DETECCIÓN DE ALERTAS PRE-ERUPTIVAS VOLCÁNICAS BASADA EN APRENDIZAJE INCREMENTAL

JOSÉ EDUARDO GÓMEZ DAZA

Tesis de Maestría en Ingeniería Telemática

Director: PhD. David Camilo Corrales

Co-Director. PhD. Juan Carlos Corrales

Universidad del Cauca

Facultad de Ingeniería Electrónica y Telecomunicaciones

Departamento de Telemática - Grupo de Ingeniería Telemática

Línea de Investigación e-@mbiente

Popayán, febrero de 2019

JOSE EDUARDO GÓMEZ DAZA

DETECCIÓN DE ALERTAS PRE-ERUPTIVAS VOLCÁNICAS BASADA EN APRENDIZAJE INCREMENTAL

Tesis presentada a la Facultad de Ingeniería
Electrónica y Telecomunicaciones de la
Universidad del Cauca para la obtención del
Título de

Magíster en:
Ingeniería Telemática

Director:
PhD. David Camilo Corrales
Co-Director.
PhD. Juan Carlos Corrales

Popayán
2019

La gloria siempre a Dios

A mis padres, por ser el motor de mi vida y la inspiración de todas mis mañanas

A mis hermanas, por su comprensión y cariño

Agradecimientos

Además de agradecer, también quiero dedicar este trabajo a mis padres, Bolívar Gómez y Mireya Daza, por el ser los motores que me impulsan cada día, ellos siempre están en mi corazón. Gracias por creer en mí y por apoyarme en los momentos más difíciles.

A mis hermanas, Elcy y Nina Gómez, las cuales me brindan su apoyo incondicional en cada paso que doy y a quienes quiero demasiado.

A mis tutores, David Camilo y Juan Carlos Corrales, quienes me brindaron todo su conocimiento y dedicación durante estos años. Gracias por sus consejos, ayuda y comprensión. Camilo amigo gracias por tanto.

A todos y cada uno de los integrantes del Servicio Geológico Colombiano; quienes me brindaron desde un comienzo ayuda y asesoría en todo momento. Gracias por hacerme el trabajo tan agradable.

A todo el grupo de profesores y compañeros del grupo de Telemática de la Universidad del Cauca, en especial a Iván López, Jesús Cerón y Emmanuel Lasso.

Un agradecimiento especial a mi compañero y amigo Oscar Manzo, por ser una parte fundamental en el desarrollo de este trabajo de grado, al transmitirme conocimiento y todo el apoyo necesario que requerí en todos estos años.

A todos mis amigos, les agradezco por cruzarse en mi camino y ser una parte muy importante de mi vida.

A mi tutor en mi estancia en Madrid (España), el profesor José Antonio Iglesias. Gracias por hacerme parte de su grupo de investigación y por haberme enseñado tanto.

A todos, ¡un millón de gracias!

Resumen estructurado

Antecedentes: Los volcanes son estructuras geológicas que generan situaciones de emergencia para quienes viven en su entorno. Los riesgos a los que esta expuesta una población son terremotos, flujos, explosiones, emisiones de gases y cenizas, etc; causando morbilidad y alta mortalidad debido al tamaño de grandes erupciones. Indirectamente, los eventos vulcanológicos pueden causar deterioro socioeconómico, daño de líneas de transporte vitales e infraestructuras y, en general, alterar las condiciones de vida de las poblaciones comprometidas. Por esta razón, la vigilancia de volcanes es una tarea clave para detectar anomalías volcánicas en tiempo real y actuar en consecuencia. Por otro lado, de las ciencias de la computación se tiene que las técnicas de aprendizaje automático se han posicionado como herramientas para dar solución a diversos problemas de la vida real, como la clasificación y detección de intrusos, monitorización de procesos industriales, entre otros. Estas técnicas tienen un buen desempeño cuando se cuenta con datos del dominio a priori, donde algún algoritmo de clasificación automática es entrenado con el conjunto de datos y se obtiene un modelo capaz de obtener una clasificación o predicción con un alto porcentaje de precisión.

Objetivos: Desarrollar un sistema que permita la detección de alertas pre-eruptivas a partir de la detección de anomalías volcánicas y que sea capaz de tratar con flujos de datos provenientes de sus estaciones de monitorización, permitiendo que dicha detección mantenga una precisión aceptable.

Métodos: Se propone usar un algoritmo de detección de valores atípicos que implemente aprendizaje incremental para no almacenar todos los ejemplos del flujo de datos provenientes de las estaciones de deformación y geoquímica volcánica y que actualice la función modelo cada vez que ocurran cambios. De esta forma, se pretende calcular los valores atípicos en tiempo real y generar las alertas respectivas, las cuales serán clasificadas por los expertos de la vigilancia volcánica según sea el caso.

Resultados: La presente propuesta entregó como resultados una serie de conjuntos de datos que involucran información volcanológica perteneciente a las áreas de monitorización volcánica de geoquímica y deformación. Estos datos fueron recolectados a través de estaciones de inclinometría y dióxido de carbono, ubicadas en cercanías del volcán Puracé (departamento del Cauca), además, se entrega un prototipo capaz de detectar en tiempo real las diferentes anomalías generadas en el volcán.

Conclusiones: El dominio de aplicación utilizado en la presente investigación demostró que usar el algoritmo RDE (Recursive Density Estimation) en monitorización volcánica es una buena opción para encontrar valores atípicos y generar alertas que permitan a los expertos conocer las anomalías que están ocurriendo en el volcán en tiempo real.

Palabras Clave: Outlier, RDE, anomalías, deformación, geoquímica, entorno dinámico, anomalías, flujo de datos, monitorización volcánica.

ABSTRACT

Background: Volcanoes are geological structures that generate emergency situations for those who live in their environment. The risks to which the population is exposed (earthquakes, flows, explosions, emissions of gases and ashes, etc.) cause morbidity and high mortality due to the size of large eruptions. Indirectly, volcanic events can cause socioeconomic deterioration, damage to vital transport lines and infrastructures, in general, alter the living conditions of the populations involved. For this reason, volcano monitoring is a key task to detect volcanic anomalies in real time and act accordingly. On the other hand, from the computer science we know that the automatic learning techniques have been positioned as tools to solve various real-life problems, such as classification and detection of intruders, monitoring of industrial processes, among others. These techniques have better performance when data from the a priori domain is available, where an automatic classification algorithm is trained with the data set obtaining a model capable of creating a classification or prediction with a high percentage of accuracy.

Objectives: To develop a system that allows the detection of pre-eruptive alerts from the detection of volcanic anomalies, which is able to deal with data flows coming from its monitoring stations, allowing this detection to maintain an acceptable precision

Methods: It is proposed to use an atypical value detection algorithm that implements incremental learning so as not to store all the examples of the data flow coming from the deformation and volcanic geochemical stations and to update the model function every time changes occur. In this way, it is intended to calculate the outliers in real time and generate the respective alerts, which will be classified by experts of volcanic monitoring as the case may be.

Results: The present proposal delivered as results a series of data sets that involve vulcanological information pertaining to the areas of volcanic monitoring of geochemistry and deformation. These data were collected through inclinometry and carbon dioxide stations, located near the Puracé volcano (department of Cauca), in addition, a prototype is delivered capable of detecting in real time the different anomalies generated in the volcano.

Conclusions: The domain of application used in the present investigation showed that using the RDE algorithm (Recursive Density Estimation) in volcanic monitoring is a good option to find outliers and generate alerts that allow experts to know the anomalies that are occurring in the volcano in real time.

Key words: Outlier, RDE, anomalies, deformation, geochemistry, dynamic environment, anomalies, data flow, volcanic monitoring.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ECUACIONES

Capítulo 1

En este capítulo se aborda en primera instancia, el planteamiento del problema que permite conocer el estado actual del conocimiento en cuanto a la vigilancia volcánica, seguidamente se planea el escenario de motivación que evidencia la brecha principal a la hora de detectar alertas volcánicas. Se continúa exponiendo el objetivo general y los específicos a desarrollar en esta tesis de grado. Finalmente se muestran las contribuciones que se generan al término de este proyecto de Maestría.

Introducción

1.1. Planteamiento del problema

Las condiciones geológicas y climáticas de América Latina y el Caribe propician la ocurrencia de desastres naturales. Los Andes y las montañas del Caribe y América Central se asientan en las áreas de interacción de las placas tectónicas, una característica que determina la alta sismicidad de la región [1]. A través de los años, los volcanes han generado situaciones de emergencia provocando eventos con una gran capacidad de destrucción. Son muchas las poblaciones asentadas en áreas próximas a volcanes que conviven con una compleja combinación de beneficios y riesgos. En el primer caso, los beneficios son varios: agrícolas, turísticos, terapéuticos, etc. Los riesgos en un volcán activo pueden afectar la salud de una población de forma directa, a causa de sus sismos, flujos de lodo, explosiones, emisiones de gases y cenizas, entre otros, ocasionando morbilidad por diferentes patologías y mortalidad alta por la exposición al trauma. Indirectamente, pueden ocasionar el deterioro socioeconómico, daño de líneas vitales o de infraestructuras y en general, alterar las condiciones de vida de las poblaciones comprometidas por la actividad volcánica [2].

Teniendo en cuenta lo anterior, es importante saber que los volcanes son monitoreados permanentemente en su gran mayoría por entidades del sector público tales como el USGC (Servicio Geológico de los Estados Unidos) y el OVDAS (Observatorio Vulcanológico de los Volcanes del Sur). Dichas entidades brindan a la comunidad reportes del nivel de actividad y fenómenos ocurridos en el interior de un volcán. En Colombia, la entidad oficial encargada de la monitorización de volcanes es el SGC (Servicio Geológico Colombiano – http://www.sgc.gov.co), el cual se divide en los observatorios vulcanológicos y sismológicos de Popayán (OVSPo), Pasto (OVSP) y Manizales (OVSM). El personal experto de dichas entidades se encarga en muchas ocasiones de analizar manualmente los datos de las distintas técnicas y estaciones correlacionando sus variables, con el objetivo de generar alertas (obtenidas empíricamente) que notifiquen acerca de un posible

evento volcánico en caso de ser necesario. Las alertas en vulcanología implican vigilancia en tiempo real de los posibles escenarios que suelen ocurrir antes de una erupción (alertas pre-eruptivas), los cuales describen cambios relevantes de un volcán permitiendo analizar y tomar decisiones acerca del nivel de peligrosidad y el posterior impacto para las comunidades aledañas al volcán [3].

Desde las ciencias de la computación se han abordado diferentes problemas descritos en el párrafo anterior como: clasificación y localizaciones automáticas de eventos sísmicos de origen volcánico para determinar fuentes sismogénicas, detección de emisiones de gases y deformación del cono volcánico, tal como lo evidencian los autores en [4]–[7], los cuales abordan dichas situaciones haciendo uso de algoritmos de aprendizaje supervisado (AS) tradicional. Básicamente, el AS se caracteriza por construir un clasificador a partir de un conjunto de ejemplos también llamados datos de entrenamiento, con el objetivo de detectar o predecir un valor [8]. Sin embargo, un aspecto importante a considerar en este entorno (volcánico) es que estos datos deben analizarse en tiempo real para producir una respuesta rápida y específica.

A pesar de la importancia de los algoritmos de AS en vulcanología, las investigaciones adelantadas hasta el momento presentan las siguientes limitaciones:

- **Obsolescencia en el clasificador**:
  Dejan de lado la inclusión de nuevos ejemplos una vez construido el modelo [9], lo cual genera obsolescencia en el clasificador, ya que los volcanes son considerados sistemas dinámicos (continuos cambios a través del tiempo) al presentar constantemente una serie de interacciones de la corteza terrestre, específicamente de las placas tectónicas. Además, a medida que los volcanes avanzan en sus distintos niveles de actividad, su línea base de referencia cambia totalmente.

- **Carencia de alertas volcánicas:**
  No consideran detectar anomalías volcánicas que permitan generar alertas sobre riesgos potenciales a las comunidades aleñadas de un volcán.

- **Falta de integración de información entre áreas de vigilancia volcánica:**
  Sólo utilizan parámetros o variables de una misma área de monitorización volcánica (Sismología, Deformación, Geoquímica) y no consideran correlacionar variables de diferentes áreas para obtener resultados más precisos. Correlacionar distintos parámetros de vigilancia volcánica permite obtener una visión global de lo que está ocurriendo en un volcán. Cada una de estas áreas es la causa/efecto de la otra, originando de esta forma movimientos volcánicos a causa de la desgasificación, rompimiento de rocas por la presión del magma, explosiones de gases y procesos de deformación del sistema a causa de la presión que ejerce el magma.

Teniendo en cuenta las consideraciones descritas anteriormente se hace necesario generar alertas pre-eruptivas a partir de la detección de anomalías volcánicas que

permitan identificar el nivel de actividad de un volcán a través del aprendizaje incremental, el cual es un enfoque que permite actualizar automáticamente los modelos al detectar un cambio de contexto. En este sentido, la presente propuesta de maestría plantea la siguiente pregunta de investigación:

¿Cómo detectar alertas (anomalías) pre-eruptivas de un volcán, teniendo en cuenta las características de un sistema dinámico?

Para responder la anterior pregunta de investigación, se parte de la siguiente hipótesis:

Las técnicas de aprendizaje incremental pueden ser usadas para realizar una adecuada detección de alertas pre-eruptivas a través del tiempo, permitiendo la inclusión de nuevas instancias que son adquiridas en tiempo real mediante estaciones de vigilancia volcánica.

1.2. Escenario de motivación

Las técnicas de aprendizaje automático se han convertido en opciones para solucionar diversos problemas de la vida real, como la clasificación y detección de intrusos [10], virus [11], monitorización de procesos industriales [12], entre otros. Esto funciona correctamente cuando se tienen los datos de dominio a priori, donde algún algoritmo de clasificación es entrenado con su correspondiente conjunto de datos y se obtiene un modelo capaz de obtener el resultado ideal con un alto porcentaje de precisión en la mayoría de los casos. Ahora bien, cuando los datos provienen de flujos continuos, principalmente de entornos dinámicos los cuales presentan diferentes cambios a medida que van llegando los datos y si se pretenden usar con AS tradicional, se presentan los siguientes inconvenientes:

- Los clasificadores dejan de lado la inclusión de nuevos ejemplos una vez construido el modelo, lo cual genera obsolescencia en el clasificador ya que los volcanes son considerados sistemas dinámicos [13].

- La única forma de evitar que los clasificadores no sean obsoletos, es re-entrenar el/los algoritmos de AS por cada instancia nueva o cada nuevo conjunto de datos, lo cual generaría un consumo excesivo de recursos computacionales al entrenar el algoritmo con todos los datos nuevamente. Por otro lado, la demanda de tiempo no permitiría obtener una respuesta inmediata, lo cual es crucial en la vigilancia de volcanes [14] [15].

1.3. Objetivos

1.3.1. Objetivo general

Proponer un mecanismo que genere la detección de alertas pre-eruptivas volcánicas haciendo uso de técnicas de aprendizaje incremental.

1.3.2. Objetivos específicos

- Establecer alertas pre-eruptivas que soporten la monitorización volcánica.
- Definir un conjunto de datos de entrenamiento para la clasificación de alertas pre-eruptivas volcánicas.
- Adaptar un algoritmo incremental que permita clasificar alertas pre-eruptivas volcánicas.
- Evaluar experimentalmente un prototipo que implemente los algoritmos de aprendizaje incremental adaptados.

1.4. Contribuciones

Las principales contribuciones de este trabajo de maestría son:

- Seis (6) conjuntos de datos abstraídos de las técnicas de geoquímica y deformación volcánica.
- Nueve (9) conjuntos de datos organizados por la orientación de las estaciones de monitorización, que comprenden los grupos cardinales: oriental, occidental y central.
- Veintiún (21) conjuntos de datos correspondientes a siete (7) estaciones de monitorización (Cocuy3, Cráter, Agua Blanca, Cocuy2, Curiquinga, Guañarita y Lavas Rojas) durante tres distintos periodos de evaluación.
- Adaptación de un algoritmo de aprendizaje incremental que detecta en tiempo real anomalías pre-eruptivas de un volcán.
- Un prototipo llamado VOD (Volcano Outlier Detector), el cual es utilizado por los investigadores del Observatorio Vulcanológico y Sismológico de Popayán para monitorear el volcán Puracé.
- Los siguientes artículos de investigación:
    - *Monitoring of vulcano Puracé through seismic signals: Description of a real dataset.* Este artículo fue expuesto en la conferencia: “Evolving and Adaptative Intelligents Systems (EAIS)”, llevado a cabo desde el 31 de mayo al 2 de junio de 2017 en Ljubljana, Slovenia. Las memorias del evento se encuentran en IEEE Xplore. (URL: https://ieeexplore.ieee.org/document/7954838)

- Un artículo aceptado en la conferencia “17thMexican International Conference on Artifitial Intelligence – MICAI 2018”, titulado “Volcanic anomalies detection through Recursive Density Estimation”. Este artículo está incluido en los Proceedings de la LNAI categorizado en Publindex como A2. (URL: https://link.springer.com/book/10.1007/978-3-030-04491-6)
- Un artículo aceptado en la conferencia “17thMexican International Conference on Artifitial Intelligence - MICAI 2018”, titulado “Incremental versus non-incremental learning in volcano monitoring tasks: A systematic review”. Este artículo está incluido en las memorias de IEEE Xplore.

1.5. Contenido de la monografía

La monografía se encuentra estructurada a partir de las fases de la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) [16], [17], la cual ofrece una descripción del ciclo de vida para proyectos de minería de datos. Como se puede observar en la figura 1, el ciclo de vida de CRISP-DM contiene seis fases (las flechas indican las dependencias entre fases).

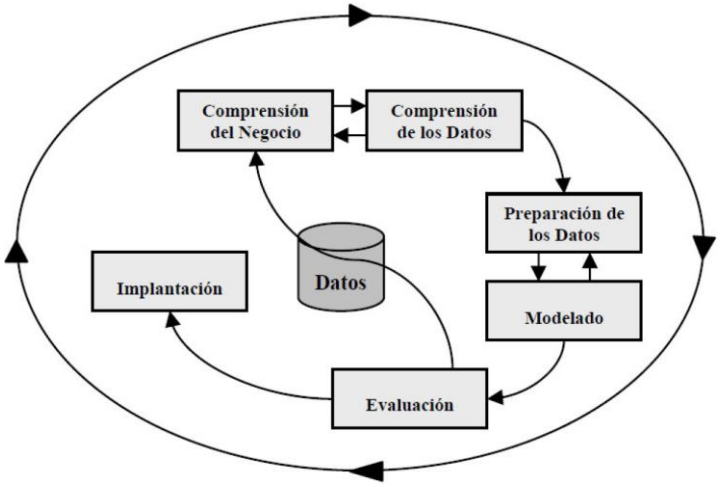


Figura 1. Fases de CRISP-DM. Tomado de [17]

Cada una de estas fases se presenta en los siguientes capítulos del presente documento como se describe a continuación:

- **Capítulo II. Estado actual del conocimiento y comprensión del negocio**

  En este capítulo se presentan los conceptos y trabajos relacionados que giran alrededor del problema de investigación declarado.

- **Capítulo III. Comprensión y preparación de los datos**

  En este apartado se explican en detalle los conjuntos de datos construidos, lo cual implica la exploración de estos mediante gráficas y tablas. Posteriormente, se aplican enfoques, métodos o técnicas para la preparación de los datos (fusión de datos de las estaciones de monitorización, eliminación o sustitución de valores en blanco o ausentes de los campos norte, este y temperatura en el caso de los inclinómetros y concentración de gas y temperatura en las estaciones monitoras de dióxido de carbono).

- **Capítulo IV. Modelado**

  Expone los datos provenientes de las estaciones de monitorización, métodos y algoritmos usados para la construcción del sistema los cuales interactúan entre sí para lograr detección de alertas volcánicas.
- **Capítulo V. Experimentación y Evaluación**

  Presenta el proceso de evaluación y las pruebas ejecutadas sobre los algoritmos utilizados en el capítulo de Modelado, con el fin de analizar los resultados y rendimiento.

- **Capítulo VI. Despliegue: Prototipo**

  Dentro de la metodología CRISP-DM se conoce como implementación. En este capítulo se presenta el proceso de desarrollo de software llevado a cabo para construir el prototipo que contiene los algoritmos de aprendizaje incremental adaptados para la detección de alertas (anomalías) pre-eruptivas.

- **Capítulo VII. Conclusiones y Trabajos Futuros**

  Finalmente, se analizan los resultados del trabajo realizado, se detallan las principales contribuciones obtenidas en la ejecución del proyecto y se expone un conjunto de recomendaciones importantes para el desarrollo de trabajos futuros.

Capítulo 2

Estado actual del conocimiento / Comprensión del negocio

En este capítulo se presentan las bases teóricas para comprender la temática del presente trabajo de maestría, el cual propone detectar alertas (anomalías) de origen volcánico basado en técnicas de aprendizaje incremental, haciendo uso de los datos adquiridos por estaciones de monitorización volcánico de deformación y geoquímica. Posteriormente, se exponen los trabajos relacionados respecto al problema de investigación declarado. Finalmente, se realiza un resumen que presenta los principales aportes de este capítulo.

2.1 Conceptos generales

2.1.1 Sistema dinámico

Un sistema dinámico se refiere al conjunto de elementos y relaciones existentes entre sí y su entorno, cuyo comportamiento depende del tiempo, donde existe acumulación de energía o información de estados anteriores que condiciona la respuesta ante los estímulos actuales [18]. Basado en la anterior definición, un volcán se considera un sistema dinámico debido a que su línea base y sus condiciones internas están en constante cambio con el paso del tiempo [13].

2.1.2 Vigilancia volcánica

El continuo transporte de magma debajo de un volcán activo produce cambios que pueden ser detectados con instrumentos tecnológicos y observaciones por parte de analistas expertos. Así, la vigilancia volcánica se encarga de la monitorización de diferentes parámetros en estas estructuras para detectar cambios en caso de ocurrir un evento. Cada volcán tiene su propio comportamiento y una monitorización permanente e independiente que permite caracterizarlo y definir líneas base en los diferentes parámetros que se vigilan: sismicidad; geoquímica de gases y deformación, las cuales son las áreas más importantes de la vigilancia volcánica y son objeto de estudio de las áreas de sismología; geoquímica y geodesia [19]–[21]. A continuación, se describe cada una de estas áreas.

Sismología

La sismología volcánica permite conocer aspectos del sistema volcánico y representa la herramienta principal para monitorizar el proceso de erupción de un determinado volcán [3], [22], [23]. A continuación, son presentados algunos de los tipos de sismos volcánicos más conocidos:

- **Vulcano Tectónicos (VT):** asociados a fracturamiento de roca a causa de la presión ejercida por el magma en el interior del volcán.
- **Largo periodo (LP):** generados por la interacción de gas y el movimiento de fluidos desde la cámara magmática hasta la superficie.
- **Tremores (TR):** señal característica de erupciones volcánicas. Describen una señal en forma de pulsos o señales de corta duración. Su origen y frecuencia es similar a los sismos LP y en su espectro de frecuencias se distingue con claridad una banda relacionada con la componente de fluidos y otra con el fenómeno de fractura.
- **Tornillos (TO):** eventos sísmicos de largo periodo (LP) que se caracterizan por presentar un decaimiento suave a través del tiempo y por tener espectros de frecuencia monocromáticos.

Entre las fuentes que originan sismos volcánicos se encuentran: movimiento de magmas, desgasificación, rompimiento de rocas por presión del magma o de los gases, explosiones de gases en el cráter, oscilaciones de columnas de gases, procesos de hundimiento en las calderas volcánicas, etc. [24].

Geoquímica

Es la ciencia geológica que estudia la química del planeta; en vulcanología, la geoquímica se aboca a la recolección y análisis de fluidos volcánicos. La composición química de las aguas o de los gases presentes en las emisiones de un volcán es un reflejo de su actividad. Los componentes principales de gases volcánicos son [25] [26]:

- **Vapor de agua (H_2O):** el vapor de agua es el gas formado cuando el agua pasa del estado líquido a gaseoso. A nivel molecular esto sucede cuando las moléculas de H_2O logran liberarse de las uniones que las mantienen juntas.

- **Dióxido de carbono (CO_2):** el dióxido de carbono (CO_2) es un gas incoloro, denso y poco reactivo. Forma parte de la composición de la tropósfera (capa de la atmósfera más próxima a la Tierra) actualmente en una proporción de 350 ppm (partes por millón). Su ciclo en la naturaleza está vinculado al oxígeno.

- **Dióxido de azufre (SO_2):** el dióxido de azufre, es un gas incoloro y altamente reactivo. Cuando se libera, el dióxido de azufre puede reaccionar con otros contaminantes del aire para formar material particulado fino, que son pequeñas partículas sólidas o líquidas suspendidas en el aire [27].

Alrededor del volcán Puracé existen alrededor de 15 campos de fuentes termales, muchos de los cuales son aprovechados turística y recreacionalmente, ya que presentan temperaturas comprendidas entre 20 - 90 ° C.

Para monitorizar un volcán activo se estudian los cambios de los anteriores componentes a través del tiempo. Los niveles base exponen las características del sistema volcánico y con repetidas mediciones se pueden identificar fluctuaciones y algunas veces pronosticar la actividad eruptiva.

