Creación de algoritmos inteligentes basados en la teoría de Machine Learning tradicional para la clasificación de los eventos sísmicos en el volcán Llaima (multiclase)

Cachipuendo Yacelga, César Ariel y Tutillo Moyón, Javier Alejandro

Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Telecomunicaciones

Trabajo de integración curricular, previo a la obtención del título de Ingeniero en Telecomunicaciones

Ing. Lara Cueva, Román Alcides PhD.

18 de septiembre del 2023

Copyleaks

Plagiarism report

O3_Cachipuendo-Tutillo_Escrito_Reco...

Scan details

Scan time: August 25th, 2023 at 18:56 UTC

Total Pages: 72

Total Words: 17974

Plagiarism Detection

9.7%

Types of plagiarism		Words
Identical	1.9%	349
Minor Changes	0.4%	63
Paraphrased	7.4%	1332
Omitted Words	0%	0

AI Content Detection

N/A

Text coverage
- AI text
- Human text

Plagiarism Results: (34)

Your File — 4.2%

Javier Tutillo

No introduction available.

Your File — 1.8%

EDWIN ALEXANDER CASTILLO TIPANTU♦A

No introduction available.

T-ESPE-044263.pdf — 1.1%

https://repositorio.espe.edu.ec/jspui/bitstream/21000/23743...

core i5

1 Sistema de reconocimiento de microterremotos en tiempo real del volcán Cotopaxi aplicando aprendizaje supervisado. Altamirano Rodrígu...

Firmado electrónicamente por: ROMAN ALCIDES LARA CUEVA

Ing. Lara Cueva, Román Alcides PhD.

Director

Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Telecomunicaciones

Certificación

Certifico que el trabajo de integración curricular: "**Creación de algoritmos inteligentes basados en la teoría de Machine Learning tradicional para la clasificación de los eventos sísmicos en el volcán Llaima (multiclase)**" fue realizado por los señores **Cachipuendo Yacelga, César Ariel y Tutillo Moyón, Javier Alejandro,** el mismo que cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, además fue revisado y analizada en su totalidad por la herramienta de prevención y/o verificación de similitud de contenidos; razón por la cual me permito acreditar y autorizar para que se lo sustente públicamente.

Sangolquí, 18 de septiembre del 2023

Firmado electrónicamente por:
ROMAN ALCIDES LARA CUEVA

.....

Ing. Lara Cueva, Román Alcides PhD.

C. C. 1713988218

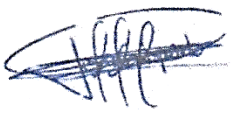
Departamento de Eléctrica, Electrónica y Telecomunicaciones

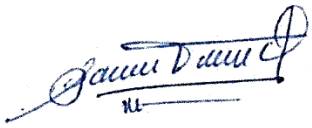
Carrera de Telecomunicaciones

Responsabilidad de Autoría

Nosotros, **Cachipuendo Yacelga, César Ariel y Tutillo Moyón, Javier Alejandro**, con cédulas de ciudadanía 1726237132 y 1726298167, declaramos que el contenido, ideas y criterios del trabajo de integración curricular: **Creación de algoritmos inteligentes basados en la teoría de Machine Learning tradicional para la clasificación de los eventos sísmicos en el volcán Llaima (multiclase)** es de nuestra autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos, y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 18 de septiembre del 2023





.....	.....
Cachipuendo Yacelga, César Ariel	**Tutillo Moyón, Javier Alejandro**
C.C.: 1726237132	C.C.: 1726298167

Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Telecomunicaciones

Autorización de Publicación

Nosotros **Cachipuendo Yacelga, César Ariel y Tutillo Moyón, Javier Alejandro**, con cédulas de ciudadanía 1726237132 y 1726298167, autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de integración curricular: **Creación de algoritmos inteligentes basados en la teoría de Machine Learning tradicional para la clasificación de los eventos sísmicos en el volcán Llaima (multiclase)** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de nuestra responsabilidad.

Sangolquí, 18 de septiembre de 2023

.....	.....
Cachipuendo Yacelga, César Ariel	**Tutillo Moyón, Javier Alejandro**
C.C.: 1726237132	C.C.: 1726298167

Dedicatoria

Este proyecto de titulación está dedicado al incansable esfuerzo diario de mi familia, quienes me han brindado su apoyo constante para perseguir y alcanzar mis sueños. A mi querida mamá Blanca, cuya forma de ser y cariño incondicional están presentes en cada paso importante que doy en la vida. A mi admirable papá Nelson, por brindarme su apoyo y ser una fuente inagotable de motivación en mi vida, lo cual ha sido fundamental para que logre avanzar en la realización de mis metas. A mi estimada hermana Viviana, cuya presencia infunde alegría y respalda cada uno de mis pasos con confianza y apoyo. Todos esenciales en mi vida, sin su presencia, no habría logrado llegar hasta este punto.

Javier Alejandro Tutillo Moyón

Este proyecto de titulación rinde homenaje a la incansable dedicación de mi familia, quienes con su amor y estímulo han sido la fuerza impulsora en esta travesía académica y personal. Su constante apoyo ha sido mi viento en popa hacia el logro de mis metas. Quiero expresar mi profundo agradecimiento a mi madre, Ligia, y a mi padre, Luis, cuya orientación y estímulo han sido pilares esenciales en cada paso de mi trayectoria. Reconozco con gratitud y amor la importancia de mis hermanas, Arianna y Nathalia, cuya presencia ha robustecido mi camino y ha enriquecido mi crecimiento. Agradezco también a mi leal compañero, Lobo, quien me ha acompañado en innumerables jornadas de estudio, brindándome compañía y alegría.

César Ariel Cachipuendo Yacelga

Agradecimientos

Doy gracias a Dios por haberme concedido una familia trabajadora y llena de alegría, quienes me han brindado la oportunidad de perseguir mis sueños y convertirme en alguien valioso y beneficioso para la sociedad. Quiero expresar mi gratitud al Dr. Román Lara por su notable paciencia y valiosa contribución al compartir tanto su conocimiento como su experiencia. Su apoyo ha sido fundamental para llevar a cabo la culminación exitosa de este proyecto. Quiero reconocer a Cesar, Alex, Xavi y Oscar por ser verdaderos amigos quienes me han invitado a vivir el día a día como si fuera el último. Su amistad y compañía ha hecho que esta etapa universitaria sea inolvidable. Y no menos importante quiero agradecerme por creer en mí, por dar lo mejor y nunca rendirme.

Javier Alejandro Tutillo Moyón

Quiero expresar mi gratitud a las personas que me acompañaron durante mi etapa universitaria, especialmente a mis compañeros de aula quienes se convirtieron en mis amigos, con los cuales hemos superado desafíos académicos. A toda mi familia quienes me brindaron su apoyo y cuidados incondicionales sin importar la situación, en especial a mis hermanas y mascota quienes son mi motivo de inspiración a seguir adelante. También agradezco a mis mejores amigos por los concejos y habladas. Por último, a nuestro tutor por compartir sus conocimientos y ser participe activo en el desarrollo de este trabajo.

César Ariel Cachipuendo Yacelga

Índice de contenidos

Agradecimientos **7**

Resumen **16**

Abstract **17**

Capítulo I **18**

Introducción 18

Antecedentes 20

Justificación e importancia 24

Objetivo 25

Actividades 25

Capítulo II **26**

Métodos y materiales 26

Conjunto de datos del clasificador 27

Volcán Llaima 27

Microsismos 28

Tectónico 29

Base de datos 32

Preprocesamiento 33

Filtrado 34

Recorte 34

Procesamiento 35

PSD por el Método Welch 35

Extracción de 84 características 37

Normalización 39

RFE por el método de envoltura 40

Bases de datos a entrenar 42

Balanceo de la base de datos 43

Entrenamiento de algoritmos de ML 44

Árbol de decisión 45

k Vecino más Cercano 47

Máquina de Vectores de Soporte 49

Métricas de desempeño 51

Exactitud 52

Precisión 52

Especificidad 53

Recall o Sensibilidad 53

BER 53

Generación de modelos 53

Prueba de modelos 54

Votación 54

Selección de los mejores modelos por microsismo 54

Comparación de predicciones entre modelos 55

Votación por mayoría 55

Capítulo III 56

Resultados 56

Preprocesamiento 56

Filtrado 56

Recorte 58

Procesamiento 60

Densidad espectral de potencia: Método Welch 60

Extracción de 84 características 60

PSD Welch RFE por método de envoltura 63

84 características RFE por método de envoltura 67

Bases de datos a entrenar 70

Balanceo de datos 71

Sistema de votación 75

Entrenamiento de algoritmos ML 75

Validación y pruebas de DT 77

Validación y pruebas de k-NN 78

Validación y prueba de SVM 79

Generación de modelos y pruebas 80

Modelo óptimo de DT 81

Modelo óptimo de k-NN 85

Modelo óptimo de SVM 88

Algoritmo de Votación 91

Pruebas con señales sintéticas 93

Microsismos por Bootstrap 93

Microsismos por CGAN 94

Interfaz Gráfica 96

Capítulo IV 103

Conclusiones 103

Trabajos Futuros 105

Bibliografía 108

Apéndices 113

Índice de tablas

Tabla 1 *84 Características entre tiempo, frecuencia y escala* 38

Tabla 2 *Matriz de confusión* 51

Tabla 3 *Obtención de 84 características* 61

Tabla 4 *Evaluación del barrido PSD Welch RFE con DT* 64

Tabla 5 *Evaluación del barrido PSD Welch RFE con k*-NN 65

Tabla 6 *Evaluación del barrido PSD Welch RFE con SVM* 65

Tabla 7 *Evaluación del barrido 84 características RFE con DT* 68

Tabla 8 *Evaluación del barrido 84 características RFE con k*-NN 68

Tabla 9 *Evaluación del barrido 84 características RFE con SVM* 69

Tabla 10 *Bases de datos y número de características utilizada para entrenamientos* 70

Tabla 11 *Rendimiento de la variación de entrenamiento del modelo DT para PSD Welch RFE* 71

Tabla 12 *Rendimiento de la variación de entrenamiento del modelo k*-NN *para PSD Welch RFE* 72

Tabla 13 *Rendimiento de la variación de entrenamiento del modelo SVM para PSD Welch RFE* 72

Tabla 14 *Rendimiento de la variación de pruebas del modelo DT para PSD Welch RFE* 73

Tabla 15 *Rendimiento de la variación de pruebas del modelo k*-NN *para PSD Welch RFE* 73

Tabla 16 *Rendimiento de la variación de pruebas del modelo SVM para PSD Welch RFE* 74

Tabla 17 *Evaluación de rendimiento en con datos de validación del modelo DT.* 77

Tabla 18 *Evaluación de rendimiento con datos de prueba del modelo DT* 77

Tabla 19 *Evaluación de rendimiento con datos de validación del modelo k*-NN 78

Tabla 20 *Evaluación de rendimiento con datos de prueba del modelo k*-NN 78

Tabla 21 *Evaluación de rendimiento con datos de validación del modelo SVM* 79

Tabla 22 *Evaluación de rendimiento con datos de prueba del modelo SVM* 79

Tabla 23 *Resultado de los modelos óptimos obtenidos* 80

Tabla 24 *Evaluación de rendimiento en validación por microsismo del modelo DT* 83

Tabla 25 *Evaluación de rendimiento en pruebas por microsismo del modelo DT*..... 84

Tabla 26 *Evaluación de rendimiento en validación por microsismo del modelo k*-NN 86

Tabla 27 *Evaluación de rendimiento en pruebas por microsismo del modelo k*-NN..... 88

Tabla 28 *Evaluación de rendimiento en validación por microsismo del modelo SVM* 90

Tabla 29 *Evaluación de rendimiento en pruebas por microsismo del modelo SVM* 91

Tabla 30 *Evaluación de rendimiento en pruebas del algoritmo de votación* 92

Tabla 31 Evaluación de rendimiento en pruebas por microsismo del algoritmo de votación 93

Tabla 32 *Evaluación de rendimiento en pruebas Bootstrap sintéticas del algoritmo de votación* 94

Tabla 33 *Evaluación de rendimiento en pruebas CGAN del algoritmo de votación*..... 95

