



REPÚBLICA DEL ECUADOR

# Escuela Politécnica Nacional

" E SCIENTIA HOMINIS SALUS "

La versión digital de esta tesis está protegida por la Ley de Derechos de Autor del Ecuador.

Los derechos de autor han sido entregados a la "ESCUELA POLITÉCNICA NACIONAL" bajo el libre consentimiento del (los) autor(es).

Al consultar esta tesis deberá acatar con las disposiciones de la Ley y las siguientes condiciones de uso:

- Cualquier uso que haga de estos documentos o imágenes deben ser sólo para efectos de investigación o estudio académico, y usted no puede ponerlos a disposición de otra persona.
- Usted deberá reconocer el derecho del autor a ser identificado y citado como el autor de esta tesis.
- No se podrá obtener ningún beneficio comercial y las obras derivadas tienen que estar bajo los mismos términos de licencia que el trabajo original.

El Libre Acceso a la información, promueve el reconocimiento de la originalidad de las ideas de los demás, respetando las normas de presentación y de citación de autores con el fin de no incurrir en actos ilegítimos de copiar y hacer pasar como propias las creaciones de terceras personas.

***Respeto hacia sí mismo y hacia los demás.***

# **ESCUELA POLITÉCNICA NACIONAL**

## **FACULTAD DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA**

### **IMPLEMENTACIÓN DE UN MÉTODO DE AGRUPACIÓN DE SEÑALES SÍSMICAS GENERADAS POR EL VOLCÁN COTOPAXI BASADO EN APRENDIZAJE AUTOMÁTICO NO SUPERVISADO UTILIZANDO EL MODELO DE MEZCLA GAUSSIANA**

**TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE  
INGENIERO EN ELECTRÓNICA Y TELECOMUNICACIONES**

**DIEGO ALEJANDRO RÍOS CARRILLO**

**DIRECTOR: PhD. FELIPE LEONEL GRIJALVA ARÉVALO**

**CODIRECTOR: PhD. ANA MARÍA ZAMBRANO VIZUETE**

**Quito, febrero 2022**

## **AVAL**

Certificamos que el presente trabajo fue desarrollado por Diego Alejandro Ríos Carrillo, bajo nuestra supervisión.



---

**PhD. FELIPE LEONEL GRIJALVA ARÉVALO**  
**DIRECTOR DEL TRABAJO DE TITULACIÓN**



---

**PhD. ANA MARÍA ZAMBRANO VIZUETE**  
**CODIRECTOR DEL TRABAJO DE TITULACIÓN**

# DECLARACIÓN DE AUTORÍA

Yo, DIEGO ALEJANDRO RÍOS CARRILLO, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración dejo constancia de que la Escuela Politécnica Nacional podrá hacer uso del presente trabajo según los términos estipulados en la Ley, Reglamentos y Normas vigentes.



---

DIEGO ALEJANDRO RÍOS CARRILLO

# DEDICATORIA

A mis padres y hermana.

*Diego Ríos*

# AGRADECIMIENTO

A todas las personas que fueron parte del proceso para culminar mi carrera universitaria.

*Gracias.*

***Diego***

# ÍNDICE DE CONTENIDO

AVAL .....	I
DECLARACIÓN DE AUTORIA.....	II
DEDICATORIA .....	III
AGRADECIMIENTO.....	IV
ÍNDICE DE CONTENIDO.....	V
RESUMEN .....	VII
ABSTRACT .....	VIII
1. INTRODUCCIÓN.....	1
1.1. OBJETIVO GENERAL.....	2
1.2. OBJETIVOS ESPECÍFICOS.....	2
1.3. ALCANCE.....	3
1.4. MARCO TEÓRICO .....	3
1.4.1. SISMICIDAD VOLCÁNICA .....	4
1.4.1.1. Volcán Cotopaxi .....	6
1.4.2. APRENDIZAJE AUTOMÁTICO.....	7
1.4.2.1. Aprendizaje supervisado.....	9
1.4.2.2. Aprendizaje no supervisado .....	9
1.4.3. MÁQUINAS DE VECTOR SOPORTE.....	10
1.4.4. K-MEANS .....	11
1.4.5. MODELO DE MEZCLA GAUSSIANA.....	13
1.4.6. TÉCNICAS PARA EL PROCESAMIENTO DE DATOS .....	16
1.4.6.1. Filtro Butterworth .....	16
1.4.6.2. Densidad espectral de potencia método de Burg .....	16

1.4.6.3.	Normalización de datos con <i>z-score</i> .....	17
1.4.6.4.	Análisis de componentes principales .....	18
1.4.6.5.	t-Distributed Stochastic Neighbor Embedding (t-SNE) .....	18
1.4.7.	MATLAB .....	19
2.	METODOLOGÍA .....	21
2.1.	BASES DE DATOS .....	21
2.2.	MÉTODO PROPUESTO .....	22
2.3.	PREPROCESAMIENTO .....	23
2.4.	APRENDIZAJE DE UN EXTRACTOR DE CARACTERÍSTICAS A PARTIR DE LA BASE DE DATOS NO ETIQUETADA .....	24
2.5.	CLASIFICADOR .....	25
2.6.	CONFIGURACIÓN DE LA ETAPA NO SUPERVISADA .....	25
2.7.	CONFIGURACIÓN DE LA ETAPA SUPERVISADA .....	26
2.8.	MÉTRICAS DE EVALUACIÓN .....	26
2.9.	LÍNEA DE BASE .....	27
3.	RESULTADOS Y DISCUSIÓN .....	28
3.1.	RESULTADOS .....	28
3.2.	DISCUSIÓN .....	31
4.	CONCLUSIONES Y RECOMENDACIONES .....	34
4.1.	CONCLUSIONES .....	34
4.2.	RECOMENDACIONES .....	34
5.	REFERENCIAS BIBLIOGRÁFICAS .....	36
	ANEXOS .....	42



# RESUMEN

El creciente número de investigaciones en el área de aprendizaje automático, demuestra su utilidad para aplicarlo en diferentes campos de estudio. Por ejemplo, se puede agrupar diferentes tipos de sismos volcánicos para conocer con mayor detalle la dinámica de un volcán y su comportamiento. De esta forma se lograría mitigar los riesgos que una posible erupción volcánica podría ocasionar, en beneficio de las personas que habitan en las inmediaciones de un volcán.

En el presente trabajo se realizó una agrupación de señales sísmicas volcánicas utilizando enfoques no supervisados como el modelo de mezcla gaussiana. La agrupación de señales se la utilizó para la obtención de características. Las señales o eventos sísmicos utilizados pasaron por varios pasos antes de ingresar a una etapa supervisada. Se descubrió que los enfoques no supervisados aprenden una amplia gama de características relevantes que están altamente correlacionadas con características dadas por expertos (línea de base) a lo largo de los años. Por último, los resultados demostraron que los enfoques no supervisados que se compararon (modelo de mezcla gaussiana y k-means) lograron rendimientos estadísticamente similares con respecto a la línea de base. Aún más, el modelo de mezcla gaussiana superó a las características de la línea de base en señales ruidosas.

**PALABRAS CLAVE:** Aprendizaje automático, no supervisado, agrupación, modelo de mezcla gaussiana, sismos volcánicos, volcán, Cotopaxi

# ABSTRACT

The growing number of investigations in the machine learning area demonstrates its usefulness to apply it in various fields of study. For example, different types of volcanic earthquakes can be cluster to learn more about the dynamics of a volcano and its behavior. In this context, it would be possible to mitigate the risks that a possible volcanic eruption could cause, for the benefit of the people who live in the vicinity of a volcano.

In the present work, clustering of volcanic seismic signals was carried out using unsupervised approaches such as the Gaussian Mixture Model. The clustering of signals was employed to obtain features. The seismic events that were used went through several steps before entering a supervised stage. Unsupervised approaches were found to allow learning a wide range of important features that are highly correlated with features provided by experts (baseline) over the years. Consequently, the results demonstrated that the unsupervised approaches that were compared (Gaussian Mixture Model and k-means) achieved statistically similar performances with respect to the baseline. Finally, the Gaussian Mixture Model outperformed the baseline features in noisy conditions.

**KEYWORDS:** Machine learning, unsupervised, clustering, Gaussian Mixture Model, volcanic earthquakes, volcano, Cotopaxi

# 1. INTRODUCCIÓN

La erupción de un volcán desencadena varios efectos adversos para las personas que habitan en las inmediaciones del mismo como: las emisiones de gases, caída de ceniza, lahares por los flancos del volcán, corrientes de lava, etc. Desde el año 1500 se han producido aproximadamente 280 mil muertes por efecto de las erupciones volcánicas como se indica en [1] alrededor de todo el mundo. Situándose específicamente en Ecuador según los registros del Instituto Geofísico de la Escuela Politécnica Nacional (IG EPN) existen 97 volcanes que se clasifican en: extinto o en reposo, potencialmente activo, activo y en erupción [2].

Entre los volcanes más representativos del país y monitoreados por el IG EPN están: Antisana (ubicado en la Cordillera Real del Ecuador a unos 50 kilómetros al sur este de la ciudad de Quito), Chiles y Cerro Negro (en la frontera de Ecuador), Cotopaxi (se encuentra sobre la Cordillera Real a 35km al noroeste de la ciudad de Latacunga y a 45km al sureste de Quito), Cuicocha (en la Cordillera Occidental de los Andes del Ecuador a 10km de la ciudad de Cotacachi), Reventador (a 90km al este de Quito), Pichincha (ubicado a 12km al oeste de Quito), Sangay (situado en la provincia de Morona Santiago), Tungurahua (está a 8km al norte de la ciudad de Baños, a 33km al sureste de la ciudad de Ambato y a 120km al sur de Quito) [2]. Cabe señalar que, en las Islas Galápagos, en el registro del IG EPN se muestran alrededor de 20 volcanes.

Por los datos propuestos, existen varios volcanes ubicados en un radio de decenas de kilómetros de algunas ciudades del Ecuador, por lo cual el peligro de que en algún momento uno de estos volcanes pase a estado de erupción está latente. Una forma de conocer el comportamiento de un volcán y prevenir a la población ante cualquier eventualidad, es monitorearlo constantemente, para ello se utilizan sensores que detectan y miden los diferentes parámetros de un volcán como: saltos de gas, medición de deformaciones y análisis de señales sísmicas con sismómetros [3].

Los volcanes día tras día generan un sin número de eventos sísmicos de diferentes tipos, al existir muchos eventos por volcán y teniendo en cuenta que cada tipo de señal puede tener varias características que la hacen única, se hace importante contar con un sistema de agrupamiento de eventos computarizado que considere todas las características de la señal para situarlas en una clase de sismo volcánico. Actualmente existen técnicas eficientes para clasificar y agrupar señales como el aprendizaje automático, de ahí se derivan diferentes métodos para cumplir con el objetivo de agrupar señales [3]. Una de las técnicas de

aprendizaje automático es el no supervisado, que consiste en encontrar una “estructura” en los datos, sin que el algoritmo conozca los tipos de sismos que hay en una base de datos. Este proceso reduciría el alto gasto económico en recursos humanos, para etiquetar los tipos de sismos y el tiempo de procesamiento de las señales. También se podría descubrir nuevas dinámicas en cuanto al comportamiento del volcán vinculadas a diferentes tipos de sismos ya que la actividad de un volcán cambia con el tiempo, esta afirmación se la puede contrastar analizando los informes del volcán Cotopaxi proporcionados por el IG EPN [2].

En este contexto, originalmente se planteó utilizar solo una etapa no supervisada con el modelo de mezcla gaussiana para agrupar tipos de sismos. Se decidió ir más allá para incluir una etapa supervisada para realizar la clasificación de los eventos sísmicos. Se empleará análisis de componentes principales para reducir la redundancia entre las características de las señales a utilizar y se validarán los resultados comparando el modelo de mezcla gaussiana y el modelo k-means como algoritmos de aprendizaje no supervisado.

El presente capítulo inicia con la descripción de los objetivos general y específicos en las Secciones 1.1 y 1.2, respectivamente. Se detalla el alcance del trabajo (Sección 1.3), y se culmina con el marco teórico en la Sección 1.4.

## **1.1. OBJETIVO GENERAL**

Implementar un método de agrupación de señales sísmicas generadas por el volcán Cotopaxi basado en aprendizaje automático no supervisado utilizando el modelo de mezcla gaussiana.

## **1.2. OBJETIVOS ESPECÍFICOS**

- Estudiar los conceptos básicos de sismicidad volcánica y de aprendizaje automático.
- Investigar sobre el modelo de mezcla gaussiana.
- Codificar el método de agrupación de señales sísmicas del volcán Cotopaxi.
- Comparar la exactitud del modelo de mezcla gaussiana con el modelo k-means para la agrupación de señales sísmicas.

### **1.3. ALCANCE**

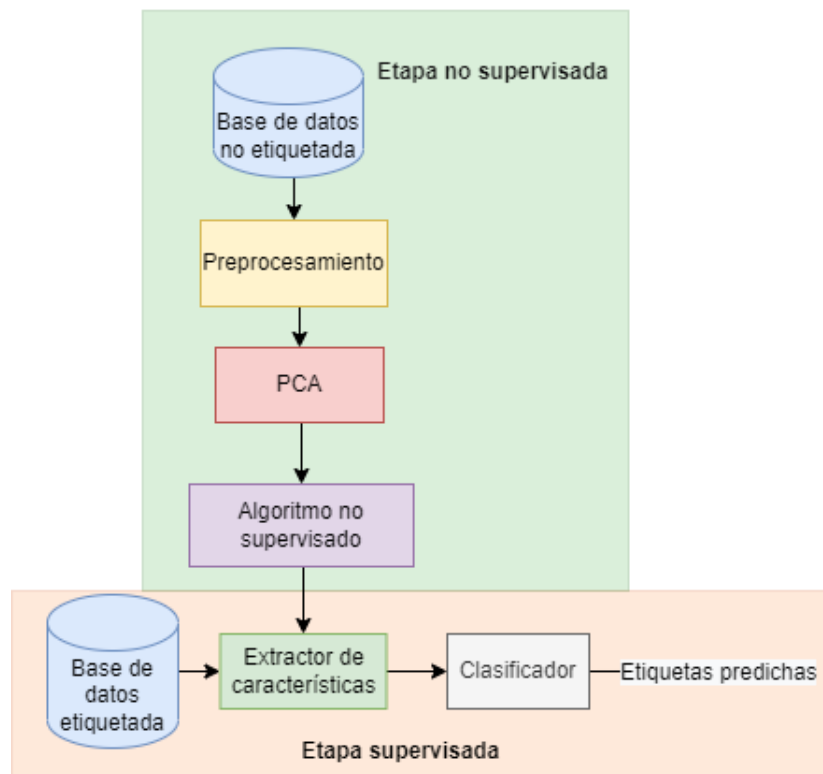
Para el efecto del proyecto técnico se utilizará aprendizaje automático no supervisado con el cual se quiere agrupar los tipos de sismos VT y LP del volcán Cotopaxi. Además, se consideró añadir una etapa supervisada para clasificar los eventos. Se utilizará el modelo de mezcla gaussiana como método de agrupación de señales. El modelo tiene la ventaja de arrojar una probabilidad de pertenencia hacia un tipo de sismo, es decir, el modelo puede devolver al usuario una probabilidad alta de pertenencia a un grupo lo que daría la certeza de que la señal pertenece a ese grupo. De lo contrario, cuando el algoritmo da una probabilidad de pertenencia baja de una señal a un grupo quiere decir que no pertenece a ese tipo de sismo, lo que ayuda a la agrupación de señales sísmicas del volcán.