Geodesia

En vulcanología, la geodesia es definida como la variación que experimentan las dimensiones de un volcán, debido a la acumulación y ascenso de magma, lo cual ejerce presión sobre la estructura del volcán, cambiando su forma tal y como se aprecia en la Figura 2 [19]. Las técnicas o metodologías aplicadas en este campo son variables y van desde las líneas de nivelación de alta precisión hasta las redes geodésicas más complejas. Otros métodos son los inclinómetros, extensómetros, mareógrafos, niveles de líquidos de vasos comunicantes o de burbuja y clinómetros de péndulo horizontal. La deformación volcánica en el Servicio Geológico Colombiano es monitorizada a través de inclinómetros electrónicos, los cuales registran los cambios en un plano de una componente axial y radial que permite determinar los movimientos de una pequeña región del edificio volcánico [28].

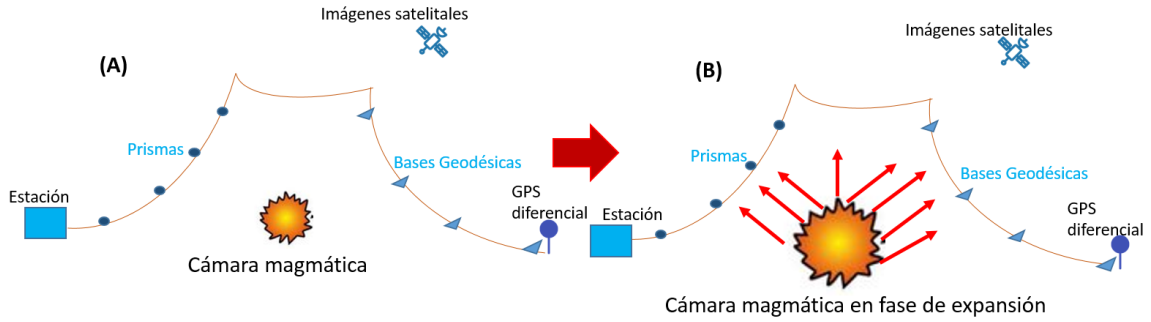


Figura 2. Esquema que ilustra la deformación de un volcán. A: edificio volcánico previo al proceso deformativo, B: edificio volcánico deformado

El interior de un volcán está constituido por una estructura geológica formada por la acumulación de material volcánico, dentro de la cual se encuentra un reservorio de diferentes elementos que se hallan en estado sólido, líquido y gaseoso, a los cuales se le denomina magma. La acumulación y el ascenso del magma hacia la superficie, genera un aumento en la presión, desencadenando el fracturamiento de rocas y el movimiento de fluidos dentro del volcán, generándose o no eventos sísmicos. Adicionalmente, la salida del magma hacia la superficie causa una presión y empuje constante en la estructura interna del volcán, deformando el edificio volcánico, fenómeno que se manifiesta con levantamientos o hundimientos de la superficie volcánica que se conocen como inflación o deflación volcánica. Por otra parte, los cambios en el estado de esfuerzos de la corteza bajo el volcán pueden ser otra causa de deformación volcánica [29].

Tipos de deformación

- **Deformación Elástica**

Consiste en la recuperación de la forma inicial del edificio volcánico una vez termine el proceso que está causando la deformación. Este tipo de deformación es constante y se manifiesta durante el tiempo en que está actuando la fuerza interna que la genera. Cuando esta fuerza deja de actuar, la deformación cesa. Esta deformación se caracteriza por la recuperación del espacio o de la forma inicial del volcán.

- **Deformación Plástica**

Consiste en una deformación permanente del edificio volcánico una vez termine el proceso que la esté causando.

2.1.3 Alertas pre - eruptivas volcánicas

Las alertas en general son mensajes cortos o notificaciones que informan acerca de la ocurrencia de un evento que potencialmente sucederá; en el contexto de prevención de desastres son indispensables para que las personas sigan instrucciones específicas, ya sea extremando las precauciones o incrementando la vigilancia hacia situaciones atípicas que puedan presentarse y así evitar cualquier contingencia [30]. En vulcanología, una alerta pre-eruptiva representa manifestaciones de cambios graduales o abruptos en un determinado volcán; estas alertas son difundidas al público mediante informes técnicos y boletines semanales, mensuales y anuales [31], [32]. Por estas razones se requiere que un volcán sea monitorizado en tiempo real y así poder alertar al personal técnico y a las comunidades de los distintos escenarios de riesgos, como lo son: las emisiones de gases volcánicos, los cuales pueden llegar a ser precursores de movimiento de fluidos y emisiones de ceniza; deformaciones considerables que pueden desencadenar flujos de lodo; y sismicidad distal, proximal, enjambres sísmicos, entre otros [33].

2.1.4 Aprendizaje incremental

Los flujos de datos en línea (real-time dataflow) y los entornos dinámicos son dos líneas de investigación comúnmente usadas en muchas de las aplicaciones actuales [34]–[37]. A diario se generan gran cantidad de flujos de datos procedentes de diferentes dispositivos, sensores, internet, etc. Estos flujos de datos son almacenados en bases de datos y generalmente no siguen distribuciones de datos uniformes (describen un comportamiento estocástico), ya que están influenciados por parámetros estadísticos que cambian rápidamente, dependiendo del fenómeno en estudio [36]. Por estas razones, el uso de técnicas tradicionales de aprendizaje

supervisado para clasificación y agrupamiento genera problemas de obsolescencia, cuya única solución es el re-entrenamiento constante de estos sistemas. Sin embargo, esta actualización se convierte en una tarea muy costosa e ineficiente por la gran cantidad de datos que son necesarios procesar.

Para poder afrontar esta necesidad en los sistemas de aprendizaje, surgió el aprendizaje incremental como un paradigma del aprendizaje supervisado donde el proceso de aprendizaje se lleva a cabo siempre que exista una nueva instancia (también llamada ejemplo), ajustando el proceso de acuerdo con lo que estos nuevos ejemplos aportan. Así, la diferencia más importante del aprendizaje supervisado no incremental respecto al incremental es que el primero no considera el proceso de entrenamiento como una tarea continua.

Matemáticamente, el aprendizaje incremental se define del siguiente modo [38]:

Sea $T = \{(x^{\rightarrow}, y)\}: y = f(x^{\rightarrow})\}$, el conjunto de ejemplos de entrenamiento disponibles en un instante $t \in < 1,2,3, \ldots >$. Un algoritmo de aprendizaje se dice incremental si a partir de una secuencia de $< T_1, T_{2,\ldots..} T_i >$ produce una secuencia de hipótesis $\mathrm{H} = < H_1, H_{2,\ldots..} H_i >$, donde la hipótesis actual H_i es función de la hipótesis anterior H_{i-1} y el conjunto de ejemplos leídos en el instante actual T_i.

En la actualidad existen diversos algoritmos que permiten implementar un enfoque de aprendizaje incremental. Algunos de ellos tienen como base los algoritmos de aprendizaje supervisado (no incremental), tales como ISVM (Incremental Super Vector Machine) [39] cuya base son las máquinas de vector soporte, Learn++ [15] fundamentado en las redes neuronales artificiales, CVFDT (Concept- Adapting Very Fast Decision Tree) [40] basado en árboles de decisión, entre otros. Por otro lado, se encuentran los métodos incrementales basados en la densidad de datos usando un enfoque recursivo, los cuales se pueden adaptar al problema propuesto debido a que no necesitan almacenar todos los datos y no imponen limitaciones de memoria. Estos enfoques han sido objeto de estudios en los últimos años y fueron introducidos en el año 2001 [41], [42] y patentados en 2016 [43].

Teniendo en cuenta que el dominio de aplicación es un sistema dinámico, donde las estaciones de monitorización obtienen un dato cada minuto y que a lo largo del tiempo son generados diversos cambios abruptos y graduales, se determina que los algoritmos incrementales basados en los enfoques de aprendizaje supervisado no representan la mejor solución al problema declarado en el presente trabajo de maestría, ya que estos tienen problemas para detectar los tipos de cambios o algunos de ellos no son sensibles al ruido. Además, estos algoritmos tienen problemas al detectar cambios abruptos en cortos periodos de tiempo, ya que es necesario ajustar nuevamente la función matemática que representa los modelos [44]. Por este motivo, la presente tesis se enfoca en el uso de los algoritmos de estimación de la densidad de forma adaptativa, que son la base de los sistemas auto-adaptativos o su término en inglés *Evolving Systems*.

Sistemas auto-adaptativos

Los sistemas auto-adaptativos o Evolving Systems (ES) [45]–[47] son una iniciativa reciente para abordar problemas conceptuales de flujos de datos continuos. Nacen bajo las necesidades de los sistemas incrementales (grandes cantidades de datos continuos, flujos de datos en tiempo real, detección de cambios en una secuencia de datos).

A diferencia del aprendizaje incremental tradicional, los ES poseen una estructura capaz de modificarse en función de los datos recibidos ya que no mantienen una estructura fija. Así, los ES son capaces de modificar su estructura, su funcionalidad y su representación interna del conocimiento para adaptarse a los cambios en el entorno. El núcleo de este tipo de enfoques son los sistemas basados en estimaciones de densidad. Las estimaciones de densidad, se enfocan en la detección de anomalías mediante algoritmos que requieren identificar la Función de Densidad de Probabilidad (FDP) calculada mediante la proximidad espacial entre un ejemplo y el resto de los ejemplos representados en un determinado espacio de características. Algunos de los algoritmos más populares para la detección de anomalías basados en el cálculo de estimaciones de densidad son LOF (Local Outlier Factor)[48] y RDE (Recursive Density Estimation). Por otro lado, cabe mencionar que también existen otros algoritmos tales como ABOD (Angle-Based Outlier Detection in High-dimensional Data) [49] que también permiten detectar anomalías basándose en la distancia entre ángulos. Estos algoritmos son explicados en detalle en el Capítulo 4 (Modelado).

2.2 Trabajos relacionados

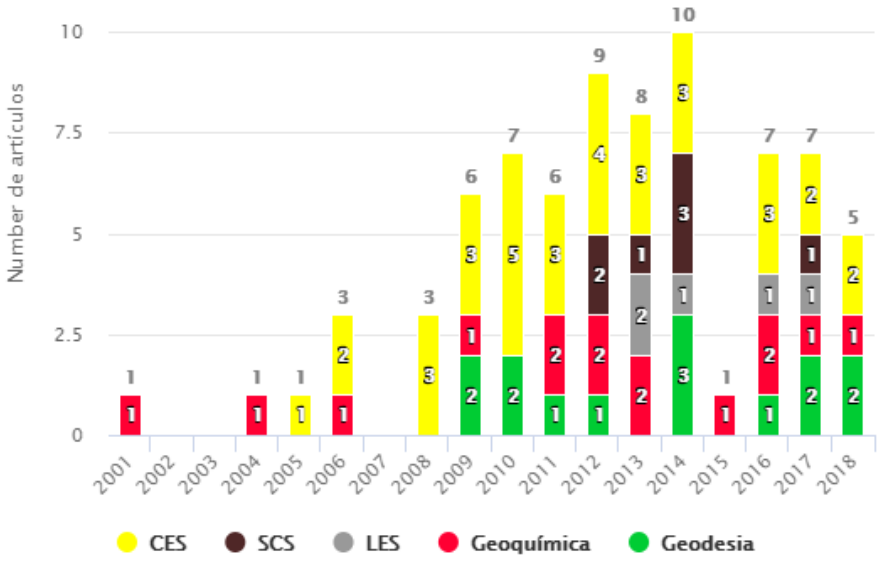
En esta sección, inicialmente se presenta una revisión literaria de las técnicas de aprendizaje no-incrementales e incrementales aplicadas en vulcanología. En este sentido, se revisaron documentos como artículos de investigación y trabajos de grado, haciendo uso de las bases de datos bibliográficas: IEEE Xplore (https://ieeexplore.ieee.org), Springer (https://www.springer.com) y Google Scholar (https://scholar.google.es/).

Los resultados de la revisión literaria indican que en la actualidad los problemas de origen volcánico son objeto de estudio de aprendizaje supervisado no incremental. Por este motivo, esta revisión fue dividida entre aprendizaje supervisado no incremental en vulcanología y aprendizaje incremental + sistemas auto-adaptativos en otros dominios de investigación. La totalidad de las figuras que abarcan la revisión sistemática se encuentran en el anexo E.

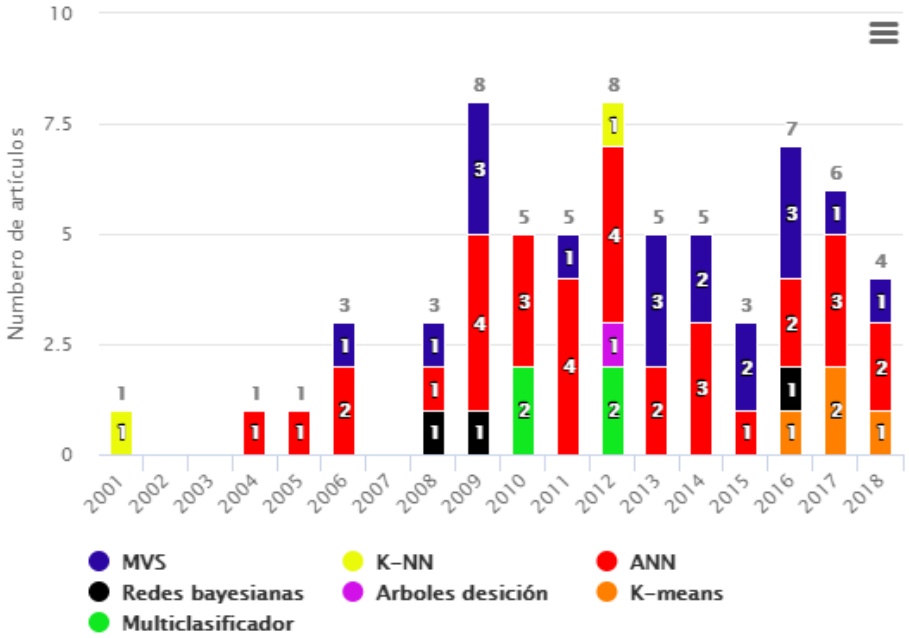
La figura 3 (a) explica las principales áreas aplicadas a la monitorización de volcanes haciendo uso de aprendizaje supervisado tradicional entre los años 2001 y 2018

- Clasificación de eventos sísmicos volcánicos o tectónicos (CES),
- Selección de las características más relevantes de las señales sísmicas (SCS),
- Localización de eventos sísmicos donde se encuentra su hipocentro y epicentro (LES),
- Geoquímica: emisión de gases volcánicos.
- Geodesia: deformación del subsuelo y cono volcánico.

Por otra parte, la figura 3 (b) presenta los principales algoritmos de aprendizaje supervisado tradicional usados en vulcanología. Los resultados de la revisión literaria indican que la clasificación automática de eventos sísmicos presenta la mayor cantidad de trabajos con 32 estudios, seguido de trabajos de investigación del de área de geoquímica (14 estudios). Adicionalmente, se observa que las redes neuronales son los enfoques más usados para vulcanología en aprendizaje supervisado.



3 (a) Número de estudios en sub áreas de monitorización volcánica



3 (b) Número de estudios clasificados por aprendizaje supervisado

Figura 3. Aprendizaje no incremental

Como se mencionó anteriormente, de acuerdo con la revisión literaria desarrollada, hasta la fecha no se han encontrado trabajos de investigación aplicados a problemas de vulcanología con algoritmos aprendizaje incremental. Sin embargo, diferentes algoritmos incrementales han sido utilizados en diversos dominios de aplicación como lo presenta la Figura 4a. Así mismo, la Figura 4b presenta los algoritmos de aprendizaje incremental usados en los trabajos de investigación durante los años 2000 al 2018.

Los resultados exponen que la visión artificial es el área con mayor número de estudios (18 trabajos de investigación), especialmente, en el reconocimiento facial, teniendo en cuenta que los gestos faciales cambian a medida que transcurre el tiempo. Desde la perspectiva del uso frecuente de algoritmos incrementales tradicionales, se tiene que el algoritmo “Incremental Super Vector Machine” (ISVM) es el más utilizado (27 trabajos de investigación), seguido por otros algoritmos tales como PECS, FACIL, IADEM Y SEA (15) cuyo incremento en investigaciones ha venido desde comienzos del 2006 hasta el presente año.

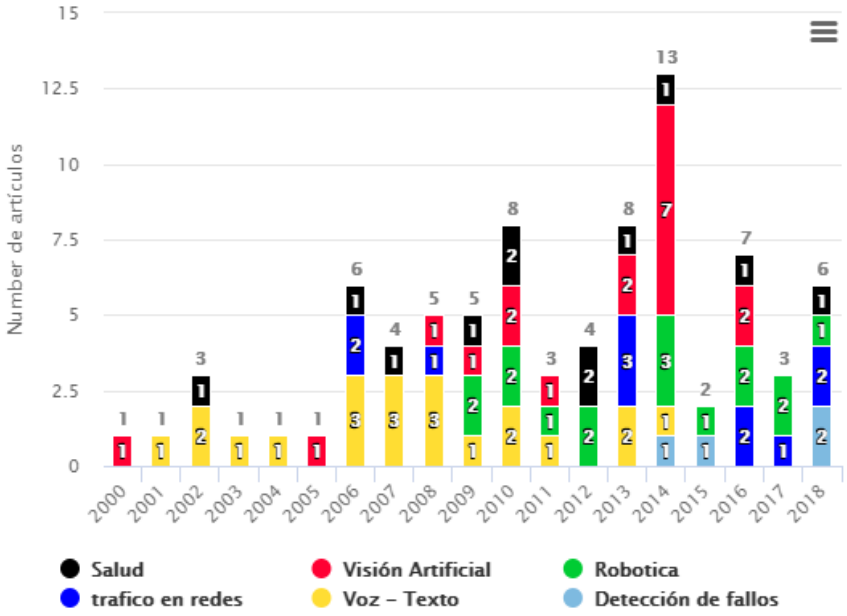


Figura 4 (a). Principales dominios de aplicación donde el aprendizaje incremental es aplicado

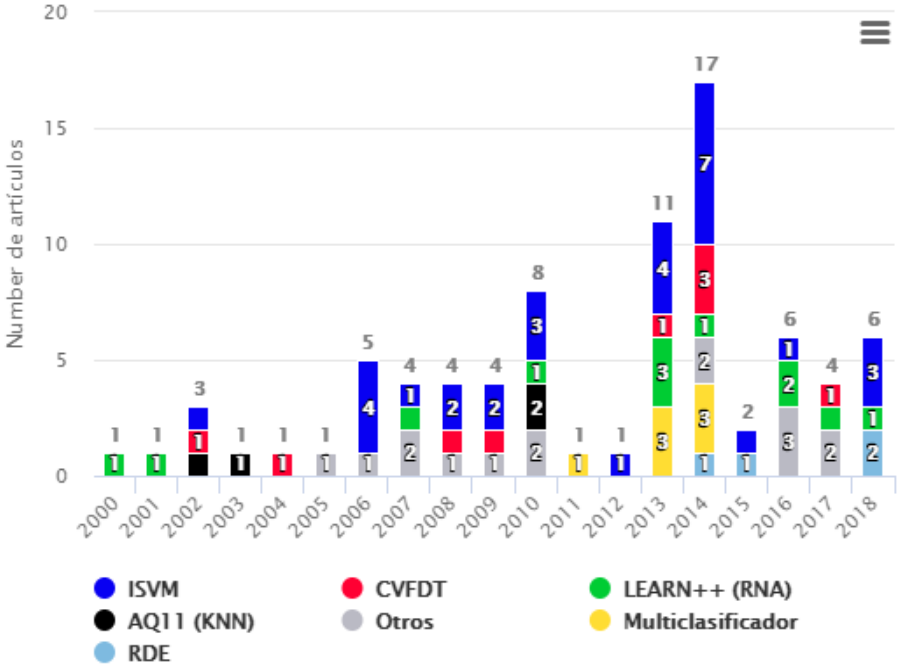


Figura 4 (b). Algoritmos de aprendizaje incremental aplicados

Figura 4. Aprendizaje incremental

En resumen, la revisión literaria presentada anteriormente, permitió identificar el uso de algoritmos de aprendizaje incremental, más precisamente, el enfoque de sistemas auto-adaptativos adecuado para abordar el problema propuesto en el presente trabajo de maestría. Así, esta búsqueda permitió verificar que hasta el momento no se han encontrado trabajos para la detección de alertas (anomalías) pre-eruptivas volcánicas usando algoritmos auto-adaptativos. A continuación, se presentan estudios en los cuales han sido implementados ese tipo de algoritmos en diferentes dominios de aplicación.

2.2.1 Sistemas auto - adaptativos para la detección de anomalías

El reconocimiento de la actividad humana ha sido un campo de investigación de gran interés en las últimas dos décadas. Sin embargo, pocos investigadores han considerado el aspecto dinámico o evolutivo de las actividades humanas, pues la manera en que una persona ejecuta determinada actividad puede cambiar y evolucionar al pasar el tiempo debido a sus hábitos, habilidad para realizar determinada actividad, envejecimiento natural o aparición de alguna enfermedad y a su estado de humor. Por esa razón, en [50] es propuesto un enfoque basado en Evolving Fuzzy Systems llamado Evolving Activities of Daily Living Classifier (EvAClass). El enfoque parte de la premisa de que los flujos de datos provenientes de diferentes sensores ubicados en un hogar inteligente pueden ser transformados en secuencias ordenadas de datos útiles para representar cada Actividad de la Vida Diaria (ADL). EvAClass crea y actualiza desde cero una librería que contiene los prototipos de cada ADL (los modelos de una ADL representada por una distribución de subsecuencias). Para clasificar una nueva instancia, EvAClass calcula la similitud de esa nueva instancia con los prototipos existentes en la librería por medio de la similitud de cosenos y simultáneamente calcula el potencial que tiene esa nueva instancia para ser incluida en la librería usando una función de densidad que relaciona la distancia acumulada entre esa instancia y todas las que componen un prototipo. Cuando es necesario incluir una nueva instancia en la librería significa que un cambio (anomalía) ha sido detectado, lo que da lugar a la creación de alertas o recomendaciones según el contexto de aplicación, por ejemplo, en el contexto del cuidado de adultos mayores, podría significar un deterioro en su capacidad motora. La evaluación de este enfoque fue realizada comparando su nivel de precisión con otros enfoques basados en aprendizaje automático incremental sobre un dataset de actividad humana. Como resultado obtuvieron que EvAClass supera la precisión de los demás enfoques incrementales con precisión de 94.2%.

En [51] es presentado un sistema para la clasificación del comportamiento de usuarios del sistema UNIX, este sistema usa el enfoque eClass, el cual es un algoritmo de clasificación incremental basado en Evolving Fuzzy Systems y funciona de la siguiente manera: el primer paso para la clasificación es la creación de un perfil de usuario basado en la frecuencia de comandos, es decir, en el número de veces que el usuario escribe un comando UNIX en su teclado. Una vez la frecuencia para cada comando ha sido calculada, el modelo del usuario es representado por un vector de comandos que contiene la distribución de esos valores. Cada modelo está compuesto por varias reglas difusas para cada clase (que representan un tipo usuario). Las reglas son creadas desde cero usando un enfoque de clustering evolutivo que decide cuando crear nuevas reglas. Este enfoque es el mismo que se describió en el estudio anterior, en el cual se calcula la similitud de cada nueva instancia (en este caso una secuencia de comandos) y su potencial de convertirse en un prototipo de un modelo. De esta manera, por medio de la detección de pequeñas similitudes (anomalías), es posible la detección de intrusos que intentan usar la computadora de un determinado usuario. La evaluación de esta propuesta fue realizada por medio de la comparación del porcentaje de instancias correctamente clasificadas (exactitud) dentro de su correspondiente usuario usando algoritmos convencionales basados en aprendizaje automático tales como C4.5, PART, Nearest Neighbor, Naive Bayes y

Support Vector Machine. Aunque la exactitud de eClass fue alrededor de 89% y la de SVM y Naive Bayes estuvo cerca al 95%, a diferencia de estos dos últimos métodos, eClass no necesita almacenar grandes flujos de datos en memoria y mantiene siempre los modelos actualizados.

Denis Kolev y otros autores [52] presentan un método para la detección de anomalías/averías durante un vuelo. En primer lugar, es realizada la identificación de las fases y sub-fases de vuelo por medio del método Evolving Clustering (eClustering), el cual obtiene el número de sub-fases del vuelo sin el ajuste de ningún tipo de parámetro. En segundo lugar, es utilizado el concepto de Recursive Density Estimation (RDE) y eClass para detectar anomalías o fallos en las diferentes fases de vuelo con base en los clusters creados en el paso anterior. RDE es basado en la función de Cauchy la cual tiene propiedades similares que la función Gaussiana, pero a diferencia de esta última, esta puede ser actualizada recursivamente. La evaluación de este método consistió en usar dos datasets con el fin de obtener el número de falsos positivos y falsos negativos con respecto a los resultados de un sistema convencional para el análisis de datos del vuelo: un dataset contiene datos de 38 vuelos realizados por tres aeronaves TUpolev-204 y el otro contiene datos de 175 vuelos hechos por un avión Boeing-737. Los resultados indican que RDE obtiene buenos resultados para el primer dataset (6.81% de falsos negativos y 5.21% de falsos positivos), mientras que para el segundo dataset son obtenidos mejores resultados con eClass (8.25% falsos negativos y 2.51% falsos positivos). La hipótesis planteada por los autores indica que los resultados obtenidos en el primer dataset se deben a que la mayoría de las anomalías eran técnicas, mientras que para el segundo dataset las anomalías correspondían a errores de los pilotos.