Índice de figuras

Figura 1 *Diagrama de bloques general del sistema de clasificación* 26

Figura 2 *Ejemplo LP y su espectro en frecuencia* 29

Figura 3 *Ejemplo TC y su espectro en frecuencia* 29

Figura 4 *Ejemplo TR y su espectro en frecuencia* 31

Figura 5 *Ejemplo VT y su espectro en frecuencia* 31

Figura 6 *Base de datos* 33

Figura 7 *Procedimiento realizado en método de envoltura* 41

Figura 8 *Balanceo de datos* 43

Figura 9 *Proceso realizado en aprendizaje de máquina* 44

Figura 10 *Estructura de un árbol de decisión* 46

Figura 11 *Ejemplo de clasificación algoritmo k*-NN 48

Figura 12 *Representación gráfica de SVM* 50

Figura 13 *Sistema de Votación* 54

Figura 14 *PSD no filtrada de microsismo LP* 56

Figura 15 *Filtro digital FIR* 57

Figura 16 *PSD filtrada de microsismo LP* 58

Figura 17 *Microsismo LP filtrado en el dominio del tiempo* 59

Figura 18 *Microsismo LP filtrado y recortado en el dominio del tiempo* 59

Figura 19 *Obtención PSD Welch de microsismo, con una resolución de 257 puntos* 60

Figura 20 *Mesh de 84 características antes de normalizar* 62

Figura 21 *Mesh de 84 características después de normalizar* 62

Figura 22 *Histograma de características seleccionadas por método de envoltura* 63

Figura 23 *Representación de características sobre señales sismo volcánicas* 66

Figura 24 *Histograma de características seleccionadas por método de envoltura* 67

Figura 25 *Representación de características sobre señales sismo volcánicas* 70

Figura 26 *Balanceo de datos óptimo*..... 75

Figura 27 *Votación entre modelos* 76

Figura 28 *Mejor BER al entrenar el algoritmo DT*..... 81

Figura 29 *Árbol de decisión del Mejor BER al entrenar el algoritmo DT*..... 81

Figura 30 *Cuadro de confusión por microsismo del mejor BER en validación del algoritmo DT* 82

Figura 31 *Cuadro de confusión por microsismo del mejor BER en pruebas del algoritmo DT* 84

Figura 32 *Mejor BER al entrenar el algoritmo k*-NN..... 85

Figura 33 *Cuadro de confusión por microsismo del mejor BER en validación del algoritmo k*-NN 86

Figura 34 *Cuadro de confusión por microsismo del mejor BER en pruebas del algoritmo k*-NN..... 87

Figura 35 *Mejor BER al entrenar el algoritmo SVM*..... 88

Figura 36 *Cuadro de confusión por microsismo del mejor BER en validación del algoritmo SVM* 89

Figura 37 *Cuadro de confusión por microsismo del mejor BER en pruebas del algoritmo SVM* 90

Figura 38 *Cuadro de confusión por microsismo del mejor BER en pruebas del algoritmo votación* 92

Figura 39 *Cuadro de confusión por microsismo sintético Bootstrap del algoritmo de votación*..... 94

Figura 40 *Cuadro de confusión por microsismo sintético CGAN del algoritmo de votación* 95

Figura 41 *Interfaz gráfica: presentación de autores* 96

Figura 42 *Interfaz gráfica: configuración de operación automática* 98

Figura 43 *Interfaz gráfica: presentación de señal ingresada*..... 99

Figura 44 *Interfaz gráfica de procesamiento de señal ingresada*..... 100

Figura 45 *Interfaz gráfica: registro de señal detectada y clasificada* 101

Figura 46 *Interfaz gráfica: Acerca de* 102

Resumen

El estudio sismológico ha proporcionado información sumamente valiosa acerca del comportamiento de distintos volcanes, lo cual ha posibilitado llevar a cabo un análisis exhaustivo de su actividad volcánica. La monitorización constante de estos volcanes ayuda a realizar estimaciones de eventos futuros, lo que ha impulsado al desarrollo de métodos automáticos para reconocer microsismos, los cuales son esenciales para la emisión de alertas tempranas, contribuyendo así a la protección de vidas.

En este contexto, se plantea implementar algoritmos inteligentes basados en la teoría de Aprendizaje de Máquina (ML, del inglés *Machine Learning*) tradicional para la clasificación de los microsismos del volcán Llaima, centrándonos en cuatro señales de microsismos como: Volcano Tectónicos, Largo Período, Tremor y Tectónico. La propuesta presenta tres algoritmos de aprendizaje supervisado como: Árbol de Decisión (DT, del inglés *Decision Tree*), *k*-Vecinos más Cercanos (*k*-NN, del inglés *k-Nearest Neighbors*) y Máquina de Vectores de Soporte (SVM, del inglés *Support Vector Machine)*. Por último, presentamos un algoritmo de votación que opera entre los algoritmos de ML mencionados. Para evaluar el desempeño de estos algoritmos se consideran métricas como Exactitud, Precisión, Sensibilidad, Especificad y la tasa de error balanceado (BER, del inglés *Balance Error Rate)*. Los resultados obtenidos al evaluar el sistema de clasificación propuesto presentan una Exactitud de 92% para DT, 94% para *k*-NN, 96% para SVM y 96% al aplicar el enfoque de votación y un BER de 0.05, 0.04, 0.03 y 0.03 respectivamente.

Palabras clave: Aprendizaje supervisado, Llaima, clasificación, microsismos, características espectrales.

Abstract

The seismological study has provided extremely valuable information about the behavior of different volcanoes, which has made it possible to carry out an exhaustive analysis of their volcanic activity. The constant monitoring of these volcanoes helps to estimate future events, which has led to the development of automatic methods for recognizing micro-earthquakes, which are essential for the issuance of early warnings, thus contributing to the protection of lives.

In this context, it is proposed to implement intelligent algorithms based on traditional Machine Learning (ML) theory for the classification of the Llaima volcano micro-earthquakes, focusing on four micro-earthquake signals such as: Volcano Tectonic, Long Period, Tremor and Tectonic. The proposal presents three supervised learning algorithms such as: Decision Tree (DT), *k*-Nearest Neighbors (*k*-NN) and Support Vector Machine (SVM). Finally, we present a voting algorithm that operates between the aforementioned ML algorithms. To evaluate the performance of these algorithms, metrics such as Accuracy, Accuracy, Sensitivity, Specificity and the Balanced Error Rate (BER) are considered. The results obtained when evaluating the proposed classification system show an Accuracy of 92% for DT, 94% for *k*-NN, 96% for SVM and 96% when applying the voting approach and a BER of 0.05, 0.04, 0.03 and 0.03 respectively.

Key words: Supervised learning, Llaima, classification, microseismic, spectral characteristics.

Capítulo I

Introducción

Ecuador, un país que alberga un total de 90 volcanes en su territorio de los cuales se han identificado 27 como potencialmente activos, enfrenta una constante sensación de alerta entre su población debido a la posibilidad de verse afectado por riegos naturales de considerable magnitud. Este contexto ha propiciado un estado de cautela y preparación en la sociedad. La investigación focalizada en volcanes emblemáticos como el Cotopaxi en Ecuador y el Llaima en Chile ha permitido a los científicos adquirir un entendimiento más amplio y diversificado acerca del comportamiento volcánico previo a sus episodios eruptivos.

Las actividades volcánicas más actuales registradas del volcán Llaima fueron durante el 2007 al 2009, donde su ciclo eruptivo estuvo caracterizado por varias fases bien definidas. La primera fase ocurrida entre el 26 mayo al 31 diciembre del 2007, la actividad sísmica del volcán se inicia con la observación de explosiones menores y la expulsión de cenizas. Las señales sísmicas captadas por dichas explosiones corresponden a microsismos de Largo Periodo (LP) y ocasionalmente Tremores (TR). En su segunda fase comienza la erupción mayor del ciclo con explosiones más vigorosas, donde se alcanzó una cantidad de energía sísmica liberada de aproximadamente 5000 unidades promedio de la amplitud absoluta por minuto, provocado por un TR altamente energético. Entre las fases 3 y 6, ocurrida durante el 02 enero hasta el 01 julio del 2008, la actividad volcánica fue caracterizada por emisiones esporádicas de cenizas y gases, explosiones laterales aisladas, explosiones menores con proyección de salpicaduras conocidos como *spatters* y una reducción de su actividad sísmica para la fase 6. En la fase 7, se registró una destacada reactivación sísmica que se manifestó a través de 5 episodios eruptivos. En un principio, se observó una predominancia de microsismos LP, seguida por la aparición de tremores energéticos y concluye con un retorno a microsismos LP. A lo largo de las fases 8, 9 y 10, se pudo notar un ciclo

eruptivo y sísmico decreciente, caracterizado por explosiones de cenizas de naturaleza débil y esporádica. Esta actividad se mantuvo hasta la fecha del 07 abril del 2009, (Moreno, 2009).

En la actualidad, la monitorización y reconocimiento de microsismos provenientes del volcán Llaima se lleva a cabo mediante una red instrumental de monitorización que incluye estaciones sismológicas, estaciones GPS, inclinómetros electrónicos, cámaras de observación fija y sensores de gases por absorción espectroscópica. Estas estaciones están distribuidas en distancias que varían entre 1km y 21km con relación con el cráter activo (Franco Marín, 2019). La información recopilada de la detección de los microsismos es almacenada para su posterior clasificación, a menudo con el uso de método tradicional como la clasificación visual realizada por un experto.

La aplicación de algoritmos de aprendizaje automático (ML, del inglés *Machine Learning*) surge como una solución para la detección y clasificación de microsismos. Estos algoritmos entrenados permiten automatizar estas tareas. Entre las técnicas más utilizadas para la detección son de promedio a corto plazo (STA, del inglés *Short-Time-Average Through*) y promedio a largo plazo (LTA, del inglés *Long-Time-Average Trigger*), estos algoritmos procesan ondas sísmicas en ventanas dinámicas de tiempo de corto y largo periodo. El STA mide amplitud en un momento específico de una señal sísmica con el propósito de identificar eventuales terremotos. Por otro lado, LTA determina la amplitud promedio del ruido sísmico presente en la señal, (Trnkoczy, 1999).

El proceso de STA/LTA involucra varios pasos clave para determinar la presencia de eventos sísmicos. En primera instancia, funciona al calcular la amplitud absoluta de cada punto de datos en una señal de entrada y luego se obtiene el promedio de estas amplitudes absolutas en dos ventanas diferentes. La relación entre estos dos valores STA/LTA se compara con un umbral predefinido por el usuario. Si la relación STA/LTA supera el umbral, se activa el canal y comienza la grabación de señales sísmicas.

Sin embargo, superar el umbral no garantiza que se detecte un evento sísmico. Los sistemas sísmicos suelen contar con un mecanismo de "disparo de votación" que define cuántos canales deben activarse antes de que la grabación comience. Esto se debe a que a veces los ruidos aleatorios pueden hacer que la relación STA/LTA supere el umbral. El mecanismo de disparo de votación ayuda a evitar que esto suceda. Una vez que la actividad sísmica disminuye gradualmente, el canal se detiene. Esto ocurre cuando la relación STA/LTA desciende por debajo de otro umbral establecido por el usuario. Es evidente que el nivel de umbral de activación STA/LTA debe ser menor o igual que el nivel de umbral de disparo STA/LTA, (Ginez, 2019). Otro algoritmo de detección es detección de actividad de voz (VAD, del inglés *Voice Activity Detection*) para identificar señales sísmicas incluso en presencia de niveles de ruido, lo que facilita la identificación de microsismos en los conjuntos de datos capturados por las estaciones sismológicas.

La detección de microsismos da paso a algoritmos de ML como Árboles de Decisión (DT, del inglés *Decision Tree*), *k*-Vecinos Más Cercanos (*k*-NN, del inglés *k-Nearest Neighbors*) y Máquina de Vectores de Soporte (SVM, del inglés *Support Vector Machine)*, para a llevar a cabo la clasificación eficiente y precisa de estos microsismos. Si bien el balanceo de la base de datos puede presentar una limitación al utilizar ciertos algoritmos de ML, existe una solución que involucra el uso de técnicas de aprendizaje profundo. Estas técnicas pueden resultar especialmente beneficiosas en este contexto, ya que no requieren modificar la base de datos existente, lo que simplifica significativamente el proceso de entrenamiento.

Antecedentes

El estudio realizado por varios investigadores sobre la clasificación de microsismos entrega una perspectiva interesante al emplear diferentes técnicas de clasificación basados en la teoría de ML tradicional, los cuales pueden ser de gran utilidad para determinar el comportamiento intrínseco de un volcán.