Se utilizará la base de datos pública de señales sísmicas del volcán Cotopaxi ESeismic [1] que contiene 1187 eventos VT y LP. De forma adicional se considera el uso de dos bases de datos extras una perteneciente al volcán Cotopaxi recientemente liberada por el IG EPN, con 22640 eventos sísmicos de varios tipos, y otra base de datos correspondiente al volcán Llaima ubicado en la Región de La Araucanía (Chile) con 1614 eventos VT y LP, de igual forma recientemente liberada, por el Observatorio Vulcanológico de los Andes Sur (OVDAS) [4]. Se seguirá el esquema presentado en la Figura 1.1. Para que la agrupación tenga una eficiencia alta primero se debe realizar el pre-procesamiento de la base de datos a utilizar, luego se realizará el análisis de componentes principales para utilizar el resultado sobre un algoritmo no supervisado. Consecuentemente, se realiza la obtención de características.

Adicionalmente, se utiliza una etapa supervisada para clasificar los eventos. Para finalizar, los resultados obtenidos con el modelo de mezcla gaussiana serán comparados con los obtenidos en k-means. También se considera realizar una comparación con características bien conocidas obtenidas del volcán Cotopaxi disponibles en ESeismic (línea de base). Para el desarrollo del trabajo se utiliza MATLAB, es importante señalar que el trabajo de titulación no tiene producto final demostrable.

### **1.4. MARCO TEÓRICO**

En el presente apartado se dan a conocer los conceptos principales de los temas involucrados en el desarrollo del trabajo.



**Figura 1.1:** Esquema de implementación del método de agrupación de señales sísmicas del volcán Cotopaxi. Considerando de forma adicional una etapa supervisada

### 1.4.1. SISMICIDAD VOLCÁNICA

Se conoce como sismicidad volcánica al estudio de los eventos sísmicos que tienen un origen volcánico, dichos sismos pueden ocurrir en o cerca del volcán a un radio aproximado de 15 km, e incluso podrían estar relacionados con un proceso volcánico [5]. A diferencia de los sismos tectónicos que tienen una profundidad aproximada de 15 km en la corteza, los volcánicos ocurren en gran cantidad a profundidades entre 1 a 9 km, además en su mayoría se dan en enjambres (grupos de muchos eventos sísmicos) y se cree que representan varios procesos físicos al interior de un volcán [5].

En la actualidad aproximadamente 200 volcanes son monitoreados sísmicamente, un valor cercano al tercio de los volcanes que han entrado en erupción históricamente (514) [5]. De acuerdo a la explicación de [6], los sismos volcánicos están en el escalón más bajo en cuanto a peligrosidad se refiere, esto se puede contrastar debido a que en ninguna erupción previa del volcán Cotopaxi los eventos sísmicos produjeron daños en las construcciones o afectaciones hacia la población. Al contrario, y como beneficio, los sismos volcánicos pueden ayudar a entender el comportamiento de un volcán e incluso anticipar una probable erupción. De forma general, la reactivación de un volcán en su mayoría se precede por un aumento

en la cantidad de sismos en su interior [6].

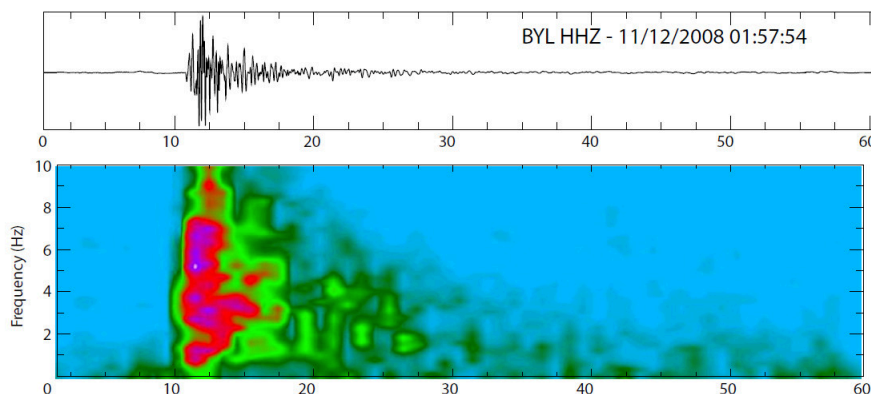
Cuando ocurre un sismo el 99 % de la energía que se libera es en forma de calor y deformación, solamente el 1 % representa a la energía liberada en forma de ondas elásticas [5]. Es decir que los instrumentos de medición sísmica (sismómetros) censan el 1 % de la energía liberada en el sismo. Las ondas elásticas registradas pueden ser de diferentes tipos y se leen en un sismograma (representación del sismo en papel u otro medio) [7]. Una vez que se conoce la señal sísmica se debe determinar a qué tipo de sismo pertenece, para luego construir modelos físicos para determinar las causas probables de las ondas [5]. La forma tradicional de clasificar los eventos se la realiza observando la apariencia del sismograma [5].

De acuerdo a varias fuentes [5], [8], [9] y [10]. Existen varios tipos de sismos volcánicos, tales como: Volcano Tectónico (VT), Largo Período (LP), híbrido (HYB), regional (REG), tremores (TR), tectónicos (TC), explosión (EXP), *icequake* (ICE), tremor espasmódico (TRESP), muy largo período (VLP), tremor de baja frecuencia (TRBA) y tremor armónico (TREMI). Estos eventos sísmicos regularmente son clasificados manualmente por expertos.

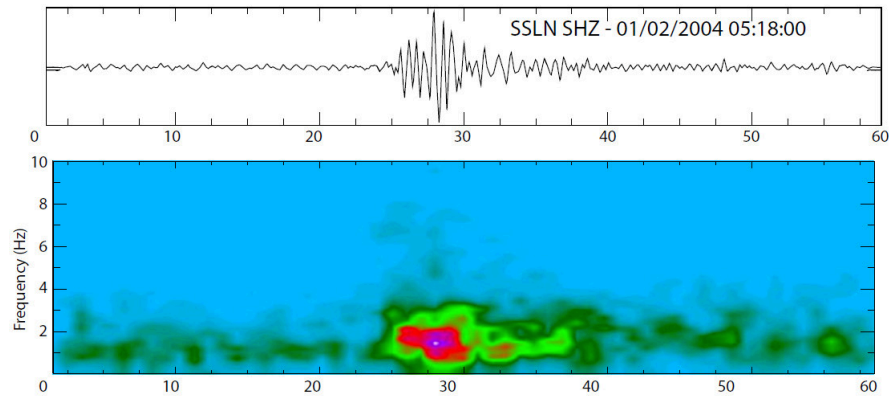
Se definen los principales tipos de sismos a utilizar en el presente trabajo:

- **VT (Volcano-Tectonic, Volcano Tectónico):** Eventos producidos por fracturas internas de roca en el volcán inducidas por movimiento del magma [11]. El espectro de frecuencia puede llegar hasta los 15 Hz [3].
- **LP (Long-Period, Largo Período):** Relacionado con el movimiento de fluidos [3]. Su espectro es de baja frecuencia, comparado con los eventos VT, entre 2 a 5 Hz [12].

En las Figuras 1.2 y 1.3 se muestran ejemplos de los eventos VT y LP, respectivamente.



**Figura 1.2:** Evento VT, forma de onda (arriba) y espectrograma (abajo). Fuente [5]



**Figura 1.3:** Evento LP, forma de onda (arriba) y espectrograma (abajo). Fuente [5]

#### 1.4.1.1. Volcán Cotopaxi

El volcán Cotopaxi es un estratovolcán; volcán de forma cónica, con pendientes mayores a  $20^\circ$  en sus flancos, formado por capas de escoria emanadas en anteriores erupciones y el apilamiento de mantos de lava [6]. Sus coordenadas geográficas son: Latitud  $0^\circ 38'$  Sur; Longitud  $78^\circ 26'$  Oeste. Situado en la Cordillera Oriental de los Andes ecuatorianos a 45 km al norte de Latacunga y a 60 km al sureste de la capital del Ecuador Quito [6], es uno de los volcanes activos cubierto por glaciares más altos del mundo con una elevación de 5876 metros sobre el nivel del mar [13]. Tiene un relieve entre 2000 a 3000 metros desde su base (16 x 19 km) a la cima [6]. A pesar de que en las últimas décadas se ha visto una disminución importante en el área del casquete glaciar que rodea al volcán, los lahares son una consideración importante como un fenómeno peligroso en el caso de que una erupción llegue a suceder.

A lo largo de la historia el volcán Cotopaxi ha presentado varias erupciones, siendo cinco sus grandes procesos eruptivos. Las fuentes bibliográficas inician el registro de erupciones en el año 1533 en el cual se destruyeron seis o siete pueblos indígenas [6]. Entre los años de 1742 a 1744 se produjeron varias erupciones siendo la del 9 de diciembre de 1744 la de mayor magnitud, causando graves daños en la ciudad de Latacunga [14]. En el período de 1766 a 1768, según [14], se dio una gran explosión con el lanzamiento de piedras, material incandescente, ceniza y flujos piroclásticos provocando considerables estragos en el Valle de los Chillos y nuevamente en Latacunga. Entre 1853 y 1854 se presentó una erupción de forma similar a la ocurrida de 1766 a 1768 con varios efectos adversos para las personas que vivían cerca del volcán [6]. En 1877 se da la mayor erupción histórica documentada a detalle del volcán Cotopaxi, el proceso eruptivo se dio a inicios del año, y el 26 de junio



aproximadamente a las 10 am el volcán arrojó lava derritiendo los glaciares y generando grandes lahares que dejarían cuantiosas pérdidas económicas y alrededor de 1000 personas fallecidas [14].

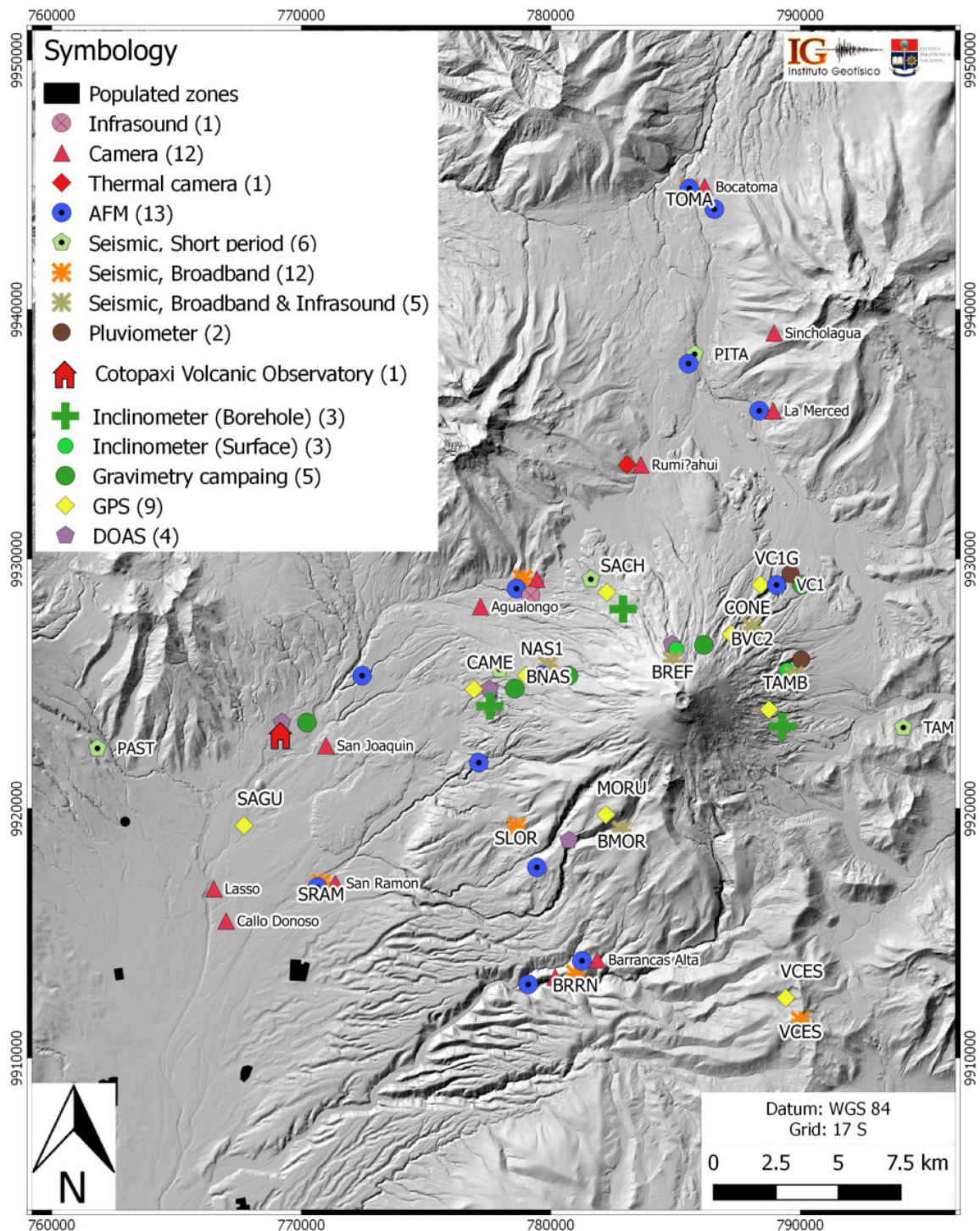
Enfocándose en los eventos sísmicos del Cotopaxi, como se afirma en [5], en las erupciones descritas anteriormente las personas que habitaban en las inmediaciones del volcán sintieron algunos sismos de origen volcánico. Los sismos no produjeron daños en las construcciones ni afectaciones en la salud de la población, más bien fueron una advertencia ante el daño que causaría el volcán [6]. Como formas de prevención y estudio, desde 1977 el volcán Cotopaxi es monitoreado de forma permanente por el Instituto Geofísico de la Escuela Politécnica Nacional (IG EPN), que posee una red de monitoreo en las inmediaciones del volcán. La red está compuesta por cámaras, pluviómetros, inclinómetros, sismómetros, entre otros instrumentos de medición [15] cuya ubicación se puede visualizar en la Figura 1.4.