Por otra parte, desde el análisis de redes sociales, en [53] es presentado un enfoque para crear modelos de comportamiento de usuario con base en las propiedades de los perfiles de usuario de comunidades específicas en Twitter. El enfoque es basado en Evolving Fuzzy Systems, lo que posibilita no solo el análisis de gran cantidad de perfiles de usuario en tiempo real sino obtener un conocimiento de los perfiles de usuario actualizado. Luego de obtener las propiedades del perfil de usuario por medio de la API de Twitter (por ejemplo: número de seguidores, número de twits escritos, colores del perfil, etc), son obtenidos los modelos de usuario, cada uno de ellos representados por un conjunto de reglas difusas. Los modelos son obtenidos por medio del uso del algoritmo eClustering y la detección de outliers (perfiles de usuario que se relacionan con los modelos creados) es realizada por el algoritmo RDE.

Para este proyecto, ha sido seleccionado el uso de los sistemas auto-adaptativos puesto que las técnicas tradicionales de aprendizaje incremental son adecuadas para representar sistemas que sufren pequeños cambios en su estructura, sin embargo, para el manejo de sistemas complejos con múltiples modos de operación o cambios drásticos en sus características como son los eventos de origen volcánico, estas técnicas convencionales suelen necesitar mucho tiempo para aprender los nuevos parámetros del modelo. Por su parte, los sistemas auto-adaptativos como se ha mencionado anteriormente, son capaces de modificar tanto sus parámetros como su estructura en función de los datos recibidos, permitiendo

que estos puedan desarrollarse y aprender por sí mismos. De esta forma, la estructura se adapta al entorno dinámico de los volcanes conforme a su variación.

2.3 Resumen

Con el fin de comprender la temática del presente trabajo de maestría, este capítulo presentó los conceptos teóricos relacionados con la vigilancia volcánica, tales como productos volcánicos, sismología, geoquímica y geodesia. Seguidamente, se explicaron los tipos de alertas pre-eruptivas volcánicas y la definición formal de aprendizaje incremental tradicional y sistemas auto-adaptativos. Finalmente, fueron expuestos los trabajos relacionados respecto al problema declarado mediante una revisión literaria divida entre aprendizaje no incremental en dominios de vulcanología y aprendizaje incremental incluyendo el enfoque de sistemas auto - adaptativos en la detección de anomalías en tiempo real. Como conclusión, serán utilizados los sistemas auto-adaptativos, ya que estos permiten adaptarse rápidamente a cualquier tipo de cambio gradual o abrupto producido por eventos volcánicos.

Capítulo 3

Comprensión y preparación de los datos

En este capítulo se describe el proceso de comprensión de datos, que abarca el área vulcanológica estudiada; los instrumentos de monitorización vulcanológica y su ubicación; además de la descripción del conjunto de datos utilizado. En la segunda parte de este capítulo se presenta el proceso de preparación de datos (pre-procesamiento), con el objetivo de adaptar el conjunto de datos a los modelos auto-adaptativos que serán explicados en el Capítulo 4.

3.1 Comprensión de los datos

A continuación, se explica en detalle el área vulcanológica de estudio, los instrumentos de monitorización y los datos utilizados en el presente trabajo de maestría.

3.1.1 Área de estudio: volcán Puracé

El área de estudio del presente trabajo de maestría es el volcán Puracé, el cual se encuentra ubicado en el departamento del Cauca (ubicación geográfica: latitud 2.32°N, longitud 76.40°W). Este volcán hace parte de la cadena volcánica de “los Coconucos”, la cual está compuesta por 15 centros eruptivos alineados con una orientación al noroeste de 40º, siendo el Puracé el más septentrional de la cadena [54], como lo presenta la Figura 5.

Figura 5a. Volcán Puracé. Imagen captada durante sobrevuelo del 22 de octubre de 2011

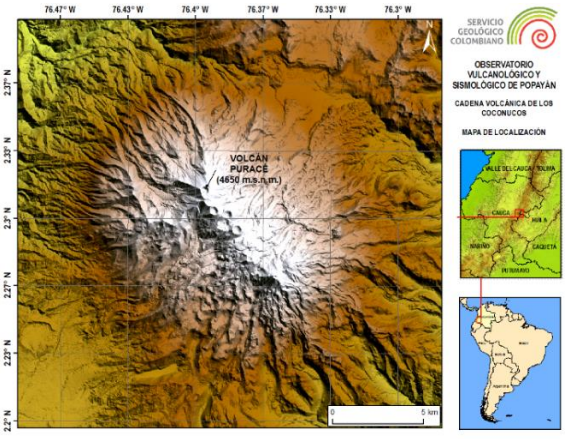


Figura 5b. Mapa de localización del volcán Puracé (Imagen suministrada por el OVSPo).

Figura 5. Ubicación volcán Puracé

El volcán Puracé es monitorizado por el Observatorio Vulcanológico y Sismológico de Popayán desde el año 1993 hasta la fecha. Desde que se ha monitorizado, se han presentado cambios importantes en la actividad sísmica que produjeron fractura en las fuentes distales al noreste de la fuente sismogénica de la vereda San Rafael, como la ocurrida a finales del mes de diciembre del año 2015. En marzo del año 2013, debido a la actividad superficial, se desencadenó un cambio de nivel por deslizamientos de tierra cerca de la mina de azufre cercana al volcán, lo que ocasionó un colapso de un tramo de la mina. Además, en febrero del año 2016 se registraron cambios significativos en deformación y geoquímica, lo que ocasionó un enjambre sísmico días más tarde, tal y como se observa en el informe de actividad de ese mismo año [55]–[57]. Es importante mencionar que la última erupción conocida del volcán Puracé ocurrió en el año 1977

3.1.2 Red de vigilancia del volcán Puracé

El volcán Puracé cuenta actualmente con una red de vigilancia compuesta por 43 estaciones telemétricas y 59 no telemétricas. Estas estaciones se emplean para medir diferentes parámetros de vigilancia como sismología, geodesia, geofísica, geoquímica, climatología y la actividad superficial del volcán. Los datos utilizados en el presente trabajo de maestría provienen de siete (7) estaciones telemétricas que cuentan con el mantenimiento adecuado para la monitorización volcánica de las áreas de geoquímica y deformación volcánica.

Para analizar los procesos geoquímicos, se usaron los datos de dos (2) estaciones de dióxido de carbono (CO_2), mientras que para realizar la monitorización de los procesos de deformación se utilizaron los datos de cinco (5) inclinómetros electrónicos. A continuación, se describen en detalle los instrumentos de medición utilizados:

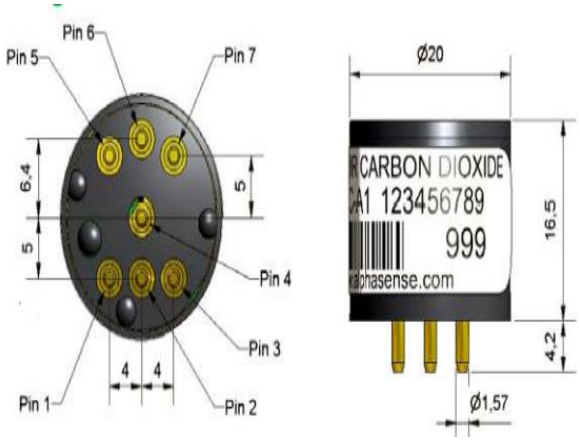
Instrumentos de medición: geoquímica

Las mediciones geoquímicas en el volcán Puracé se realizan en toda el área de influencia de la Cadena Volcánica de “los Coconucos” y tienen como objetivo obtener datos que permitan detectar variaciones en la temperatura y la concentración del dióxido de carbono (CO_2) [58].

El dióxido de carbono (CO_2) es un gas incoloro e inholoro, ligeramente soluble en agua, y tiene una densidad de 1.799 g/L [59]. Su origen en sistemas geotermales se debe a las reacciones químicas derivadas de la actividad magmática que se dan en algunos minerales, en rocas carbonatadas y sedimentarias no carbonatadas y en la materia orgánica presente en sedimentos, la cual representa la principal fuente de emisión en sistemas volcánicos [60]. Debido a esta relación con los procesos volcánicos, y con el fin de comprender la interacción en la superficie, este gas es monitoreado en el volcán Puracé. En las estaciones de muestreo de gases CO_2 se recolecta información de la concentración (flujo) de dióxido de carbono durante un minuto, la cual se registra mediante la unidad de medida de pico curios sobre litros (pCi/L) y temperatura interna del instrumento en grados centígrados

(°C). La Figura 6 presenta una estación que mide el flujo de gas CO_2 emanado por la estación en un determinado tiempo y la temperatura interna del instrumento.

6ª. Estación real de monitorización de gas dióxido de carbono



6b. funcionamiento interno de estación monitora de dióxido de carbono

Figura 6. Estación monitora de dióxido de carbono

En este sentido, para el presente trabajo de maestría, se utilizaron datos de dos (2) estaciones de geoquímica las cuales están compuestas a su vez por dos (2) atributos como se expone en la Tabla 1.

Estación		**Concentración de gas (flujo)**	**Temperatura**
1	Cráter	Cra_C	Cra_T
2	Cocuy3	Co3_C	Co3_T

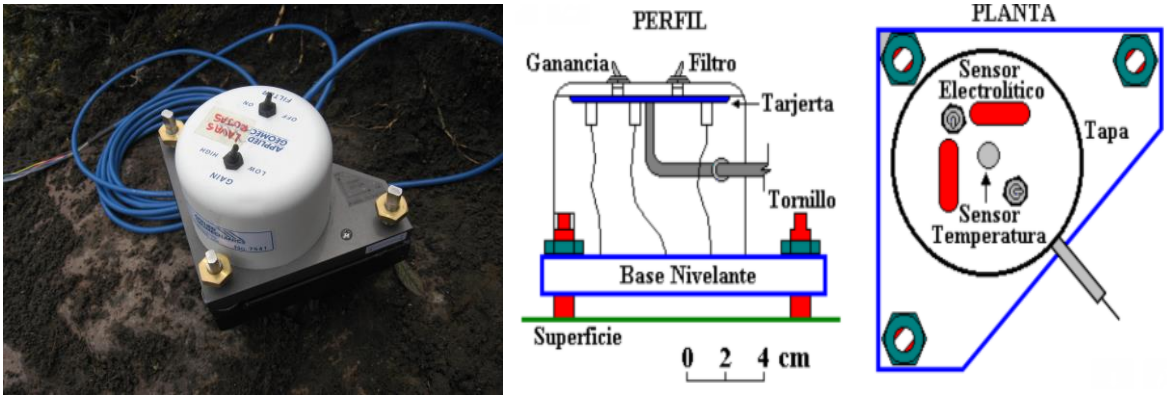
Tabla 1. Atributos de dióxido de carbono

Instrumentos de medición: deformación

La monitorización en deformación tiene como objetivo estudiar los cambios de la forma de la superficie del volcán, para determinar las variaciones de las dimensiones del edificio volcánico (inflación/deflación) debido a la interacción de un cuerpo magmático. La monitorización en deformación volcánica permite cuantificar y cualificar las deformaciones volcánicas, mediante el procesamiento, sistematización, análisis e interpretación de los datos tomados en campo y adquiridos telemétricamente para determinar las características de la fuente causante de la deformación [54].

Para medir los cambios horizontales y verticales que ocurren en un volcán, se emplean métodos geodésicos de alta precisión y métodos electrónicos de alta sensibilidad, ya que las deformaciones volcánicas son del orden de los milímetros, décimas y centésimas de milímetro y en micro radianes.

La monitorización de la deformación en el OVSPo se lleva a cabo a través de un inclinómetro electrónico, el cual es un instrumento que mide las inclinaciones de la superficie del volcán (en micro radianes) por medio de una plataforma de nivelación triangular, permitiendo medir los cambios de una pendiente sobre la cual tiene dos sensores de inclinación (de niveles electrolíticos) orientados ortogonalmente (norte–sur y este–oeste). La mayoría de las estaciones de inclinometría cuentan con un sensor de temperatura para control de cambios térmicos [29]. En la Figura 7 se presenta un inclinómetro, el cual hace parte de una estación de deformación perteneciente a la red de vigilancia del OVSPo. En el inclinómetro se incorporan dos (2) sensores de inclinación paralelos a lados del triángulo de la plataforma y un sensor de temperatura.



7ª. Inclinómetro electrónico estación OVSPo *7b. Funcionamiento interno de Inclinómetro electrónico*

Figura 7. Inclinómetro electrónico (Fuente: OVSPo)

De esta forma, se utilizaron los datos de cinco (5) inclinómetros que monitorizan el volcán Puracé. La Tabla 2 presenta las variables medidas por cada estación de inclinometría: componente norte (N), este (E) y temperatura interna. Los dos primeros parámetros son obtenidos en micro radianes, mientras que la temperatura es obtenida en grados centígrados.

Estación		Componente N	Componente E	temperatura
1	Curiquinga	CUR_N	CUR_E	CUR_T
2	Cocuy2	CO2_N	CO2_E	CO2_T
3	Agua Blanca	ABL_N	ABL_E	ABL_T
4	Guañarita	GUA_N	GUA_E	GUA_T
5	Lavas Rojas	LAR_N	LAR_E	LAR_T

Tabla 2. Atributos de los inclinómetros

Ubicación de los instrumentos de medición

En la Figura 8 se presenta el mapa con la ubicación cartográfica de las estaciones usadas en este trabajo de maestría, en donde se identifican con círculos de color amarillo las estaciones de dióxido de carbono (geoquímica) y las estaciones de inclinometría (deformación) mediante triángulos azules.

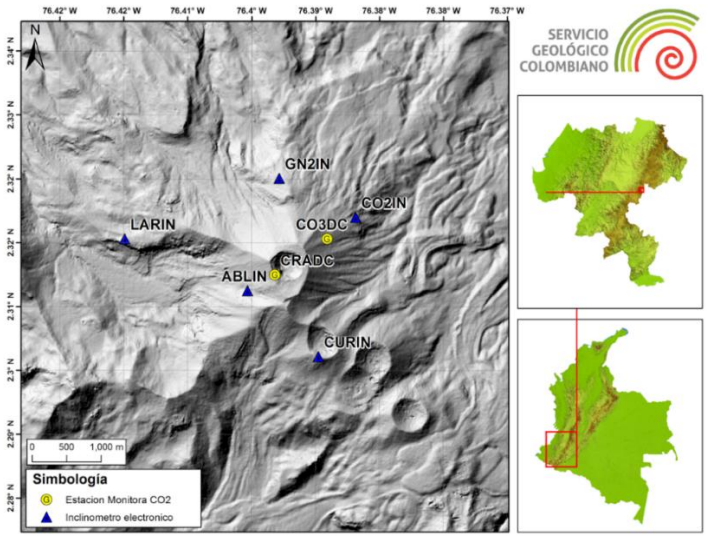


Figura 8. Mapa de localización de estaciones monitoras CO2 e inclinómetros electrónicos. Fuente OVSPo

Los códigos de cada una de las estaciones son representados de la siguiente manera: Cráter (CRADC), Cocuy (CO3DC), Guañarita (GN2IN), Lavas Rojas (LARIN), Curiquinga (CURIN), Cocuy (CO2IN) y Agua Blanca (ABLIN).

3.1.3 Descripción del conjunto de datos

El conjunto de datos construido consta de 388.000 instancias, correspondientes a la monitorización del volcán Puracé desde el **2017-01-01 hasta 2017-09-30**. Los datos son obtenidos por los instrumentos explicados en las secciones 3.2.1.1 y 3.2.1.2, pertenecientes a las áreas de geoquímica y deformación volcánica con una frecuencia de muestreo de un (1) dato por minuto. El periodo de tiempo del conjunto de datos fue elegido teniendo en cuenta que en los tres primeros semestres del año 2017 se presentaron diversos tipos de eventos volcánicos en el volcán Puracé que involucraban directamente las áreas de monitorización de geoquímica, deformación y sismología.

La tabla 3 presenta una descripción general del conjunto de datos utilizado.

Estación	Vigilancia	Atributos Numéricos	Valores perdidos
Cocuy3	Geoquímica	2	4.3 %
Cráter	Geoquímica	2	5.1 %
Agua Blanca	Deformación	3	2.9 %
Curiquinga	Deformación	3	0.3 %
Guañarita	Deformación	3	2.8 %
Cocuy2	Deformación	3	5.2 %
Lavas Rojas	Deformación	3	4.7 %

Tabla 3. Descripción general del conjunto de datos

Una vez analizado el conjunto de datos, se procede a realizar la preparación de estos, como se explica a continuación.

3.2 Preparación de los datos

En esta sección se presentan las técnicas utilizadas para el procesamiento de los datos. Para realizar la correcta preparación de los datos, seguimos el enfoque propuesto en [61], [62]. En este orden de ideas, se realizó un tratamiento de valores perdidos y detección de valores erróneos sobre el conjunto de datos.

3.2.1 Tratamiento de valores perdidos

En el conjunto de datos utilizado se ha identificado estaciones de vigilancia volcánica con problemas de transmisión de datos en un determinado instante de tiempo, lo cual genera valores perdidos. Las Tablas 4 y 5 presentan un resumen de valores perdidos encontrados en las variables que componen las estaciones de geoquímica y deformación respectivamente.

Estación	Atributo	Valores perdidos (%)
Cráter	Flujo de gas (concentración)	3.1%
	Temperatura interna	7.1%
Cocuy3	Flujo de gas (concentración)	5.5%
	Temperatura interna	3.1%

Tabla 4. Valores perdidos en las estaciones de geoquímica

Estación	Atributo	Valores perdidos (%)
Agua Blanca	Componente Norte	3.4%
	Componente Este	1.1%
	Temperatura	4.2%
Curiquinga	Componente Norte	1.7%
	Componente Este	0%
	Temperatura	0.9%
Guañarita	Componente Norte	0.8%
	Componente Este	2.3%
	Temperatura	4.0%
Cocuy2	Componente Norte	4.7%
	Componente Este	3.8%
	Temperatura	7.1%
Lavas Rojas	Componente Norte	1.9%
	Componente Este	0.9%
	Temperatura	12.3%

Tabla 5. Valores perdidos en las estaciones de deformación

Durante las últimas décadas se han propuesto distintas metodologías para sustituir datos faltantes, los más populares son eliminaciones en forma de listas (listwise) o de pares (pairwise) [63]. Actualmente Listwise Deletion (LD) es el método que se

usa con mayor frecuencia para la imputación de datos. Este realiza un análisis de datos completos omitiendo las instancias que contengan al menos un valor perdido, es decir, toma en cuenta únicamente las observaciones que disponen de información completa para todas las variables. Por su parte, Pairwise Delection utiliza distintos tamaños de una muestra correlacionado los atributos para obtener una información más completa. La elección de un método sobre el otro depende exclusivamente del dominio de aplicación. En consecuencia, para el caso de la detección de alertas de origen volcánico, los valores perdidos han sido tratados mediante el método LD, ya que la decisión de evaluar una anomalía como verdadera o falsa depende de todas las variables de un registro emitido por una estación.

3.2.2. Detección de valores erróneos

Las estaciones de vigilancia volcánica en algunas ocasiones presentan fallas, transmitiendo datos alfanuméricos que cambian completamente el significado de los valores. Estos valores alfanuméricos generan comportamientos erróneos en los modelos auto-adaptativos que no reflejan la realidad de los datos, por lo cual deben ser detectados y eliminados.
Los valores erróneos fueron detectados mediante el uso de diagramas de caja [64] y los umbrales de Tukey [65]. Los diagramas de cajas son gráficos basados en cuartiles y describen características importantes tales como dispersión y simetría. Para su realización se representan tres cuartiles y los valores máximo y mínimo. Una gráfica de este tipo consiste en una caja rectangular donde los lados más alargados representan el rango intercuartílico (RI). El rectángulo está dividido por un segmento vertical que indica dónde se posiciona la mediana (segundo cuartil: Q2) y su relación con el primer y tercer cuartil, es decir Q1 y Q3 respectivamente. Los diagramas de cajas presentan además información sobre la tendencia central, dispersión y simetría de los datos en estudio permitiendo identificar observaciones que se alejan del resto de los datos. Estos valores representan los datos erróneos emitidos por las estaciones. El primer cuartil (Q1) es un valor en el conjunto de datos que contiene el 25% de los valores que se encuentran debajo de él. El tercer cuartil (Q3) es el valor que sobrepasa al 75% de los valores de la distribución. Finalmente, es importante mencionar que un rango intercuartílico (RI) se define como la diferencia entre el cuartil tres y el cuartil uno (Q3 - Q1).

La figura 9 representa un ejemplo de un diagrama de cajas en la cual se pueden visualizar los tres cuartiles y los valores máximo y mínimos de un diagrama de caja.

Figura 9. Diagrama de cajas

El cálculo de cada cuartil es calculado como se expone en la ecuación 1. El valor de “i” en la ecuación representa el número del cuartil a calcular y “n” es el número

de instancias del conjunto de datos. Tener en cuenta que los datos deben ser ordenados antes de proceder con los cálculos.

$$Q_i = \frac{i\ x\ n}{4}\ ;\quad i\ = 1, 2, 3$$

Ecuación 1. Calculo de cuartiles

Por su parte, los umbrales de Tukey así como el rango intercuartílico pertenecen a uno de los métodos usados para determinar valores atípicos en una presentación dada (en este proyecto, se usan para eliminar valores erróneos). Técnicamente, los umbrales de Tukey son los valores inferiores a Q1 – 1.5 (Q3 – Q1) o los valores superiores a Q3 + 1.5 (Q3 – Q1).

La Figura 10, presenta un ejemplo del diagrama de caja para la variable flujo de gas (concentración) de la estación monitora de gas CO_2.

Ejemplo	1	2	3	4	5	6	7	8	9	10	11	12	13
Dato	0.104	0.423	0.431	0.45	0.475	0.481	0.486	0.534	0.535	0.548	0.57	0.786	0.816

Figura 10. Ejemplo de diagrama de caja CO2

Los valores de los cuartiles son presentados a continuación:

$$Q_1 = \frac{1\ x\ 13}{4} = 3.25\ ;\ Q_2 = \frac{2\ x\ 13}{4} = 6.5\ ;\quad Q_3 = \frac{3\ x\ 13}{4} = 9.75\ ;$$

El valor real de cada cuartil se encuentra promediando los valores izquierda – derecha de cada índice de cada cuartil como se expone a continuación:

$$q_1 = \frac{0.431 + 0.45}{2} = 0.440;\ q_2 = \frac{0.481 + 0.486}{2} = 0.483\ ;\quad q_3 = \frac{0.535 + 0.548}{2} = 0.541;$$

Finalmente, se obtiene los valores de los umbrales de Tukey

- Inferiores: 0.440 – 1.5 (0.541-0.440) = 0,2885
- Superiores: 0.541 + 1.5 (0.541-0.440) = 0,6925

Se encuentra que los ejemplos 12 y 13 están por encima de los límites de los umbrales de Tukey, y por lo tanto son datos erróneos que se deben eliminar.

Al aplicar el método presentado anteriormente sobre los conjuntos de datos se detectaron los valores erróneos (alfanuméricos) en las variables que componen las estaciones de geoquímica y deformación, presentados en las tablas 6 y 7.

Estación	Atributo	Valores erróneos (%)
Cráter	Flujo de gas (concentración)	1.2%
	Temperatura interna	0%
Cocuy3	Flujo de gas (concentración)	0.77%
	Temperatura interna	0.2%

Tabla 6. Valores erróneos en las estaciones de geoquímica

Estación	Atributo	Valores erróneos (%)
Agua Blanca	Componente Norte	2.3%
	Componente Este	0.28%
	Temperatura	1.1%
Cocuy2	Componente Norte	0.65%
	Componente Este	0.65%
	Temperatura	0.71%
Curiquinga	Componente Norte	2.3%
	Componente Este	2.3%
	Temperatura	1.8%
Guañarita	Componente Norte	0.2%
	Componente Este	2.29%
	Temperatura	0%
Lavas Rojas	Componente Norte	1.6%
	Componente Este	1.6%
	Temperatura	1.4%

Tabla 7. Valores erróneos en las estaciones de deformación

3.3 Resumen

Este capítulo tuvo como finalidad describir los procesos llevados a cabo para la comprensión y preparación de los conjuntos de datos creados. En primera instancia se presenta la información básica del volcán tal como su ubicación y antecedentes del mismo. De forma seguida, se menciona la red de vigilancia del volcán Puracé, la cual la componen estaciones sismológicas, geoquímicas y de deformación volcánica, entre otras. Además, cada uno de los instrumentos de medición es descrito en detalle. Para terminar la comprensión de los datos, se describen los conjuntos de datos usados para llevar a cabo el presente proyecto de grado. Finalmente, son mencionados los métodos llevados a cabo para preparar los conjuntos de datos antes de ser usados en el proceso de modelado.