La investigación realizada por (Lara-Cueva, Carrera, Morejon, & Benitez, 2016) presenta un análisis comparativo de clasificadores con microsismos como Volcano Tectónico (VT) y LP, a partir de una base de datos proporcionada por el Instituto Geofísico de la Escuela Politécnica Nacional (IGEPN), referentes a microsismos ocurridos en el volcán Cotopaxi, donde a partir de cada microsismo se extrajeron 79 características basadas en análisis en tiempo y frecuencia, utilizadas para la implementación de los algoritmos de DT, *k*-NN y Redes Neuronales, con el fin de determinar aquel algoritmo que presente mejor desempeño en la clasificación; *k*-NN con un valor de k=29, mostró que sus métricas de desempeño superan el 95% con un tiempo de procesamiento del sistema (tp) de 143 ms al ser la mejor respecto a los otros algoritmos, además de mostrar mejoras al utilizar el método *wrapper*, donde se encontró que las características más relevantes se reducían a 16 de las cuales al utilizar 3 de estas se alcanzó un desempeño más alto con un 97% y una reducción de su tp a 7ms.

La investigación sobre microsismos en tiempo real es un tema a discutir por la importancia que este presenta para identificar posibles erupciones mediante la monitorización y detección de la actividad de microsismos en un volcán como lo indica (Altamirano, 2021) refiriéndose al volcán Cotopaxi, donde realiza procesos de segmentación de la señal en ventanas, extracción de características de tiempo, frecuencia y escala, selección de características, entre otros; para aplicar modelos de aprendizaje supervisado como DT, *k*-NN y SVM. Los resultados obtenidos en el clasificador con una tasa de error de balance (BER, del inglés *Balanced Error Rate*) de 0.115, el cual está por encima del requerimiento del IGEPN de 0.01, considera que las pruebas realizadas son en escenarios desproporcionados, puesto que se tienen 282 microsismos LP y 50 VT.

En otro enfoque (Iglesias & Rosero, 2023) presentan el uso de algoritmos de aprendizaje profundo para clasificar microsismos de tipo LP y VT. Este análisis se apoya en datos proporcionados por el IGEPN, así como en bases de datos de datos sintéticas y combinadas que fusionan ambas fuentes de

información. Además, la incorporación de los coeficientes de dos transformadas *Wavelet* discretas, *Daubechies* y *Symlets*, enriquece aún más la metodología al generar dos nuevas bases de datos mixtas.

Este proceso de generación de bases de datos mixtas ha demostrado un impacto positivo en la generalización y robustez de los clasificadores. Este enfoque ha permitido obtener modelos de mayor precisión al abordar la clasificación de microsismos. Los resultados arrojados por la implementación del algoritmo *Autoencoder* Apilado (SA, del inglés *Stacked Autoencoder*) han superado de manera notable los resultados mediante la utilización de la Red Neuronal Profunda (DNN, del inglés *Deep Neural Network),* incluso en comparación con los objetivos establecidos por el IGEPN.

Mientras que en (Lara, Lara-Cueva, Larco, Carrera, & León, 2021) entregan un enfoque de aprendizaje profundo en un sistema de reconocimiento automático de microsismos, en la cual tanto para las etapas de detección y clasificación la utilización de Redes Neuronales Convolucionales (CNN, del inglés *Convolutional Neuronal Network*) proporcionan una mejora en los resultados de las métricas de desempeño, por lo cual mediante el uso de espectrogramas generados a partir de periodo gramas con tamaño de ventana variable se aplica el método *Transfer Learning* a modelos como *AlexNet*, *GoogLeNet*, *SqueezeNet*, etc., el sistema diseñado obtiene una precisión del 99% en la etapa de detección y un 97% en la etapa de clasificación.

Además (Minango, 2022) propone un clasificador de microsismos al usar características psicoacústicas con la capacidad de detectar tres clases de microsismos LP, VT y otros, de este último se encuentran microsismos regionales (RG), híbridos (HB) y deslizamiento de glaciares (IC), al utilizar la base de datos proporcionada por el IGEPN de la cual se extrajo 11 características espectrales representadas en las escalas psicoacústicas Mel, Bark y Ancho de banda rectangular equivalente (ERB, del inglés *Equivalent Rectangular Bandwidth*), aplicadas a un análisis de componentes principales cuyos resultados indican que las características con mayor precedencia son Disminución, Planitud y

Atenuación espectral. Una vez obtenidas estas características al utilizar técnicas basadas en aprendizaje supervisado como DT, *k*-NN y en aprendizaje no supervisado con el *Autoencoder*, las métricas de desempeño muestran que tanto para *k*-NN y *Autoencoder* es del 98.41% con un BER del 0.018, superior al 96.29% obtenido en DT con un BER del 0.042, resultados validados por cada modelo de clasificación con la base de datos del volcán Llaima en Chile proporcionado por el Observatorio Vulcanológico de los Andes del Sur (OVDAS).

Estos logros se reflejan en las métricas de desempeño, donde el BER se posiciona en 0.0020 para SA, 0.0027 para DNN. Estos resultados resaltan la efectividad y superioridad del enfoque basado en SA, lo que a su vez sugiere un avance prometedor en la capacidad de identificación y clasificación de microsismos con aplicaciones significativas en la sismología y la investigación volcánica.

La propuesta que indica (Galarza & Vega, 2022) en diseñar un modelo basado en la Red Neuronal Adversario Generativo Condicional (CGAN, del inglés *Conditional Generative Adversarial Network*), tiene la capacidad de generar señales sintéticas representativas de microsismos de tipo LP y VT, en este proceso utiliza la base de datos suministrada por la IGEPN. Durante el proceso de entrenamiento, se observó que, al aumentar el número de filtros y capas en la CGAN, surgieron problemas inherentes al entrenamiento, mientras que la reducción de filtros resultó en señales sintéticas que eran altamente similares a las originales. La clasificación de estas señales sintéticas arrojó resultados precisos: una exactitud del 90.4% en el sistema de reconocimiento, un 97.5% en el *Autoencoder* y un 90.5 % en la clasificación visual.

En relación a este tema, (Zapata, 2022) propone la generación de microsismos empleando la técnica de remuestreo mediante *bootstrapping*, con el objetivo de obtener microsismos que presenten similitudes en sus características con los microsismos reales LP y VT registrados en el volcán Cotopaxi. Este enfoque implica la reiterada selección de muestras de las señales originales, lo que permite generar

nuevas señales con propiedades semejantes. Para lograr esto, se hace uso de los espectrogramas derivados de las señales originales, los cuales proporcionan una representación en el dominio tiempo-frecuencia. La evaluación de los resultados obtenidos a través de esta técnica arroja resultados prometedores, confirmados tanto por una inspección visual como por la métrica de evaluación conocida como distancia de *Frechet*.

Justificación e importancia

Tradicionalmente, la vigilancia y monitorización de un volcán se lo hace de forma visual a través de los observatorios vulcanológicos y de personal capacitado en el área como lo son expertos en geofísica, donde el estudio de la sismicidad volcánica es una herramienta importante para entender el comportamiento de dicho fenómeno, por estas razones existe una gran necesidad de estudiar métodos que permitan precautelar las vidas humanas mediante sistemas de alerta temprana frente a posibles erupciones mediante la monitorización y reconocimiento de microsismos de un volcán, con el fin de anticipar procesos eruptivos que podrían ser devastadores para un territorio.

En la actualidad, se han desarrollado investigaciones que proponen la clasificación automática de los microsismos, basadas en técnicas de ML tradicional, con el fin de determinar el tipo microsismo registrado. No obstante, a nuestro entender aún no existen métodos concretos de clasificación multiclase que permitan obtener el valor requerido del BER indicado por el IGEPN establecido en 0.01.

El propósito de este trabajo es colaborar en el desarrollo de un clasificador multiclase basándonos en investigaciones previas realizadas con el volcán Cotopaxi y Llaima. Este sistema de clasificaciones etiqueta microsismos como: VT, LP, TR y Tectónicos (TC). Además, se busca avanzar con la generalización de un clasificador de microsismos con diferentes volcanes. Esto permitirá una aplicación más amplia y confiable de los resultados obtenidos en diferentes contextos volcánicos, al contribuir en el entendimiento más completo de los microsismos en estos entornos.

Objetivo

Crear algoritmos inteligentes basados en la teoría de Machine Learning tradicional para la clasificación de los eventos sísmicos en el volcán Llaima (multiclase).

Actividades

- Actividad 1. - identificación de las propiedades y variables del problema a tener en cuenta.
- Actividad 2. - reunión de la base de datos representativa, con datos provistos por el Observatorio Volcanológico de los Andes del Sur.
- Actividad 3. - creación de una estructura de datos (temporales, espaciales, otros.) con un soporte común.
- Actividad 4. - pruebas con los algoritmos basados en la teoría de Machine Learning especializados.
- Actividad 5. - pruebas y evaluación del desempeño.

Capítulo II

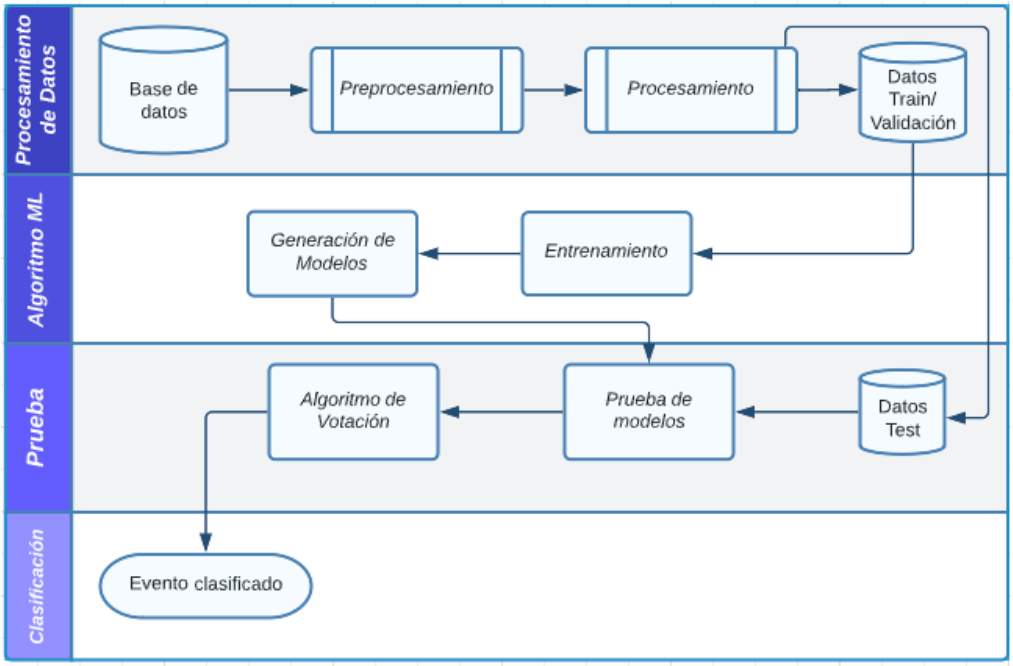
Métodos y materiales

En este capítulo se presenta una metodología basada en un enfoque experimental. Aquí se describen los procedimientos empleados para el preprocesamiento, procesamiento, entrenamiento de algoritmos de ML, generación de modelos, pruebas y uso de un algoritmo de votación, tal y como se detalla en la Figura 1. Este proceso se aplica a la clasificación de los microsismos LP, TC, TR y VT, extraídos de la base de datos del volcán Llaima.

Para la aplicación de los mencionados métodos se crearon códigos en el software matemático MATLAB® versión R2022b.

Figura 1

Diagrama de bloques general del sistema de clasificación



La Figura 1 muestra un esquema que detalla de manera simple el sistema de clasificación. Está dividido en cuatro etapas generales, cada etapa tiene una serie de procesos específicos y organizados de manera lineal.

La etapa de preprocesamiento de la base de datos se encarga de filtrar y recortar las señales, es decir, eliminar ruido, frecuencias no deseadas y seleccionar solamente la señal de interés. A continuación, en procesamiento se divide en cuatro subetapas: extracción de la densidad espectral de potencia (PSD, del inglés *Power Spectral Density*) mediante el método Welch, extracción de características en los dominios de tiempo, frecuencia y escala, aplicación del algoritmo de eliminación de características recursivas (RFE, del inglés *Recursive Feature Elimination*) por el método de envoltura, normalización y balanceo de la base de datos.