#### **1.4.2. APRENDIZAJE AUTOMÁTICO**

El avance tecnológico ha hecho posible recopilar una gran cantidad de datos que se generan día tras día, para que todos estos datos sean útiles se los debe procesar y lograr un significado. El aprendizaje automático o más conocido como *Machine Learning*, en [16], se define como un método computacional donde la capacidad para resolver un determinado problema se construye mediante ejemplos anteriores. El aprendizaje automático busca predecir comportamientos o generar estructuras en determinadas situaciones a partir de una base de datos. *Machine Learning* utiliza la teoría de la estadística con el fin de construir modelos matemáticos para realizar inferencias a partir de una muestra [17].

De [17] se extraen algunos ejemplos en donde se puede aplicar el aprendizaje automático. En medicina, para el diagnóstico médico. En el sector financiero, los bancos analizan datos para crear modelos que se utilizarán en aplicaciones de mercado de valores, detección de fraudes y análisis de créditos. En telecomunicaciones, para mejorar la calidad del servicio y optimizar la red. En física, astronomía y biología solo las computadoras son capaces de analizar grandes cantidades de datos. En el presente trabajo se utiliza el aprendizaje automático para analizar señales sísmicas de origen volcánico, con el fin de agrupar sismos, esto podría desembocar en una posible prevención y mitigación de riesgos relacionados a una eventual erupción volcánica.

El aprendizaje automático se divide en: aprendizaje supervisado, semi-supervisado, no su-



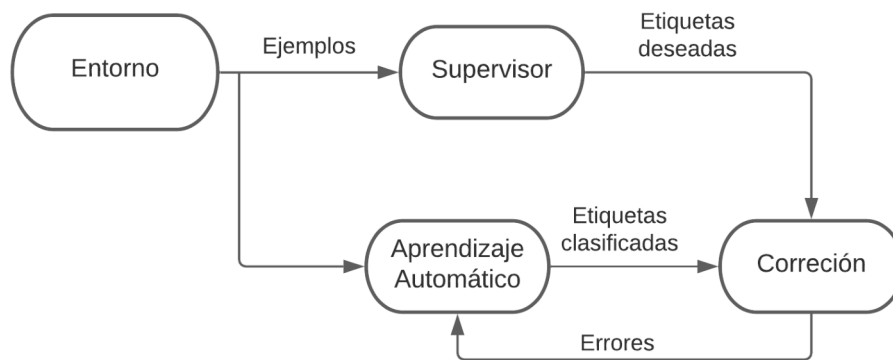
**Figura 1.4:** Red de monitoreo del volcán Cotopaxi, IG EPN. Fuente [15]

pervisado y reforzado. En las Secciones 1.4.2.1 y 1.4.2.2 se describe los aprendizajes supervisado y no supervisado, respectivamente. Los aprendizajes semi-supervisado y reforzado no se incluyen en el presente trabajo. Cada uno de estos tipos de aprendizaje se compone de diferentes algoritmos, dichos algoritmos se diferencian por los modelos que utilizan, los

criterios de rendimiento que optimizan o la forma en que se ajustan los parámetros en la optimización [18].

#### 1.4.2.1. Aprendizaje supervisado

El aprendizaje supervisado es un modelo de aprendizaje automático en donde los parámetros se optimizan para minimizar la diferencia entre el resultado objetivo y el resultado calculado [16]. El objetivo de este tipo de aprendizaje es enseñarle a la computadora cómo realizar algo después de darle las respuestas correctas [19]. En [18] se indica que, en el aprendizaje supervisado existe un supervisor que proporciona el resultado deseado. Por ejemplo, el tipo de señal sísmica (VT o LP) al que pertenece un evento. En la Figura 1.5 se ilustra la definición de aprendizaje supervisado.



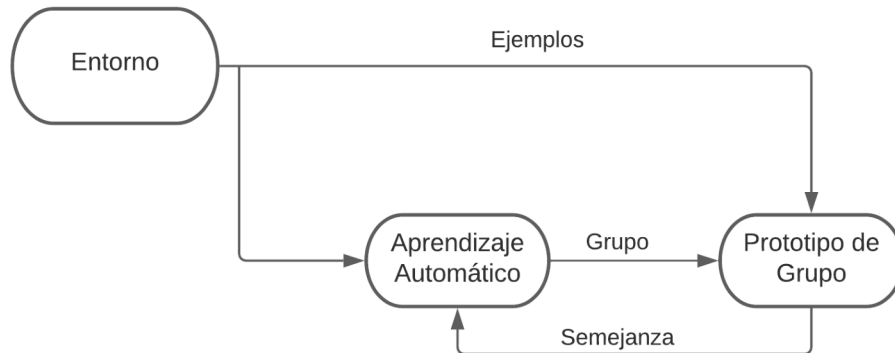
**Figura 1.5:** Definición ilustrada de aprendizaje supervisado. Adaptada de [16]

A la tarea de clasificación se aplica algoritmos de aprendizaje supervisado, cada algoritmo de aprendizaje realiza un conjunto de suposiciones sobre los datos para lograr clasificarlos [18]. Puntualmente, en el presente trabajo se utiliza máquinas de vector soporte como algoritmo de aprendizaje supervisado.

#### 1.4.2.2. Aprendizaje no supervisado

El aprendizaje no supervisado, en el cual se realiza un agrupamiento o *clustering*, es otro modelo de aprendizaje automático que a diferencia del supervisado no utiliza un supervisor y solo se tiene datos de entrada [17]. De la expresión *clustering* se deriva la palabra *cluster* que significa grupo o estructura. En este tipo de aprendizaje las estructuras de agrupación se inicializan al azar y los ejemplos de formación no están etiquetados [16]. El objetivo del aprendizaje no supervisado es encontrar una estructura, semejanza o característica en los

datos para agruparlos entre sí, es decir que la computadora aprenda por sí misma [19]. Por ejemplo, se tiene varios eventos de sismos volcánicos entre los cuales no se conoce el tipo al que pertenece cada evento, pero cada observación guarda similitudes con otros eventos. Debido a las similitudes que tienen los eventos entre sí se los puede agrupar en diferentes *clusters*. En la Figura 1.6 se ilustra la definición de aprendizaje no supervisado.



**Figura 1.6:** Definición ilustrada de aprendizaje no supervisado. Adaptada de [16]

Existen varios tipos de algoritmos basados en aprendizaje no supervisado. En el presente trabajo se utilizan k-means y el modelo de mezcla gaussiana.

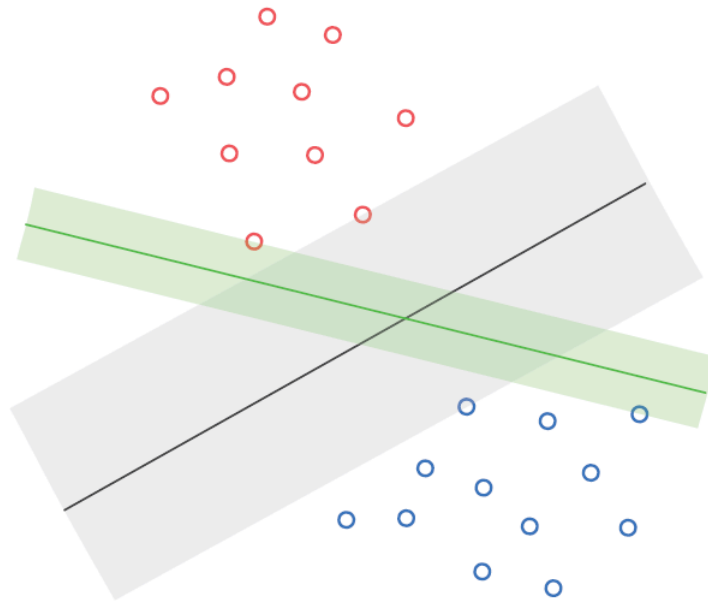
### 1.4.3. MÁQUINAS DE VECTOR SOPORTE

Máquinas de vector soporte (*Support Vector Machine*, SVM) es un algoritmo de aprendizaje automático supervisado, es decir que se utiliza como un clasificador de datos binarios. SVM dibuja márgenes como decisión de límites entre clases en un conjunto de datos proporcionado [20]. Estos márgenes se crean de forma en que la distancia entre cada clase y el margen más cercano sea la máxima posible, de esta forma se consigue el objetivo de obtener el menor error posible.

Además, los márgenes se definen como la distancia entre dos vectores soporte separados por un hiperplano [16]. Dado que entre las observaciones de dos clases puede haber un número infinito de hiperplanos, SVM encuentra el hiperplano que maximiza el margen entre las clases, por lo que las nuevas observaciones tienden a clasificarse con mayor precisión.

En la Figura 1.7, se muestra una representación de un clasificador utilizando SVM, se observa que el separador de color negro divide el espacio de una forma más uniforme con relación a los conjuntos de datos, es notorio que tiene un margen más grande. Con el hiperplano de

color negro se tenderá a clasificar los puntos futuros con mayor precisión. Ahora, analizando el separador verde (con margen más pequeño) se puede ver que divide el espacio de una forma incómoda y tenderá a clasificar erróneamente los puntos futuros. Por lo tanto, se puede deducir que un hiperplano con un mayor margen distinguirá mejor las clases, ya que divide el espacio de una forma más equitativa de acuerdo a la distribución de los datos, como se afirma en [16].



**Figura 1.7:** Representación de un clasificador utilizando máquinas de vector soporte. Fuente [21]

#### 1.4.4. K-MEANS

K-means es el algoritmo más popular y utilizado en aprendizaje automático no supervisado [19]. El algoritmo ha sido utilizado en diferentes estudios para crear *clusters* en bases de datos sin etiquetar según la distancia media entre los *clusters* [20]. El parámetro,  $K$ , en k-means corresponde al número de *clusters*; dicho número debe decidirse con antelación utilizando la heurística del codo [16], la cual se explica en esta Sección.

El algoritmo k-means se puede resumir en los siguientes pasos, [22–24].

**Paso 1:** Inicializar aleatoriamente los  $K$  centroides del *cluster*  $\mu$ , Ecuación 1.1.

$$(\mu_1, \mu_2, \mu_3, \dots, \mu_K) \tag{1.1}$$

**Paso 2:** Asignar observaciones al centro del *cluster* más cercano de acuerdo a la Ecuación 1.2

$$\ell_i \leftarrow \arg \min_k \|\mu_k - \mathbf{x}_i\|_2^2, \quad (1.2)$$

donde  $\ell_i$  es la etiqueta inferida (es decir, *cluster*) dada la  $i$ -ésima observación  $\mathbf{x}_i$ .

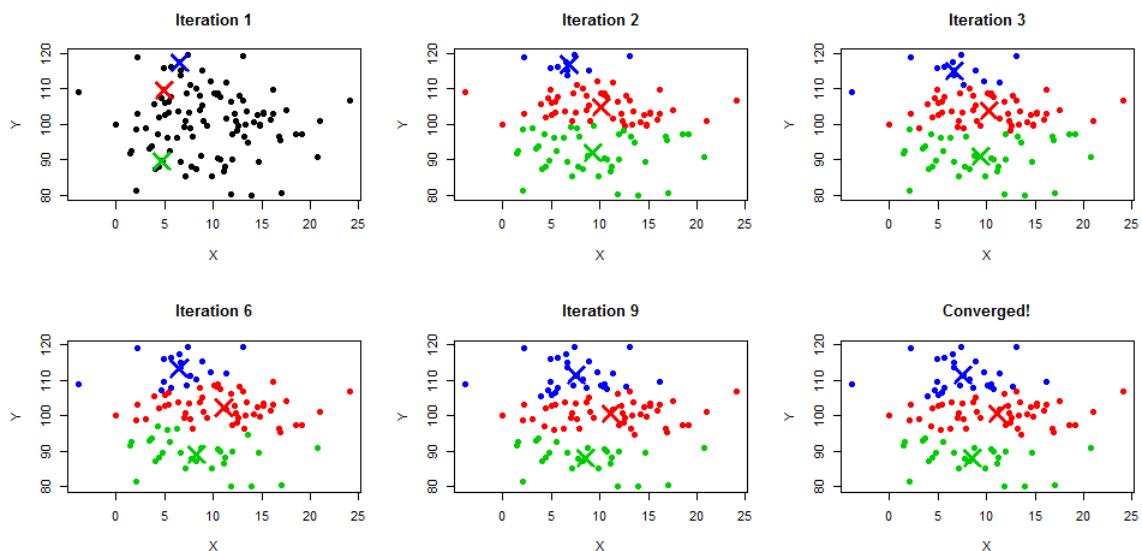
**Paso 3:** Actualizar la posición de los centroides según las observaciones, [24], que se asignaron a los *clusters* según la Ecuación 1.3

$$\mu_k = \frac{1}{n_k} \sum_{i:\ell_i=j} \mathbf{x}_k, \quad (1.3)$$

donde,  $n_k$  es el número de observaciones en el *cluster*  $k$ . Básicamente, en este paso se calcula la media de las observaciones en ese *cluster*, que a su vez será el nuevo centro del *cluster* [24].

**Paso 4:** Los pasos 2 y 3 se repiten hasta alcanzar la convergencia, es decir, cuando no hay más cambios en la asignación de observaciones a los *clusters*.