Capítulo 4

Modelado

Tal y como se mencionó en el capítulo 2 (estado actual del conocimiento/ comprensión del negocio), la orientación que se le dio a este trabajo para la detección de eventos pre – eruptivos de origen volcánico estuvo dirigida hacia el uso de los sistemas auto - adaptativos. Retomando los análisis realizados en el estado del arte, los algoritmos más populares para la detección de anomalías son ABOD (Angle-Based Outlier Detection in High-dimensional Data), RDE (Recursive Density Estimation) y LOF (Local Outlier Factor).

A continuación, serán descritos en detalle los algoritmos mencionados.

4.1. Angle-Based Outlier Detection in High-dimensional Data (ABOD)

En [49], Kriegel propone un enfoque llamado ABOD (Angle-Based Outlier Detection in High-dimensional Data), el cual es un algoritmo que tiene como finalidad la detección de valores atípicos basados en ángulos entre flujos de datos de alta dimensión, donde la varianza es evaluada mediante la diferencia de los ángulos entre puntos en un espacio vectorial. Un punto se define como un valor atípico si la mayoría de los otros puntos se ubican en direcciones distintas. Esto se puede cuantificar utilizando el factor de valor atípico basado en el ángulo de la observación. Si el valor es pequeño, la observación se identifica como un valor atípico; en caso contrario es un punto normal. Una ventaja de este método es que no depende de ninguna selección de parámetros que influya en la calidad de la clasificación obtenida. Sin embargo, ABOD solo considera las relaciones entre cada punto y sus vecinos y no considera las relaciones entre vecinos, causando la identificación incorrecta de valores atípicos.

4.2 Local Outlier Factor (LOF)

LOF [48] [66] fue el primer algoritmo de detección de anomalías de tipo incremental. Este es un algoritmo creado para la detección de valores atípicos en un flujo de datos, soportado en el cálculo de densidad de vecindades que arroja como resultado un valor (Local Outlier Factor) de un objeto “p” que representa el grado en que “p” es una anomalía. Los puntos con alto valor de LOF tienen densidades locales más pequeñas que sus vecinos y representan valores atípicos más fuertes. El algoritmo LOF calcula un valor para cada registro de datos insertado en el conjunto de datos, con el propósito de determinar si el registro de datos insertado es atípico. Los valores de LOF para los registros de datos existentes se actualizan si es necesario y su desempeño depende de las estructuras de indexación propias

del algoritmo, las cuales son limitaciones en relación a la dimensionalidad de los datos y no es aplicable cuando se requiere procesar un conjunto de datos de grandes dimensiones.

La expresión matemática que define al algoritmo LOF está dada por la ecuación 2:

$$LOF_{MinPts}(p) = \frac{\sum_{MinPts(p)} \frac{lrd_{MinPts}(0)}{lrd_{MinPts}(p)}}{|N_{MinPts\ (p)}|}$$

Ecuación 2. LOF

Donde LOF es la media de los coeficientes de la densidad local de accesibilidad (lrd= La densidad local de accesibilidad de p es el inverso de la distancia media entre P y los objetos en su k-vecindad.) de “p” y los de puntos vecinos más cercanos. Intuitivamente, los valores de LOF de “p” van a ser muy altos si su densidad local de accesibilidad (lrd) son mucho más bajos que los de sus vecinos. El parámetro MinPts, es el valor que define el vecindario a crear alrededor de una fila y contra cuyos integrantes se van a realizar las mediciones para determinar el valor de la anomalía.

4.3 Recursive Density Estimation (RDE)

La densidad de datos es una medida clave en la detección de valores atípicos y otros problemas relacionados [67]. Sin embargo, la estimación (común) de la densidad de datos es muy compleja ya que requiere almacenar todo el conjunto de datos en cada análisis, encontrando limitaciones en memoria y potencia de cálculo. Para abordar estas restricciones, en el año 2001 se introdujo un enfoque recursivo [41], [42] llamado RDE (Recursive Data Estimation) y se patentó en 2016 [43].

Como se detalla en [68], utilizando RDE es posible obtener el resultado de la expresión de densidad de datos almacenando (y actualizando) solo una cantidad muy pequeña de datos (media de todas las muestras, μ_K y la cantidad promedio de productos escalares $\sum_K$ en cada momento k). Por lo tanto, este cálculo recursivo se puede llevar a cabo muy rápido, en tiempo cercano al real. Además, es importante señalar que el valor obtenido es exactamente el mismo (no una aproximación) en comparación con la estimación común. Mediante el uso de RDE, la tarea de detección de valores atípicos puede funcionar en modo en línea y en tiempo cercano al real.

Los sistemas auto-adaptativos se basan en el concepto de potencial para definir la proximidad espacial entre un ejemplo y el resto de ejemplos representados en un Espacio de Características, agrupando cada uno de los ejemplos a partir de su potencial. El potencial se puede calcular por diferentes métodos y funciones. RDE calcula recursivamente el potencial de un ejemplo utilizando como fórmula la distancia coseno [69], [70]. La ecuación del cálculo de potencial para un ejemplo

X_k, se presenta en la ecuación 3, donde se observa que el cálculo del potencial de un ejemplo necesita el cálculo de la suma de las distancias de todos ejemplos del espacio de características. Esto implica la necesidad de almacenar todos los ejemplos, por lo que es necesaria una cantidad de memoria elevada y el cálculo de un número considerable de operaciones. Con la finalidad de no almacenar todos los ejemplos recibidos.

$$P_k\,(x_k) = \frac{1}{1 + \frac{\sum_{i=1}^{k-1} distancia^2\,(x_k, x_i)}{k-1}}$$

Ecuación 3. Cálculo de potencial.

Angelov y Zhou [71] proponen un método recursivo capaz de calcular el potencial de un ejemplo. La importancia de este cálculo reside en que el resultado de dicha expresión es exactamente el mismo que si se almacenaran todos los ejemplos y se aplicara la Ecuación 4. La diferencia entre el método no recursivo y la ecuación recursiva, es que ésta última sólo necesita almacenar un conjunto muy reducido de datos

Ahora bien, a partir de las ecuaciones 4 y 5 se genera la ecuación 6, haciendo un reemplazo de la fórmula de la distancia coseno

$$d_{cos}(x_k, x_i) = 1 - \frac{\sum_{j=1}^{n} x_k^j x_i^j}{\sqrt{\sum_{j=1}^{n}(x_k^j)^2 \ \sum_{j=1}^{n}(x_i^j)^2}}$$

Ecuación 4. Simplificación de distancia coseno.

$$P_k(x_k) = \frac{1}{1 + [\frac{1}{k-1}\sum_{i=1}^{k-1}\left[1 - \frac{\sum_{j=1}^{n} x_k^j x_i^j}{\sqrt{\sum_{j=1}^{n}\left(x_k^j\right)^2 \ \sum_{j=1}^{n}\left(x_i^j\right)^2}}\right]^2]}$$

Ecuación 5. Distancia coseno

Finalmente, luego de simplificar la ecuación, el potencial de un ejemplo puede ser calculado como se explica en la Ecuación 6 y dicho cálculo no requiere almacenar todos los elementos, sino actualizar en cada caso únicamente el vector de datos.

$$\frac{1}{2 - \frac{1}{k-1}\frac{1}{\sqrt{\sum_{j=1}^{n}\left(x_k^j\right)^2}}\sum_{j=1}^{n} x_k^j\, b_k^j}; k = 2,3,\ldots;\ P_1(x_1) = 1$$

Ecuación 6. Distancia coseno recursivo.

A pesar de que RDE no necesita almacenar ninguna instancia ni realizar configuraciones adicionales, se recomienda especificar un número considerable de instancias iniciales para que el modelo sobreajuste las variables y parámetros con el objetivo de evitar el arranque en frio del sistema. De esta manera, a medida que se va procesando más instancias, RDE se vuelve más confiable y preciso. Por otro

lado, es importante definir el grado de rigurosidad del algoritmo de detección de anomalías, esto se logra mediante el parámetro *Sigma*, el cual tiene tres posibles valores desde uno hasta tres, siendo el parámetro uno (1) el menos riguroso y tres (3) el más riguroso.

Como se ha mencionado en capítulos anteriores, la vigilancia volcánica es un dominio de aplicación sensible para la generación de alertas, por lo tanto, en este proyecto de maestría se tomó la decisión junto a los expertos del OVSPo de tener calibrado la variable *Sigma* con valor igual a tres (3) con el fin de obtener una mayor confiabilidad de precisión para las anomalías detectadas.

4.4. Características de los algoritmos

La tabla 8 describe las principales ventajas y desventajas de los tres algoritmos evaluados. En la primera columna están enmarcados los atributos más relevantes en la detección de cambios en algoritmos incrementales, los cuales fueron seleccionados de acuerdo a las investigaciones realizadas en [72]–[74]; mientras en las siguientes columnas se encuentran los algoritmos ABOD, RDE y LOF. Cada celda de la tabla perteneciente a un algoritmo es señalizada cuando cumple con un criterio de descripción.

Descripción	ABOD	RDE	LOF
Incremental		✓	✓
Recursivo		✓	
Múltiple dimensionalidad	✓	✓	
Cambios Abruptos	✓	✓	✓
Cambios graduales		✓	✓
Sensibilidad al ruido	✓	✓	
Auto actualizable		✓	
Ventanas deslizantes	✓		✓
Configuración manual	✓		✓

Tabla 8. Tabla de ventajas y desventajas de algoritmos auto-adaptativos

De acuerdo a la tabla anterior, las características del algoritmo RDE son las más deseadas y representan la mejor opción para realizar vigilancia volcánica teniendo en cuenta las siguientes características de la monitorización en un volcán:

- Los volcanes se consideran sistemas dinámicos debido a que sufren continuamente cambios a través del tiempo a causa de interacciones y choques de las placas tectónicas.
- Los datos de los volcanes son obtenidos por estaciones a manera de flujo de datos y tienen una frecuencia de muestreo de una instancia por minuto.
- Los cambios detectados dentro de la actividad volcánica pueden ser abruptos, graduales y sensibles al ruido.
- Cada estación mide distintos parámetros del lugar donde esté instalado.
- Los cambios pueden ocurrir en cualquier momento por lo que las alertas deben ser generadas en tiempo real.

Si bien es cierto que los algoritmos descritos cumplen con algunas de las características necesarias, la principal justificación del uso de RDE para el problema de detección de alertas volcánicas radica principalmente en que ABOD y LOF necesitan almacenar en vectores una gran cantidad de ejemplos y sus respectivos atributos para luego procesarlos y re-ajustar las formulas en el momento de actualizar el modelo. Esto implicaría una demanda de memoria y tiempo de procesamiento significativo, que evitaría responder con rapidez ante un evento o serie de eventos dados en un instante de tiempo corto.

4.5 Enfoques

Con ayuda de los expertos del OVSPo (ver sección 5.2) se han identificado tres enfoques diferentes (técnica, orientación y estación) con el fin de ofrecerle al experto una mayor variedad de respuestas frente a cualquier anomalía detectada en la actividad volcánica dada la monitorización en tiempo real el volcán Puracé, a través de las áreas de geoquímica y deformación. Los enfoques tienen asociados tres conjuntos de datos correspondientes con tres periodos de tiempo de vigilancia volcánica.

- Periodo 1: desde el 1 de enero hasta 31 de marzo de 2017
- Periodo 2: desde el 1 de abril hasta 30 de junio de 2017
- Periodo 3: desde el 1 de julio hasta 30 de septiembre de 2017

Estos periodos fueron elegidos teniendo en cuenta que trimestralmente son realizadas revisiones manuales de la actividad volcánica por parte del personal experto. Por otro lado, para realizar la validación de los resultados obtenidos es preciso contar con el tiempo de las dos personas expertas de cada área, las cuales accedieron a colaborar con dicha validación en secciones no muy extensas que abarcaran posibles periodos de cambios en el volcán; esto debido a sus múltiples ocupaciones dentro del instituto y a labores de campo fuera del OVSPo.

Cada conjunto de datos (dataset) corresponde a un trimestre de vigilancia volcánica (aproximadamente 129.600 instancias). En este orden de ideas, se obtiene un total de 36 conjuntos de datos correspondientes a las dos técnicas (deformación geoquímica), los tres flancos del volcán (oriental, occidental y central) y las siete estaciones de monitorización, tal y como se aprecia en la Figura 11.

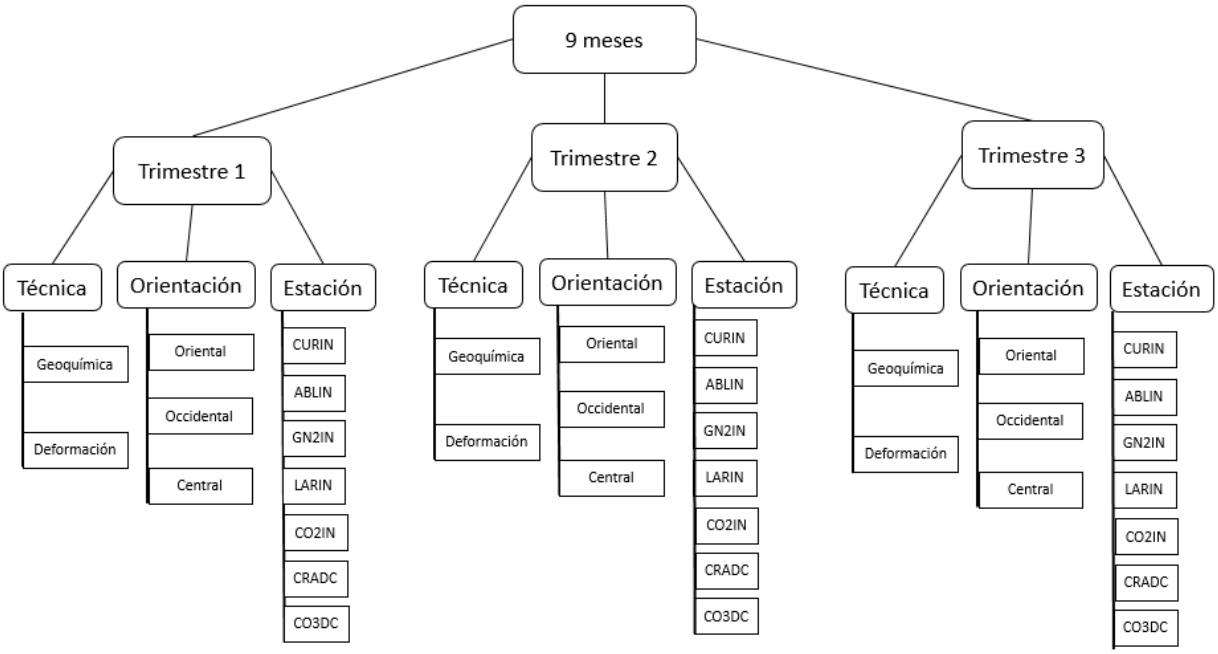


Figura 11. Conjunto de datos de 9 meses de monitorización del volcán Puracé

En este proyecto los resultados fueron evaluados por los mismos expertos del OVSPo, los cuales revisaron una a una las anomalías detectadas por el algoritmo. Posteriormente, los expertos clasificaron los eventos según la alerta pre-eruptiva identificada, basándose en su conocimiento y experiencia.

4.5.1 Enfoque por técnica

La vigilancia volcánica en el enfoque por técnica se divide en inclinometría y monitorización de gases de dióxido de carbono. Este enfoque permite tener como referencia solo estaciones de una misma técnica y así determinar si una anomalía es generada por un cambio de inclinación del cono volcánico o por desgasificación de gas CO_2. A continuación, se relacionan las estaciones de inclinometría y dióxido de carbono que fueron usadas en este proyecto.

Estaciones Inclinometría: Agua blanca, Guañarita, Cocuy2, Curiquinga, Lavas Rojas.

Estaciones Dióxido de carbono: Cráter, Cocuy.

Debido a que un inclinómetro obtiene atributos de las componentes norte, este y la temperatura ambiente, una instancia de la técnica de inclinometría está compuesta por 15 atributos. De igual forma, una instancia de la técnica de dióxido de carbono está compuesta por 4 atributos ya que una estación envía datos de flujo de gas CO_2 y temperatura.

4.5.2 Enfoque por orientación

El segundo enfoque es agrupado de acuerdo a los tres flancos principales de actividad histórica del volcán Puracé (flanco oriental, occidental y central), donde han existido algunos cambios físicos en el cono volcánico (esto por recomendación del personal experto del OVSPo). En las figuras 12,13 y 14 se expone la ubicación de las estaciones para el enfoque por orientación (estas figuras pueden ser visualizadas en un tamaño mayor en el anexo C).

Figura 12. Flanco Oriental

Figura 13. Flanco Occidental

Figura 14. Flanco Central

4.5.3 Enfoque por estación

Finalmente, el tercer enfoque tiene en cuenta cada estación de manera independiente ya sea del área de deformación o de geoquímica. Este enfoque ha sido elegido con el objetivo de tener la certeza de cuál es la estación que generó una alerta pre-eruptiva. La desventaja de este enfoque es que no todo cambio anómalo en una estación es una alerta volcánica, ya que, generalmente, más de una estación debe tener manifestaciones similares.

4.6 Aplicación de RDE a los enfoques para la detección de alertas volcánicas

Ahora bien, como resultado de aplicar RDE a cualquiera de los enfoques propuestos en este documento, se generan gráficas similares a las Figuras 15, 16 y 17 (a modo de ejemplo se presenta una por cada enfoque, sin embargo, en el anexo A se presenta todas las figuras de cada enfoque y sus respectivos periodos en un mayor tamaño). En cada una de las figuras obtenidas tras aplicar RDE, en el eje X horizontal se presentan el número de instancias (120000) correspondientes a un periodo de tres meses de monitorización volcánica del volcán Puracé (para este ejemplo el enfoque visualizado es por técnica para el área de deformación volcánica); y en el eje Y vertical se representa la densidad obtenida por el algoritmo (rango de valores comprendidos entre cero y uno). Adicionalmente, dos curvas pueden ser identificadas: la línea superior de color rojo, que representa el promedio de los datos; la segunda (línea color amarilla) que delimita el umbral de los valores

atípicos, donde los valores que están por debajo de la segunda curva son las anomalías detectadas.

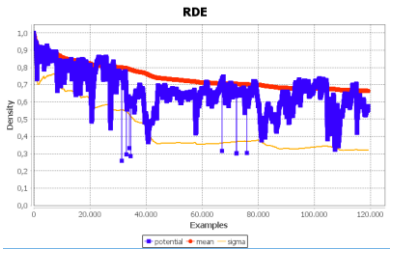


Figura 15. Outliers RDE – Enfoque técnica - deformación periodo III

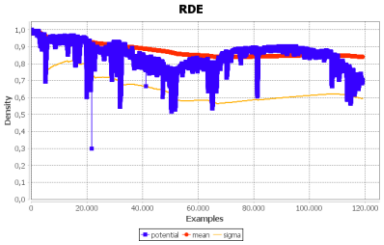


Figura 16. Outliers RDE – Enfoque Orientación - Flanco occidental periodo I

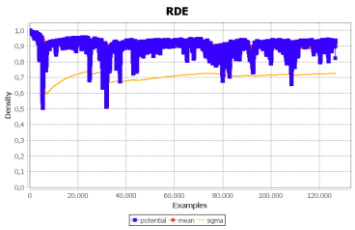


Figura 17. Outliers RDE – Enfoque Estación - Guañarita 2 - periodo II

En los ejemplos anteriores, el algoritmo detectó 14, 10 y 6 valores atípicos respectivamente. Los cuales fueron sometidos a validación por parte de los expertos del OVSPo para encontrar los verdaderos positivos, falsos positivos y verdaderos negativos. Estos valores atípicos detectados por el algoritmo, corresponden a una fecha y hora específica. Por ejemplo, los valores atípicos identificados en la figura 9 se corresponden con el primer periodo y en el rango de instancias de 20434 a 20970 se aprecia un conjunto de anomalías, correspondiente a las fechas 16/07/2017 16:30:00 y 17/07/2017 2:11:00. El análisis de la variación de comportamiento en varias de las estaciones de deformación permite identificar que se están generando fenómenos volcánicos fuera de lo normal y son contrastados con la opinión del experto correspondiente como se mencionó anteriormente.

Siguiendo con lo anterior, y tomando como base la figura 13, es decir las anomalías detectadas para el enfoque técnica – área deformación con RDE para la monitorización de volcanes sobre el III periodo; la Figura 18 presenta el comportamiento de las componentes Norte, Este y temperatura para cada una de las estaciones del área de deformación (Agua blanca, Cocuy, Curiquinga, Guañarita y Lavas Rojas) en el periodo de estudio. La inclinación en la componente X (este), y Y (norte) de cada inclinómetro es medida en unidades de micro radianes y es representada por líneas de color azul y rojo, respectivamente, mientras que la temperatura, cuya unidad es medida en grados centígrados se puede visualizar con línea de color verde. Finalmente, se puede observar que a cada una de estas estaciones se le ha adicionado barras verticales de color gris, que indican las anomalías detectadas en un instante de tiempo; el grosor de cada barra indica la repetición de anomalías durante un rango de tiempo. Es importante mencionar que todas las figuras relacionadas con la ubicación de las anomalías sobre cada estación se encuentran en el anexo D.

Figura 18. Anomalías por estación en el enfoque por técnica (deformación)

La Figura 19 muestra los resultados de la detección de anomalías volcánicas para el enfoque por orientación, periodo I – Flanco occidental (Figura 14) Para este enfoque se presentan las gráficas de las estaciones de deformación: Guañarita y Cocuy2.

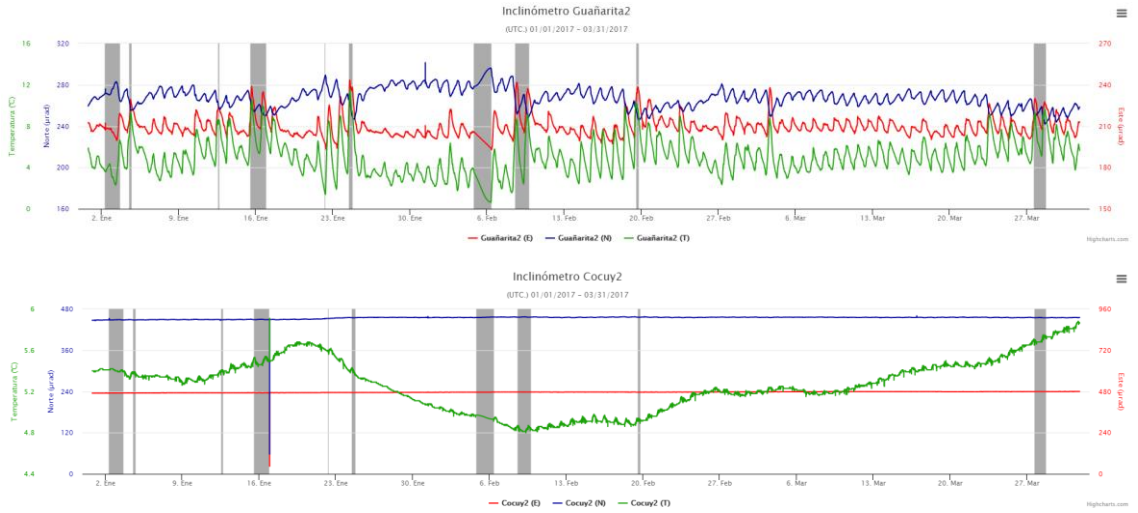


Figura 19. Anomalías por estación en el enfoque por orientación (flanco occidental)

Finalmente, para el enfoque por estación, la figura 20 expone los resultados obtenidos de la estación Guañarita, correspondientes a la detección de anomalías volcánicas del segundo periodo (Figura 11).

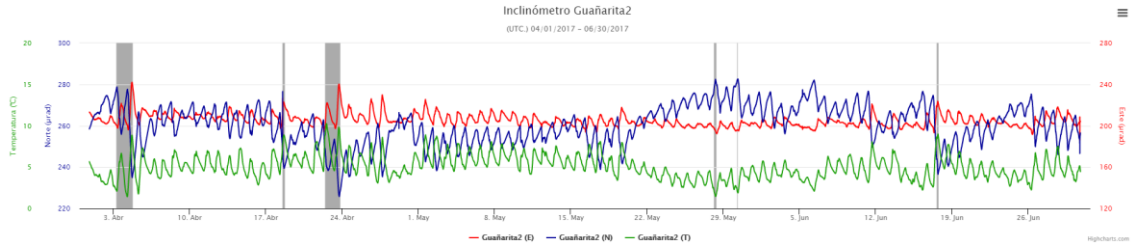


Figura 20. Anomalías por estación

4.7 Resumen

En este capítulo se identificaron tres de los algoritmos relacionados con la detección de anomalías en tiempo real, sobre los cuales se enmarcaron las principales ventajas y desventajas mediante una tabla comparativa. A partir de ello, fue justificado el uso de RDE para la detección de alertas pre–eruptivas volcánicas. Se continuó con la explicación de los tres enfoques (técnica, orientación y estación) generados para obtener una mejor respuesta al detectar anomalías de origen volcánico. Finalmente, se presenta el uso de RDE para cada uno de los enfoques de detección de anomalías mediante ejemplos.

Capítulo 5

Experimentación y evaluación

En este capítulo se presenta el proceso de experimentación y las pruebas ejecutadas sobre el sistema para la detección de alertas pre–eruptivas volcánicas basadas en aprendizaje incremental.