En la etapa de aplicación de algoritmos ML, se entrenan los diferentes conjuntos de datos obtenidos en la etapa de procesamiento, se comparan las métricas de desempeño como: Exactitud (A, del inglés *Accuracy*), Precisión (P, del inglés *Precision*), Especificidad (S, del inglés *Specificity*), Sensibilidad (R, del inglés *Recall*) y BER de los diferentes modelos y se generan los tres mejores modelos para maximizar los resultados en el sistema de clasificación.

Por último, en la etapa de pruebas, se utilizan los modelos generados para clasificar los datos de prueba y evaluar su rendimiento en base a las métricas ya mencionadas, lo que permite determinar qué algoritmo de ML es el óptimo para implementar un sistema de votación basado en prioridades.

Conjunto de datos del clasificador

Volcán Llaima

El Volcán Llaima se encuentra ubicado en la región de Araucanía, en el centro-sur de Chile, es uno de los volcanes más activos de Sudamérica. Este volcán se encuentra en la cordillera de los Andes a una altitud de aproximadamente 3.125 metros por encima del nivel del mar. Su edificio volcánico tiene

una forma cónica que consta de dos cimas principales, la cumbre norte revela un cráter abierto de 350 m de diámetro y más de 300 m de profundidad, mientras que en su cumbre sur se observa rastros de un cono de escorias de menor tamaño ubicado dentro de un cráter parcialmente bloqueado con una longitud aproximada de 200 metros (SERNAGEOMIN, 2023).

Microsismos

Los volcanes no solo liberan gases y material volcánico, sino que producen microsismos que son indicadores del comportamiento y cambios que presenta un volcán. Estos microsismos pueden ser producidos por múltiples factores como el movimiento del magma a través de fracturas nuevas o preexistentes generadas por la inyección de magma, desgasificación del magma, fracturación de la roca subterránea debido a cambios bruscos de temperatura, entre otras, (OVSICORI, 2023). Los microsismos que están relacionados con la actividad volcánica son:

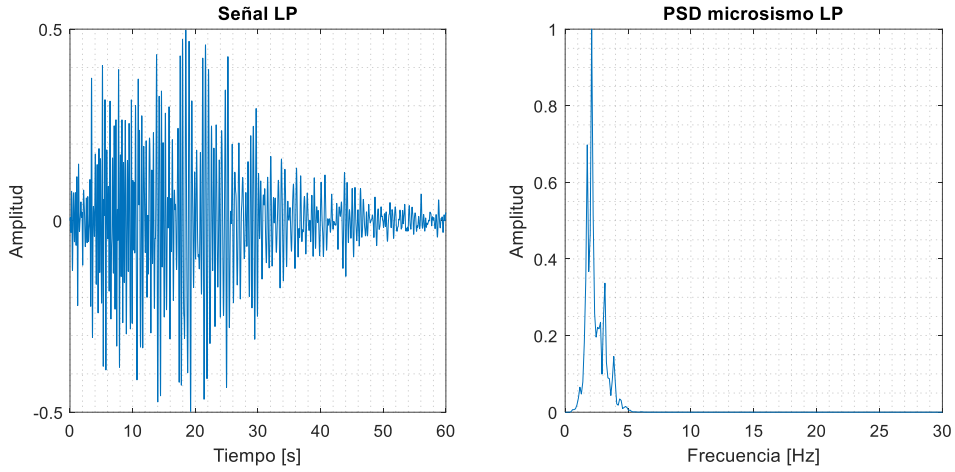
Largo periodo

También conocido como sismo de baja frecuencia presente en la Figura 2, se originan cerca de la superficie, a profundidades inferiores a 1km. Estos microsismos se consideran de naturaleza transitoria y no destructiva y se atribuyen a la resonancia de grietas o conductos llenos de fluidos en el sistema volcánico.

Los LP se caracterizan en tener una duración que oscila entre algunos segundos hasta algo más de 1 minuto. Además, presentan magnitudes muy pequeñas y un contenido espectral limitado. Por lo general, sus frecuencias se encuentran en el rango de 1 a 5 Hz, aunque algunos microsismos se pueden concentrar el 95% de la energía espectral entre los 5 y 10 Hz (CENAPRED, 2018).

Figura 2

Ejemplo LP y su espectro en frecuencia

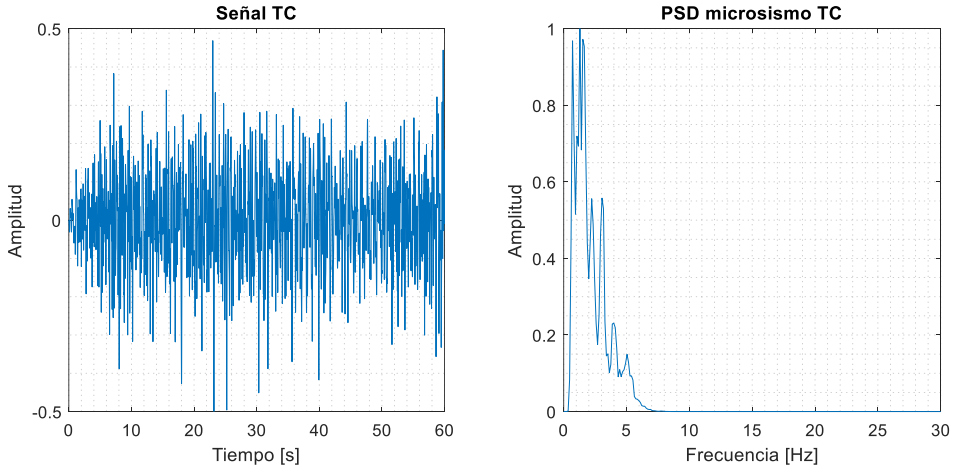


Tectónico

Un ejemplo de microsismos tectónicos se presenta en la Figura 3.

Figura 3

Ejemplo TC y su espectro en frecuencia



Los TC son fenómenos sísmicos de pequeña magnitud. Estos microsismos son resultado de la interacción de las placas tectónicas y las tensiones acumuladas en la corteza terrestre en las cercanías de un volcán activo. El rango de frecuencias se extiende desde alrededor de 0.1 Hz hasta varios Hz.

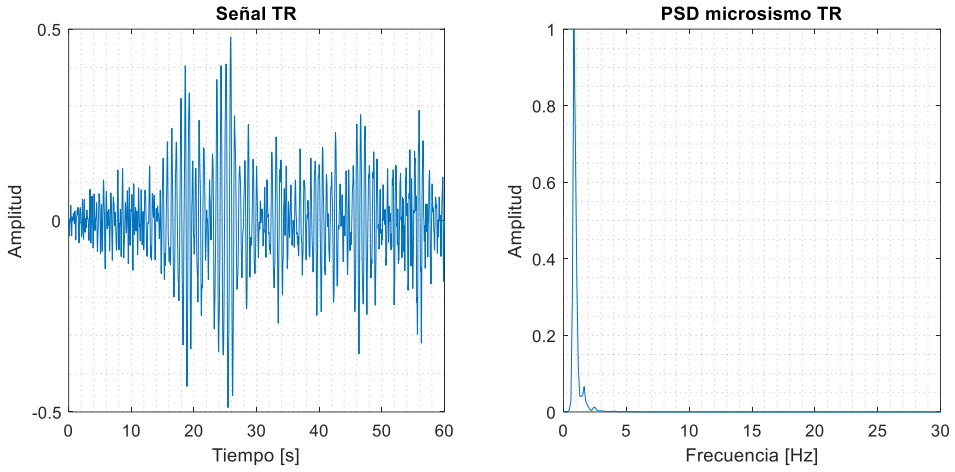
Tremor

Los TR son procesos de fuente no destructiva y de carácter persistente, caracterizada por generar microsismos con amplitud constante durante un largo periodo de tiempo que puede oscilar entre varios minutos y horas (Yugsi, 2022). El TR está asociado con movimiento de fluidos en el interior del volcán y su característica en frecuencia es dentro del rango de 0.5 y 3 Hz, esto se puede apreciar en la Figura 4. En la clasificación de tremores existen diferentes tipos como: armónicos, espasmódicos, episódico, volcánico de banda ancha.

- *Tremores armónicos*. - son aquellos que tienen una frecuencia fundamental y sus armónicos bien definidos, muestra una periodicidad regular y una amplitud constante.
- *Tremores espasmódicos*. - no mantienen una frecuencia fundamental constante y sus amplitudes son variables.
- *Tremores episódicos*. - se presenta en intervalos definidos de tiempo claramente definidos, al alternar entre periodos de actividad sísmica con una mayor amplitud y frecuencia, y periodos de inactividad o actividad mínima sísmica.
- *Tremor volcánico de banda ancha*. - abarca un amplio rango de frecuencias, bajas y altas, y un cambio de su amplitud a lo largo de la frecuencia. (CENAPRED, 2018)

Figura 4

Ejemplo TR y su espectro en frecuencia

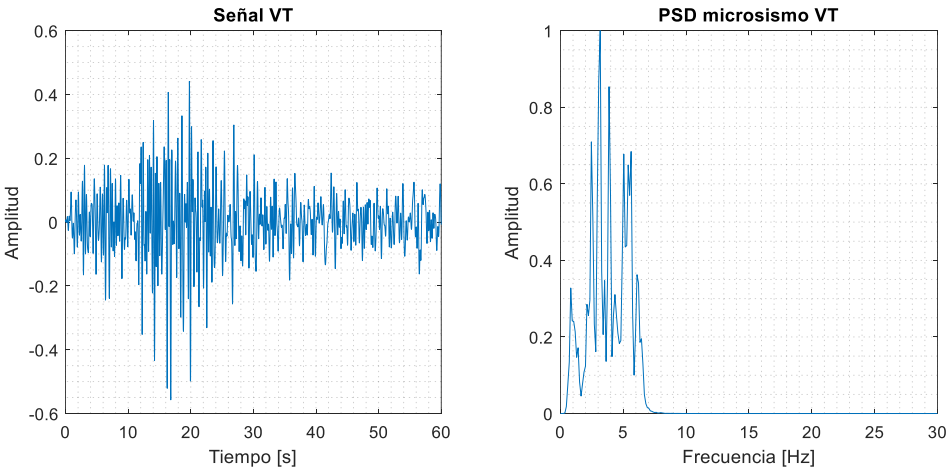


Volcano Tectónicos

Los VT también llamados de alta frecuencia, se producen debido a la fracturación de rocas adyacente a los conductos volcánicos, (Yugsi, 2022).

Figura 5

Ejemplo VT y su espectro en frecuencia



Estos microsismos se caracterizan por tener frecuencias predominantes en el rango de 5 a 15 Hz, aunque ocasionalmente estas pueden ser más altas, como se puede apreciar en la Figura 5. Además, presentan un amplio contenido espectral y son microsismos asociados con el rompimiento de rocas, y la apertura de grietas en la estructura volcánica.

Base de datos

La base de datos proviene del volcán Llaima[1]. Estas señales fueron recopiladas de la estación sísmica LAV, una de las siete estaciones encargadas de la monitorización constante del volcán Llaima, bajo la supervisión del Observatorio Vulcanológico de los Andes del Sur (OVDAS). La base de datos contiene un total de 3592 señales registradas entre los años 2010 y 2016, con una frecuencia de muestreo de 100 Hz, 6000 muestras por cada señal, se utiliza un filtro de paso de banda Butterworth de décimo orden en el rango de 1 a 10 Hz para preservar el ancho de banda que contiene el rango de interés. La base de datos está clasificada en cuatro categorías principales: VT con 304 señales, LP con 1310 señales, TR con 490 señales y TC con 1488 señales (Canário, Fernandes, Curilem, Huenupan, & Araujo, 2020).

La estructura de la base de datos utilizada en la presente investigación se presenta en la Figura 6. Esta base de datos está compuesta por 11 columnas, las cuales contienen información relevante sobre los microsismos. Las columnas de mayor interés para el clasificador corresponden a: *SampleRate*, *Type*, *StartPoint*, *EndPoint* y *Data*. Estas se describen a continuación:

- *SampleRate*. - indica la tasa de muestreo utilizada por la estación "LAV" para capturar la información sobre el microsismo.

[1] El conjunto de datos fue proporcionado cortesía del Observatorio Vulcanológico de los Andes Sur (OVDAS) está disponible en http://dx.doi.org/10.17632/dv8nwdd36k

- *Type*. - indica el tipo de microsismo al que pertenece cada conjunto de dato. Esta puede ser LP, TC, TR y VT.
- *StartPoint y EndPoint.* - indica el punto de inicio y final del microsismo en el conjunto de muestras de *Data.*
- *Data.* - almacena la información del microsismo capturada por la estación "LAV".