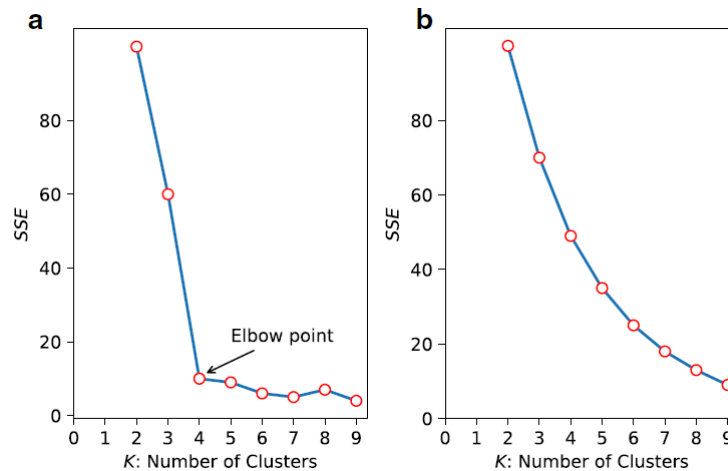
En la Figura 1.8 se presenta gráficamente el proceso descrito en los pasos anteriores. En la iteración 1 la inicialización de los centroides, en la iteración 2 la asignación de observaciones, en el paso 3 se actualizan las posiciones de los centroides para luego asignar nuevamente observaciones al centro del *cluster* más cercano y finalmente la repetición de los pasos 2 y 3 en las demás iteraciones hasta que el modelo converja.



**Figura 1.8:** Algoritmo k-means ilustrado. Fuente [25]

### Heurística del codo para elegir el número de *clusters*

Los algoritmos que se utilizan en aprendizaje automático no supervisado requieren de una selección del número de *clusters* que se van a emplear. La elección de este número es algo complicada cuando no se tiene un conocimiento extenso de la base de datos [26]. La heurística del codo es una alternativa para seleccionar el número de *clusters* a utilizar, es una técnica gráfica en la cual se visualiza el codo de una curva [27]. En la Figura 1.9 **a**, se observa el punto del codo explícitamente, y en **b**, el codo de la curva es ambiguo por lo cual se puede usar una heurística para elegir el número de *clusters*.



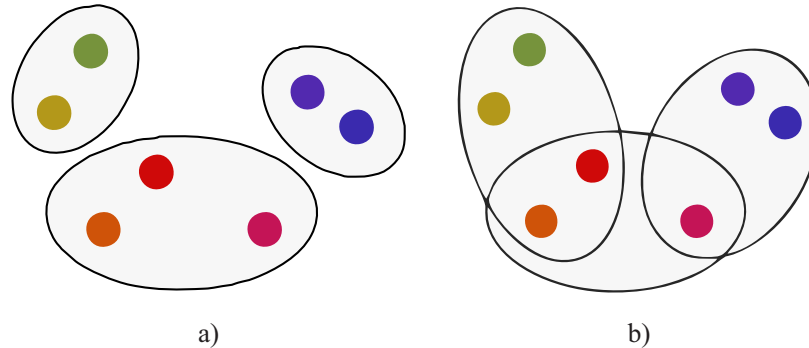
**Figura 1.9:** Curvas heurística del codo. Fuente [26]

En el eje  $x$  de la Figura 1.9 se especifica el número de *clusters*, el eje  $y$  corresponde a la métrica del algoritmo que se implemente sea k-means, modelo de mezcla gaussiana u otro. La heurística del codo se utiliza en el presente trabajo tanto para k-means como para el modelo de mezcla gaussiana.

#### 1.4.5. MODELO DE MEZCLA GAUSSIANA

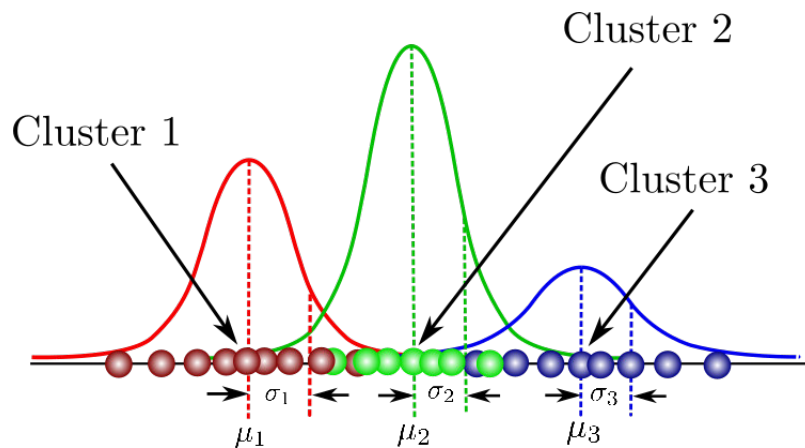
El modelo de mezcla gaussiana (*Gaussian Mixture Model*, GMM) es un algoritmo de aprendizaje no supervisado basado en probabilidades. Para explicar la forma en que GMM realiza las asignaciones de las observaciones a los *clusters* primero se debe mencionar que k-means lleva a cabo una asignación dura. Es decir que, k-means simplemente asigna observaciones a los *clusters*, es un modelo determinista [28]. Se conoce como asignación dura cuando el algoritmo no duda sobre la pertenencia de un evento a un *cluster*, simplemente se destina la observación al *cluster* [29]. GMM realiza una asignación suave, esto quiere decir que además de los *clusters* que se generan en k-means, también entrega probabilidades de pertenencia de un evento a cada uno de los *clusters*. Por ello, se puede apreciar si la

pertenencia de una observación a un *cluster* no está bien definida [28]. En la Figura 1.10 se observa la diferencia entre la asignación dura realizada por k-means y la asignación suave que efectúa GMM.



**Figura 1.10:** a) Asignación dura (k-means) y b) asignación suave (GMM)

Un modelo de mezcla gaussiana está formado por dos o más distribuciones gaussianas. En la Figura 1.11 se observa cómo ocurre la mezcla entre distribuciones gaussianas para tres *clusters*.

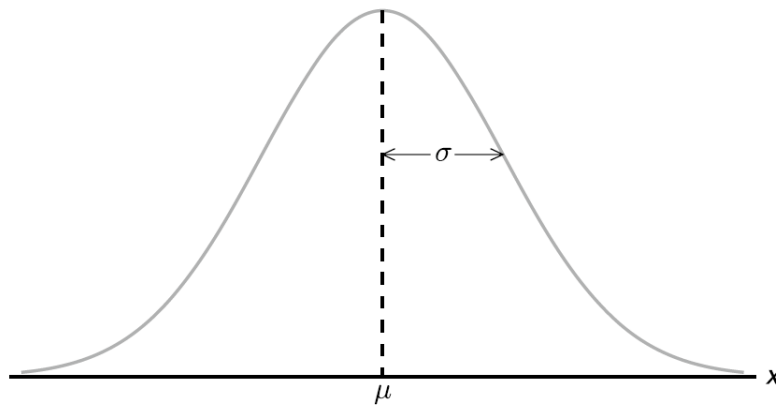


**Figura 1.11:** Mezcla de distribuciones gaussianas para tres *clusters*. Fuente [30]

### Distribución Gaussiana

La distribución normal o gaussiana es una distribución de probabilidad en forma de campana donde la probabilidad es máxima en el promedio [16]. Los parámetros de la distribución gaussiana escalar son la media o promedio  $\mu$  y la varianza  $\sigma^2$ , donde  $\sigma$  es la desviación estándar [28]. La distribución gaussiana se ubica en  $\mu$ , que indica el centro de la distribución. En tanto que  $\sigma$  muestra el esparcimiento de la normal, como se visualiza en la Figura 1.12. GMM asigna una observación a un determinado *cluster* con ciertas probabilidades de pertenencia. Estas probabilidades asumen una distribución gaussiana multivariante  $N(\mathbf{x}|\mu, \Sigma)$  que está completamente especificada por un vector medio  $\mu$  y una matriz de covarianza  $\Sigma$ .





**Figura 1.12:** Curva de la distribución gaussiana o normal

El modelo de mezcla gaussiana es un caso específico de los modelos de mezcla [16], por lo cual, cuando se tiene una mezcla de distribuciones gaussianas, es necesario definir el peso  $\pi$  que está asociado a cada distribución gaussiana. La suma de los pesos de todos los *clusters* es 1. Por lo tanto, cada componente de la mezcla representa un *cluster* único  $k$  especificado por los parámetros del modelo  $\{\pi_k, \mu_k, \Sigma_k\}$ , que se estiman mediante el algoritmo de maximización de expectativas (*Expectation Maximization*, EM).

Según el modelo, la probabilidad previa (*prior probability*) de que la etiqueta inferida sea  $k$  viene dada por la Ecuación 1.4

$$p(\ell_i = k) = \pi_k, \quad (1.4)$$

donde  $\ell_i$ , como en k-means, es la asignación de *cluster* para la observación  $\mathbf{x}_i$  [20]. Además, se sabe que la probabilidad (*likelihood*) de cierta observación  $\mathbf{x}_i$  dada la etiqueta inferida  $\ell_i$  es  $k$ , sigue la Ecuación 1.5

$$p(\mathbf{x}_i | \ell_i = k; \mu_k, \Sigma_k) = N(\mathbf{x}_i | \mu_k, \Sigma_k) \quad (1.5)$$

GMM produce un conjunto de asignaciones suaves por observación  $\mathbf{x}$ , es decir, las probabilidades de que una observación pertenezca a cada uno de los  $k$  *clusters*. Una vez que el algoritmo EM aprende los parámetros del *cluster*  $\{\pi_k, \mu_k, \Sigma_k\}$ , las asignaciones suaves se cuantifican mediante el vector de responsabilidad  $\mathbf{r}_{ik} \in \mathbb{R}^k$ . Concretamente,  $\mathbf{r}_{ik}$  describe la probabilidad de que la etiqueta inferida sea  $k$  dada una observación  $\mathbf{x}_i$  y los parámetros del *cluster* [16] de acuerdo con la Ecuación 1.6

$$\mathbf{r}_{ik} = p(\ell_i = k | \{\pi_j, \mu_j, \Sigma_j\}_{j=1}^K, \mathbf{x}_i), \quad (1.6)$$

donde  $K$  es el número total de *clusters* en el modelo de mezcla.

Finalmente, aplicando la regla de Bayes se calcula el vector de responsabilidad para una nueva instancia arbitraria  $\mathbf{x}_i \in \mathbb{R}^d$ , Ecuación 1.7

$$\mathbf{r}_{ik} = \frac{\pi_k N(\mathbf{x}_i | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(\mathbf{x}_i | \mu_j, \Sigma_j)}, \quad (1.7)$$

### Algoritmo de maximización de expectativas

El algoritmo de maximización de expectativas (*Expectation Maximization*, EM) es el encargado de obtener los parámetros, en este caso para el modelo de mezcla gaussiana. De forma general, el algoritmo empieza con una estimación inicial aleatoria de los parámetros. Luego continua con el paso E, en donde se calcula una distribución de probabilidad sobre posibles casos de pertenencia u observaciones “esperadas” con los parámetros iniciales [31]. El algoritmo sigue con el paso M, en el cual se reestiman los parámetros del modelo, se conoce como paso M ya que la reestimación del modelo se puede considerar como una “maximización” de la probabilidad esperada de los datos [32]. Después de varias repeticiones de los pasos E y M de forma alternada, el algoritmo converge.

## 1.4.6. TÉCNICAS PARA EL PROCESAMIENTO DE DATOS

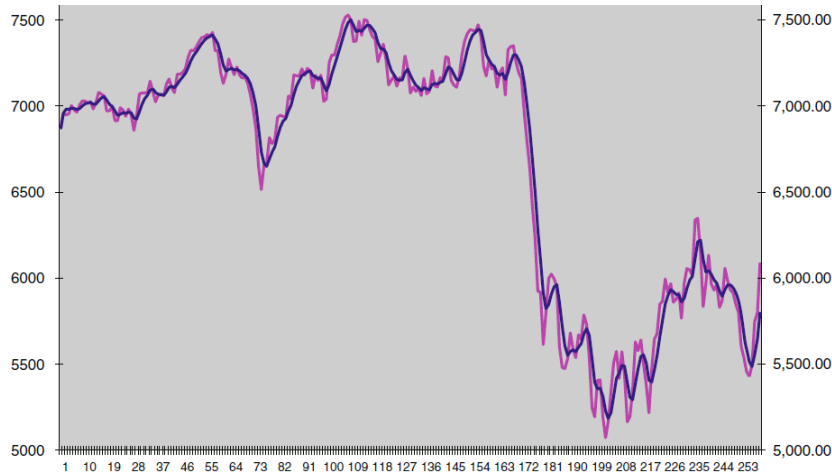
Para que exista una correcta implementación del método de agrupación de señales, es necesario procesar los datos antes de utilizar los algoritmos de k-means o GMM. En el procesamiento de los datos se aplican diferentes técnicas que se explican a continuación.

### 1.4.6.1. Filtro Butterworth

Steven Butterworth, ingeniero inglés, en el año 1930 enfrentó y resolvió el problema de encontrar un filtro plano en su máxima expresión [33]. Encontrar un filtro plano en su máxima expresión, quiere decir que todas las frecuencias que se encuentran dentro de la banda de paso sean tratadas de la misma manera [34]. La respuesta que se espera obtener del filtro Butterworth se observa en la Figura 1.13.

### 1.4.6.2. Densidad espectral de potencia método de Burg

Estimar la densidad espectral de potencia (*Power Spectral Density*, PSD) es una de las técnicas de procesamiento de señales más importantes, consiste en identificar el contenido



**Figura 1.13:** Respuesta de un filtro Butterworth (en azul) sobre una señal (en morado). Fuente [33]

espectral de una cantidad variable en el tiempo (señal) [35]. Experimentalmente, se utilizan analizadores de espectro para observar el contenido espectral de una señal. Los métodos habituales de análisis espectral se pueden dividir en dos clases principales:

- Métodos no paramétricos
- Métodos paramétricos

Existen varios algoritmos para realizar la estimación de la PSD. Entre los algoritmos no paramétricos está la Transformada de Fourier que no es un método óptimo porque sufre efectos de fuga espectral debido a la ventana de muestreo [36].

Por otro lado, uno de los algoritmos que se encuentra dentro de los métodos paramétricos es el método de Burg. El método de Burg se utiliza para el procesamiento de señales tanto en el dominio del tiempo, como en el dominio de la frecuencia. La ventaja de Burg, es que puede suavizar el espectro en comparación con métodos basados en frecuencia para encontrar las frecuencias características ajustando un modelo autorregresivo (AR) de un orden dado [37], para mayor detalle de AR referirse a [38].