5.1 Métricas de evaluación

5.1.1 Matriz de confusión

La matriz de confusión (tabla 9) permite observar, mediante una tabla de contingencia la distribución de los errores cometidos por un clasificador. Cada columna de la matriz representa el número de clasificaciones de cada clase (valor predicho), mientras que cada fila representa la clasificación real, interpretada por el experto en el área indicada de la monitorización volcánica (valor real), es importante mencionar que en una matriz de confusión, los verdaderos positivos y verdaderos negativos son los valores deseados [75].

		Valor predicho	
		Anomalía	No anomalía
Valor real	Anomalía	**VP**	FN
	No anomalía	FP	**VN**

Tabla 9. Matriz de confusión

A partir de la matriz de confusión se pueden extraer conceptos básicos (en la sección 5.3 serán descritos estos conceptos aplicados en el contexto volcánico) y a continuación son presentados:

Se tiene la variable objetivo c, la cual contiene las clases $c_1, c_2, c_3, \dots, c_n$, de este modo por cada clase se deben aplicar los siguientes conceptos:

- ✓ **Falsos Positivos (FP):** número de instancias incorrectamente clasificadas en la clase c_x.
- ✓ **Falsos Negativos (FN):** instancias de la clase c_x que fueron incorrectamente clasificadas en otra clase.
- ✓ **Verdaderos Positivos (VP):** número instancias correctamente clasificadas en la clase c_x.
- ✓ **Verdaderos Negativos (VN):** todas las instancias restantes correctamente clasificadas diferente a la clase c_x.

Una vez obtenido el número de falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos por cada clase, posteriormente se calculan dos métricas empleadas en la medida del rendimiento de los sistemas de búsqueda, reconocimiento de información y reconocimiento de patrones [76], a continuación son expuestas cada una de ellas:

- **Precisión**: está definido como la proporción de instancias verdaderas del conjunto de instancias predichas como positivas (capacidad del clasificador para evitar el ruido) [77], y es calculada de la siguiente forma:

$$p = \frac{|VP|}{|VP + FP|}$$

Ecuación 7. Precisión

- **Exhaustividad**: esta medida conocida en Ingles como Recall, es la encargada de calcular la proporción de verdaderos positivos predichos entre todos los positivos (instancias relevantes clasificadas) [75]. La ecuación se expresa como:

$$e = \frac{|VP|}{|VP + FN|}$$

Ecuación 8. Exhaustividad

- **Medida-F:** es un balance de la precisión y la exhaustividad, en otras palabras es considerada una media ponderada de la precisión y exhaustividad donde la puntuación alcanza su mejor valor en 1 y el peor en 0 [78]. En la Ecuación 9 es presentada:

$$F = 2\ x\ \frac{p\ x\ e}{p + e}$$

Ecuación 9. Medida F

Vale la pena resaltar que esta es conocida como la medida f_1, ya que la precisión y la exhaustividad son pesados uniformemente, con $\beta = 1$ siendo la formula general:

$$f_\beta = \frac{\left(1 + \beta^2\right) x\ p\ x\ e}{\beta^2\ x\ p + e}\ ;\ \ \beta\ > 0$$

Ecuación 10. Formula general Medida F

Aplicando la definición anterior en el contexto del actual proyecto, podríamos definir que:

- Un Verdadero Positivo (**VP**) es una anomalía detectada por el sistema implementado, la cual es validada por un experto del SGC.
- Un Falso Positivo (**FP**) es una anomalía detectada por el sistema implementado que no es considerada como tal por un experto.
- Falsos negativos (**FN**) son las anomalías que eventualmente ocurrieron en el volcán Puracé, pero que el sistema no lo detecto.
- Verdaderos negativos (**VN**) son todos los ejemplos que el sistema no detecto y que tampoco hacen parte de anomalías volcánicas.

5.2 Plan de pruebas

Las anomalías detectadas por RDE fueron analizados por expertos del OVSPo tanto del área de Geoquímica como de Deformación volcánica. A continuación, son presentados los datos personales de los expertos en geoquímica y deformación volcánica, que ayudaron con la interpretación de las anomalías para poder etiquetar los resultados.

- **Geoquímica**: Luisa Fernanda Mesa, e-mail: lmesa@sgc.gov.co, Licenciada en Biología y Química - Maestría en Química – Universidad de Caldas.

- **Deformación**: Jorge Armando Alpala, e-mail: jalpala@sgc.gov.co, Ingeniero Civil - Especialista en SIG (Universidad de Caldas) y MSc (c) en Geomática (Universidad del Cauca)

La tabla 10 resume el plan de pruebas llevado a cabo para realizar la evaluación del sistema. En la primera columna se encuentran los diferentes enfoques de evaluación (cada uno debe evaluar 3 periodos); las siguientes columnas conforman la red de estaciones analizadas en este proyecto. Cada una de las celdas marcadas con una letra “X” identifica las estaciones analizadas dentro de un enfoque de evaluación.

Para cada periodo en un enfoque de monitorización, la evaluación del sistema se realiza de la siguiente manera:

1. El experto encargado recibe anomalías.
2. De cada anomalía se abstrae la fecha completa de ocurrencia.
3. Se obtiene las estaciones involucradas en el enfoque según la tabla 10.
4. El experto verifica si la anomalía detectada por el algoritmo pertenece a un cambio volcánico en cualquiera de las estaciones obtenidas.
5. Si la anomalía es detectada por alguna de las estaciones involucradas entonces se obtendrá un verdadero positivo, en caso contrario será un falso positivo.

Enfoques de evaluación	Estaciones						
	CO3DC	**CRADC**	**ABLIN**	**CO2IN**	**CURIN**	**GUAIN**	**LARIN**
Deformación			X	X	X	X	X
Geoquímica	X	X					
Oriental			X		X		X
Occidental				X		X	
Central	X	X					

Tabla 10. Plan de pruebas

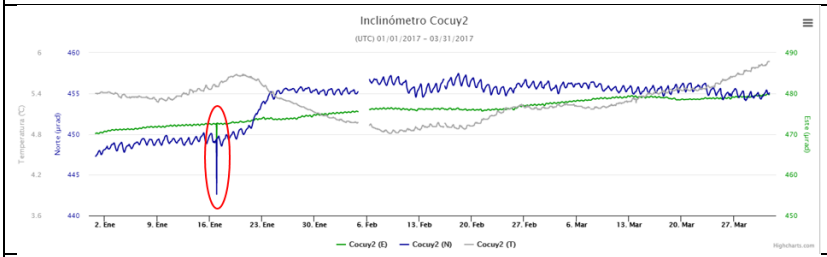
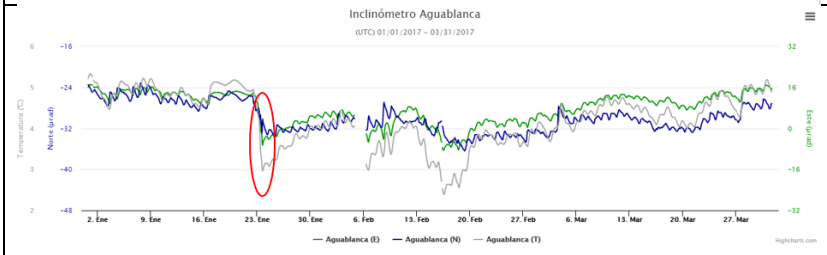
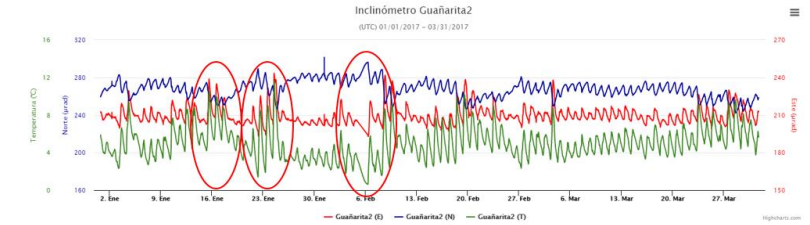
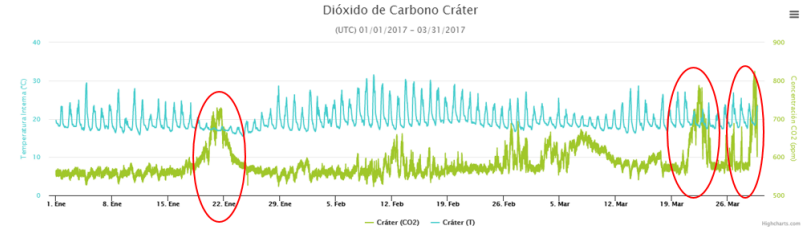
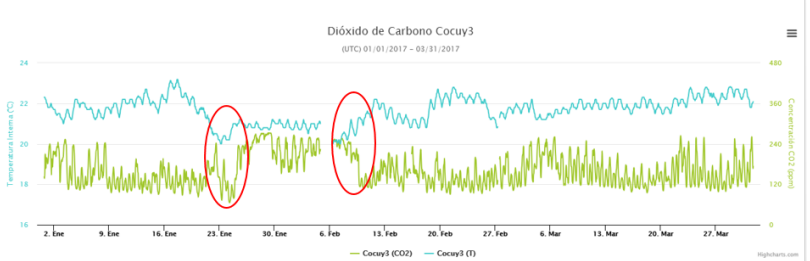
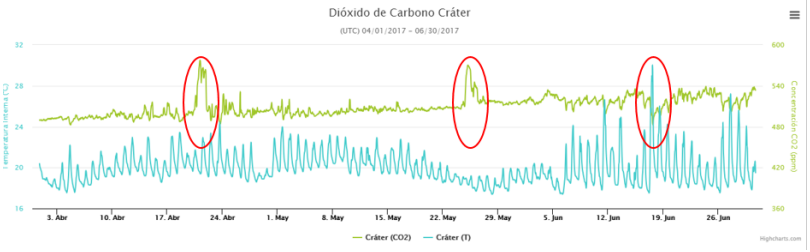
5.3 Resultados

La detección de anomalías por el algoritmo RDE ha sido cuantificada y analizada desde la relación entre verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN). En las siguientes subsecciones se presentará el análisis y relación de los resultados obtenidos en cada enfoque. Por otro lado, la tabla 11 contiene el resumen de los resultados que se obtienen al ejecutar el algoritmo propuesto para todos los conjuntos de datos mencionados en apartados anteriores. En cada uno de estos resultados se identifican el número de anomalías detectadas (verdaderos positivos), no detectadas (falsos negativos) y anomalías que fueron detectadas, pero realmente no lo eran (falsos positivos). Bajo estas variables, fueron calculas las medidas de desempeño de precisión (PRE), exhaustividad (EXH) y medida F (M-F). Cada una de las interpretaciones de los expertos se realizó en base a los reportes telemétricos de cada estación, donde el experto confirma o no la veracidad de una anomalía generada por el prototipo creado para este proyecto. Ahora bien, dentro de cada gráfica de transmisión de datos de una estación, son adicionadas barras verticales que representan las anomalías detectadas por el algoritmo en el/los instantes precisos de la ocurrencia.

	VP	FP	FN	PRE	EXH	M-F
Técnica						
Deformación	21	6	5	**0,78**	**0,81**	**0,8**
Geoquímica	20	4	4	**0,83**	**0,83**	**0,83**
Promedios	21	5	4,5	0,81	0,82	0,81
Grupo						
Flanco Oriental	16	3	3	0,84	0,84	0,84
Flanco occidental	21	2	4	0,91	0,84	0,88
Flanco Central	20	4	4	**0,83**	**0,83**	**0,83**
Promedios	19	3,33	3,7	0,86	0,84	0,85
Estación						
CO3DC	16	5	0	0,76	0,99	0,86
CRADC	23	7	1	0,78	0,96	0,86
ABLIN	8	4	1	0,67	0,89	0,76
CO2IN	16	5	3	0,76	0,84	0,80
CURIN	19	5	1	0,79	0,96	0,86
GUAIN	18	4	1	0,82	0,95	0,88
LARIN	23	7	0	0,76	0,99	0,87
Promedios	17,9	5,3	1	0,76	0,94	0,84

Tabla 11. Resumen de resultados

En la tabla 12 son presentados algunos ejemplos de las clasificaciones con las anomalías interpretadas por los expertos y detectadas por el algoritmo. En cada uno de los ejemplos de las clasificaciones las barras verticales se han cambiado por círculos en color rojo para identificar con mayor facilidad el momento exacto de la ocurrencia de la anomalía detectada.

Anomalía	Interpretación
	Detección de anomalía por pequeña deformación volcánica. La componente Norte y Este varían bruscamente, mientras la temperatura crece gradualmente.
	Detección de anomalía por dilatación del efecto térmico de bajas temperaturas.
	Cambios recurrentes generados en dilatación y contracción de la roca volcánica.
	Desgasificación alta en la estación cráter – dióxido de carbono con temperatura normal.
	Desgasificación media en la estación de medición de gas dióxido de carbono Cocuy2.
	En la estación cráter en el periodo II (1 abril de 2017 – 30 junio de 2017) se presenta tres anomalías detectadas también por el prototipo, las dos primeras hacen referencia a desgasificación media en los niveles de flujo de gas CO_2.

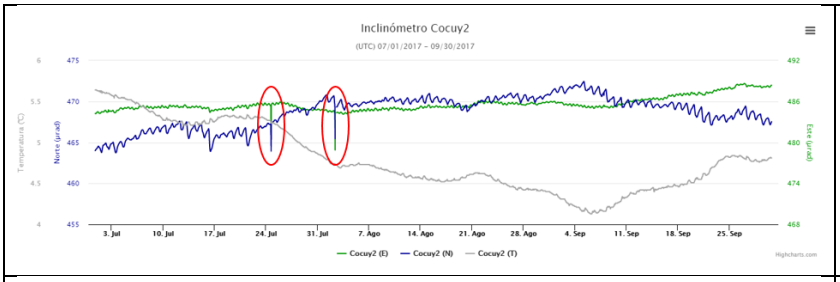
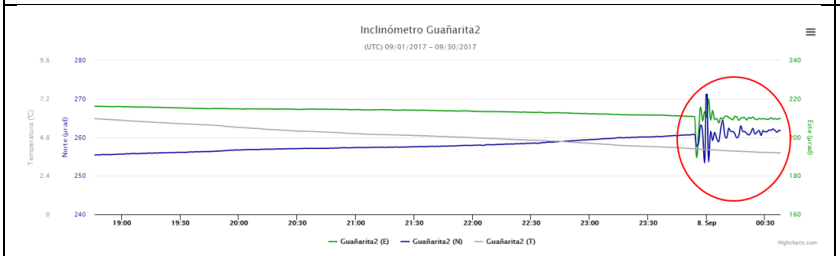
	Anomalía característica de problemas de digitalización, ya que las tres variables tienen un cambio brusco incluido la temperatura. En este caso registrada en la estación de inclinometría Cocuy3 para el III periodo.
	Detección de una anomalía por evento sísmico distal que fue registrado por la estación guañarita2. El sismo corresponde al terremoto con epicentro en Ecuador ocurrido el 9 se septiembre del año 2017.

Tabla 12. Ejemplos de anomalías identificadas y etiquetadas por los expertos

En la siguiente sección se encuentran los resultados obtenidos por cada enfoque, sobre los cuales se ha generado una matriz de confusión y la respectiva clasificación de los verdaderos positivos de acuerdo a la interpretación realizada por los expertos. Las clasificaciones realizadas se distinguen a continuación:

- **Contracción**: causada por bajas temperatura afectando el edificio volcánico.
- **Dilatación**: Efecto térmico por calor.
- **Dilatación y contracción:** causada por fenómenos recurrentes día-noche.
- **Digitalización** causada por errores en la telemetría de las estaciones.
- **Deformación volcánica baja:** causada por pequeñas inyecciones de magma o por deflación del cono volcánico por desgacificación.
- **Deformación pronunciada:** Deformación considerable del cono volcánico.
- **Eventos sísmicos**: detección de eventos sísmicos distales.
- **Altas temperaturas** en una estación geoquímica potencialmente precursora de actividad en procesos sísmicos y fumarólicos.
- **Desgasificación continúa:** Emanación de gas CO_2 en cortos periodos de tiempo.
- **Desgasificación media:** Emanación pulsos de gas CO_2 medianamente pronunciados.
- **Desgasificación baja:** Emanación de pequeños pulsos de gas CO_2.

Finalmente, por cada uno de los enfoques se presenta la gráfica de valores obtenidos con las métricas evaluadas junto a su interpretación.

5.3.1 Enfoque por técnica

Análisis experto deformación

La matriz de confusión y la clasificación de los verdaderos positivos de los resultados obtenidos por RDE en el enfoque de deformación son presentadas en las tablas 13 y 14.

Número de instancias=393120		Valor predicho	
		Anomalía	No anomalía
Valor real	Anomalía	21 (VP)	5 (FN)
	No anomalía	6 (FP)	393087 (VN)

Tabla 13. Matriz de confusión para Geoquímica

Cantidad de verdaderos positivos	Clasificación
10	Dilatación y contracción
2	Contracción
4	Dilatación
3	Digitalización
2	Eventos sísmicos distales

Tabla 14. Clasificación de verdaderos positivos para el enfoque Deformación

Análisis experto geoquímica

La matriz de confusión y la clasificación de los verdaderos positivos de los resultados obtenidos por RDE para el enfoque geoquímica son presentadas en las tablas 15 y 16.

Número de instancias=393120		Valor predicho	
		Anomalía	No anomalía
Valor real	Anomalía	20 (VP)	4 (FN)
	No anomalía	4 (FP)	393092 (VN)

Tabla 15. Matriz de confusión para Deformación

Cantidad de verdaderos positivos	Clasificación
3	Desgasificación baja
2	Altas temperaturas
9	Desgasificación continua
4	Desgasificación media
2	Digitalización

Tabla 16. Clasificación de verdaderos positivos para el enfoque Geoquímica

En la figura 21 se presenta la comparación de las medidas de desempeño: Precisión, Exhaustividad y Medida F. El eje horizontal (x) presenta el tipo de técnica elegido y el eje vertical (Y) representa el valor obtenido desde cero a uno.

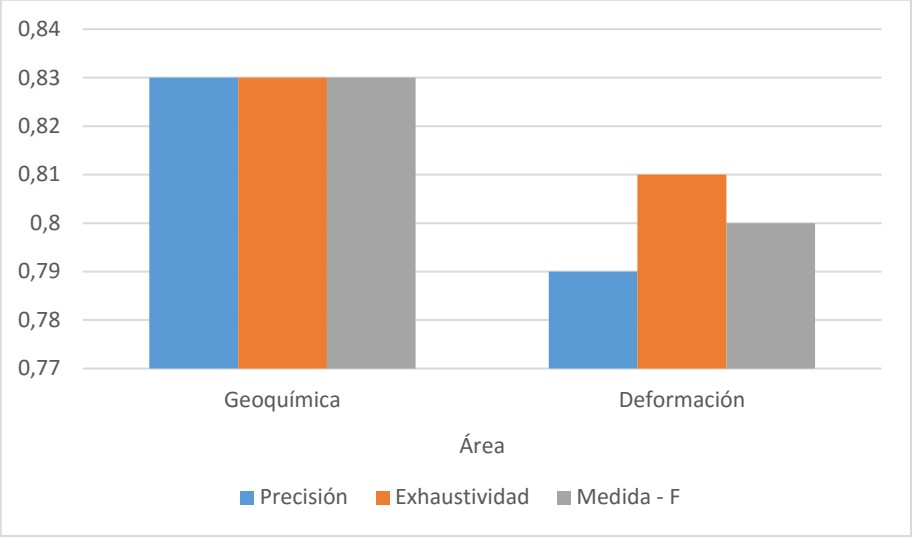


Figura 21. Comparación del rendimiento del algoritmo RDE entre geoquímica y deformación.

Los resultados para el área de Geoquímica presentan mejores valores en las tres medidas calculadas en comparación con el área de Deformación, donde se el número de falsos positivos es mayor. La Exhaustividad en ambas técnicas obtiene un buen valor, lo que significa que el algoritmo tiene un buen número de instancias relevantes recuperadas (en ambas técnicas está por encima del 81 %). Teniendo en cuenta que ambas técnicas son complementarias en el estudio de la vigilancia volcánica, es posible afirmar que los tres valores de las medidas obtenidas son aceptables para la detección de anomalías y constituyen una base para posteriores estudios de investigación hacia eventos volcánicos haciendo uso de las Tecnologías de la Información y Comunicaciones (TIC).

5.3.2 Enfoque por orientación

Análisis experto flanco oriental

La matriz de confusión y la clasificación de los verdaderos positivos de los resultados obtenidos por RDE en el enfoque del flanco oriental son presentadas en las tablas 17 y 18.

Número de instancias=393120		Valor predicho	
		Anomalía	No anomalía
Valor real	Anomalía	16 (VP)	3 (FN)
	No anomalía	3 (FP)	393099 (VN)

Tabla 17. Matriz de confusión para el flanco oriental

Cantidad de verdaderos positivos	Clasificación
2	Dilatación y contracción
4	Contracción
4	Dilatación
5	Digitalización
2	Sismicidad distal

Tabla 18. Clasificación de verdaderos positivos para el enfoque flanco oriental

Análisis experto flanco occidental

La matriz de confusión y la clasificación de los verdaderos positivos de los resultados obtenidos por RDE en el enfoque del flanco occidental son presentadas en las tablas 19 y 20.

Número de instancias=393120		Valor predicho	
		Anomalía	No anomalía
Valor real	Anomalía	21 (VP)	4 (FN)
	No anomalía	2 (FP)	393090 (VN)

Tabla 19. Matriz de confusión para el flanco occidental

Cantidad de verdaderos positivos	Clasificación
2	Contracción
5	Deformación baja
2	Deformación pronunciada
5	Dilatación
4	Digitalización
3	Sismos distales

Tabla 20. Clasificación de verdaderos positivos para el enfoque flanco occidental

Análisis experto flanco central

La matriz de confusión y la clasificación de los verdaderos positivos de los resultados obtenidos por RDE en el enfoque del flanco central son presentadas en las tablas 21 y 22.

Número de instancias=393120		Valor predicho	
		Anomalía	No anomalía
Valor real	Anomalía	20 (VP)	4 (FN)
	No anomalía	4 (FP)	393092 (VN)

Tabla 21. Matriz de confusión para el flanco central

Cantidad de verdaderos positivos	Clasificación
3	Desgasificación baja
2	Altas temperaturas
9	Desgasificación continua
4	Desgasificación media
2	Digitalización

Tabla 22. Clasificación de verdaderos positivos para el enfoque flanco central

Adicionalmente, de las matrices de confusión presentadas anteriormente, se genera la figura 22, en la cual son comparadas las medidas de desempeño: Precisión, Exhaustividad y Medida F. El eje horizontal (X) presenta el tipo de técnica elegida y el eje vertical (Y) representa el valor obtenido desde cero a uno.

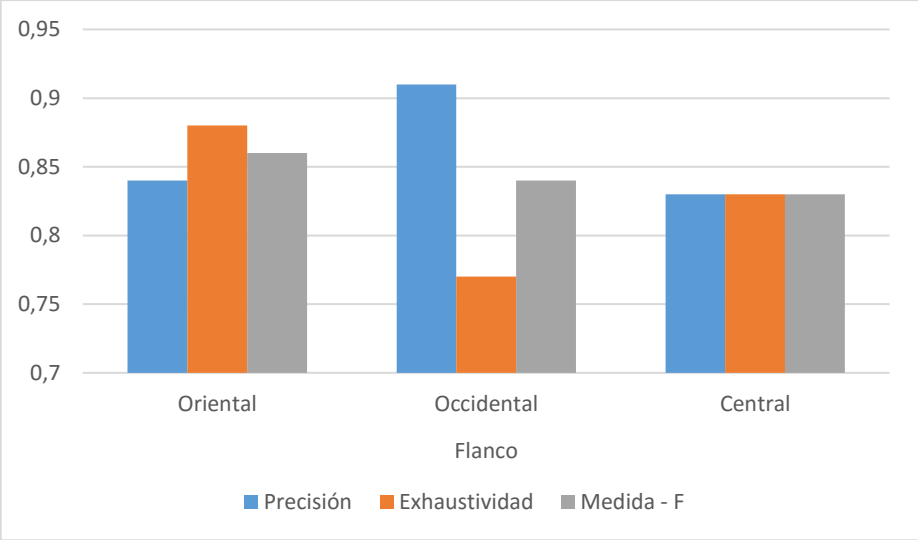


Figura 22. Comparación del rendimiento del algoritmo RDE entre los flancos oriental, occidental y central.

Se puede observar que el flanco occidental obtiene mejores resultados en Precisión. Esto se debe a que en este flanco se encuentra la estación Cocuy, las cual es una estación de referencia para el volcán Puracé. Por otro lado, en las fechas fijadas para las pruebas, el flanco occidental tuvo un enjambre de actividad sísmica en la fuente sismogénica de la región San Rafael que queda al occidente del volcán.

5.3.3. Enfoque por estación

Análisis experto estación CO3DC

La matriz de confusión y la clasificación de los verdaderos positivos de los resultados obtenidos por RDE en el enfoque de la estación CO3DC son presentadas en las tablas 23 y 24.