Figura 6

Base de datos

	1 Network	2 Station	3 SampleRate	4 Component	5 Year	6 Month	7 Type	8 Duration	9 StartPoint	10 EndPoint	11 Data
1	"Chile"	"LAV"	100	"SHZ"	"2010-2016"	1	"LP"	38.1100	1	3812	*1x6000 double*
2	"Chile"	"LAV"	100	"SHZ"	"2010-2016"	1	"LP"	39.6700	1	3968	*1x6000 double*
3	"Chile"	"LAV"	100	"SHZ"	"2010-2016"	1	"LP"	47.5300	1	4754	*1x6000 double*
4	"Chile"	"LAV"	100	"SHZ"	"2010-2016"	1	"LP"	30.2500	1	3026	*1x6000 double*
5	"Chile"	"LAV"	100	"SHZ"	"2010-2016"	1	"LP"	22.9900	1	2300	*1x6000 double*
6	"Chile"	"LAV"	100	"SHZ"	"2010-2016"	1	"LP"	34.3200	1	3433	*1x6000 double*
7	"Chile"	"LAV"	100	"SHZ"	"2010-2016"	1	"LP"	27.2200	1	2723	*1x6000 double*
8	"Chile"	"LAV"	100	"SHZ"	"2010-2016"	1	"LP"	37.9400	1	3795	*1x6000 double*
9	"Chile"	"LAV"	100	"SHZ"	"2010-2016"	1	"LP"	39.3200	1	3933	*1x6000 double*
10	"Chile"	"LAV"	100	"SHZ"	"2010-2016"	1	"LP"	30.6000	1	3061	*1x6000 double*
11	"Chile"	"LAV"	100	"SHZ"	"2010-2016"	1	"LP"	43.2100	1	4322	*1x6000 double*
12	"Chile"	"LAV"	100	"SHZ"	"2010-2016"	1	"LP"	33.5300	1	3354	*1x6000 double*

Nota. Distribución de la base de datos. Columnas relevantes para el clasificador: *SampleRate*, *Type*, *StartPont*, *EndPoint* y *Data*.

Preprocesamiento

En este proceso se inicia realmente la etapa del procesamiento de la base de datos para tener como resultado los mejores modelos para el sistema de clasificación de los microsismos. En este punto se selecciona el conjunto de las señales *Data* que sufre una serie de cambios como el filtrado, recorte, extracción de características y normalización. Las cuales nos permiten obtener varias bases de datos para la posterior aplicación del algoritmo RFE por envoltura y el entrenamiento de los algoritmos de aprendizaje automático.

Filtrado

El filtrado de señales es una técnica esencial que permite mejorar la calidad de una señal, separar componentes y eliminar ruido no deseado, lo que facilita su análisis, interpretación y aplicación en diversas áreas. En este contexto, para eliminar cualquier distorsión causada por el retraso de fase y las frecuencias asociadas al ruido de fondo oceánico, se aplicará un filtrado digital bidireccional de respuesta al impulso finito (FIR, del inglés *Finite Impulse Response*). El operador $h_M\{s_M\}$ representa matemáticamente el proceso de filtraje aplicado al conjunto de señales de la matriz S indicado por la ecuación (1) , donde s_M representa una señal de este conjunto y M representa las 3592 señales contenidas en la base de datos.

$$S = \{s_1^T, s_2^T, s_3^T, \dots, s_M^T\}^T , \tag{1}$$

El resultado se almacena en la matriz H cómo se indica en la ecuación (2), donde h_M representa la aplicación del filtro a cada señal.

$$H = \{h_1^T, h_2^T, h_3^T, \dots, h_M^T\}^T. \tag{2}$$

Recorte

Una vez filtrada la base de datos nos enfocamos en recortar la señal, esto significa seleccionar específicamente el microsismo del conjunto de muestras total. Este recorte se lo realiza mediante la aplicación de los puntos de inicio y fin otorgados por la misma base de datos. El operador $r_M\{h_M\}$ representa el proceso de recorte aplicado a las señales r_M, almacenadas en la matriz R como se indica en la ecuación (3).

$$R = \{r_1^T, r_2^T, r_3^T, \dots, r_M^T\} . \tag{3}$$

Procesamiento

El procesamiento de las señales de microsismos desempeña un papel crucial en la extracción de información significativa y relevante. La aplicación de técnicas como la extracción de la PSD mediante el método Welch, así como la extracción de características en tiempo, frecuencia y escala, basándose en trabajos previos que han identificado 84 características para cada señal, permiten descomponer y analizar los microsismos en diferentes dominios. Esto proporciona una comprensión más profunda de las características y patrones presentes en los microsismos. Además, la normalización, la aplicación de RFE y el balanceo de la base de datos garantizan una representación equilibrada y coherente de los datos, se evita los sesgos y se asegura un rendimiento óptimo en los algoritmos de aprendizaje automático.

PSD por el Método Welch

Para realizar un análisis espectral de los microsismos, es necesario aplicar diferentes técnicas que permitan obtener información relevante en frecuencia. En este sentido, la PSD desempeña un papel fundamental. Este es una medida que describe cómo se distribuye la energía de una señal en diferentes frecuencias. En este caso, se emplea el método de Welch para calcular la PSD.

Welch es una técnica utilizada para estimar la PSD de una señal. Consiste en dividir la señal en segmentos de tiempo y calcular el periodograma en ventanas para cada segmento. Luego, se promedian los periodogramas para obtener una estimación más precisa de la PSD. Este enfoque de segmentación y promediado en ventanas tiene como objetivo reducir el sesgo y la varianza en el cálculo de la PSD (Alessio, 2016).

La elección de este método se basa en investigaciones previas, como el trabajo realizado por (Lara-Cueva, y otros, 2020), donde se implementó un sistema de reconocimiento de microsismos en el volcán Cotopaxi al utilizar distintos métodos de densidad espectral, como Yule-Walker, Burg y Welch.

Los resultados demostraron que el método de Welch alcanzó una exactitud del 90% y un BER de 0.05, superaron los métodos de Yule-Walker y Burg en términos de métricas de rendimiento. Estos hallazgos respaldan la importancia y eficacia del método de Welch en el análisis de microsismos al permitir una mayor precisión en la caracterización y reconocimiento de microsismos.

En este estudio, implementamos la técnica de PSD Welch por medio de la metodología propuesta por (Lara-Cueva, D.S., Carrera, Ruiz, & Rojo-Álvarez, 2016). Según esta referencia, se adopta un enfoque que asegura una resolución equilibrada en frecuencia y tiempo al restringir el cálculo de la Transformada Rápida de Fourier (FFT, del inglés *Fast Fourier Transform*) a 512 puntos. Además, se mantienen fijos los siguientes parámetros:

- *Ventana de Hamming.* - proporciona una representación más precisa del contenido de frecuencia de la señal y evita distorsiones en el análisis de frecuencia.
- *Resolución de 512 puntos.* - indica el número de puntos en la FFT.
- *Solapamiento al 50%.* - indica el solapamiento entre ventanas, lo que permite que se superpongan para reducir los efectos de borde en el cálculo de la PSD.

Por tanto, con estos parámetros se logra obtener un equilibrio entre la resolución de características en frecuencia y el costo computacional asociado al obtener modelos de ML y como resultado se proporcionan 257 características en frecuencia. El operador $p_M\{r_M\}$ representa la aplicación de la PSD Welch a las señales recortadas r_M, estas se almacenan en una matriz P, donde p_M representa la PSD Welch de cada señal como se indica en la ecuación (4).

$$P = \{p_1^T, p_2^T, p_3^T, \dots, p_M^T\}^T . \tag{4}$$

Extracción de 84 características

La extracción de características desempeña un papel fundamental en el procesamiento y comprensión de un diverso conjunto de señales. Este algoritmo destaca por su capacidad para representar señales en diversos dominios, como el temporal, frecuencial y de escala. Estas características resultan especialmente relevantes y contribuyen significativamente a la comprensión y análisis profundo de las señales de estudio. (Pérez, y otros, 2020).

Cada dominio tiene componentes distintivos y un enfoque específico para una función particular. La incorporación de características en tiempo, frecuencia y escala acelera el proceso, permite un estudio más detallado, un análisis más completo de las características y una mejor comprensión, lo que mejora los resultados de rendimiento del aprendizaje automático.

En la Tabla 1 se presenta de manera detallada el conjunto completo de 84 características que se calculan para cada microsismo en los tres dominios mencionados. Estas características se obtienen al aplicar fórmulas relacionadas con el trabajo de investigación de (Pérez, y otros, 2020).

El operador $c_M\{r_M\}$ presenta la aplicación de la PSD Welch a las señales r_M. Este resultado se almacena en una matriz C cómo se indica en la ecuación (5), donde c_M representa la extracción de las 84 características por señal.

$$C = \{c_1^T, c_2^T, c_3^T, \dots, c_M^T\}^T . \tag{5}$$

Tabla 1

84 Características entre tiempo, frecuencia y escala

Dominio del Tiempo		**Dominio de Frecuencia**		**Dominio de Escala**	
ID	*Características*	*ID*	*Características*	*ID*	*Características*
f1	Media	f29	Potencia RMS	f57	Porcentaje de energía para D1
f2	Desviación Estándar	f30	Densidad de picos en RMS.	f58	Porcentaje de energía para D2
f3	Varianza	f31	2do Pico con valor más alto	f59	Porcentaje de energía para D3
f4	Entropía	f32	Frec. de 2do pico más alto	f60	Porcentaje de energía para D4
f5	Kurtosis	f33	3er Pico con valor más alto	f61	Porcentaje de energía para D5
f6	Entropía Multiescala (MSE).	f34	Frec. de 3er pico más alto	f62	Porcentaje de energía para D6
f7	Tiempo hasta máximo pico	**Dominio de Escala**		f63	A6 RMS en tiempo-dominio
f8	Valor RMS	f35	A6 Max. pico en frec-dominio	f64	D1 RMS en tiempo-dominio
f9	Valor pico a pico.	f36	D1 Max. pico en frec-dominio	f65	D2 RMS en tiempo-dominio
f10	Pico a RMS porción	f37	D2 Max. pico en frec-dominio	f66	D3 RMS en tiempo-dominio
f11	Energía	f38	D3 Max. pico en frec-dominio	f67	D4 RMS en tiempo-dominio
f12	Ratio del cero	f39	D4 Max. pico en frec-dominio	f68	5 RMS en tiempo-dominio
f13	Density de picos above RMS	f40	D5 Max. pico en frec-dominio	f69	6 RMS en tiempo dominio
Dominio de Frecuencia		f41	D6 Max. Pico en frec-dominio	f70	A6 Pico a pico en tiempo-dominio
f14	Frecuencia de máximo pico.	F42	A6 Frec. De max. Pico	f71	D1 Pico a pico en tiempo-dominio
f15	Bandwidth de 90% energía	f43	D2 Frec. De max. Pico	f72	D2 Pico a pico en tiempo-dominio
f16	Entropía	f44	D3 Frec. De max. Pico	f73	D3 Pico a pico en tiempo-dominio
f17	Media	f45	D4 Frec. De max. Pico	f74	D4 Pico a pico en tiempo-dominio
f18	Desviación Estándar	f46	D5 Frec. De max. Pico	f75	Porcentaje de energía para A6
f19	Varianza	f47	D6 Frec. De max. Pico	f76	D6 Pico a pico en tiempo-dominio

Dominio del Tiempo		Dominio de Frecuencia		Dominio de Escala	
ID	*Características*	*ID*	*Características*	*ID*	*Características*
f20	Energía	f48	A6 Media en frec.-dominio	f77	A6 Pico a RMS porción en tiempo-dominio
f21	Kurtosis	f49	D1 Media en frec.-dominio	f78	D1 Pico a RMS porción en tiempo-dominio
f22	Entropía Multiescala	f50	D2 Media en frec.-dominio	f79	D2 Pico a RMS porción en tiempo-dominio
f23	Máximo pico en 10-20 Hz banda	f51	D3 Media en frec.-dominio	f80	D3 Pico a RMS porción en tiempo-dominio
f24	Frec. De max. Pico en 10-20 Hz banda	f52	D4 Media en frec.-dominio	f81	D4 Pico a RMS porción en tiempo-dominio
f25	Máximo pico en 20-30 Hz banda	f53	D5 Media en frec.-dominio	f82	D5 Pico a RMS porción en tiempo-dominio
f26	Frec. De max. Pico en 20-30 Hz banda	f54	D6 Media en frec.-dominio	f83	D6 Pico a RMS porción en tiempo-dominio
f27	Valor RMS	f55	Media energía de componentes	f84	Media energía de wavelet coeficientes
f28	Pico a RMS porción	f56	Porcentaje de energía para A6		

Nota. Características obtenidas de (Pérez, y otros, 2020).