#### 1.4.6.3. Normalización de datos con z-score

La normalización de los conjuntos de datos se la realiza para trabajar con valores comparables entre si, es decir que sean lo más parejos posibles. En el presente trabajo se utiliza z-score como método de normalización. En donde, z-score sigue la Ecuación 1.8

$$\mathbf{z} = \frac{\mathbf{x} - \mu}{\sigma}, \quad (1.8)$$

donde,  $\mathbf{x}$  es la observación,  $\mu$  la media y  $\sigma$  la desviación estándar. Las columnas de la matriz normalizada se centran a una media 0 y una escala para tener una desviación estándar de 1 como se indica en [39].

#### **1.4.6.4. Análisis de componentes principales**

El análisis de componentes principales (*Principal Component Analysis*, PCA), es una técnica multivariante [19]. PCA analiza varias variables dependientes que se relacionan entre sí [40], es una transformación de base ortogonal en la que los primeros componentes principales tienen la mayor parte de la varianza del conjunto de datos [41]. De [42], se rescata que PCA es una técnica clásica en la compresión de datos, el análisis de datos estadísticos y la extracción de características. Cuando se maneja una gran cantidad de características es necesario reducirlas y filtrarlas para elegir las más significativas, esto se contrasta en [43], en donde se realiza una selección de características y luego se aplica el modelo de mezcla gaussiana justamente para agrupar eventos sísmicos del volcán Cotopaxi.

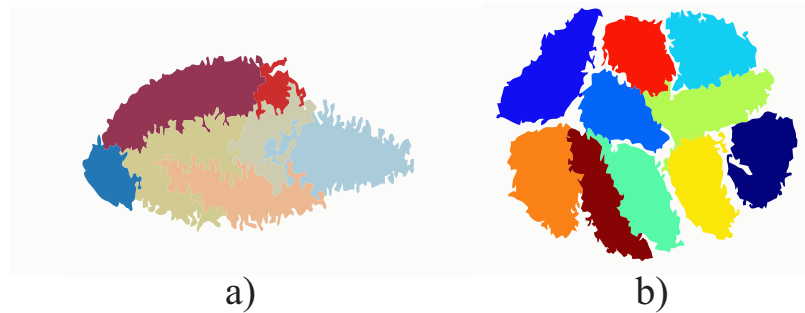
Los objetivos de PCA se resumen en los siguientes [40]:

1. Extraer la información más importante de una matriz o tabla de datos.
2. Comprimir o reducir el tamaño de los datos.
3. Analizar la estructura de las variables y observaciones.

PCA calcula nuevas características (componentes principales) que se obtienen como combinaciones lineales de los datos originales [40]. El primer componente principal tendrá la mayor varianza posible [19]. El segundo componente también debe tener la mayor varianza posible y como requisito tiene que ser ortogonal al primer componente [41]. Los componentes que faltan se calculan de forma similar, reteniendo la mayor cantidad de varianza en cada componente. El proceso matemático para calcular los componentes principales se encuentra con mayor detalle en [43].

#### **1.4.6.5. t-Distributed Stochastic Neighbor Embedding (t-SNE)**

La técnica t-SNE es útil para visualizar datos, especialmente cuando se quiere observar bases de datos con alta dimensionalidad en enfoques no supervisados [44]. t-SNE hace un mejor trabajo que PCA en la visualización de datos, esta afirmación se puede constatar observando la Figura 1.14 en donde para el caso a) PCA, se puede ver que los grupos no



**Figura 1.14:** Visualización de una base de datos con a) PCA y b) t-SNE

están bien separados y es difícil diferenciar cada uno. Por otro lado, en b) t-SNE, se realiza una mejor separación de grupos en donde es fácil distinguirlos.

En el caso de t-SNE se reduce la dimensionalidad de los datos a dos o tres dimensiones, conservando la mayor cantidad de información posible de la base de datos original [44].

#### 1.4.7. MATLAB

MATLAB (*MATrix LABoratory*), laboratorio de matrices, es un programa desarrollado por Math Works, Inc. especialmente para el procesamiento digital de señales [45]. El programa está disponible en diferentes sistemas operativos Windows o UNIX [46]. MATLAB es interactivo y también eficaz para procesar datos en forma de vectores o matrices. Lo que hace a MATLAB un programa enfocado a la investigación y estudio es la integración entre un ambiente de programación (*scripts*), el cálculo digital y la visualización de datos [45]. El lenguaje de código que usa MATLAB es similar al lenguaje de programación C. MATLAB contiene varias funciones matemáticas robustas [47], que permiten desarrollar aplicaciones en diferentes áreas de investigación. Adicionalmente, MATLAB tiene varios *toolboxes* que son extensiones de las funciones básicas y desempeñan tareas específicas en el procesamiento digital de señales [45]. Los *toolboxes* que se utilizan en el presente trabajo son los siguientes:

- **Signal Processing Toolbox:** Tiene funciones para el preprocesamiento, análisis y extracción de características de señales. Incluye herramientas para la estimación del espectro de potencia, filtrado, remuestreo y suavizado de señales. En cuanto a la extracción de características, mediante el *toolbox* se puede encontrar patrones y determinar similitudes entre señales.
- **Statistics and Machine Learning Toolbox:** Contiene funciones y aplicaciones para modelar, describir, y analizar datos. El *toolbox* puede llevar a cabo pruebas de hipótesis y utilizar *clustering* para el análisis exploratorio de datos. La aplicación *Classification*

*Learner*, incluida en el *toolbox*, puede realizar inferencias a partir de datos, con el fin de clasificar señales en diferentes tipos. Principalmente, el *toolbox* tiene algoritmos de aprendizaje supervisado y no supervisado, entre los cuales están máquinas de vector soporte, k-means, modelo de mezcla gaussiana, etc.

## 2. METODOLOGÍA

Este apartado está compuesto por varias Secciones. La Sección 2.1 contiene una explicación de las bases de datos utilizadas. En la Sección 2.2, se explica el método empleado para realizar la agrupación de los eventos sísmicos, los cuales pasan por un preprocesamiento (Sección 2.3). La obtención de las características se la realiza por medio de un extractor de características, el cual se detalla en la Sección 2.4. Se incluye la explicación del clasificador a usar sobre los eventos sísmicos (Sección 2.5). Las configuraciones de las dos etapas (no supervisada y supervisada) se describen en las Secciones 2.6 y 2.7, respectivamente. En la Sección 2.8 se enumeran las métricas de evaluación para cuantificar los resultados obtenidos. Se explica la línea de base para comparar los enfoques no supervisados y características bien conocidas que han sido utilizadas a lo largo de los años por expertos (Sección 2.9). Finalmente, es importante mencionar que en el ANEXO A. Código fuente, se presenta la implementación realizada en código a través de un *live script* de MATLAB.

### 2.1. BASES DE DATOS

Aunque originalmente el trabajo contemplaba solo el uso de la base de datos ESeismic [1] (D2), se consideraron dos bases de datos adicionales D1 y D3. La primera base de datos D1 contiene solamente señales sin procesar no etiquetadas. Las muestras de D1 fueron recolectadas por las estaciones sísmicas BNAS, BREF, BTAM y BVC2 pertenecientes al volcán Cotopaxi, cuya ubicación se muestra en la Figura 2.1. D1 contiene 22640 observaciones con varios tipos de sismos volcánicos como: LP, VT, REG, TR, HYB, EXPL, ICE, TRESP, VLP, TRBA y TREMI. En este punto es importante resaltar que el número de eventos en D1 es considerablemente mayor a los eventos disponibles en ESeismic, lo cual es beneficioso para el aprendizaje automático.

La segunda base de datos está compuesta por *SeisBench* y *MicSig*, que pertenecen al repositorio ESeismic [1], esta base de datos será denominada D2. ESeismic incluye eventos de dos estaciones sísmicas (VC1 y BREF), que se ubican en los flancos del volcán Cotopaxi. La distribución de los 1187 eventos que tiene ESeismic es la siguiente: 1044 LP, 101 VT, 27 REG, 8 HYB y 7 ICE. Las etiquetas de los eventos de ESeismic son verificadas por especialistas del IG EPN. *MicSig* contiene señales en tiempo sin procesar y *SeisBench* tiene 84 características diseñadas por expertos y obtenidas a partir de *MicSig*, dichas características se resumen en el ANEXO B. Resumen 84 características de la línea de base. Entre las 84

características hay 13 en el dominio del tiempo, 21 en el dominio de la frecuencia y 50 en el dominio de escala.

La tercera base de datos D3, se la utiliza de forma adicional, contiene señales en tiempo del volcán Llaima ubicado en la región de la Araucanía, Chile [4]. Las señales se recolectaron por la estación sísmica LAV. La base de datos tiene 3592 observaciones: 1488 TC, 490 TR, 1310 LP y 304 VT. D3 se utiliza para validar el comportamiento del método propuesto en diferentes volcanes. De la misma forma que D2, las etiquetas de D3 son verificadas por especialistas. Para todos los conjuntos de datos antes mencionados, las señales sin procesar provienen de la componente vertical.

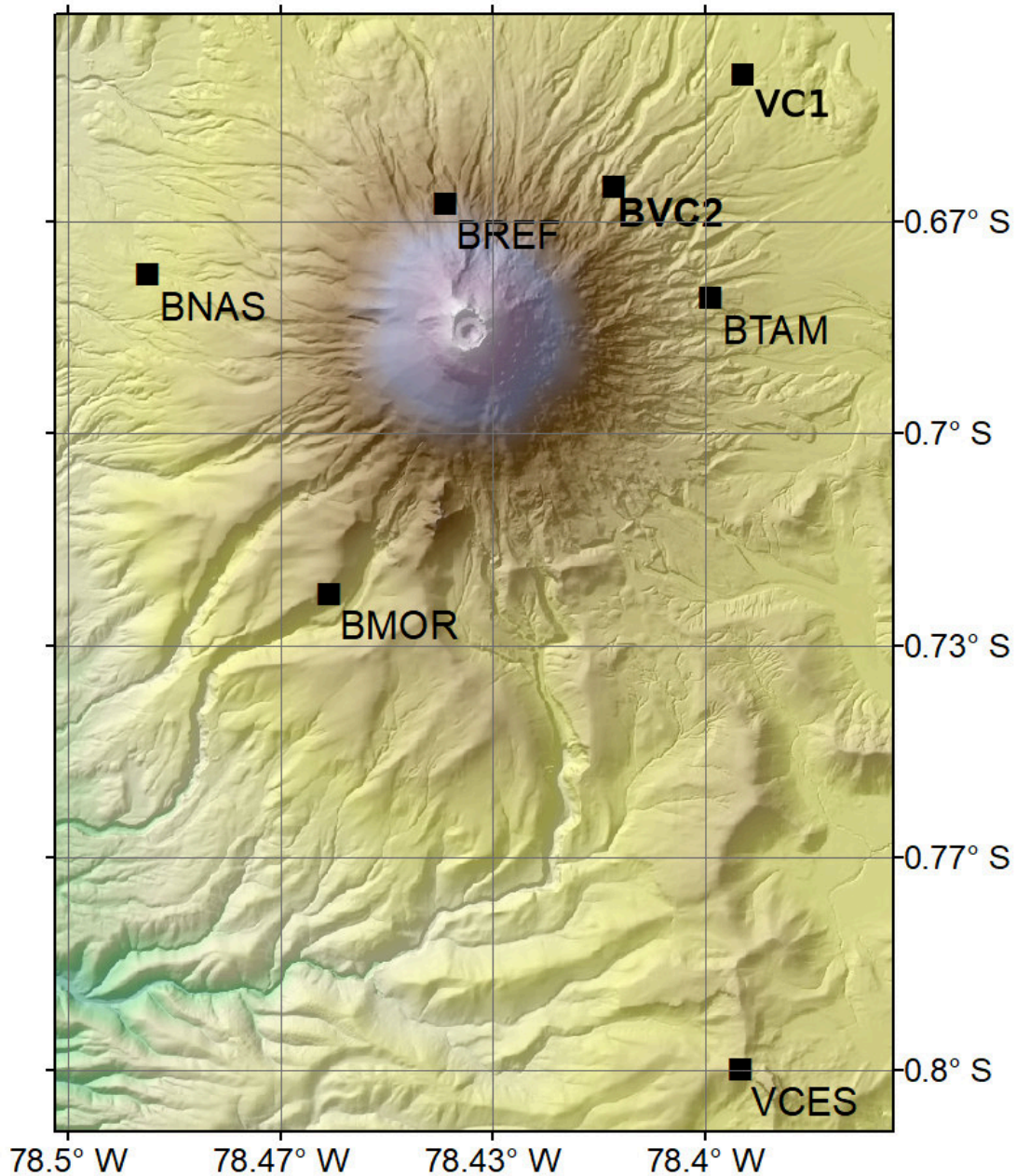
## 2.2. MÉTODO PROPUESTO

Originalmente se planteó utilizar solo una etapa no supervisada sobre la base de datos ESeismic, pero se incluye el uso de una etapa supervisada como se muestra en la Figura 1.1. La etapa no supervisada consiste en utilizar algoritmos no supervisados para aprender un extractor de características de una base de datos sin etiquetar. Después de aplicar un preprocesamiento en los datos, se usa el modelo de mezcla gaussiana o k-means como algoritmo no supervisado. Se eligió comparar estas dos técnicas no supervisadas ya que utilizan estrategias de asignación de *clusters* bien conocidas como: asignación suave (GMM) y asignación dura (k-means).

En la etapa supervisada, se utiliza una base de datos etiquetada para evaluar el rendimiento del extractor de características aprendido durante la etapa no supervisada. Específicamente, después del proceso de extracción de características, se utiliza un clasificador para predecir las etiquetas de los eventos sísmicos. A continuación, se describe a detalle los componentes del método.

El método propuesto se utiliza para las bases de datos experimentales. D1 del volcán Cotopaxi será la base de datos no etiquetada debido a que tiene un número considerablemente mayor de muestras que D2 del volcán Cotopaxi o D3 del volcán Llaima. Además, se decide usar D1 para la etapa no supervisada, ya que sus etiquetas no están verificadas como si lo están en D2 o D3. Por esta razón, D2 y D3 se emplean para la etapa supervisada. Es importante señalar que, dado que D3 proviene de un volcán diferente (Llaima) que D1 y D2, será factible analizar qué tan bien se comportan las características aprendidas durante la etapa no supervisada en diferentes volcanes. La codificación del método y los pasos para ejecutar cada parte de la implementación se presenta a detalle en el ANEXO A. Código fuente.