Número de instancias=393120		Valor predicho	
		Anomalía	No anomalía
Valor real	Anomalía	16 (VP)	0 (FN)
	No anomalía	5 (FP)	393099 (VN)

Tabla 23. Matriz de confusión para la estación Cocuy 3 – Dióxido de carbono

Cantidad de verdaderos positivos	**Clasificación**
4	Desgasificación baja
1	Altas temperaturas
8	Digitalización
3	Desgasificación media

Tabla 24. Clasificación de verdaderos positivos para el enfoque de la estación CO3DC

Análisis experto estación CRADC

La matriz de confusión y la clasificación de los verdaderos positivos de los resultados obtenidos por RDE en el enfoque de la estación CRADC son presentadas en las tablas 25 y 26.

Número de instancias=393120		Valor predicho	
		Anomalía	No anomalía
Valor real	Anomalía	25 (VP)	1 (FN)
	No anomalía	7 (FP)	393089 (VN)

Tabla 25. Matriz de confusión para la estación Cráter - Dióxido de carbono

Cantidad de verdaderos positivos	**Clasificación**
1	Desgasificación continua
2	Altas temperaturas
4	Desgasificación media
4	Desgasificación baja
14	Digitalización

Tabla 26. Clasificación de verdaderos positivos para el enfoque estación de la estación CRADC

Análisis experto estación ABLIN

La matriz de confusión y la clasificación de los verdaderos positivos de los resultados obtenidos por RDE en el enfoque de la estación ABLIN son presentadas en las tablas 27 y 28.

Número de instancias=393120		Valor predicho	
		Anomalía	No anomalía
Valor real	Anomalía	8 (VP)	1 (FN)
	No anomalía	4 (FP)	393106 (VN)

Tabla 27. Matriz de confusión para la estación Agua blanca – Inclinómetro

Cantidad de verdaderos positivos	**Clasificación**
3	Contracción
2	Deformación baja
2	Digitalización
1	Sismicidad distal

Tabla 28. Clasificación de verdaderos positivos para el enfoque de la estación ABLIN

Análisis experto estación CO2IN

La matriz de confusión y la clasificación de los verdaderos positivos de los resultados obtenidos por RDE en el enfoque de la estación CO2IN son presentadas en las tablas 29 y 30.

Número de instancias=393120		Valor predicho	
		Anomalía	No anomalía
Valor real	Anomalía	16 (VP)	3 (FN)
	No anomalía	5 (FP)	393094 (VN)

Tabla 29. Matriz de confusión para la estación Cocuy2 – Inclinómetro

Cantidad de verdaderos positivos	**Clasificación**
4	dilatación y contracción
3	Dilatación
8	Digitalización
1	Sismicidad distal

Tabla 30. Clasificación de verdaderos positivos para el enfoque de la estación CO2IN

Análisis experto estación CURIN

La matriz de confusión y la clasificación de los verdaderos positivos de los resultados obtenidos por RDE en el enfoque de la estación CURIN son presentadas en las tablas 31 y 32.

Número de instancias=393120		Valor predicho	
		Anomalía	No anomalía
Valor real	Anomalía	19 (VP)	1 (FN)
	No anomalía	5 (FP)	393093 (VN)

Tabla 31. Matriz de confusión para la estación Curiquinga - Inclinómetro

Cantidad de verdaderos positivos	Clasificación
1	Dilatación y contracción
2	Contracción
6	Dilatación
9	Digitalización
1	Sismicidad distal

Tabla 32. Clasificación de verdaderos positivos para el enfoque de la estación CURIN

Análisis experto estación GUAIN

La matriz de confusión y la clasificación de los verdaderos positivos de los resultados obtenidos por RDE en el enfoque de la estación GUAIN son presentadas en las tablas 33 y 34.

Número de instancias=393120		Valor predicho	
		Anomalía	No anomalía
Valor real	Anomalía	18 (VP)	1 (FN)
	No anomalía	4 (FP)	393097 (VN)

Tabla 33. Matriz de confusión para la estación Guañarita - Inclinómetro

Cantidad de verdaderos positivos	Clasificación
2	Dilatación y contracción
3	Contracción
6	Dilatación
3	Digitalización
1	Sismicidad distal

Tabla 34. Clasificación de verdaderos positivos para el enfoque de la estación GUAIN

Análisis experto estación LARIN

La matriz de confusión y la clasificación de los verdaderos positivos de los resultados obtenidos por RDE en el enfoque de la estación LARIN son presentadas en las tablas 35 y 36.

Número de instancias=393120		Valor predicho	
		Anomalía	No anomalía
Valor real	Anomalía	23 (VP)	0 (FN)
	No anomalía	7 (FP)	393090 (VN)

Tabla 35. Matriz de confusión para la estación Lavas rojas - Inclinómetro

Cantidad de verdaderos positivos	Clasificación
5	Contracción
4	Dilatación
14	Digitalización

Tabla 36. Clasificación de verdaderos positivos para el enfoque de la estación LARIN

Adicionalmente, de las matrices de confusión presentadas anteriormente, se genera la figura 23, en la cual son comparadas las medidas de desempeño: Precisión, Exhaustividad y Medida F. El eje horizontal (x) presenta el tipo de técnica elegido y el eje vertical (Y) representa el valor obtenido desde cero a uno.

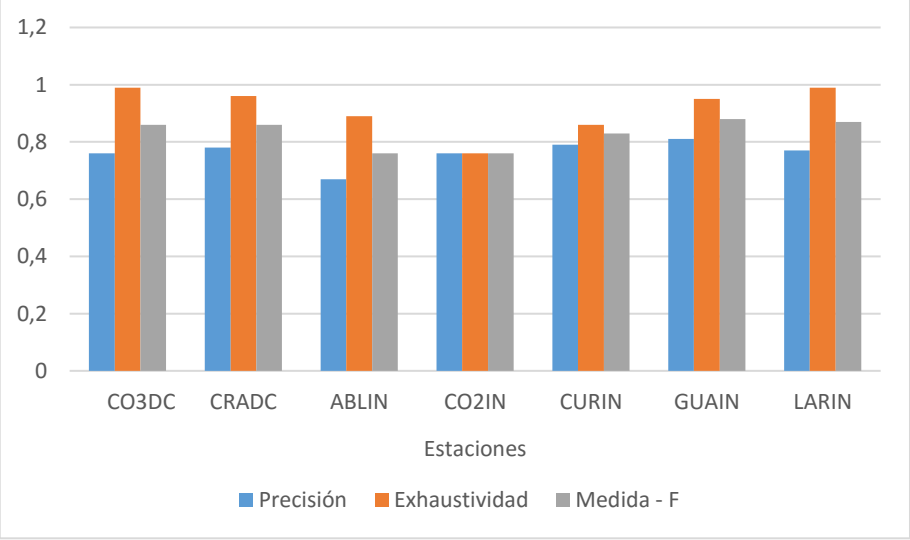


Figura 23. Comparación del rendimiento del algoritmo RDE entre geoquímica y deformación.

Se puede observar que las estaciones CRADC y GUAIN (Crater y Guañarita) obtienen los mejores resultados para áreas de Geoquímica y Deformación respectivamente. Este resultado es muy probable que se deba a la ubicación geográfica de la estación, ya que como su nombre lo indica, se encuentra exactamente en el cráter del volcán Puracé cerca de una fumarola, donde se concentra dióxido de carbono en mayores cantidades. Por su parte, Guañarita es una estación usada como línea base para proceso deformativos del volcán Puracé. Registros históricos del volcán muestran alertas volcánicas por emanación de gas CO_2 y de deformación según los boletines técnicos mensuales generados por el OVSPo [31]. Por otro lado, la estación de inclinómetria Agua Blanca (ABLIN)

obtiene los resultados más bajos de todas las estaciones evaluadas, coincidiendo con que esta misma presenta una alta cantidad de valores erróneos y perdidos.

5.4. Comparación con algoritmos LOF y ABOD

Todos los conjuntos de datos, pertenecientes a cada periodo y enfoque, fueron probados y ejecutados con los algoritmos de detección de anomalías LOF y ABOD. A continuación, los resultados obtenidos son presentados.

5.4.1. Resultados LOF

- **Detección de anomalías en el enfoque por técnica**

  Analizando los resultados obtenidos por el algoritmo LOF en el enfoque por técnica (figura 24), se observa que el área de Deformación obtiene levemente mejores resultados que Geoquímica.

  En el área de Deformación la detección de anomalías se realiza basándose en cinco estaciones (inclinómetros electrónicos), a diferencia de Geoquímica que cuenta con dos estaciones de CO_2, razón por la cual el algoritmo LOF tiene más atributos para converger en una anomalía.

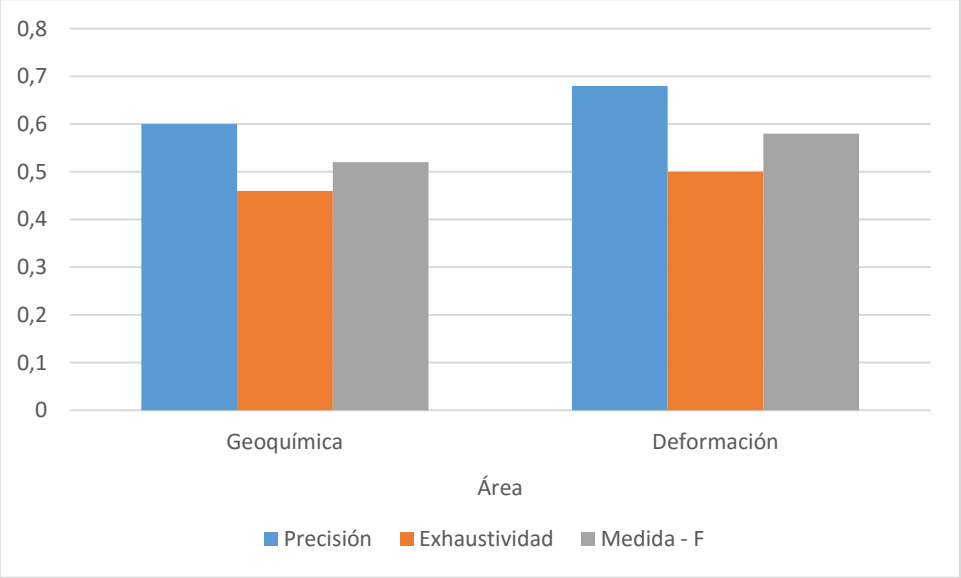


Figura 24. Anomalías por técnica usando LOF

- **Detección de anomalías en el enfoque por orientación**

  La figura 25 presenta los resultados obtenidos por el algoritmo LOF en el enfoque por orientación. La orientación central obtiene mejores resultados que las orientaciones oriental y occidental. Adicionalmente, se puede apreciar que la orientación central coincide con los resultados obtenidos en el enfoque por técnica en el área de Geoquímica, ya que las estaciones usadas en ambos

casos son las mismas. Por otro lado, las estaciones de deformación usadas para la orientación oriental (tres) y occidental (dos) no obtienen resultados que superen a la orientación central.

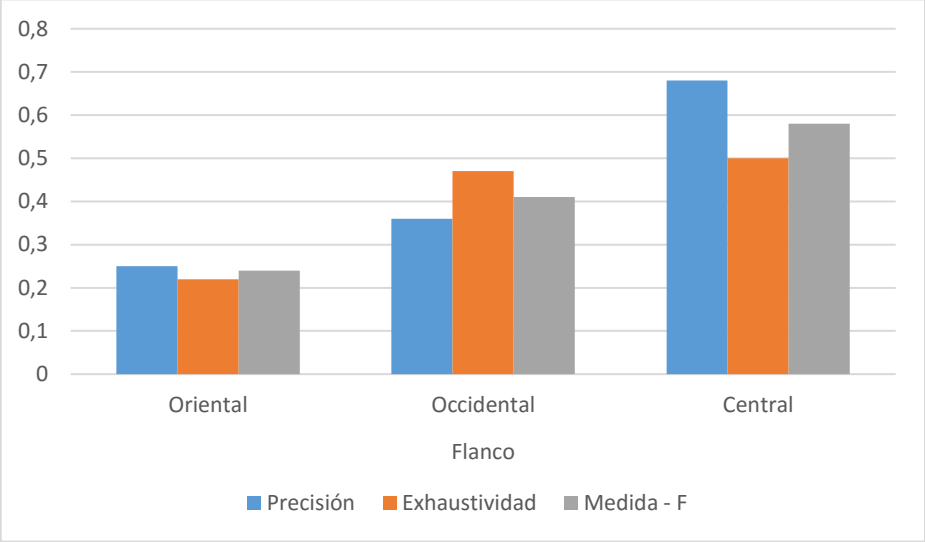


Figura 25, Anomalías por orientación usando LOF

- **Detección de anomalías en el enfoque por estación**

Finalmente, la figura 26 muestra los resultados obtenidos por el algoritmo LOF para el enfoque estación. La grafica refleja que la estación Guañarita (GUAIN) obtiene ampliamente mejores resultados para las tres métricas de evaluación que el resto de las estaciones usadas en este proyecto. En las bases de datos del OVSPo, las estaciones que más registran cambios deformativos en el volcán Puracé son las estaciones Cocuy2 y Guañarita.

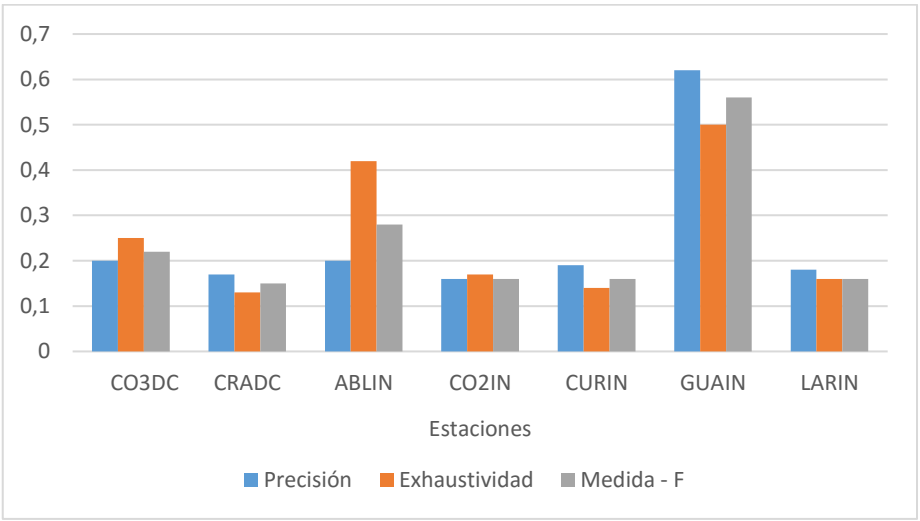


Figura 26. Anomalías por estación usando LOF

5.4.1. Resultados ABOD

- **Detección de anomalías en el enfoque por técnica**

Analizando los resultados obtenidos por el algoritmo ABOD (figura 27), se observa que el área de Geoquímica obtiene levemente mejores resultados que Deformación volcánica. Sin embargo, estos resultados no son lo suficientemente buenos para la monitorización en volcanes en comparación con el algoritmo RDE.

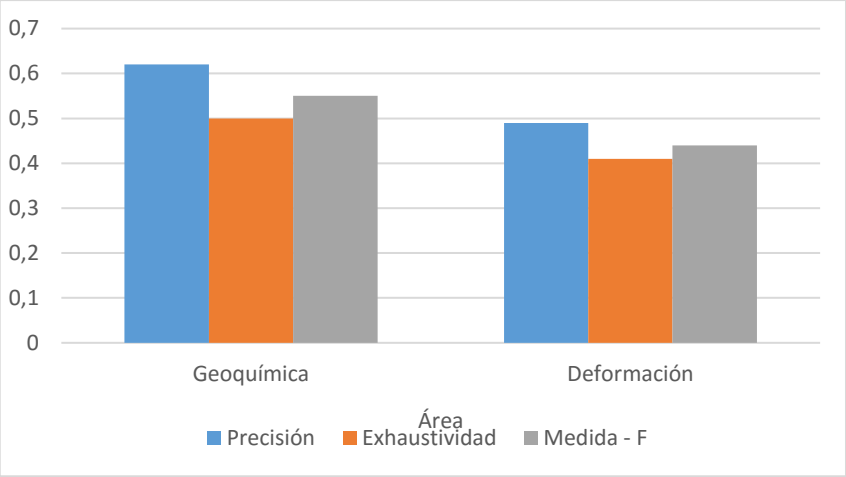


Figura 27. Anomalías por técnica usando ABOD

- **Detección de anomalías en el enfoque por orientación**

La figura 28 presenta los resultados obtenidos por el algoritmo ABOD en el enfoque por orientación, dichos resultados reflejan que la orientación central, al igual que en las pruebas por LOF, obtiene mejores resultados que las orientaciones oriental y occidental.

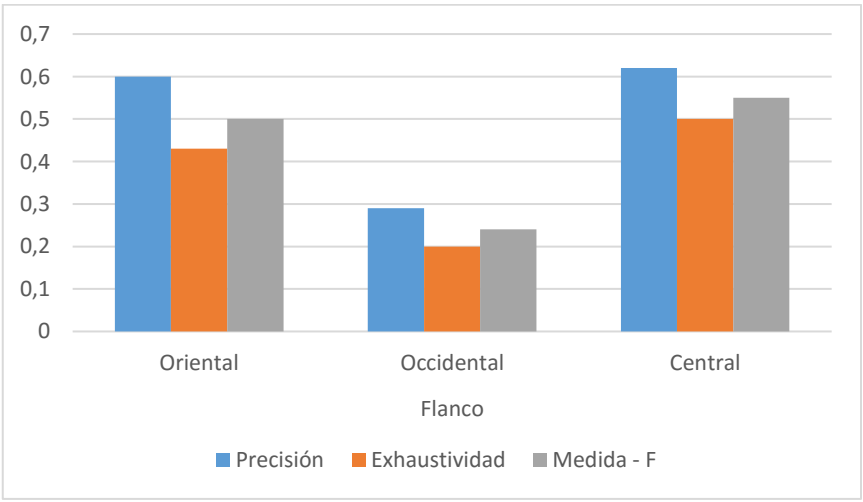


Figura 28. Anomalías por orientación usando ABOD

- **Detección de anomalías en el enfoque por estación**

La figura 29 muestra los resultados obtenidos por el algoritmo ABOD para el enfoque estación. La figura refleja que las estaciones Guañarita (GUAIN) y Cocuy2 obtienen ampliamente mejores resultados para las tres métricas de evaluación que el resto de las estaciones usadas en este proyecto. Cuatro de las cinco estaciones de Deformación presentan mejores resultados que las dos estaciones de Geoquímica, lo cual probablemente se deba a la alta dimensionalidad manejada por el algoritmo LOF y que las estaciones de Deformación tienen más atributos que las estaciones de Geoquímica.

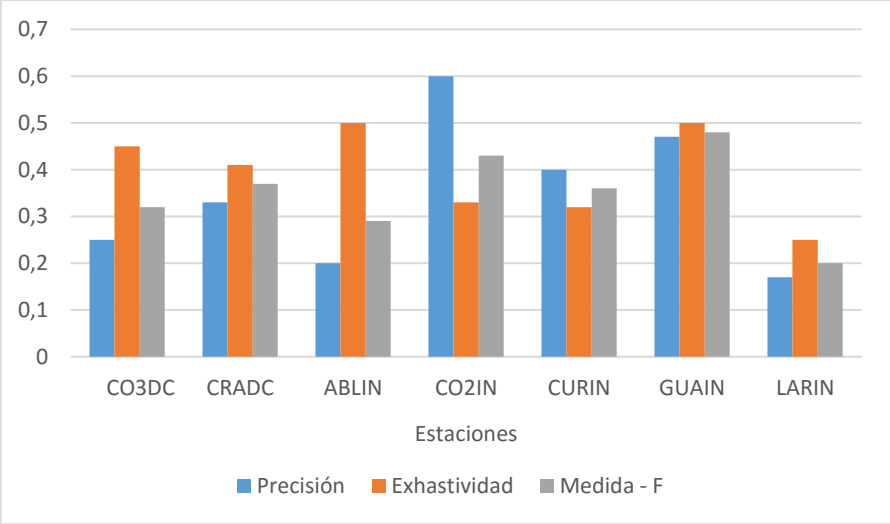


Figura 29. Anomalías por estación usando ABOD

Finalmente, se realiza una comparación que resume los resultados obtenidos por algoritmos RDE, LOF y ABOD sobre las métricas evaluadas (precisión, exhaustividad y medida F). Las tablas 37 (a) y 37 (b) muestran dicha comparación para los enfoques técnica, orientación y estación. En la tabla se puede observar que, para todos los conjuntos de datos volcánicos, RDE obtiene los mejores resultados que los demás algoritmos. De igual forma, se aprecia que el enfoque por técnica presenta los resultados más bajos.

Enfoque técnica						
Algoritmos	**LOF**		**ABOD**		**RDE**	
	Geoquímica	**Deformación**	**Geoquímica**	**Deformación**	**Geoquímica**	**Deformación**
Precisión	0.6	0.68	0.62	0.49	0.83	0.79
exhaustividad	0.46	0.5	0.50	0.41	0.83	0.81
Medida F	0.52	0.58	0.55	0.44	0.83	0.8

Enfoque orientación									
Algoritmos	**LOF**			**ABOD**			**RDE**		
	oriental	**Occidental**	**central**	**oriental**	**Occidental**	**central**	**oriental**	**Occidental**	**central**
Precisión	0.25	0.36	0.68	0.6	0.29	0.62	0.84	0.91	0.83
exhaustividad	0.22	0.47	0.5	0.43	0.2	0.5	0.88	0.77	0.83
Medida F	0.24	0.41	0.58	0.5	0.24	0.55	0.86	0.84	0.83

Figura 37. a. Resultados enfoques técnica y orientación.

Enfoque estación								
Algoritmos		**CO3DC**	**CRADC**	**ABLIN**	**CO2IN**	**CURIN**	**GUAIN**	**LARIN**
LOF	**Precisión**	0.2	0.17	0.2	0.16	0.19	0.62	0.18
	Exhaustividad	0.25	0.13	0.42	0.17	0.14	0.5	0.16
	Medida F	0.22	0.15	0.28	0.16	0.16	0.56	0.16
ABOD	**Precisión**	0.25	0.33	0.2	0.6	0.4	0.47	0.17
	Exhaustividad	0.45	0.41	0.5	0.33	0.32	0.5	0.25
	Medida F	0.32	0.37	0.29	0.43	0.36	0.48	0.2
RDE	**Precisión**	0.76	0.78	0.67	0.76	0.79	0.82	0.77
	Exhaustividad	0.99	0.96	0.89	0.76	0.86	0.95	0.99
	Medida F	0.86	0.86	0.76	0.76	0.82	0.88	0.87

Figura 37. b. Resultados enfoques estación.

Tabla 37. Comparación de algoritmos por cada enfoque

A continuación, se presenta la figura 30, la cual contiene un compilado de los resultados de la comparación de los algoritmos RDE, LOF y ABOD mostrados en las tablas 37 (a) y 37 (b) .

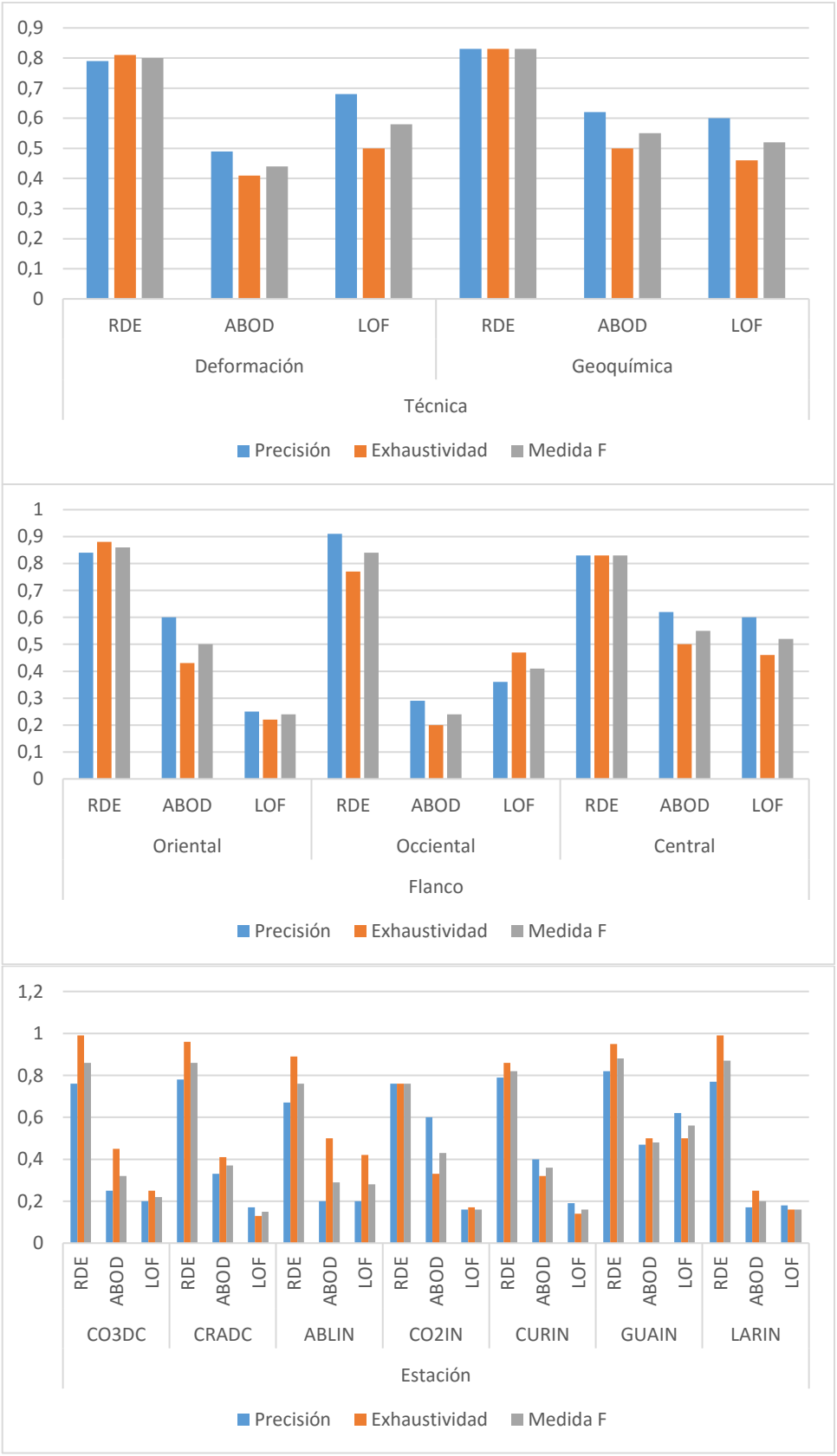


Figura 30. Comparación de los algoritmos RDE, LOF y ABOD

5.5 Resumen

En este capítulo se exponen, en primera instancia, las medidas de evaluación para las anomalías detectadas haciendo uso de RDE. Para cada uno de los enfoques propuestos se obtuvo los valores de Medida F, Precisión y Exhaustividad. Posteriormente, fueron generadas e interpretadas las matrices de confusión para cada enfoque. Los resultados obtenidos fueron determinados objetivamente por los expertos del OVSPo. De manera seguida, se mostraron gráficas comparativas de rendimiento entre las métricas y las técnicas usadas en cada enfoque. Finalmente, se realiza una comparación de los resultados con los mismos conjuntos de datos para cada enfoque aplicado a los algoritmos de detección de anomalías ABOD y LOF. En dicha comparación, se observa claramente que el sistema creado usando RDE es mejor para el dominio de aplicación volcánico.