Normalización

Después de completar los procesos de extracción de características tanto para la PSD Welch como las 84 Características. Es necesario implementar la normalización de los datos, puesto que es un proceso que tiene como objetivo transformar datos complejos en un conjunto de datos más simples, estables y compactos (MySQL, 2003). Este proceso se realiza para evitar redundancias y anomalías en los datos. Además, que el uso de esta técnica evita la creación de dependencias y relaciones innecesarias, lo que resulta en una base de datos más limpia, de menor tamaño y más sencilla, lo que acelera los tiempos de procesamiento en el aprendizaje automático (Brownlee , 2020).

La normalización empleada consiste en escalar todos los datos de forma lineal dentro de un rango de 0 a 1. La fórmula utilizada para llevar a cabo este proceso se muestra en la ecuación (6). Este proceso se aplica de forma iterativa a todas las características de la PSD Welch y 84 Características de todos los microsismos.

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{6}$$

donde:

- z. - representa el resultado normalizado de los datos.
- x. - corresponde a los datos que se desean normalizar.
- *min(x)*. - es el valor mínimo presente en los datos
- *max(x)*. - representa el valor máximo observado en los datos.

Los operadores $z1_M\{p_M\}$ y $z2_M\{c_M\}$ presentan matemáticamente la aplicación de la normalización para la extracción de la PSD Welch y las 84 características. Los resultados obtenidos se almacenan en las matrices $Z1$ y $Z2$ cómo se describe en las ecuaciones (7) y (8) correspondientemente. Aquí, $z1_M$ y $z2_M$ denotan la normalización de cada señal respectivamente.

$$Z1 = \{z1_1^T, z1_2^T, z1_3^T, \dots, z1_M^T\}^T, \tag{7}$$

$$Z2 = \{z2_1^T, z2_2^T, z2_3^T, \dots, z2_M^T\}^T. \tag{8}$$

RFE por el método de envoltura

Este es un complemento útil en el entrenamiento de modelos de aprendizaje supervisado, especialmente en el caso del algoritmo SVM. Permite seleccionar las características más relevantes de un conjunto de datos grande, lo cual reduce el tamaño de la base de datos y evita que las características adicionales introduzcan ruido en el proceso de entrenamiento del modelo. Esto tiene como objetivo mejorar el rendimiento del modelo entrenado. Algunas ventajas de aplicar este procedimiento son las siguientes:

- Reducción de sobreajuste.

- Menor costo computacional.
- Modelos de entrenamiento más simples.

El método de envoltura utiliza el rendimiento de un algoritmo de ML como criterio de evaluación para la reducción de características. Como se muestra en la Figura 7, este método sigue un procedimiento iterativo. La selección tiene un subconjunto inicial de características y el entrenamiento de un modelo de ML. Luego, se busca la característica más adecuada para el modelo y se genera un nuevo subconjunto de características. Este proceso se repite al volver a entrenar el modelo con el nuevo subconjunto, hasta que no se observen mejoras adicionales en el rendimiento del modelo.

Figura 7

Procedimiento realizado en método de envoltura

Nota. Fuente (aprendeIA, 2019).

Algunos de los métodos de envoltura son:

- *Selección hacia adelante.* - comienza con un conjunto vacío de características y se agrega una característica a la vez.
- *Eliminación hacia atrás.* – comienza con todas las características disponible y se elimina una característica a la vez.
- *Eliminación de características recursivas*. – comienza con todas las características y, en cada paso, se evalúa la importancia de estas al eliminar la menos relevante.

Para este trabajo se seleccionó el método de selección hacia adelante, donde, se utilizó un subconjunto aleatorio de la base de datos que contiene una matriz de 1000 señales, cada una compuesta de 250 señales por microsismo con un tamaño de 257 características para PSD Welch y las 84 características en tiempo, frecuencia y escala. Durante cada iteración realizada por este método y por separado para cada base de datos, se obtiene un subconjunto que influye en la selección de características, lo que resulta en un número variable de características seleccionadas.

Para obtener una visión más completa de las características determinadas por este algoritmo se utiliza el conjunto de datos completo, se realiza este método de selección de forma iterativa. Realizar este método de selección en varias ocasiones posibilita la identificación de características recurrentes con mayor frecuencia, lo que a su vez facilita tener una compresión más precisa de su relevancia.

Bases de datos a entrenar

Al finalizar la extracción y selección de características de los anteriores procesos se obtienen las siguientes bases de datos a entrenar:

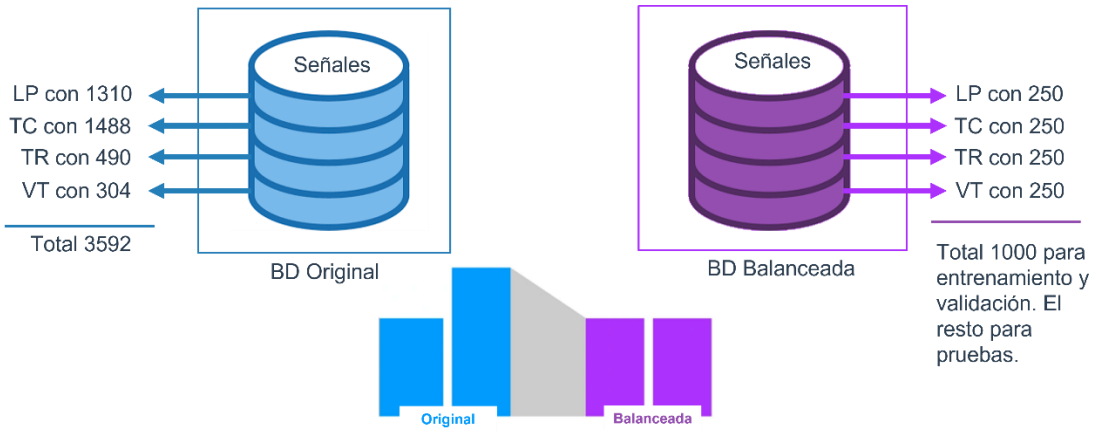
- *PSD Welch.* – 257 características.
- *PSD Welch RFE*. – número de características seleccionadas.
- *84 características*. – 13 tiempo, 21 frecuencia y 50 escala.
- *84 características RFE.* – número de características seleccionadas entre tiempo, frecuencia y escala.
- *Combinación entre PSD Welch y 84 Características.*
- *Combinación entre PSD Welch RFE y 84 Características RFE*. – número de características seleccionadas.

Balanceo de la base de datos

Al balancear la base de datos se garantiza que un modelo de ML pueda aprender patrones de todas las clases de una manera homogénea y efectiva, para así replicar predicciones precisas. Por tanto, acorde a la Figura 8, las bases de datos a emplear para el entrenamiento de los diferentes algoritmos de ML, se las balancea al observar el valor mínimo que se presenta entre los cuatro microsismos, para este caso el VT con 304 señales, se opta por seleccionar 250 señales por microsismo , es decir, 1 000 señales en total con el propósito de determinar el porcentaje adecuado para el entrenamiento y la validación de las diferentes bases de datos. Las señales restantes, que no se destinan al entrenamiento ni a la validación, se almacenan en una tabla aparte. Estas serán utilizadas posteriormente en el proceso de pruebas de los diversos modelos de clasificación generados.

Figura 8

Balanceo de datos



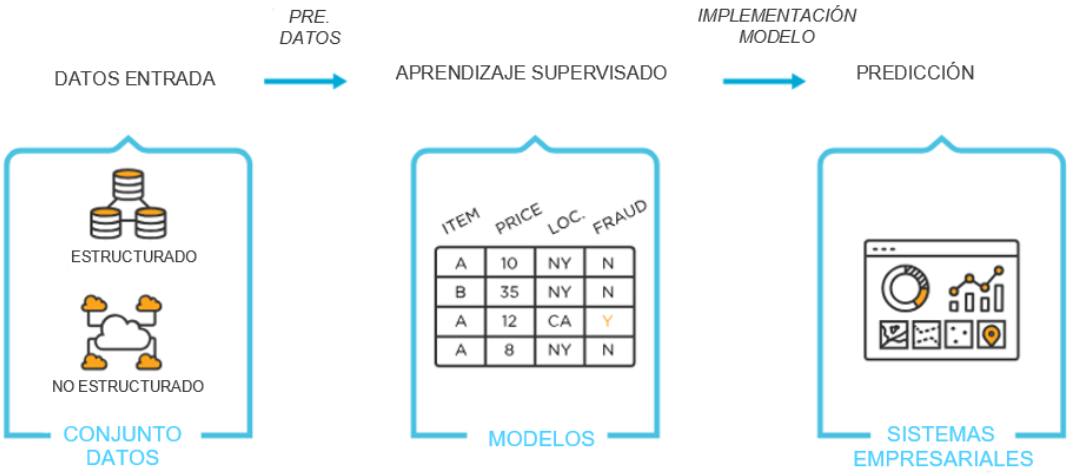
Para determinar el porcentaje de entrenamiento y validación que maximice los resultados de clasificación en pruebas, se realiza un barrido con diferentes conjuntos de datos, es decir, los valores de 50%-50%, 60%-40%, 70%-30%, 80%-20% y 90%-10% para entrenamiento y validación respectivamente.

Entrenamiento de algoritmos de ML

ML utiliza algoritmos que aprenden de manera iterativa a partir de pares de datos de entrada y salida. Esto permite a computadores encontrar información relevante sin la necesidad de una programación previa (TIBCO, 2023). Este enfoque es útil para resolver problemas de regresión lineal y clasificación, ya que proporciona resultados basados en el modelo utilizado, como se muestra en la Figura 9. Un modelo de regresión lineal se utiliza para analizar las relaciones entre dos variables continuas cuantitativas, mientras que un modelo de clasificación se utiliza para predecir una etiqueta o categoría en particular.

Figura 9

Proceso realizado en aprendizaje de máquina



Nota. Fuente (TIBCO, 2023) modificado

Por tanto, centrándonos en el contexto de clasificación por etiqueta, se generan varios modelos de clasificación. Estos modelos son programas creados a partir de patrones numéricos identificados en los datos de entrenamiento al aplicar diferentes algoritmos de clasificación. En este caso, los algoritmos de clasificación DT, *k*-NN y SVM. Para la implementación para cada uno de estos algoritmos se varia un solo parámetro que afecta directamente al resultado del BER y por ende sus métricas de desempeño.

Por tanto, para obtener el parámetro que permite maximizar los resultados de entrenamiento y prueba, se realiza un barrido entre los valores máximos y mínimos que corresponda al parámetro elegido. Este proceso de entrenamiento es repetitivo, ya que se generan modelos de clasificación y pruebas. Como se muestra en la Figura 1, los datos de entrada corresponden a las bases de datos preprocesadas y procesadas.

Los resultados y análisis de las distintas variaciones de entrenamiento se detallan en el siguiente capítulo, donde se calculan métricas de rendimiento para realizar comparaciones entre los modelos. Estas métricas se relacionan con las matrices de confusión obtenidas en cada prueba de entrenamiento de los distintos algoritmos de ML, estas incluyen A, P, S, R y BER. Los valores de estas métricas se encuentran detalladas después de la explicación de los algoritmos de clasificación a emplear.

Árbol de decisión

El algoritmo de DT se estructura mediante un nodo raíz o nodo superior, como se muestra en la Figura 10. Este nodo se ramifica en dos sub nodos llamados nodos de decisión, los cuales representan la selección de la mejor característica para dividir los datos. Estas ramificaciones continúan hasta llegar a los nodos terminales, que indica la clase a la que pertenece el dato analizado.