**Figura 2.1:** Ubicación de las estaciones sísmicas en los flancos del volcán Cotopaxi. Fuente [1]

### 2.3. PREPROCESAMIENTO

En el preprocesamiento, se sigue una serie de pasos primero para las señales de tiempo de la base de datos no etiquetada. Se recorta y se elimina los valores NaN de cada una de las señales. La frecuencia de muestreo se establece en 50 Hz. Se aplica un filtro Butterworth de orden 8 a una frecuencia de corte de 15 Hz para retener el contenido espectral más importante de las señales de tiempo. Luego, se elimina la media y se normaliza la amplitud de la señal por su valor cuadrático medio (RMS) correspondiente. Posteriormente, se utiliza

el método de Burg para estimar la densidad espectral de potencia de la señal de tiempo. Se utiliza 256 puntos en la transformada discreta de Fourier (DFT) y un orden de 40 en el modelo autorregresivo (AR) de Burg.

A partir de esta representación espectral, la componente de DC se elimina de cada señal y la amplitud se expresa en dB, dando un vector de características  $\mathbf{f}$  de longitud  $D = 128$  por señal sísmica. Estos vectores de características están ordenados en una matriz  $\mathcal{D}$  de orden  $N \times D$  compuesta por  $N$  observaciones.

Finalmente, después de aplicar una normalización *z-score* en  $\mathcal{D}$ , se usa el análisis de componentes principales (PCA) con el 99 % de varianza retenida, transformando la dimensionalidad de  $\mathcal{D}$  en una matriz de dimensionalidad reducida  $\mathbf{X}$  de tamaño  $N \times d$  en donde  $d < D$ .

Vale la pena señalar que los mismos pasos de preprocesamiento se aplican a las bases de datos etiquetadas durante la etapa supervisada. Sin embargo, la normalización de *z-score* y el análisis de componentes principales solo se aprenden en la base de datos no etiquetada durante la etapa no supervisada.

## 2.4. APRENDIZAJE DE UN EXTRACTOR DE CARACTERÍSTICAS A PARTIR DE LA BASE DE DATOS NO ETIQUETADA

La matriz  $\mathbf{X}$  se usa como entrada para el extractor de características, teniendo en cuenta que cada vector fila  $\mathbf{x} \in \mathbb{R}^d$  representa una observación. En este sentido, se compara los algoritmos de k-means y GMM como extractores de características.

Para el caso de k-means, e inspirándose en [48], se construye una matriz de banco de filtros  $\mathbf{C}$  de tamaño  $K \times d$  que contiene los centroides  $d$ -dimensionales de cada *cluster*  $k$ . Luego, dada una nueva instancia arbitraria  $\mathbf{x} \in \mathbb{R}^d$ , se la asigna a un vector de características  $\mathbf{y} \in \mathbb{R}^K$  proyectando  $\mathbf{x}$  en  $\mathbf{C}$  utilizando la Ecuación 2.1

$$\mathbf{y} = g(\mathbf{C}^T \mathbf{x}), \quad (2.1)$$

donde se elige a  $g(\cdot)$  para ser la función de activación Leaky ReLU [49], para introducir no linealidad, que viene dada por la Ecuación 2.2

$$g(\mathbf{y}) = \begin{cases} \alpha \mathbf{y}, & \text{si } \mathbf{y} < 0 \\ \mathbf{y}, & \text{otros} \end{cases}, \quad (2.2)$$

donde,  $\alpha$  es una constante pequeña.

La principal desventaja de k-means es que emplea asignaciones duras para inferir etiquetas, mientras que GMM realiza asignaciones suaves. Esto significa que GMM asigna una observación a cierto *cluster* con cierta probabilidad en lugar de la asignación de etiquetas deterministas de k-means.

Para el caso de GMM, se usa el logaritmo natural de los vectores de responsabilidad, que se obtienen con la Ec. 1.7, como vectores de características para alimentar el clasificador.

## 2.5. CLASIFICADOR

Como entrada del clasificador se utiliza una matriz compuesta por los vectores de características generados por el extractor de características y las etiquetas de cada observación obtenidas de las bases de datos etiquetadas. Se debe recalcar que el extractor de características se entrena en una base de datos no etiquetada diferente a las bases de datos etiquetadas. Para k-means se emplea como vector de características las características obtenidas de la Ec. 2.1. Por otro lado, para GMM se emplean los vectores de responsabilidad como vectores de características calculados de acuerdo a la Ec. 1.7.

Se usa el clasificador máquinas de vector soporte. Concretamente, un kernel lineal que está representado por la Ecuación 2.3

$$K(\mathbf{y}_1, \mathbf{y}_2) = \mathbf{y}_1^T \mathbf{y}_2, \quad (2.3)$$

Para el kernel mencionado, se optimiza el parámetro de penalización  $C$ , que controla el sobreajuste (*overfitting*).

## 2.6. CONFIGURACIÓN DE LA ETAPA NO SUPERVISADA

Para la etapa no supervisada, se usan los 22640 eventos sísmicos de D1 provenientes de las cuatro estaciones. Para k-means, se elige el número de *clusters*  $K$  de acuerdo con la heurística del codo evaluada en su función de costo, también se emplea una constante de  $\alpha = 0,3$  en la función de activación Leaky ReLU. Se utiliza una heurística similar para elegir  $K$  en GMM pero la métrica es AIC (*Akaike's Information Criteria*). Además, para GMM se asume una matriz de covarianza diagonal que se comparte entre todos los componentes de la mezcla.

## 2.7. CONFIGURACIÓN DE LA ETAPA SUPERVISADA

Se evalúa el extractor de características aprendido en la base de datos no etiquetada D1, en eventos sísmicos LP y VT de las bases de datos etiquetadas (D2 y D3) usando SVM. El conjunto de datos etiquetado D2 del volcán Cotopaxi tiene eventos LP y VT de las estaciones sísmicas VC1 y BREF, de acuerdo con la siguiente distribución: 480 eventos de la estación VC1 (425 LP y 55 VT), 665 eventos de la estación BREF (619 LP y 46 VT) y 1145 observaciones de ambas estaciones (1044 LP y 101 VT). En el caso de D3 del volcán Llaima, se usa 1614 eventos de la estación LAV (1310 LP y 304 VT). Los eventos sísmicos de LAV son utilizados de forma adicional para determinar qué tan expresivas son las características aprendidas por la etapa no supervisada en diferentes volcanes.

Específicamente, se usa una estrategia de validación cruzada estratificada con cinco grupos para evaluar el clasificador SVM. El uso de cinco grupos se lo realiza considerando que después se repetirán los experimentos 20 veces por cada estación, con ello se busca garantizar el teorema del límite central para tener una buena muestra estadística y lograr un nivel de confianza aceptable (estadísticamente) en los conjuntos de prueba. Además, se optimizan los hiperparámetros de SVM para cada grupo de entrenamiento utilizando una optimización Bayesiana. Como se mencionó, el experimento se repite 20 veces con diferentes números aleatorios.

## 2.8. MÉTRICAS DE EVALUACIÓN

En el presente trabajo se utilizan varias métricas de evaluación. Se genera una matriz de confusión para cada experimento, es decir para cada uno de los números aleatorios en cada una de las estaciones a experimentar. En la Figura 2.2 se muestran los elementos de una matriz de confusión de los cuales se pueden obtener diferentes métricas con su respectiva abreviatura.

Las métricas que se calculan en base a cada una de las matrices de confusión generadas son: la exactitud (*accuracy*, ACC) utilizando la Ecuación 2.4, y el *F1-score* que sigue la Ecuación 2.5

$$ACC = \frac{VP + VN}{VP + FN + VN + FP}, \quad (2.4)$$

Verdaderos positivos (VP)	Falsos positivos (FP)
Falsos negativos (FN)	Verdaderos Negativos (VN)

**Figura 2.2:** Elementos de una matriz de confusión y su abreviación

$$F1 = \frac{VP}{VP * \frac{1}{2} (FP + FN)}, \quad (2.5)$$

Además, se calcula el área bajo curva (*Area Under Curve*, AUC) a partir de la curva de la característica de funcionamiento del receptor (*Receiver Operating Characteristic*, ROC). Es importante mencionar que se calcula una curva ROC en cada experimento realizado por cada una de las estaciones y bases de datos descritas. La curva ROC se construye en base a diferentes umbrales de decisión para saber que tan bien se realiza la clasificación, en este caso entre dos tipos de eventos sísmicos. El valor que resume que tan bien funciona el clasificador es el AUC, cuando la clasificación se llevó a cabo de una forma ideal el valor de AUC es 1 y una clasificación totalmente errada tendría un valor de 0 como AUC.

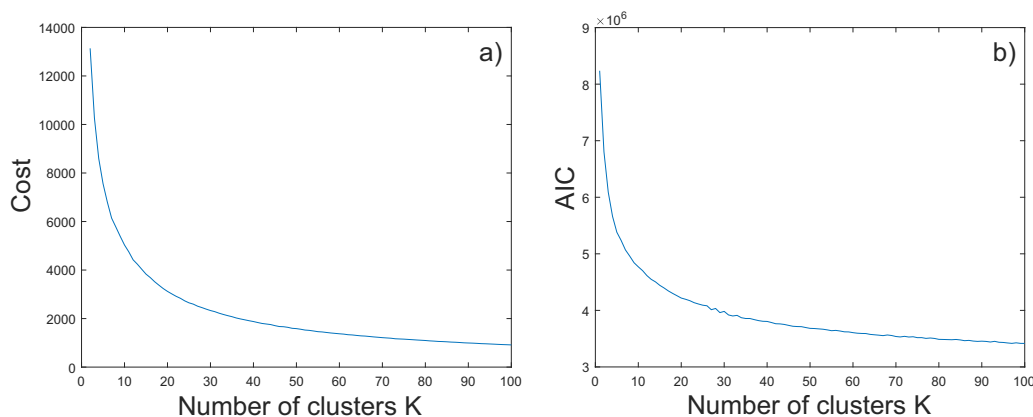
## 2.9. LÍNEA DE BASE

El clasificador SVM bajo la misma configuración experimental también se usa en las características de *SeisBench*. *SeisBench* ofrece 84 características en diferentes dominios (es decir, tiempo, frecuencia y escala) que han sido diseñadas por expertos durante muchos años. Consultar [1] para explorar estas características en detalle. El rendimiento de la clasificación utilizando las 84 características de *SeisBench*, calculadas en D2 y D3, servirá como línea de base para comparar el extractor automático de características aprendido de algoritmos no supervisados. Para probar la significancia estadística entre el método propuesto y la línea de base, se usa *t-test* a un nivel de significancia de 0.05.

### 3. RESULTADOS Y DISCUSIÓN

#### 3.1. RESULTADOS

La Figura 3.1 muestra los resultados después de aplicar la heurística del codo en D1 para los algoritmos k-means y GMM. La heurística sugiere que el número de *clusters*  $K$  es 30 para ambos algoritmos. La explicación de la heurística del codo se la encuentra con mayor detalle en la Sección 1.4.4.



**Figura 3.1:** Heurística del codo en a) k-means y b) GMM

La tabla 3.1 muestra los resultados que se obtuvieron al promediar las métricas de evaluación (ACC, AUC y F1) en las 20 ejecuciones para GMM, k-means y la línea de base. Observar que para D2, se evalúa las métricas en BREF y VC1 por separado, y también en ambas estaciones. La métrica que ofrece una mejora estadística a un nivel de significancia de 0.05 se resalta con un \*.

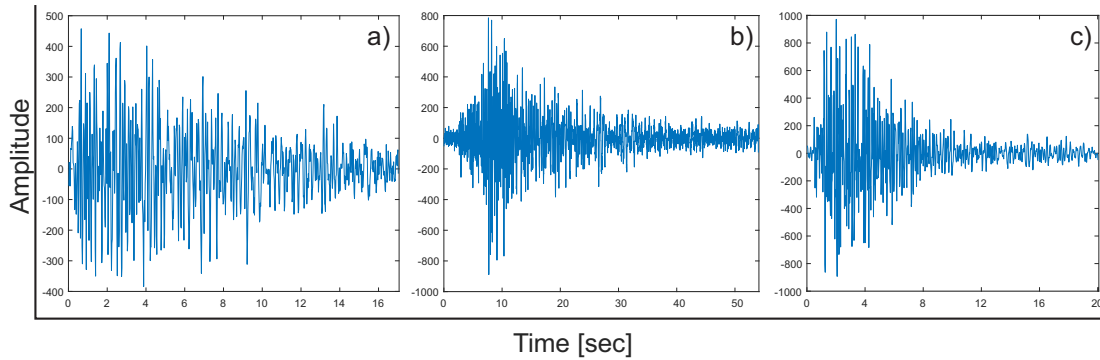
Los resultados muestran que no existe una diferencia estadística entre los algoritmos no supervisados y la línea de base en todas las estaciones y bases de datos empleadas. La única excepción es para la estación BREF de D2 donde la métrica AUC para ambos métodos no supervisados mejoró estadísticamente con respecto a la línea de base. Vale la pena señalar que se encontró que BREF es la estación más desafiante analizada en este documento, ya

**Tabla 3.1:** Resultados de las métricas de evaluación para D2 y D3.

\* representa una mejora estadística con respecto a la línea de base a un nivel de significancia de 0,05

	D2 (BREF)			D2 (VC1)			D2 (BREF y VC1)			D3 (LAV)		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
<b>GMM</b>	0.953	0.962*	0.975	0.970	0.983	0.983	0.957	0.972	0.976	0.972	0.976	0.983
<b>k-means</b>	0.952	0.952*	0.974	0.977	0.992	0.986	0.962	0.973	0.979	0.971	0.982	0.982
<b>Línea de base</b>	0.958	0.933	0.978	0.977	0.986	0.987	0.965	0.976	0.981	0.977	0.985	0.986

que la mayoría de las señales sin procesar son mucho más ruidosas que otras estaciones (consultar la Figura 3.2 para ver algunos ejemplos). Esto también se confirma con la incorporación de t-SNE (ver Figura 3.3), donde está claro que las observaciones en BREF son más difíciles de separar unas de otras. Este análisis sugiere que las características aprendidas de la etapa no supervisada son más sólidas en condiciones ruidosas.