Cabe mencionar que la totalidad de las gráficas y su respectivo análisis se encuentran en el anexo B.

Capítulo 6

Despliegue: Prototipo

En este capítulo se presenta el prototipo que implementa la detección de anomalías para la generación de alertas pre – eruptivas volcánicas para el volcán Puracé. Para la construcción del prototipo se utilizó la metodología de desarrollo SCRUM [79]. De acuerdo a esta metodología, se ha definido una serie de iteraciones que satisfacen los objetivos del proyecto.

A continuación, en la Tabla 38 se lista los productos a desarrollar en el prototipo:

Iteración	Descripción
1	Diseño de la arquitectura del sistema
	Adquisición de datos de todas las estaciones
	Creación de Conjuntos de datos
	Diseño de interfaces
2	Módulo de monitorización en tiempo real
	Adaptación de RDE al sistema
	Grafica de densidad en tiempo real
3	Módulo de Detección de anomalías
	Módulo de Clasificación
	Módulo de configuración
	Integración de Módulos

Tabla 38. Lista de productos SCRUM

El prototipo creado para el presente trabajo de maestría fue llamado VOD (Volcano Outlier Detector) en su versión *1.0* y está desarrollado en el lenguaje de programación Java 1.8, en esta aplicación solo se usó la librería postgresql-42.1.1.jre6.jar para la adquisición de datos de cada una de las estaciones de vigilancia volcánica. Las características de hardware como de software del equipo en el cual fue probado el prototipo son presentadas en la tabla 39.

Marca equipo	Hewlett-Packard EliteBook
Sistema operativo	Windows 10
Procesador	Intel ® core ™ i7 – 4600U CPU @ 2.10 GHz 2,70 GHz
RAM	8,00 GB

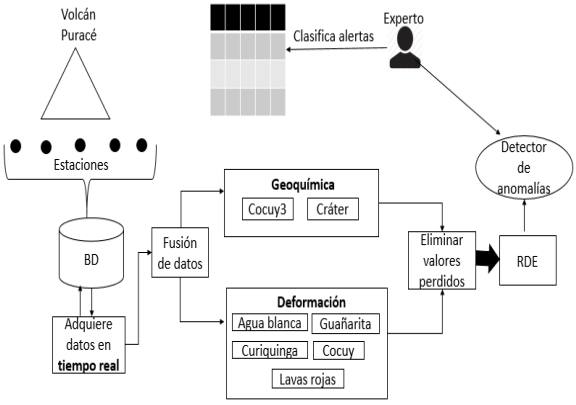
Tabla 39. Características del equipo de cómputo

Ahora bien, para el desarrollo de este prototipo se consideraron tres iteraciones, las cuales se describen a continuación con sus casos de uso y sus respectivas historias de usuario.

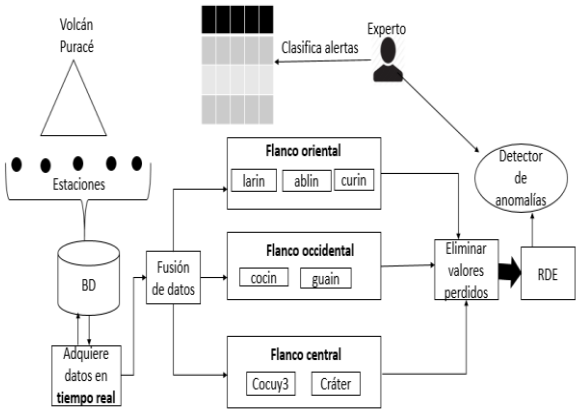
6.1 Iteración 1

Como se mencionó en el capítulo 4 (modelado), se generaron tres enfoques que permitan identificar el mejor mecanismo para la detección de alertas pre-eruptivas a partir de la detección de anomalías volcánicas. Por esta razón, la arquitectura del sistema (Figura 31) contiene tres imágenes: a, b y c (una por cada enfoque); y estas figuras pueden ser visualizadas en un tamaño mayor en el anexo C. Cada arquitectura refleja los diferentes módulos que contiene el sistema:

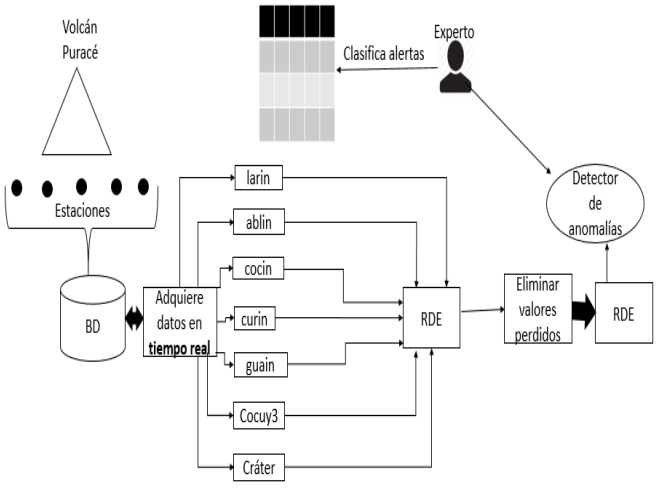
- Adquisición de datos de las estaciones en campo.
- Almacenamiento de datos en crudo en diferentes bases de datos ubicadas en el OVSPo.
- Según sea el caso, se realiza la selección de las estaciones y campos necesarios (selección de características).
- Ordenamiento en tiempo real de las instancias según el enfoque elegido.
- Filtro de datos que permite eliminar los valores nulos.
- Clasificación de la alerta volcánica según la decisión del experto.



a) Técnica (geoquímica y deformación)



b) Orientación.



Estación

Figura 31. Arquitectura por enfoque

En la Figura 32 se observa la interfaz de inicio del prototipo software. Una vez es ejecutada la aplicación el usuario final puede elegir el tipo de monitorización (en tiempo real) a realizar al volcán Puracé. Es posible elegir entre los enfoques: técnica, región y estación. Además, cuenta con una opción de configuración, cuya implementación hace parte de la segunda iteración.

Figura 32. Front End prototipo

En la Figura 33 se presenta el diagrama de caso de uso para la adquisición de datos dependiendo del enfoque elegido y donde se especifica el conjunto de funcionalidades del prototipo. La adquisición de datos dentro del sistema consite en realizar conexiones sobre el motor de base de datos y generar consultas SQL para obtener los datos necesarios de acuerdo al enfoque elegido por el usuario final.

Inicialmente, desde campo son enviados a la base de datos los datos crudos de todas las estaciones por diferentes canales en tiempo real y, de forma seguida, de acuerdo al enfoque seleccionado por el usuario, son tomados los datos de la(s) estación(es) que intervienen en la respectiva monitorización y se abstraen cada uno de los campos.

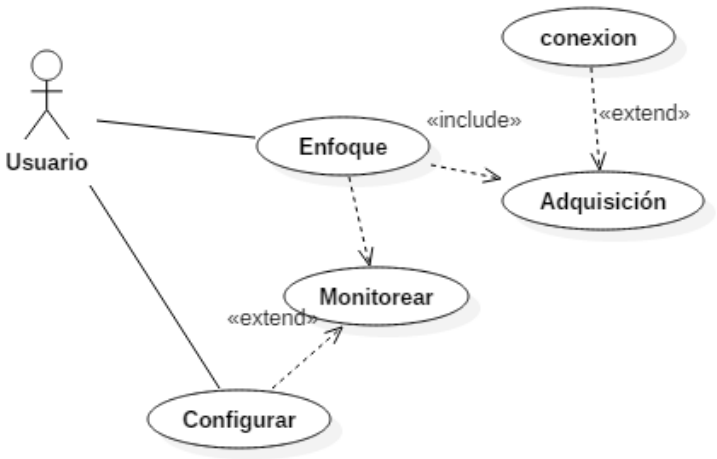


Figura 33. Caso de uso - Iteración 1

De acuerdo al flujo de la primera iteración, la Historia de Usuario se muestra en la Tabla 40.

Historia de Usuario	
Número: 1	**Usuario:** Cliente
Nombre de historia: Adquisición de datos dependiendo del enfoque elegido	
Prioridad: Alta	**Riesgo en desarrollo:** Alta
Puntos estimados: 2	**Iteración asignada:** 1
Programador responsable: José Eduardo Gómez Daza	
Descripción: Generar interfaces gráficas para cada enfoque elegido por el usuario final brindando de esta forma las posibles opciones, así mismo la interfaz para las opciones de configuración con las que contara el prototipo. De esta forma, una vez se haya elegido dicho enfoque, el sistema debe adquirir en tiempo real los datos transmitidos, conectándose y generando consultas SQL a la base de datos del OVSPo. Estos datos son organizados para generar instancias que serán evaluadas posteriormente por RDE. Finalmente, si algún valor de esta instancia es nulo, este será eliminado.	
Validación: La interfaz correspondiente a la selección del usuario debe ser mostrada y la instancia adquirida debe corresponder también al enfoque elegido.	

Tabla 40. Historia de Usuario uno.

A continuación se encuentran las pantallas del prototipo desarrollado una vez es seleccionado el enfoque a monitorear.

Si la opción elegida es la monitorización por región, se presenta una interfaz para elegir si se requiere hacer vigilancia por el flanco oriental, occidental o central del volcán. Con cada opción se exponen las estaciones que conforman cada flanco tal y como se observa en la figura 34.

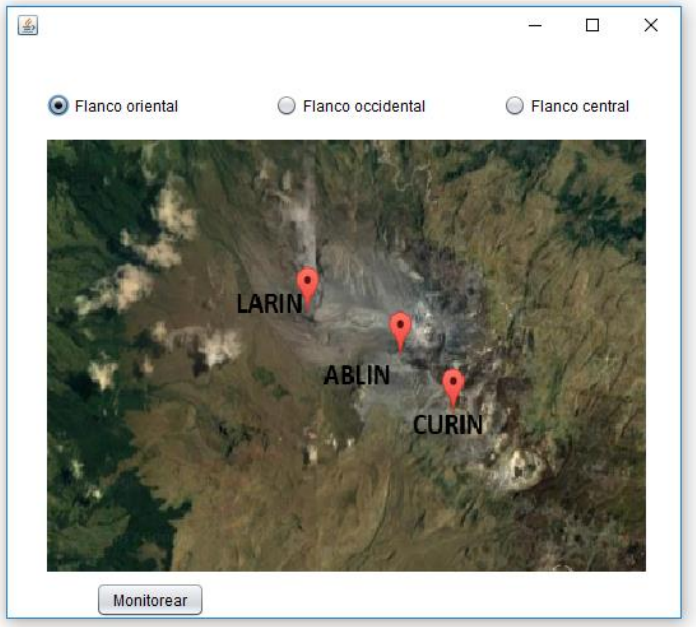


Figura 34. Enfoque orientación - prototipo

Si la opción elegida es la monitorización por el enfoque de estación, el usuario debe elegir una de las estaciones ya sea del área de geoquímica o de deformación volcánica y seguidamente presionar el botón monitorizar (figura 35).

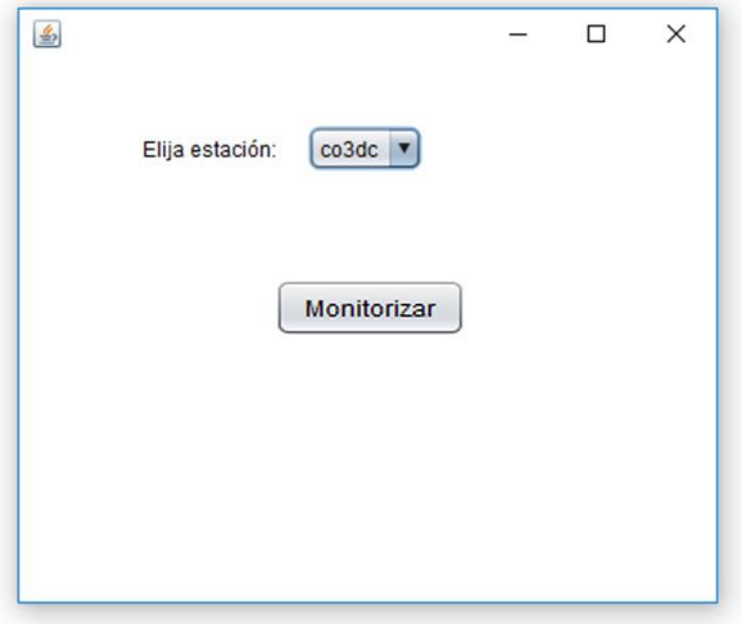


Figura 35. Enfoque estación - prototipo

Cuando la selección de enfoque es por técnica, se despliega una interfaz donde las opciones son: monitorización por Deformación (todas las estaciones de inclinometría) o Geoquímica (total de estaciones de medición de gas CO_2), como se observa en la figura 36.

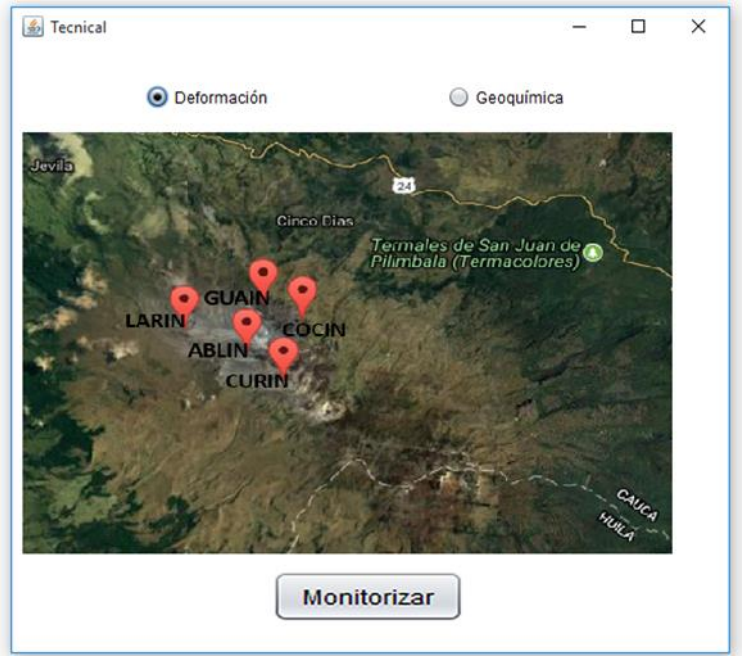


Figura 36. Enfoque por técnica - prototipo

6.2. Iteración 2

En la figura 37 se presenta el diagrama de caso de uso para la monitorización de alertas volcánicas según el enfoque elegido, donde se define el conjunto de funcionalidades más importantes que el prototipo debe cumplir.

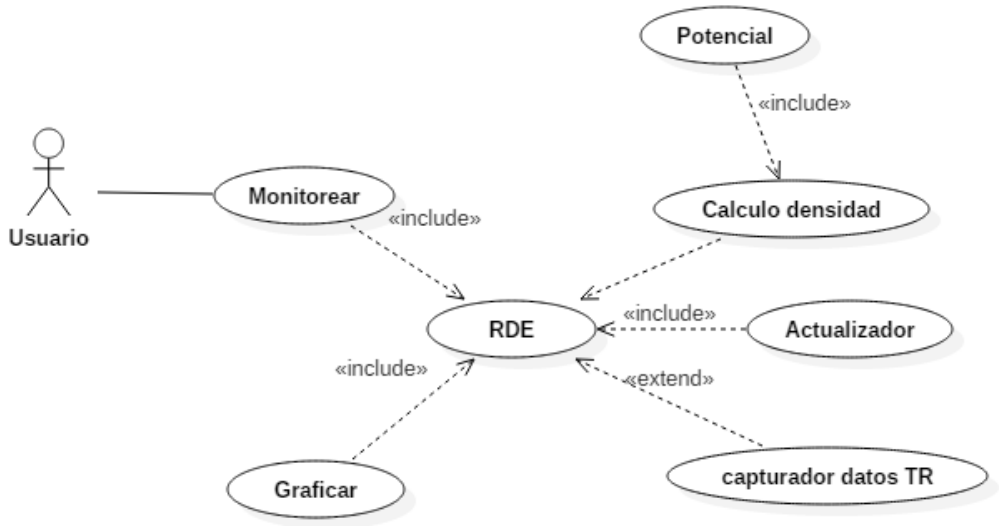


Figura 37. Caso de uso – iteración 2

Para esta segunda iteración, se genera una nueva historia de usuario (Tabla 41) a desarrollar.

Historia de Usuario	
Número: 2	**Usuario:** Cliente
Nombre de historia: Monitorización de alertas volcánicas	
Prioridad: Alta	**Riesgo en desarrollo:** Alta
Puntos estimados: 3	**Iteración asignada:** 1
Programador responsable: José Eduardo Gómez Daza	
Descripción: Consta de la adaptación del algoritmo RDE para que, mediante el cálculo de densidad, actualice la función recursiva de acuerdo a los cambios detectados, permitiendo observar en una gráfica en tiempo real dichos cambios, los cuales al final reflejaran las alertas pre-eruptivas.	
Validación: La monitorización volcánica debe realizarse para cada enfoque y sus subsecciones respectivas. Cada minuto debe llegar un nuevo dato y este será procesado por el algoritmo determinando, con el fin de determinar si es o no una anomalía.	

Tabla 41. Historia de usuario dos

En caso de no elegir ninguna configuración para monitorizar cualquiera de los enfoques, la gráfica que identifica la detección de anomalías tiene un aspecto similar a la Figura 38, donde el eje vertical (Y) identifica la densidad y el eje horizontal (X), la fecha de monitorización. Los colores de las líneas en la gráfica se

asocian con densidad (color azul), sigma (color naranja) y la media aritmética (color rojo). Ahora bien, una anomalía es detectada por el prototipo únicamente cuando la densidad está por debajo de la línea sigma.

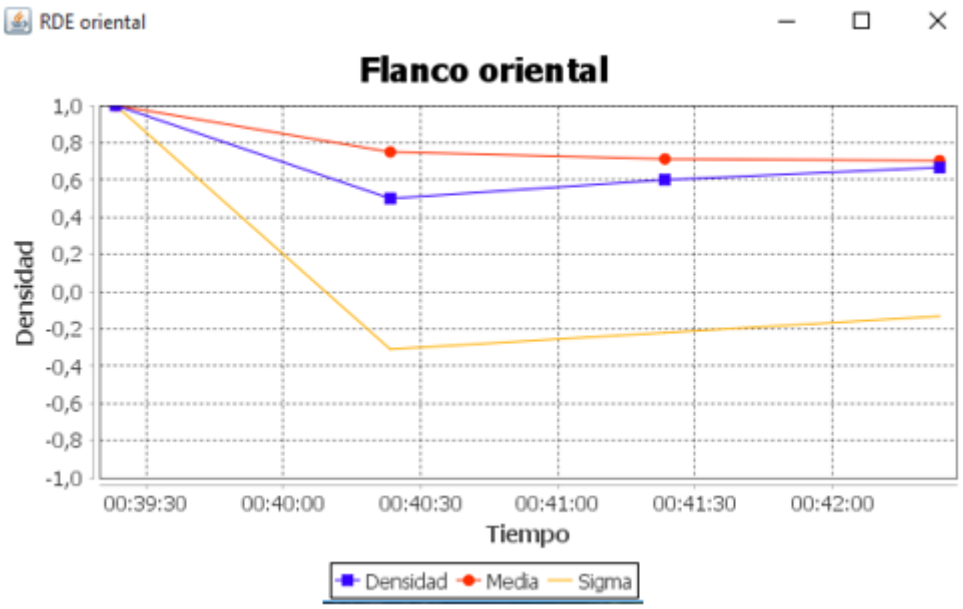


Figura 38. RDE online - prototipo

A diferencia de la figura 38, donde la monitorización es llevada a cabo sin realizar configuraciones previas al prototipo, en la figura 39 se ha configurado el sistema para que la monitorización inicie con un número previo de ejemplos evitando así el arranque en frio. En dicha figura es posible observar 7 anomalías, las cuales como se mencionó anteriormente se identifican porque están debajo de la línea de la variable sigma.

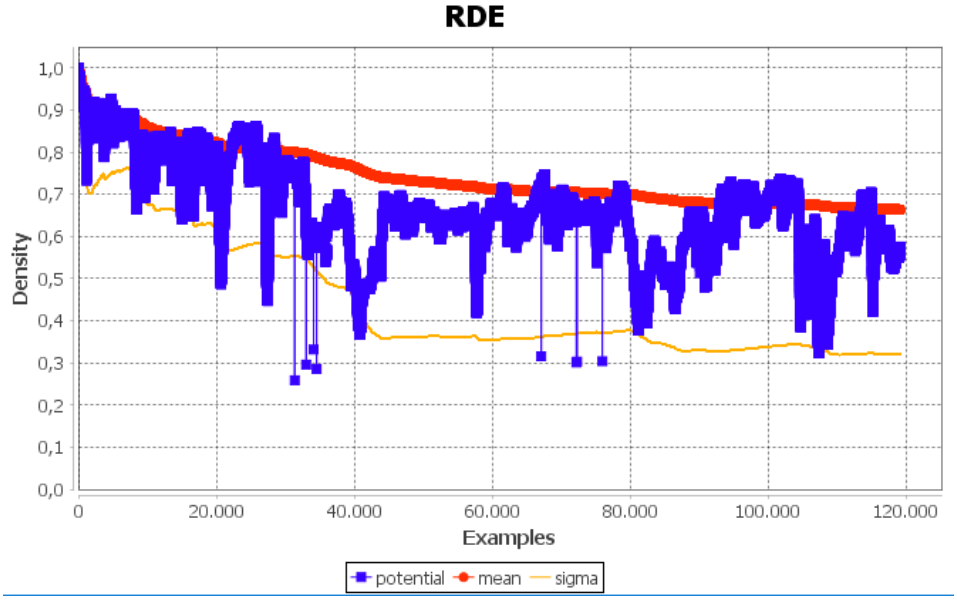


Figura 39. RDE histórico - prototipo

6.3 Iteración 3

En la figura 40 se presenta el diagrama de caso de uso para la detección y clasificación de alertas volcánicas. Adicionalmente, esta funcionalidad permite al usuario configurar los parámetros de entrada para la detección de anomalías en tiempo real.

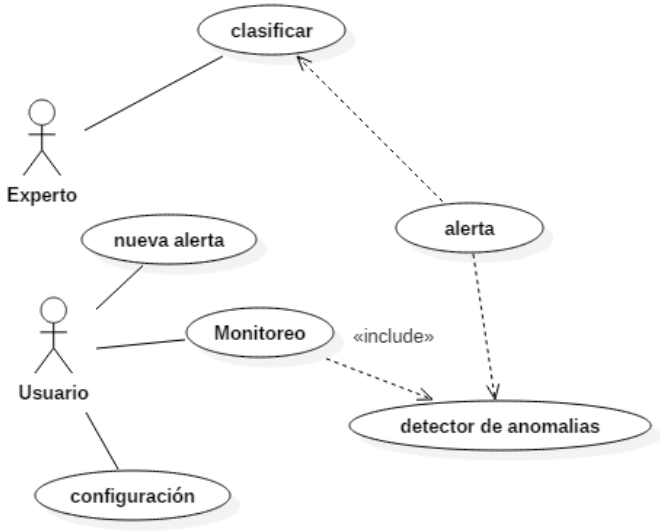


Figura 40. Caso de uso – iteración 3

En la Tabla 42 se presenta la historia de usuario: “Clasificación de alertas”, la cual corresponde a la última iteración.