Este proceso se logra al aplicar medidas de impurezas, como la entropía o el índice de Gini, definido por la ecuación (9). Estas medidas evalúan la homogeneidad de las clases en cada división y tienden a valores cercanos a cero, la proporción de una de las clases (representada por p) es extremadamente pequeña (InteractiveChaos, 2023).

$$G = 1 - \sum_{k} p_k^2 \,. \tag{9}$$

donde:

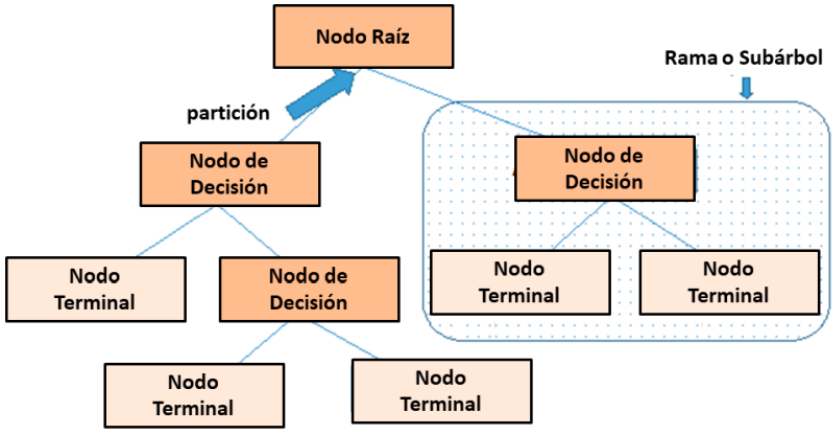
- $G.$ – índice Gini.
- $Pk.$ – probabilidad de una clase.
- $k.$ – números de términos.

Este algoritmo se caracteriza por la división de datos de entrenamiento conjuntos homogéneos basados en los valores más significativos de las variables de entrada (Altamirano, 2021). Una de las ventajas de este algoritmo es su capacidad de manejar tanto datos categóricos como numéricos, así como datos grandes y complejos. Además, los DT son resistentes al ruido y a los valores atípicos.

Sin embargo, los DT tienen la tendencia de sobre ajustarse, especialmente al ser profundos y se adaptan en exceso a los datos de entrenamiento. Esta limitación puede resultar en un rendimiento deficiente en datos nuevos o no vistos anteriormente.

Figura 10

Estructura de un árbol de decisión



Nota. Fuente (Arana, 2021)

El presente trabajo, se entrena mediante la variación de un parámetro en específico que afecta su estructura y rendimiento. Para este caso se varía el Número Máximo de divisiones (MaxNumSplits,

del inglés *Maximum Number of Splits*), el cual determina el número máximo de divisiones del árbol en sus nodos internos. Mientras que los demás parámetros configurables del algoritmo se los mantiene por defecto. Por tanto, los parámetros quedan determinados a continuación:

Parámetros de entrenamiento DT:

- *MaxNumSplits*. – Número máximo de divisiones en este caso se varía entre 1 a 100.
- *MinLeafSize*. – Tamaño mínimo de una hoja del árbol, por defecto 1.
- *MinParentSize*. - tamaño mínimo de un nodo padre antes de que realice una división, por defecto 10.
- *SplitCriterion*. – especifica el criterio para dividir un nodo, por defecto GDI (del inglés Gini's Diversity Index).
- *PruneCriterion*. – indica el criterio para la poda del árbol, por defecto error.
- *PruneAlpha*. – es el factor para controlar la poda, por defecto 0.05.
- *MergeLeaves*. – determina si se deben fusionar las hojas del árbol, por defecto off, es decir, esta desactivado en este caso.

k Vecino más Cercano

El algoritmo de *k*-NN se basa en memorizar los puntos de datos de entrenamiento y luego aprende al utilizar los datos de prueba, asigna una etiqueta a cada dato en función de las etiquetas de sus *k* vecinos más cercanos. El valor de *k* determina cuántos vecinos se consideran para llevar a cabo la clasificación. Este algoritmo se destaca por su simplicidad, alta precisión y resistencia a valores atípicos (Altamirano, 2021). Sin embargo, debido a que se utiliza toda la base de datos de datos durante el entrenamiento, requiere una mayor capacidad de memoria, lo que lo hace más adecuado para conjunto de datos pequeños con pocas características.

El cálculo de la distancia euclidiana, que sirve como base para este algoritmo, se describe mediante la ecuación (10). Esta ecuación se utiliza para determinar la distancia entre el dato que se desea clasificar y los *k* vecinos más cercanos.

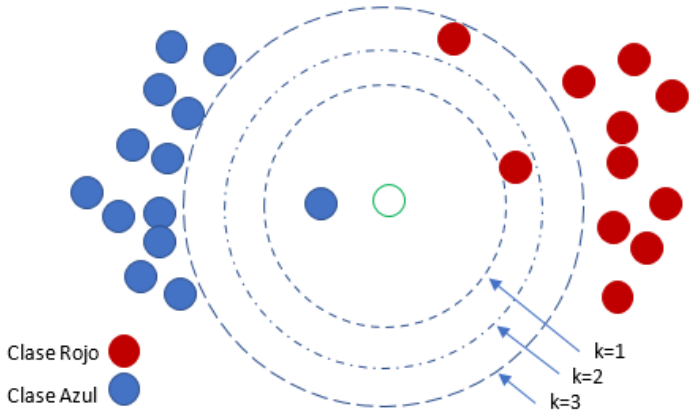
$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_2)^2}\,. \tag{10}$$

donde:

- $D.$ – distancia euclidiana.
- $x, y.$– coordenadas cartesianas del punto de interés.

Figura 11

Ejemplo de clasificación algoritmo k-NN



Nota. Realización propia del autor

La Figura 11, se ilustra un ejemplo donde el algoritmo de *k*-NN toma una decisión sobre la etiqueta asignada a un dato desconocido, se utiliza un valor de *k*=1, el algoritmo clasifica el elemento como color azul, ya que su vecino más cercano pertenece a esa clase. Con *k* =2, el algoritmo requiere un criterio para la clasificación, ya que hay un vecino por cada color. Sin embargo, al utilizar *k*=3, se determina que la clase correspondiente es de color rojo al seguir el criterio de voto mayoritario.

En el contexto de este trabajo, se realiza el barrido del número de *k* vecinos, este define la cantidad de vecinos más cercanos a considerar para la clasificación. Por tanto, el parámetro influye directamente en el comportamiento y rendimiento de los modelos generados. Los restantes parámetros de ajuste del modelo se los establece por defecto.

Parámetros de entrenamiento *k*-NN:

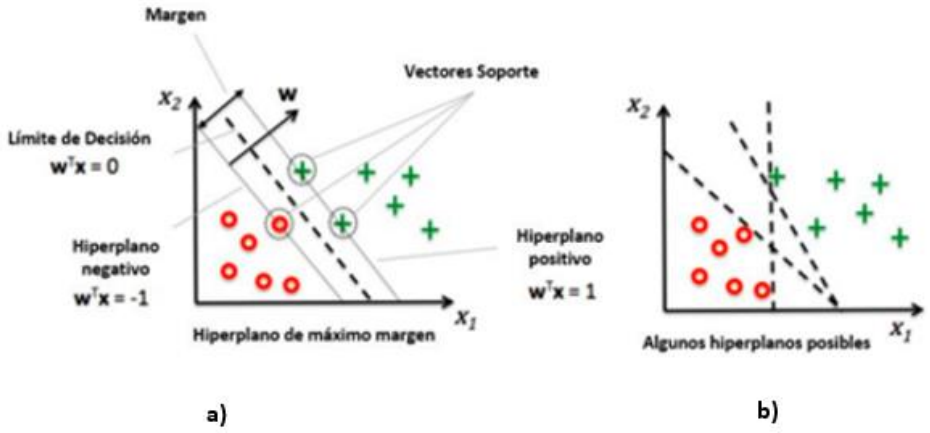
- *NumNeighbors*. – Numero de vecinos, se varia en el rango de 1 a 100.
- *Distance*. – Define la métrica de distancia que se empleará para comparar las muestras, por defecto distancia Euclidiana.
- *NSMethod*. – Es el método para resolver empates, es decir, al variar vecinos se tienen la misma cantidad de variables. Por defecto se emplea el método exhaustivo (del inglés *Exhaustive*).
- *DistanceWeight*. - determina el peso de la distancia en la votación. Por defecto igual (del inglés *Equal*).
- *Standardize*. – establece si los datos deben estandarizarse, en este caso desactivo ya que los datos están normalizados.

Máquina de Vectores de Soporte

Este algoritmo SVM utiliza hiperplanos en un espacio N-dimensional, donde N representa el número de variables independientes. Esta característica permite que el algoritmo sea preciso en la clasificación y atribuye un microsismo a cada lado del hiperplano. De esta manera, SVM es capaz de separar y clasificar eficientemente conjuntos de datos complejos, (Maisueche, 2019).

Figura 12

Representación gráfica de SVM



Nota. a) determinación del margen a partir del mejor hiperplano escogido, así como sus diferentes elementos. b) Representación de posibles hiperplanos para clasificación de dos microsismos. Fuente: (Maisueche, 2019).

La Figura 12a indica los vectores de soporte, que son aquellos puntos más cercanos al hiperplano y que influyen en su posición y orientación. Por otro lado, la Figura 12b representa los posibles hiperplanos existentes para esta clasificación. En este caso, se busca maximizar el margen, que es la distancia entre dos líneas paralelas situadas simétricamente a cada lado del límite de decisión. Un margen mayor reduce el error de generalización, mientras que un margen menor hace que los modelos entrenados sean más propensos al sobre entrenamiento (del inglés *overfitting*).

En relación con el presente trabajo, el proceso de entrenamiento se lleva a cabo al variar el parámetro *C* o restricción de caja, este controla la penalización de los errores en la clasificación durante el entrenamiento. En otras palabras, este controla el equilibrio entre la precisión y la capacidad de generalización. De la misma manera que en los anteriores algoritmos los restantes parámetros se los deja por defecto.

Parámetros de entrenamiento SVM:

- *BoxConstraint*. – Restricción de caja, este varía entre 0.1 a 100.
- *Coding*. – Enfoque de codificación para manejar problemas de multiclase. Por defecto uno versus uno (del inglés *one-vs-one*).
- *Cost*. – Es una matriz de valores que tiene costo asignado a las clasificaciones incorrectas. En este caso no se la utiliza.
- *FitPosterior*. – Se ajusta las probabilidades a posteriori de la clasificación. Por defecto desactivado.

Métricas de desempeño

Después de entrenar los diferentes algoritmos de clasificación, es crucial evaluar su rendimiento al utilizar las métricas de desempeño como la exactitud, precisión, sensibilidad o *recall*, especificidad y BER. Para calcular estas métricas, se utiliza la matriz de confusión generada para cada modelo. Esta matriz muestra los recuentos de valores predichos y reales, como se muestra en la Tabla 2. A partir de esta matriz, se pueden obtener métricas clave que proporcionan información sobre el rendimiento y la calidad de la clasificación realizada por cada algoritmo. Las columnas representan las predicciones realizadas por el modelo, mientras que las filas representan las observaciones de los datos reales.

Tabla 2

Matriz de confusión

		Predicción	
		Positivo	*Negativo*
Observación	*Positivo*	Verdadero Positivo – VP	Falso Negativo – FN
	Negativo	Falso Positivo – FP	Verdadero Negativo – VN

- VP – Verdadero Positivo. – indica la correcta clasificación de una instancia como positiva.
- VN – Verdadero Negativo. – indica la correcta clasificación de una instancia como negativa.
- FP – Falso Positivo. - indica la incorrecta clasificación de una instancia como positiva al ser en realidad negativa.
- FN – Falso Negativo. - indica la incorrecta clasificación de una instancia como negativa al ser en realidad positiva.