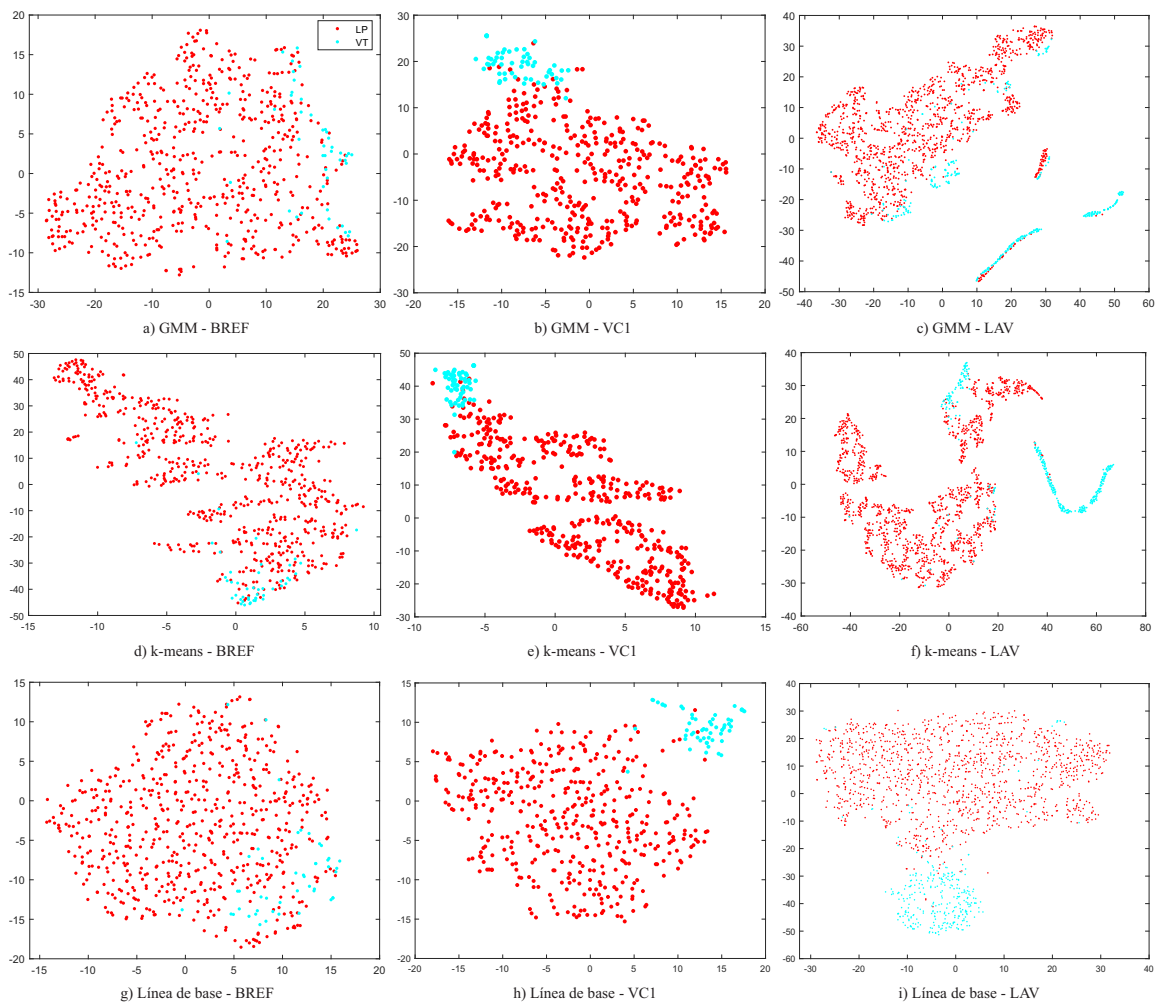


**Figura 3.2:** Ejemplos de señales VT sin procesar de las estaciones sísmicas a) BREF (base de datos D2), b) VC1 (base de datos D2) y c) LAV (base de datos D3)

Aunque para todas las estaciones, excepto BREF, el rendimiento de los algoritmos no supervisados es estadísticamente similar a la línea de base. Las técnicas no supervisadas aprendieron características relevantes automáticamente a partir de datos no etiquetados. Este es un resultado alentador que muestra que la extracción automática de características es tan eficiente como las características bien conocidas diseñadas por expertos a lo largo de los años en trabajos anteriores [1].

Con respecto a la estación LAV que proviene de otro volcán, se obtienen resultados similares al comparar la línea de base con los extractores de características implementados con aprendizaje no supervisado. Esto sugiere que la extracción automática de características realizada a partir de datos no etiquetados es una representación útil para agrupar y clasificar eventos sísmicos incluso en diferentes volcanes.

Adicionalmente para realizar una mejor comparación entre los dos algoritmos no supervisados, la Figura 3.4 muestra las matrices del valor absoluto de los coeficientes de correlación para a) GMM y b) k-means, donde se usa la convención de nomenclatura de [1] para identificar las 84 características de la línea base (es decir,  $f_1, f_2, \dots, f_{84}$ ). La correlación de las características de GMM con respecto a las características del dominio del tiempo ( $f_1$  a  $f_{13}$ ) es principalmente baja, mientras que aumenta con respecto a las características de frecuencia ( $f_{14}$  a  $f_{34}$ ). Se espera este comportamiento ya que el enfoque no supervisado aprende de la magnitud de la respuesta de frecuencia de las señales sísmicas.

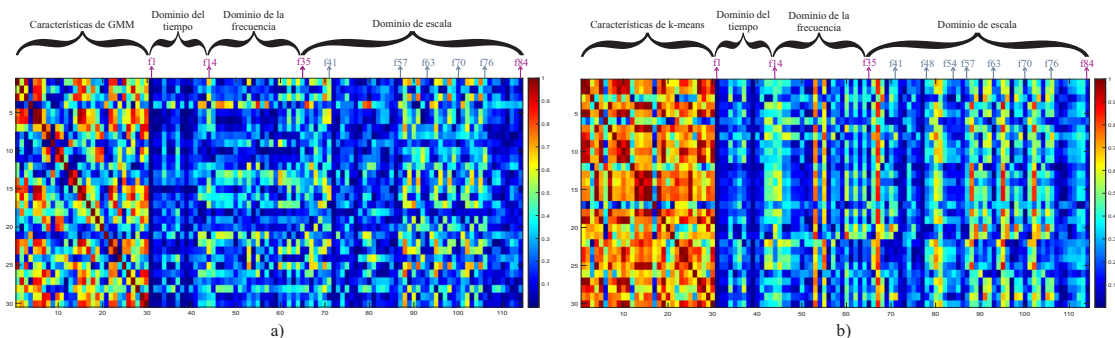


**Figura 3.3:** t-SNE de diferentes estaciones sísmicas (BREF, VC1 y LAV) para cada grupo de características (GMM, k-means y línea de base)

En las características del dominio de escala (f35 a f84) GMM tiende a correlacionarse fuertemente con las siguientes características calculadas sobre cada componente de onda: i) pico máximo en el dominio de frecuencia (f35 a f41); ii) porcentaje de energía (f57 a f62); iii) RMS en el dominio del tiempo (f63 a f69); y iv) pico a pico en el dominio del tiempo (f70 a f76).

Se observa un comportamiento similar en k-means como se muestra en la Figura 3.4 b. Además de la fuerte correlación con las mismas características de escala que GMM, k-means también tiende a correlacionarse con la media en el dominio de la frecuencia de los componentes de onda (f48 a f54). De la Figura 3.4 también es oportuno mencionar, que GMM aprende una representación menos redundante que k-means, es decir, las características de GMM están menos correlacionadas. Esto podría explicar el hecho de que GMM logra un mejor rendimiento con señales ruidosas de la estación BREF.





**Figura 3.4:** Matrices del valor absoluto de los coeficientes de correlación para a) GMM b) k-means. Dado el interés de comparar los enfoques no supervisados con las características del dominio de tiempo, frecuencia y escala con la línea de base, para una mejor visualización, se han omitido algunos coeficientes de correlación de estas matrices. Observar que se usa la misma convención de nomenclatura de [1] para identificar las 84 características de línea de base (es decir,  $f_1, f_2, \dots, f_{84}$ )

### 3.2. DISCUSIÓN

En general, se observa que los algoritmos no supervisados aprenden varias características fuertemente correlacionadas con los componentes de onda. Esta idea está respaldada por las tablas 3.2 para GMM y 3.3 para k-means, donde se presentan las 5 principales características correlacionadas entre los enfoques no supervisados y las características de la línea de base.

Específicamente, de acuerdo con las tablas 3.2 y 3.3, se puede obtener algunas ideas sobre lo que están aprendiendo los algoritmos no supervisados. Para este análisis, se debe recordar que los componentes de onda D1, D2 y D5 corresponden a bandas de frecuencia de 12.5-25 Hz, 6.25-12.5 Hz, y 0.75-1.5 Hz, respectivamente. Las características  $f_{23}, f_{25}$  y  $f_{36}$  están relacionadas con bandas de alta frecuencia (aproximadamente por encima de 10 Hz), que normalmente se encuentran en eventos VT; las características restantes están relacionadas con baja frecuencia (aproximadamente por debajo de 10 Hz), que normalmente se encuentran en eventos LP. Esto es comprensible porque la base de datos D1 tiene varios tipos de eventos sísmicos que actúan en diferentes bandas de frecuencia.

Según [1, 50–52], el aprendizaje supervisado se ha utilizado repetidamente para clasificar eventos sísmicos y, en menor medida, el aprendizaje semi-supervisado [53]. La principal desventaja de los enfoques supervisados es la necesidad de una base de datos etiquetada, que idealmente debería ser etiquetada y verificada por expertos. Además, con frecuencia es difícil generalizar modelos de aprendizaje supervisado para aplicarlos en diferentes volcanes

**Tabla 3.2:** Las cinco características principales correlacionadas entre GMM y las características de la línea de base de acuerdo con el valor absoluto del coeficiente de correlación. Los componentes de onda D1, D2 y D5 corresponden a las bandas de frecuencia de 12.5-25 Hz, 6.25-12.5 Hz, y 0.75-1.5 Hz, respectivamente

Características de la línea de base [1]	Coefficientes de correlación
f25 - Pico máximo en la banda de 20-25 Hz	0.8491
f50 - D2 Promedio en el dominio de la frecuencia	0.8203
f36 - D1 Pico máximo en el dominio de la frecuencia	0.8168
f58 - Porcentaje de energía para D2	0.7931
f61 - Porcentaje de energía para D5	0.7871

**Tabla 3.3:** Las cinco características principales correlacionadas entre k-means y las características de la línea de base de acuerdo con el valor absoluto del coeficiente de correlación. La banda de frecuencia de la componente de onda D2 es 6.25-12.5 Hz

Características de la línea de base [1]	Coefficientes de correlación
f65 - D2 RMS en el dominio del tiempo	0.8947
f37 - D2 Pico máximo en el dominio de la frecuencia	0.8670
f72 - D2 Pico-pico en el dominio del tiempo	0.8434
f23 - Pico máximo en la banda de 10-20 Hz	0.8386
f58 - Porcentaje de energía para D2	0.8337

[54].

Con respecto a los métodos de aprendizaje no supervisados, vale la pena mencionar algunos estudios que utilizan diferentes enfoques. En [55] se comparan varios algoritmos no supervisados (por ejemplo, k-means, agrupamiento espectral, entre otros) para clasificar eventos sísmicos LP y VT del volcán Cotopaxi, donde se logró el mejor desempeño de clasificación con el algoritmo BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*).

En [56] se utilizaron mapas auto organizados (*self-organized maps*) y k-means como algoritmos para clasificar los eventos sísmicos del volcán Tungurahua (Ecuador). Además, en [57] se utilizó k-means como un algoritmo no supervisado para encontrar familias de eventos sísmicos y relacionar dichas familias con el estado de la actividad interna del volcán Cotopaxi.

En [43] se usa GMM para clasificar los eventos sísmicos VT y LP del volcán Cotopaxi. Esto demuestra el potencial del aprendizaje no supervisado para clasificar eventos sísmicos y comprender de mejor manera la dinámica de un volcán.

En este contexto, los aportes de este trabajo se resumen a continuación:

i) Se propone un enfoque para aprender la representación de características a partir de datos no etiquetados para eventos sísmicos. A diferencia de los trabajos [55–57], se hace una

interpretación de las características aprendidas por los algoritmos no supervisados al correlacionar la representación no supervisada con características bien conocidas diseñadas por expertos a lo largo de los años [1].

ii) La diferencia clave con trabajos anteriores [43, 55–57] es que no se utiliza el aprendizaje no supervisado para clasificar directamente los eventos sísmicos al inferir sus etiquetas. En su lugar, se usa una etapa no supervisada para aprender un extractor de características y luego estas características se convierten en las entradas a una etapa supervisada que clasifica los eventos sísmicos correspondientes. Además, en la etapa no supervisada se comparan dos algoritmos no supervisados basados en asignaciones duras y suaves (es decir, k-means y GMM, respectivamente).

iii) A diferencia de trabajos previos [43, 55–57], la evaluación del rendimiento de la clasificación con características no supervisadas se incluye para diferentes conjuntos de datos, estaciones e incluso volcanes, para mostrar la solidez del método propuesto en diferentes escenarios.

## **4. CONCLUSIONES Y RECOMENDACIONES**

### **4.1. CONCLUSIONES**

Se implementó un método de agrupación de señales sísmicas generadas por el volcán Cotopaxi basado en aprendizaje automático no supervisado utilizando el modelo de mezcla gaussiana. En este contexto, se descubre que las técnicas no supervisadas como el modelo de mezcla gaussiana y k-means son capaces de aprender características relevantes automáticamente a partir de datos no etiquetados; la extracción automática de características es tan eficiente como las características diseñadas por expertos a lo largo de los años. Además, GMM y k-means aprenden una amplia gama de características relacionadas con eventos de alta frecuencia (por ejemplo, VT) y de baja frecuencia (como, LP).