	Historia de Usuario
Número: 3	**Usuario:** Cliente
Nombre de historia: “Clasificación de alertas”	
Prioridad: Alta	**Riesgo en desarrollo:** Alta
Puntos estimados: 3	**Iteración asignada:** 1
Programador responsable: José Eduardo Gómez Daza	
Descripción: Una vez se haya detectado una anomalía, se debe generar una opción que alerte al usuario monitor sobre la presencia de una anomalía en tiempo real. En dicho caso, el usuario experto es quien debe clasificar la anomalía presentada entre una serie de opciones y dicha clasificación debe almacenarse en un archivo de texto que debe contener los atributos y su variable objetivo. Si esta opción no se encuentra, debe existir una opción para adicionar una nueva opción de clasificación (nueva alerta pre-eruptiva). Por otro lado, la configuración realizada debe funcionar de tal forma que el algoritmo RDE debe tomar como entrada los valores de sigma, normalización y numero de datos anteriores; además de la instancia adquirida en tiempo real para realizar los cálculos y obtener las alertas en caso de ocurrir.	

Validación:
• Verificar que el archivo de clasificaciones este almacenando correctamente todas las alertas pre-eruptivas. • Verificar que los cambios del módulo configuración estén funcionando correctamente a la hora de cargar los datos anteriores.

Tabla 42. Historia de usuario tres

El prototipo cuenta con un módulo de configuración (figura 41), el cual permite elegir el grado de rigurosidad a la hora de detectar anomalías. Esta opción admite los valores 1,2 y 3; siendo la opción uno (1) la menos rigurosa. Generalmente, la opción por defecto es la número tres (3), para que cada anomalía detectada tenga mayor posibilidad de ser una verdadera anomalía y no un falso positivo.

El rendimiento del prototipo es mejor a medida que va obteniendo más instancias a lo largo del tiempo. Por tal razón, se incorporó la opción de procesar un número finito de registros y así evitar el arranque en frio (fase preparatoria). El usuario puede omitir las anomalías que se detecten en la fase de preparación e iniciar la monitorización formal cuando el sistema tenga una madurez aceptable.

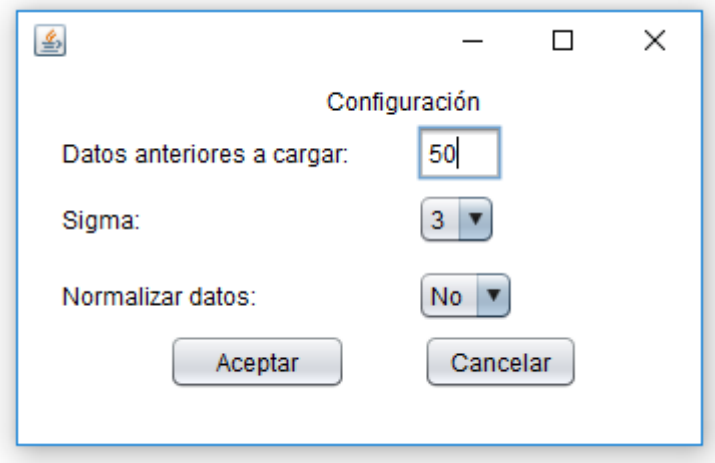


Figura 41. Configuración prototipo

Cuando el sistema detecta una anomalía, automáticamente es generada una ventana emergente que advierte al usuario final de la existencia de una eventualidad en el volcán Puracé. Seguidamente, el experto debe evaluar si efectivamente se trata de una anomalía o no lo es (Figura 42).

Figura 42. Detección de anomalía online - prototipo

Si se ha seleccionado la opción “si”, entonces se despliega una nueva ventana (figura 43), donde el experto debe clasificar a que alerta pre-eruptiva pertenece la anomalía detectada. En caso de no estar dentro del conjunto de opciones, puede crear una nueva alerta. Finalmente, la clasificación se guarda en un archivo de texto plano, que tendrá como atributos los datos adquiridos en ese momento por los sensores de las estaciones involucradas y, como etiqueta, la clasificación realizada por el experto.

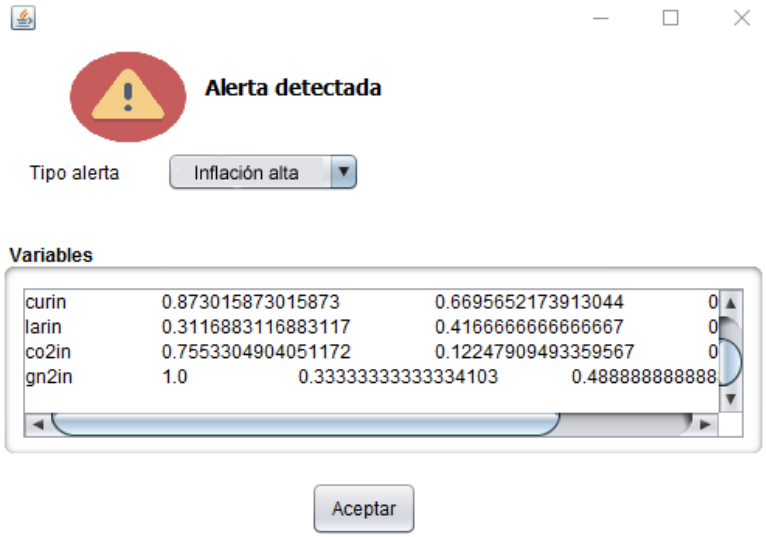


Figura 43. Clasificación de anomalía - prototipo

6.4 Resumen

En este capítulo fueron expuestos los artefactos requeridos por la metodología ágil SCRUM para el desarrollo del prototipo que implementa la detección de alertas volcánicas del volcán Puracé tales como historias de usuarios y diagramas de caso de uso para cada iteración generada. Adicionalmente, se explica la arquitectura completa del sistema creado para cada uno de los enfoques propuestos. Finalmente, se presentan las interfaces graficas de usuario que implementan el funcionamiento del prototipo creado para este proyecto de Maestría.

Capítulo 7

Conclusiones y trabajos futuros

7.1 Conclusiones

Una vez construido y evaluado el mecanismo propuesto para la detección de alertas de origen volcánico, las siguientes conclusiones fueron obtenidas:

1. El mecanismo propuesto para este proyecto de maestría ayuda a los expertos del OVSPo en la generación de alertas pre-eruptivas, dada la detección de anomalías volcánicas a partir del análisis de los datos obtenidos en alguna de las estaciones de monitorización de Geoquímica o Deformación. Esto permite un proceso de toma de decisiones ante eventos de impacto a la población de manera temprana y mejora la labor de monitorización manual a dichos expertos, al identificar las anomalías dentro de un conjunto de datos de altas dimensiones. A pesar de las ventajas que este enfoque significa en la vigilancia del volcán Puracé, este sistema cuenta con un error de precisión, en el mejor de los casos (enfoque por orientación - flanco occidental), de cerca del 9 %.

2. Los errores en los sensores y registro de datos de las estaciones generan una cantidad considerable de información errónea y faltante. Esto hace que los modelos entrenados a partir de dichos datos tengan cierta incertidumbre en la respuesta que generan y disminución de su precisión para la detección de anomalías volcánicas.

3. Abordar dos de las tres áreas más importantes de la vigilancia volcánica (Geoquímica y la Deformación) para la detección de anomalías, y con ello la generación de alertas, permite que el mecanismo desarrollado monitorice el volcán a través de cambios físicos del edificio volcánico y cambios gaseosos, los cuales son detectados por sensores ubicados en estaciones. Los cambios detectados a partir de este enfoque, al ser correlacionados con actividad sísmica, muestran relación entre sus variables.

4. El uso de RDE en el dominio de aplicación de vulcanología, tiene como ventajas:

- Permitir que el sistema este constantemente actualizado a medida que son obtenidos las registros de las estaciones en tiempo real. Esto es necesario debido a que los volcanes son sistemas dinámicos que constantemente presentan cambios y cambian sus líneas base.

- Los volcanes presentan ocasionalmente cambios abruptos, graduales y recurrentes. RDE es un algoritmo capaz de detectar cualquiera de estos cambios sin necesidad de hacer grandes cambios en sus procedimientos, debido a la estructura dinámica que este maneja.

- La monitorización volcánica requiere respuestas rápidas en la generación de alertas, debido al riesgo inminente que puede ocurrir en los alrededores de un volcán. RDE permite obtener respuestas en cortos tiempos, ya que solo almacena en memoria una pequeña cantidad de variables que se actualizan de acuerdo al último registro obtenido desde las estaciones. Por lo tanto, el sistema no requiere un re-entrenamiento del modelo.

5. El enfoque por técnica propuesto permite evaluar dos áreas de vigilancia (Geoquímica y Deformación) a partir de estaciones de su tipo ubicadas en el volcán. No obstante, no fue considerada la determinación de cuál fue la estación o estaciones involucradas en una anomalía volcánica detectada. Cabe aclarar que, en la técnica de Deformación, esta es de origen volcánico si se presenta dicha deformación en más de una estación. Por otro lado, en el enfoque por estación se conoce cuál es la estación que registra los datos que constituyen una anomalía. Sin embargo, con los datos de una sola estación no se puede determinar si un volcán tiene algún indicio de actividad que signifique un riesgo inminente para la población en sus alrededores.

7.2 Trabajos futuros

1. Integrar a la solución final presentada en este proyecto a un sistema de soporte a las decisiones tales como GDSS (Group Decision Support Systems). Esto puede ser de apoyo para orientar al usuario (experto del observatorio) sobre el tipo de alerta volcánica que está generando el enfoque propuesto, profundizando en los datos obtenidos, navegando entre ellos, y manejando dichos datos desde distintas perspectivas.

2. Usar los datos obtenidos por otras estaciones pertenecientes a distintas técnicas de monitorización volcánica de la red de vigilancia del volcán Puracé, tales como magnetómetros, estaciones monitoras de gas radón,

pluviómetros, anemómetros, potenciales eléctricos, GNSS, fuentes termales y fumarolas. Esto podría aportar al mecanismo una mejora en los resultados de las métricas de evaluación obtenidos, así como la generación de alertas de mayor complejidad, al integrar estas nuevas fuentes de información y representar de mejor manera la dinámica de un volcán.

3. Evaluar el prototipo construido en este proyecto de maestría a partir de los datos de otro volcán monitorizado por el Servicio Geológico Colombiano, con el fin de comparar los resultados. Ya que todos los volcanes tienen un comportamiento diferente, reflejado en los datos obtenidos por las estaciones y rangos de sus líneas base, la escalabilidad y generalización del prototipo pueden determinarse al hacer este tipo de pruebas. Para esto bastaría con modificar dentro del código fuente, las estaciones que harán parte de la adquisición.

4. Implementar un algoritmo auto-adaptativo de clasificación o agrupamiento como lo son: Eclass [42] o Eclustering [80] respectivamente. Esto permitiría clasificar automáticamente condiciones que generan los distintos niveles de alertas sin la intervención del experto y extraer patrones del volcán.

5. Explorar el conocimiento de expertos del área con conocimientos en deformación y geoquímica volcánica para identificar posibles variables compuestas (formadas a partir de combinaciones de las obtenidas desde los sensores de las estaciones) que estén ligadas a fenómenos vulcanológicos. Adicionalmente, en el caso de integrar mucha más información, un proceso de reducción de dimensionalidad del conjunto de datos (selección de características) permitiría identificar los atributos del conjunto de datos que no representen una ganancia de información con relación a las variables objetivo de los modelos.

8. Referencias

[1] 1. V. H. Ariscain, *Guía de Preparativos de Salud Frente a Erupciones Volcánicas*. 2015.
[2] Organización panamericana de la salud, “El sector salud, frente al riesgo volcánico, modulo II,” 2005.
[3] White R, Mc Causland W, “Distal Volcano-Tectonic Earthquakes, Usually the first reported precursor to large eruptions and indicate volume of intruding magma. VDAP-USGS. Memorias LAVAS IV, Puerto Vallarta, Mexico,” 2016.
[4] Z. Umar, B. Pradhan, A. Ahmad, M. N. Jebur, and M. S. Tehrany, “Earthquake induced landslide susceptibility mapping using an integrated ensemble frequency ratio and logistic regression models in West Sumatera Province, Indonesia,” *CATENA*, vol. 118, pp. 124–135, Jul. 2014.
[5] A. Messina and H. Langer, “Pattern recognition of volcanic tremor data on Mt. Etna (Italy) with KKAnalysis—A software program for unsupervised classification,” *Comput. Geosci.*, vol. 37, no. 7, pp. 953–961, Jul. 2011.
[6] P. A. Castro-Cabrera, M. Orozco-Alzate, A. Adami, M. Bicego, J. M. Londoño-Bonilla, and G. Castellanos-Domínguez, “A Comparison between Time-Frequency and Cepstral Feature Representations for the Classification of Seismic-Volcanic Signals,” in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, 2014, pp. 440–447.
[7] Oscar Cadena, “Detección y clasificación automática de registros sísmicos en el observatorio vulcanológico y sismologico de pasto utilizando redes neuronales artificiales.” 2011.
[8] B. Sierra, *Aprendizaje automático: conceptos básicos y avanzados*. 2006.
[9] ANDRÉS LEÓN SUÁREZ, “ESTUDIO E IMPLEMENTACIÓN EN MOA DE NUEVOS ALGORITMOS DE APRENDIZAJE INCREMENTAL BASADOS EN SUPPORT VECTOR MACHINES.” 2012.
[10] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin, “Intrusion detection by machine learning: A review,” *Expert Syst. Appl.*, vol. 36, no. 10, pp. 11994–12000, Dec. 2009.
[11] E. Aydoğan and S. Şen, “Analysis of machine learning methods on malware detection,” in *2014 22nd Signal Processing and Communications Applications Conference (SIU)*, 2014, pp. 2066–2069.
[12] S. Shah, S. Reddy, A. Sardeshmukh, B. P. Gautham, G. Shroff, and A. Srinivasan, “Application of Machine Learning Techniques for Inverse Prediction in Manufacturing Process Chains,” in *Proceedings of the 3rd World Congress on Integrated Computational Materials Engineering (ICME 2015)*, Springer, Cham, 2015, pp. 261–268.
[13] Szakacs Alexandru, *What is a volcano?* 2010.
[14] Gregory Hulley and Tshilidzi Marwala, “Evolving Classifiers: Methods for Incremental Learning.” 2008.
[15] R. Polikar, L. Upda, S. S. Upda, and V. Honavar, “Learn++: an incremental learning algorithm for supervised neural networks,” *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 31, no. 4, pp. 497–508, Nov. 2001.
[16] R. Wirth, “CRISP-DM: Towards a standard process model for data mining,” in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 2000, pp. 29–39.
[17] “CRISP-DM 1.0 step by step data mining guide.” [Online]. Available: http://www.whitepapercentral.com/browse/marketing/crisp-dm-1-0-step-by-step-data-mining-guide/. [Accessed: 17-Apr-2018].
[18] Rafael Martínez Gasca, “APLICACIÓN DE HERRAMIENTAS DE APRENDIZAJE SUPERVISADO A LA DIAGNOSIS DE SISTEMAS DINÁMICOS,” 2002.
[19] Daniel Dzurisin, “A comprehensive approach to monitoring volcano deformation as a window on the eruption cycle,” 2003.

[20] Daniel Dzurisin, “Volcano Deformation - geodetic deformation techniques,” 2003.
[21] A.T. Caselli, “Copahue volcano (argentina): a relationship between ground deformation, seismic activity and geochemical changes,” 2015.
[22] A. E. G. Winson, F. Costa, C. G. Newhall, and G. Woo, “An analysis of the issuance of volcanic alert levels during volcanic crises,” *J. Appl. Volcanol.*, vol. 3, p. 14, Sep. 2014.
[23] R. White and W. McCausland, “Volcano-tectonic earthquakes: A new tool for estimating intrusive volumes and forecasting eruptions,” *J. Volcanol. Geotherm. Res.*, vol. 309, pp. 139–155, Jan. 2016.
[24] J.M Ibañez, “Sismicidad volcánica,” 2000.
[25] L. F. M. Maldonado, S. Inguaggiato, M. T. Jaramillo, G. G. Valencia, and A. Mazot, “Volatiles and energy released by Puracé volcano,” *Bull. Volcanol.*, vol. 79, no. 12, p. 84, Dec. 2017.
[26] B. Hernández, “Gases liberados en erupciones Volcánicas, afectan a poblaciones Cercanas al área de emisión.” 2016.
[27] HealthLinkBC, “Calidad del aire exterior Dióxido de azufre (SO2).” 2017.
[28] Universidad de Colima, “Monitoreo Sismico,” Centro de etudios e investigaciones de Vulcanología, 2017.
[29] Servicio Geologico Colombiano, “instructivo primario en deformación volcanica,” 2017.
[30] Ministerior de ambiente, vivienda y desarrollo territorial, “Alertas tempranas: una estrategia para reducir impactos de los desastres naturales y preparación para el cambio climático.” 2008.
[31] “SGC - Servicio Geológico Colombiano - Boletín semanal,” 2016.
[32] “SGC - Informe Técnico,” 2018.
[33] memories conference tenerife, “cities on volcanoes 6th.” 2010.
[34] H. Bhavsar and A. Ganatra, “A Comparative Study of Training Algorithms for Supervised Machine Learning,” *Int. J. Soft Comput. Eng. IJSCE*, vol. 2, Jan. 2012.
[35] S. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques.,” *Inform. Slov.*, vol. 31, pp. 249–268, Jan. 2007.
[36] M. Sayed-Mouchaweh and E. Lughofer, Eds., *Learning in Non-Stationary Environments: Methods and Applications*. New York: Springer-Verlag, 2012.
[37] A. D. V. Govea, *Incremental Learning for Motion Prediction of Pedestrians and Vehicles*. Berlin Heidelberg: Springer-Verlag, 2010.
[38] X. Geng and K. Smith-Miles, “Incremental Learning,” in *Encyclopedia of Biometrics*, S. Z. Li and A. Jain, Eds. Springer US, 2009, pp. 731–735.
[39] J. Xu, C. Xu, B. Zou, Y. Y. Tang, J. Peng, and X. You, “New Incremental Learning Algorithm With Support Vector Machines,” *IEEE Trans. Syst. Man Cybern. Syst.*, pp. 1–12, 2018.
[40] G. Liu, H. r Cheng, Z. g Qin, Q. Liu, and C. x Liu, “E-CVFDT: An improving CVFDT method for concept drift data stream,” in *2013 International Conference on Communications, Circuits and Systems (ICCCAS)*, 2013, vol. 1, pp. 315–318.
[41] P. Angelov and R. Buswell, “Evolving rule-based models: A tool for intelligent adaptation,” in *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)*, 2001, vol. 2, pp. 1062–1067 vol.2.
[42] P. Angelov and R. Buswell, “Identification of evolving fuzzy rule-based models,” *IEEE Trans. Fuzzy Syst.*, vol. 10, no. 5, pp. 667–677, Oct. 2002.
[43] Angelov P., “Anomalous system state identification.” 2016.
[44] Angelov P., “Evolving Systems,” 2017.
[45] Plamen Angelov, “An Approach to Automatic Real-Time Novelty Detection, Object Identification, and Tracking in Video Streams Based on Recursive Density Estimation and Evolving Takagi–Sugeno Fuzzy Systems.”
[46] Anja Feldmann, Gregor Schaffrath, and Stefan Schmid, “Evolving Software,” 2012.
[47] Jose Antonio iglesias, “MODELADO AUTOMÁTICO DEL COMPORTAMIENTO DE AGENTES INTELIGENTES.” 2010.

[48] Markus M. Breunig and Hans-Peter Kriegel, “LOF: Identifying Density-Based Local Outliers.” 2000.
[49] Hans-Peter Kriegel and Matthias Schubert, “Angle-Based Outlier Detection in High-dimensional Data.” 2008.
[50] Jose Antonio Iglesias and Plamen Angelov, “Evolving Human Activity Classifier From Sensor Streams.” 2010.
[51] springer, “Emerging Paradigms in machine learning.” 2013.
[52] D. Kolev, P. Angelov, G. Markarian, M. Suvorov, and S. Lysanov, “ARFA: Automated real-time flight data analysis using evolving clustering, classifiers and recursive density estimation,” in *2013 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 2013, pp. 91–97.
[53] J. A. Iglesias, A. García-Cuerva, A. Ledezma, and A. Sanchis, “Social network analysis: Evolving Twitter mining,” in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016, pp. 001809–001814.
[54] OVSPOP, “INFORME MENSUAL DE ACTIVIDAD DE LOS VOLCANES NEVADO DEL HUILA, PURACÉ Y SOTARÁ ABRIL DE 2017.” 2017.
[55] Servicio Geologico Colombiano, “seguimiento especial de la actividad del volcán puracé.” 2016.
[56] Neil c sturchio and stalney N williams, “The hydrothermal system of Volcan Purace , Colombia.” 1993.
[57] Monica Arcila, “actividad sísmica de los volcanes nevado del Huila y Purace.” 1993.
[58] “Generalidades.” [Online]. Available: https://www2.sgc.gov.co/sgc/volcanes/VolcanPurace/Paginas/generalidades-volcan-purace.aspx. [Accessed: 02-Sep-2018].
[59] “CRC Handbook of Chemistry and Physics, 98th Edition,” *CRC Press*, 13-Jun-2017. [Online]. Available: https://www.crcpress.com/CRC-Handbook-of-Chemistry-and-Physics-98th-Edition/Rumble-Haynes/p/book/9781498784542. [Accessed: 21-Apr-2018].
[60] Fridriksson T, “Diffuse CO2 degassing through soil and geothermal exploration. Presented at “short course on surface exploration for geothermal resources.” 2009.
[61] D. Corrales, A. Ledezma, J. Corrales, D. C. Corrales, A. Ledezma, and J. C. Corrales, “From Theory to Practice: A Data Quality Framework for Classification Tasks,” *Symmetry*, vol. 10, no. 7, p. 248, Jul. 2018.
[62] D. Corrales, J. Corrales, A. Ledezma, D. C. Corrales, J. C. Corrales, and A. Ledezma, “How to Address the Data Quality Issues in Regression Models: A Guided Process for Data Cleaning,” *Symmetry*, vol. 10, no. 4, p. 99, Apr. 2018.
[63] Allison P, “Missing Data. Sage University Papers Series on Quantitative Applications in the Social Sciences,” 2001.
[64] R. Pandey, N. Srivastava, and S. Fatima, “Extending R Boxplot Analysis to Big Data in Education,” in *2015 Fifth International Conference on Communication Systems and Network Technologies*, 2015, pp. 1030–1033.
[65] G. Shevlyakov, K. Andrea, L. Choudur, P. Smirnov, A. Ulanov, and N. Vassilieva, “Robust versions of the Tukey boxplot with their application to detection of outliers,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6506–6510.
[66] H. Kuna *et al.*, “AVANCES EN PROCEDIMIENTOS DE LA EXPLOTACIÓN DE INFORMACIÓN CON ALGORITMOS BASADOS EN LA DENSIDAD PARA LA IDENTIFICACIÓN DE OUTLIERS EN BASES DE DATOS,” Jul. 2018.
[67] P. Angelov, R. Ramezani, and X. Zhou, “Autonomous novelty detection and object tracking in video streams using evolving clustering and Takagi-Sugeno type neuro-fuzzy system,” in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1456–1463.
[68] “Autonomous Learning Systems: From Data Streams to Knowledge in Real-time,” *Wiley.com*. [Online]. Available: https://www.wiley.com/en-

us/Autonomous+Learning+Systems%3A+From+Data+Streams+to+Knowledge+in+Real+time-p-9781119951520. [Accessed: 18-Apr-2018].
[69] P. P. Angelov, *Evolving Rule-Based Models: A Tool for Design of Flexible Adaptive Systems*, Softcover reprint of hardcover 1st ed. 2002 edition. Heidelberg: Physica, 2010.
[70] P. Angelov, "Typicality distribution function #x2014; A new density-based data analytics tool," in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–8.
[71] Plamen Angelov, "Evolving Fuzzy Sistem from data stream in realtime," 2006.
[72] Agustín Alejandro Ortiz, "ALGORITMO MULTICLASIFICADOR CON APRENDIZAJE INCREMENTAL QUE MANIPULA CAMBIOS DE CONCEPTOS," 2014.
[73] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv. CSUR*, vol. 46, no. 4, p. 44, Apr. 2014.
[74] I. e Žliobait\. e, "Learning under Concept Drift: an Overview," Oct. 2010.
[75] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sens. Environ.*, vol. 62, no. 1, pp. 77–89, Oct. 1997.
[76] D. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," *Mach Learn Technol*, vol. 2, Jan. 2008.
[77] search results, *Introduction to Modern Information Retrieval, Third Edition*, 3rd Edition edition. London: Facet Publishing, 2010.
[78] C. J. V. Rijsbergen, "Information Retrieval: Butterworth-Heinemann." 1979.
[79] Ken Schwaber and Jeff Sutherland, "La guia de SCRUM." 2013.
[80] ANGELOV P, "An approach to online identification of takagisugeno fuzzy models." 2004.