
drtotaloc	double precision	Desplazamiento reducido de ondas de cuerpo de todo el evento (cm^2).
drtotalor	double precision	Desplazamiento reducido onda de superficie de todo el evento (cm^2).
Observaciones	character varying(500)	Observaciones sobre el evento.

tipolectura	character(1)	
codvolcan	Integer	Código del volcán al cual se relaciona el evento.
Periodo	double precisión	Periodo de la onda de máxima amplitud del evento.
frecuencia	double precisión	Frecuencia de la onda de máxima amplitud del evento.
Codlocalizacion	Integer	Código de la localización asociada al evento.
Magnitudcoda	double precisión	
Energiacoda	double precisión	Energía calculada teniendo en cuenta la amplitud y duración del evento.
Magnitudcodal	double precisión	
Energiacodal	double precisión	

Los anexos 2 hasta los anexos 16 se encuentran en el CD.