La codificación del método de agrupación de señales sísmicas del Volcán Cotopaxi fue evaluada en un algoritmo de aprendizaje supervisado, demostrando que, para la mayoría de las estaciones y volcanes analizados, no existe una diferencia estadística entre los enfoques no supervisados y las características de la línea base. Vale la pena señalar que este es un logro importante considerando que el enfoque propuesto aprende características automáticamente a partir de datos no etiquetados mientras mantiene un desempeño de clasificación estadísticamente similar con respecto a la línea de base. Aún más, las características del modelo de mezcla gaussiana superaron las características de la línea de base en señales ruidosas provenientes de la estación BREF en el volcán Cotopaxi. Además, se demostró que los algoritmos no supervisados aprenden características útiles incluso para agrupar eventos de diferentes volcanes.

Los resultados demuestran que no existe una diferencia estadística entre los algoritmos no supervisados (k-means y el modelo de mezcla gaussiana) en todas las estaciones y bases de datos utilizadas.

### **4.2. RECOMENDACIONES**

En trabajos futuros, se podría utilizar la fase de las señales durante el entrenamiento del algoritmo no supervisado para tener en cuenta aspectos importantes de los eventos sísmicos, como su forma de onda. Además, un trabajo futuro podría utilizar algoritmos de aprendizaje no supervisados más avanzados, como autoencoders [58].

Al tratarse del estudio de eventos sísmicos que podría desembocar en el descubrimiento de nuevas dinámicas del volcán, sería ideal implementar un sistema en tiempo real que agrupe y clasifique los tipos de eventos sísmicos inmediatamente luego de que fueron generados por el volcán. Esto ayudaría a mitigar los riesgos que una eventual erupción volcánica podría ocasionar, ya que se tomarían medidas preventivas antes de que ocurra la emisión de flujos piroclásticos o la formación de lahares.

El uso de algoritmos no supervisados para obtener características a partir de datos no etiquetados, podría ser extendido en diferentes áreas de estudio en las que se tenga bases de datos a disposición. Un ejemplo es el área de la medicina en donde sería factible usar el método propuesto para predecir posibles enfermedades catastróficas como el cáncer. En [59] ya se hace el uso del modelo de mezcla gaussiana para segmentar estructuras dentro del pulmón y luego se usa máquinas de vector soporte para diferenciar entre nódulos y no nódulos.

## 5. REFERENCIAS BIBLIOGRÁFICAS

- [1] N. Pérez, D. Benítez, F. Grijalva, R. Lara-Cueva, M. Ruiz, and J. Aguilar, “Eseismic: Towards an ecuadorian volcano seismic repository,” *Journal of Volcanology and Geothermal Research*, 2020.
- [2] S. Manosalvas, “Red de observatorios vulcanológicos,” <https://www.igepn.edu.ec/red-de-observatorios-vulcanologicos-rovig>, Apr. 2021. [Online]. Disponible en: <https://www.igepn.edu.ec/>
- [3] M. Malfante, M. Dalla Mura, J.-P. Métaxian, J. I. Mars, O. Macedo, and A. Inza, “Machine learning for volcano-seismic signals: Challenges and perspectives,” *IEEE Signal Processing Magazine*, vol. 35, no. 2, pp. 20–30, 2018.
- [4] P. J. Canário, R. Fernandes de Mello, M. Curilem, F. Huenupan, and R. Araujo, “Llaima volcano dataset: In-depth comparison of deep artificial neural network architectures on seismic events classification,” *Data in brief*, vol. 30, p. 105627–105627, 2020.
- [5] S. McNutt and D. Roman, *The Encyclopedia of Volcanoes (Second Edition)*. Academic Press, 2015.
- [6] D. Andrade, M. Hall, P. Mothes, L. Troncoso, J.-P. Eissen, P. Samaniego, J. Egred, P. Ramón, D. Rivero, and H. Yepes, *Los peligros volcánicos asociados con el Cotopaxi*. Corporación Editora Nacional, 2005.
- [7] V. M. Zobin, *Introduction to Volcanic Seismology*. Oxford: Elsevier, 2016.
- [8] T. Minakami, *Fundamental research for predicting volcanic eruptions*. Earthquake Research Institute, the University of Tokyo, 1960.
- [9] —, *The 1962 eruption of Miyake-sima, one of the seven Izu islands, Japan*. Earthquake Research Institute, the University of Tokyo, 1962.
- [10] —, *Seismology of volcanoes in Japan, Physical Volcanology*. Amsterdam: Elsevier, 1974.
- [11] J. Lahr, B. Chouet, C. Stephens, J. Power, and R. Page, “Clasificación de terremotos, ubicación y análisis de errores en un entorno volcánico: implicaciones para el sistema magmático de las erupciones de 1989-1990 en el volcán redoubt,” *Journal of Volcanology and Geothermal Research*, vol. 62, no. 1–4, pp. 137–151, 1994, 1994.

- [12] P. Venegas, N. Pérez, D. Benítez, R. Lara-Cueva, and M. Ruíz, “Building machine learning models for long-period and volcano-tectonic event classification,” *IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*, 2019.
- [13] I. Molina, H. Kumagai, A. García-Aristizábal, M. Nakano, and P. Mothes, “Source process of very-long-period events accompanying long-period signals at cotopaxi volcano, ecuador,” *Journal of Volcanology and Geothermal Research*, vol. 167, no. 1, pp. 119–133, 2008.
- [14] E. Aguilera and T. Toulkeridis, *El Volcán Cotopaxi, una amenaza que acecha*. Sotavento Ed., 2005.
- [15] “Red de monitoreo del volcán cotopaxi igept,” <https://www.igept.edu.ec/images/portal/observatorios-volcanicos/volcanes/cotopaxi/Monitoreo-Instrumental-Volcan-Cotopaxi.jpg>, Jul. 2021. [Online]. Disponible en: <https://www.igept.edu.ec/>
- [16] T. Jo, *Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning*. Cham: Springer International Publishing AG, 2021.
- [17] E. Alpaydm, *Introduction to machine learning*, 3rd ed. Cambridge, Massachusetts: MIT Press, 2014.
- [18] —, *Machine learning: the new AI*. Cambridge, Massachusetts: MIT Press, 2016.
- [19] A. Y.-T. Ng, “Machine learning,” <https://www.coursera.org/learn/machine-learning>, Jul. 2021. [Online]. Disponible en: <https://www.coursera.org/>
- [20] M. W. Berry, A. Mohamed, and B. W. Yap, *Supervised and Unsupervised Learning for Data Science*, 1st ed. Cham: Springer International Publishing, 2020.
- [21] J. Watt, R. Borhani, and A. K. Katsaggelos, *Machine Learning Refined: Foundations, Algorithms, and Applications*. Cambridge University Press, 2006.
- [22] A. Likas, N. Vlassis, and J. J. Verbeek, “The global k-means clustering algorithm,” *Pattern recognition*, vol. 36, no. 2, p. 451–461, 2003.
- [23] Y.-M. Cheung, “K-means: A new generalized k-means clustering algorithm,” *Pattern recognition letters*, vol. 24, no. 15, p. 2883–2893, 2003.

- [24] K. Wagstaf, C. Cardie, S. Rogers, and S. Schroedl, "Constrained k-means clustering with background knowledge," in *International Conference on Machine Learning*, 2001, vol. 1, pp. 577–584.
- [25] W. Johnson, "Learn by marketing," <https://www.learnbymarketing.com/methods/k-means-clustering/>, Jul. 2021. [Online]. Disponible en: <https://www.learnbymarketing.com/>
- [26] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *EURASIP journal on wireless communications and networking*, vol. 2021, no. 1, pp. 1–16, 2021.
- [27] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP conference series. Materials Science and Engineering*, 2018, vol. 336, no. 1, pp. 12 017–.
- [28] E. Fox and C. Guestrin, "Machine learning: Clustering & retrieval," <https://www.coursera.org/learn/ml-clustering-and-retrieval>, Jul. 2021. [Online]. Disponible en: <https://www.coursera.org/>
- [29] N. Bouguila and W. Fan, *Mixture Models and Applications*, 1st ed. Cham: Springer International Publishing, 2020.
- [30] O. Contreras, "Gaussian mixture models explained," <https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95>, Jun. 2019. [Online]. Disponible en: <https://towardsdatascience.com/>
- [31] C. B. Do and S. Batzoglou, "What is the expectation maximization algorithm?" *Nature biotechnology*, vol. 26, no. 8, pp. 897–899, 2008.
- [32] B. C. Levy, *Principles of Signal Detection and Parameter Estimation*. New York, NY : Springer US, 2008.
- [33] R. Di Lorenzo, "Butterworth filter," in *Trading Systems*. Milano: Springer Milan, 2012, ch. 35, pp. 205–208.
- [34] S. C. Dutta Roy, *Topics in Signal Processing Analog and Digital*, 1st ed. Singapore: Springer Singapore, 2020.



- [35] A. Quinquis, *Digital signal processing using MATLAB*. London : ISTE, 2008.
- [36] W.-L. Chen, C.-D. Kan, C.-H. Lin, and T. Chen, "Evaluation of arteriovenous shunt stenosis in hemodialysis patients by using the burg method for autoregressive modeling," in *2012 International Symposium on Information Technologies in Medicine and Education*, 2012, vol. 2, pp. 560–564.
- [37] M. Akay, J. L. Semmlow, W. Welkowitz, M. D. Bauer, and J. B. Kostis, "Detection of coronary occlusions using autoregressive modeling of diastolic heart sounds," *IEEE Transactions on Biomedical Engineering*, vol. 37, no. 4, pp. 366–373, 1990.
- [38] P. S. R. Diniz, E. A. B. da Silva, and S. L. Netto, *Digital Signal Processing: System Analysis and Design*. Cambridge: Cambridge University Press, 2010.
- [39] "Mathworks zscore," [https://la.mathworks.com/help/stats/zscore.html?s\\_tid=doc\\_ta](https://la.mathworks.com/help/stats/zscore.html?s_tid=doc_ta), 2021. [Online]. Disponible en: [https://la.mathworks.com/help/stats/zscore.html?s\\_tid=doc\\_ta](https://la.mathworks.com/help/stats/zscore.html?s_tid=doc_ta)
- [40] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews. Computational statistics.*, vol. 2, no. 4, pp. 433–459, 2010.
- [41] G. R. Naik, *Advances in Principal Component Analysis Research and Development*, 1st ed. Singapore: Springer Singapore, 2018.
- [42] E. Oja, A. Hyvärinen, and J. Karhunen, "Principal component analysis and whitening," in *Independent Component Analysis*. New York: USA: John Wiley & Sons, Inc, 2001, ch. 6, pp. 125–144.
- [43] P. Venegas, N. Pérez, D. Benítez, R. Lara-Cueva, and M. Ruíz, "Combining filter-based feature selection methods and gaussian mixture model for the classification of seismic events from cotopaxi volcano," *IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING*, vol. 12, no. 6, 2019.
- [44] C. Brusil, F. Grijalva, R. Lara-Cueva, M. Ruiz, and B. Acuna, "A semi-supervised approach for microseisms classification from cotopaxi volcano," in *2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, 2019, pp. 1–6.
- [45] A. Quinquis, *Digital signal processing using MATLAB*, 2nd ed. London: ISTE, 2008.
- [46] K. Kim, *Conceptual Digital Signal Processing with MATLAB*, 1st ed. Singapore: Springer Singapore, 2021.

- [47] J. Leis, *Digital signal processing using MATLAB for students and researchers*. Hoboken, New Jersey : Wiley, 2011.
- [48] A. Coates and A. Y. Ng, "Learning feature representations with k-means," in *Neural Networks: Tricks of the Trade*. Berlin: Heidelberg: Springer Berlin Heidelberg, 2012, pp. 561–580.
- [49] M. Nielsen, *Neural networks and deep learning*. Determination Press, 2015.
- [50] R. Lara-Cueva, E. V. Carrera, J. F. Morejon, and D. Benitez, "Comparative analysis of automated classifiers applied to volcano event identification," in *2016 IEEE Colombian Conference on Communications and Computing (COLCOM)*, 2016, pp. 1–6.
- [51] M. Curilem, J. Vergara, C. San Martin, G. Fuentealba, C. Cardona, F. Huenupan, M. Chacon, M. Salman Khan, W. Hussein, and N. B. Yoma, "Pattern recognition applied to seismic signals of the llaima volcano (chile); an analysis of the events' features," *Journal of volcanology and geothermal research*, vol. 282, pp. 134—147, 2014.
- [52] D. S. Parihar, R. Ghosh, A. Akula, S. Kumar, and H. K. Sardana, "Machine learning based comparative analysis for the classification of earthquake signals," in *Proceedings of the International Conference on Computing and Communication Systems*. Singapore: Springer Singapore, 2018, pp. 369–376.
- [53] C. Brusil, F. Grijalva, R. Lara-Cueva, M. Ruiz, and B. Acuña, "A semi-supervised approach for microseisms classification from cotopaxi volcano," in *2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, 2019, pp. 1–6.
- [54] G. Cortes, R. Carniel, M. A. Mendoza, and P. Lesage, "Standardization of noisy volcanoseismic waveforms as a key step toward station-independent, robust automatic recognition," *Seismological research letters*, vol. 90, no. 2A, pp. 581—590, 2019.
- [55] A. Duque, K. Gonzalez, N. Perez, D. Benitez, F. Grijalva, R. Lara-Cueva, and M. Ruiz, "Exploring the unsupervised classification of seismic events of cotopaxi volcano," *Journal of volcanology and geothermal research*, vol. 403, pp. 107 009—, 2020.
- [56] J. A. Reyes and C. J. Jimenez Mosquera, "Non-supervised classification of volcanic-seismic events for tungurahua-volcano ecuador," in *2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM)*, 2017, pp. 1–6.

- [57] J. Anzieta, H. Ortiz, G. Arias, and M. Ruiz, "Finding possible precursors for the 2015 co-topaxi volcano eruption using unsupervised machine learning techniques," *International Journal of Geophysics*, 2019.
- [58] X. Peng, J. Feng, S. Xiao, W.-Y. Yau, J. T. Zhou, and S. Yang, "Structured autoencoders for subspace clustering," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5076–5086, 2018.
- [59] A. M. Santos, A. O. de Carvalho Filho, A. C. Silva, A. C. de Paiva, R. A. Nunes, and M. Gattass, "Automatic detection of small lung nodules in 3d ct data using gaussian mixture models, tsallis entropy and svm," *Engineering applications of artificial intelligence*, vol. 36, pp. 27–39, 2014.

## **ANEXOS**

Los siguientes anexos se encuentran en formato digital:

ANEXO A. Código fuente

ANEXO B. Resumen 84 características de la línea de base

## **ORDEN DE EMPASTADO**