Las métricas de desempeño se las define de la siguiente forma:

Exactitud

Métrica que brinda una medida general del rendimiento del modelo al evaluar su capacidad para clasificar correctamente tantos los casos positivos como negativos entre el total de casos, como se muestra en la ecuación (11). Esta métrica permite tener una visión general de la precisión del modelo en la clasificación global de los datos.

$$A(\%) = \frac{VP + VN}{VP + FP + FN + VN} \times 100 \tag{11}$$

Precisión

Métrica que evalúa la capacidad del modelo para clasificar correctamente los casos positivos. Se calcula al dividir los verdaderos positivos VP y los falsos positivos, como se indica en la ecuación (12). Esta métrica proporciona información sobre la proporción de predicciones positivas que son realmente correctas, lo que es útil para evaluar la precisión del modelo en la identificación de casos positivos.

$$P(\%) = \frac{VP}{VP + FP} \times 100 \tag{12}$$

Especificidad

Métrica utilizada para evaluar la capacidad de un modelo para identificar correctamente los casos negativos. Se calcula al dividir los verdaderos negativos VN entre la suma de verdaderos negativos y falsos positivos FP, como se indica en la ecuación (13).

$$S(\%) = \frac{VN}{VN + FP} \times 100 \tag{13}$$

Recall o Sensibilidad

Métrica que evalúa la capacidad del modelo para identificar correctamente los casos positivos. Se calcula al dividir los verdaderos positivos VP y los falsos negativos FN, como se indica en la ecuación (14).

$$R(\%) = \frac{VP}{VP + FN} \times 100 \tag{14}$$

BER

El BER es una métrica que proporciona una medida promedio de muestras mal clasificadas por cada clase. Se calcula al utilizar la ecuación (15).

$$BER = 1 - \frac{R + S}{200} \tag{15}$$

Generación de modelos

Una vez entrenados los diferentes modelos con las variaciones de las bases de datos propuestas, en este proceso se seleccionan los tres mejores modelos de clasificación, uno por cada algoritmo de clasificación utilizado. Se los guarda para posteriormente ser utilizados en el algoritmo de votación propuesto en este trabajo de titulación.

Prueba de modelos

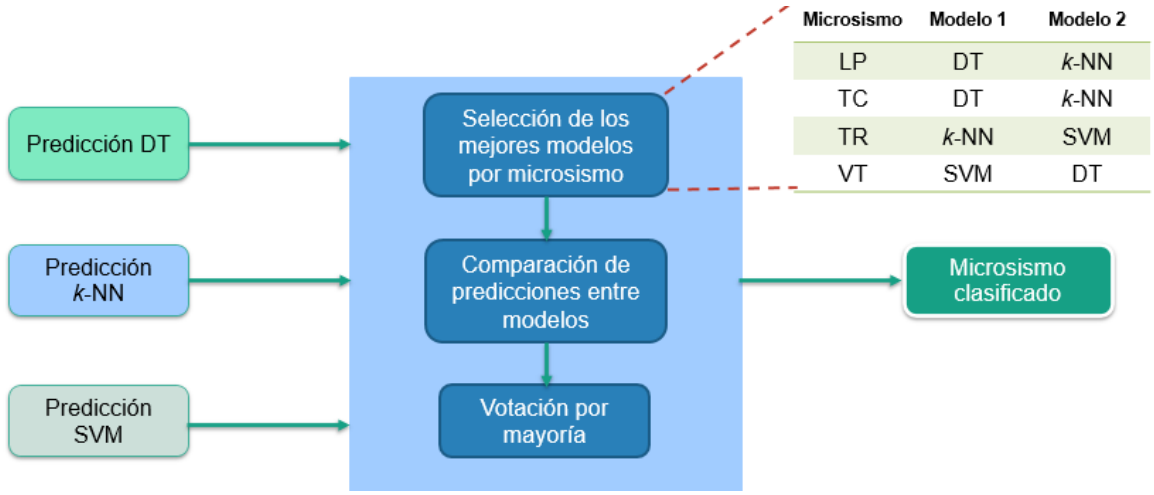
Con el fin de evaluar el rendimiento de los modelos generados, es necesario utilizar nuevos datos que no se han utilizado en el entrenamiento. En esta sección, se emplean los datos de prueba que se muestran en la Figura 8. Estos datos fueron separados al balancear la base de datos y para evaluar el rendimiento del clasificador en las pruebas, se utilizan las mismas métricas de desempeño que se emplearon al analizar las matrices de confusión durante la validación de entrenamiento.

Votación

Se utiliza el sistema de votación para determinar el tipo de microsismos resultante a partir de una nueva señal. Este sistema se basa en tres bloques como se muestra en la Figura 13 con: selección de los mejores modelos por microsismo, comparación de predicciones entre modelos y votación por mayoría.

Figura 13

Sistema de Votación



Selección de los mejores modelos por microsismo

La importancia del modelo se determina de acuerdo con la precisión alcanzada durante el entrenamiento respecto a cada microsismo. Esta se resume a una tabla donde muestra al modelo 1 y

modelo 2 como los dos mejores modelos entrenados para la clasificación de un microsismo en particular. El modelo 1 tiene mayor relevancia en la votación respecto al modelo 2.

Comparación de predicciones entre modelos

La votación se lleva a cabo de acuerdo a la predicción del modelo considerado modelo 1 para un determinado microsismo. En caso de existir dos modelos con la misma prioridad para microsismo diferentes, se toma una decisión de acuerdo al número de errores en la predicción de cada modelo de acuerdo a su tabla de confusión.

Votación por mayoría

Al no poder descartar algún modelo por prioridad de acuerdo con su predicción, se realiza la votación por mayoría, y se entrega como microsismo clasificado aquel con mayor presencia entre los modelos DT, *k*-NN Y SVM.

Capítulo III

Resultados

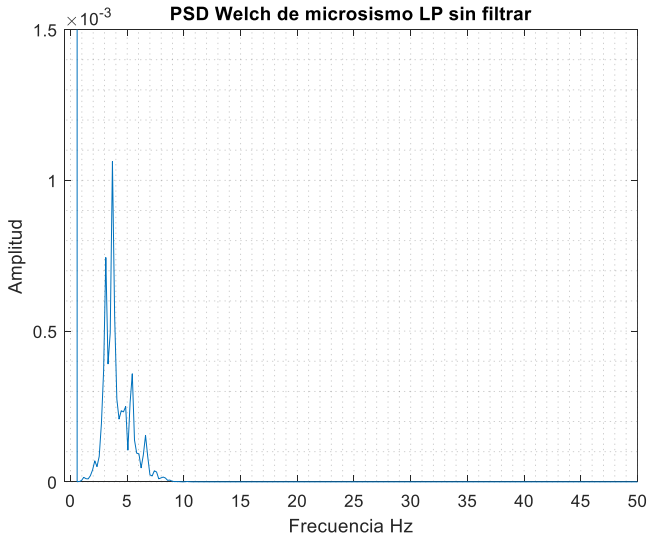
En este capítulo, se presentan los resultados obtenidos a partir del procedimiento descrito en el capítulo anterior. Se proporcionarán detalles sobre las métricas de desempeño obtenidas al utilizar los algoritmos DT, *k*-NN y SVM. Estos algoritmos fueron entrenados al utilizar diferentes parámetros junto con la selección de características para reducir el tamaño del grupo de datos utilizado, como se muestra en la Tabla 5. Se muestran las pruebas realizadas junto a sus resultados para cada modelo, y se indicarán las mejores condiciones identificadas para obtener el mejor modelo entrenado

Preprocesamiento

Filtrado

Figura 14

PSD no filtrada de microsismo LP



Nota. La presencia de un pico en 0.2 Hz es notorio independientemente de la señal de origen.

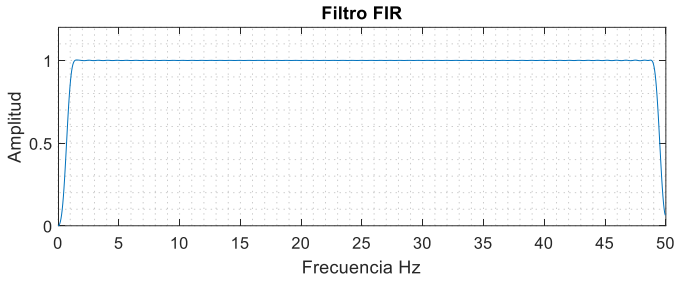
Las señales de cada microsismo se representan gráficamente en el dominio de frecuencia, como se muestra en el ejemplo del microsismo LP en la Figura 14, donde se observa que existe un pico en 0.2

Hz, debido a la presencia del ruido de fondo oceánico. La presencia del pico afecta la visualización de frecuencias distintivas que caracterizan los diferentes microsismos.

Por lo anterior, se aplicó un filtro digital bidireccional de respuesta al impulso finito (FIR) de orden 256 como se presenta en la Figura 15. El filtro se diseñó para suprimir las frecuencias en el rango de 0.7 Hz a 49.5 Hz.

Figura 15

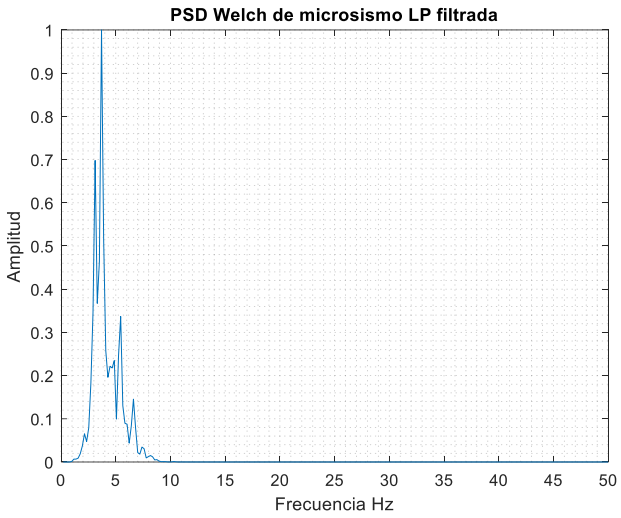
Filtro digital FIR



Como resultado del filtrado, se logró eliminar las componentes por debajo de 1 Hz, como se puede observar en la Figura 16. El filtrado digital bidireccional permitió una mejor representación de los microsismos al eliminar el ruido no deseado y resaltar las características distintivas de interés en el análisis de las señales.

Figura 16

PSD filtrada de microsismo LP



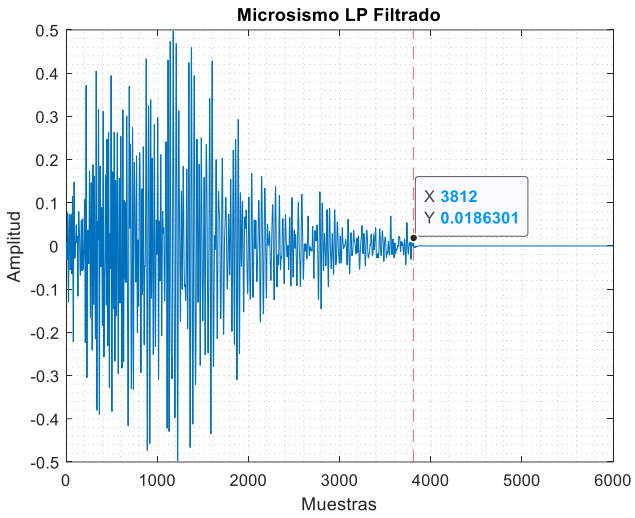
Nota. La eliminación de la componente por debajo de los 2Hz facilita la visualización de las componentes que caracterizan a la señal.

Recorte

La base de datos nos proporciona por cada microsismo 6000 muestras de señal. Pero no todas estas muestras corresponden netamente al microsismo como se muestra en la Figura 17, se representa un microsismo LP filtrado expresado en todas las 6000 muestras en el dominio del tiempo.

Figura 17

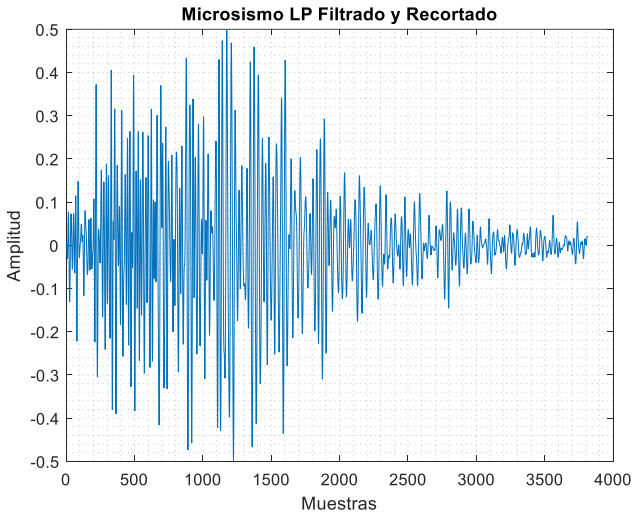
Microsismo LP filtrado en el dominio del tiempo



Dado que la base de datos contiene los puntos de inicio y fin por microsismo, se recorta la señal en la muestra 3812 y se obtiene específicamente el microsismo como se muestra en la Figura 18.

Figura 18

Microsismo LP filtrado y recortado en el dominio del tiempo



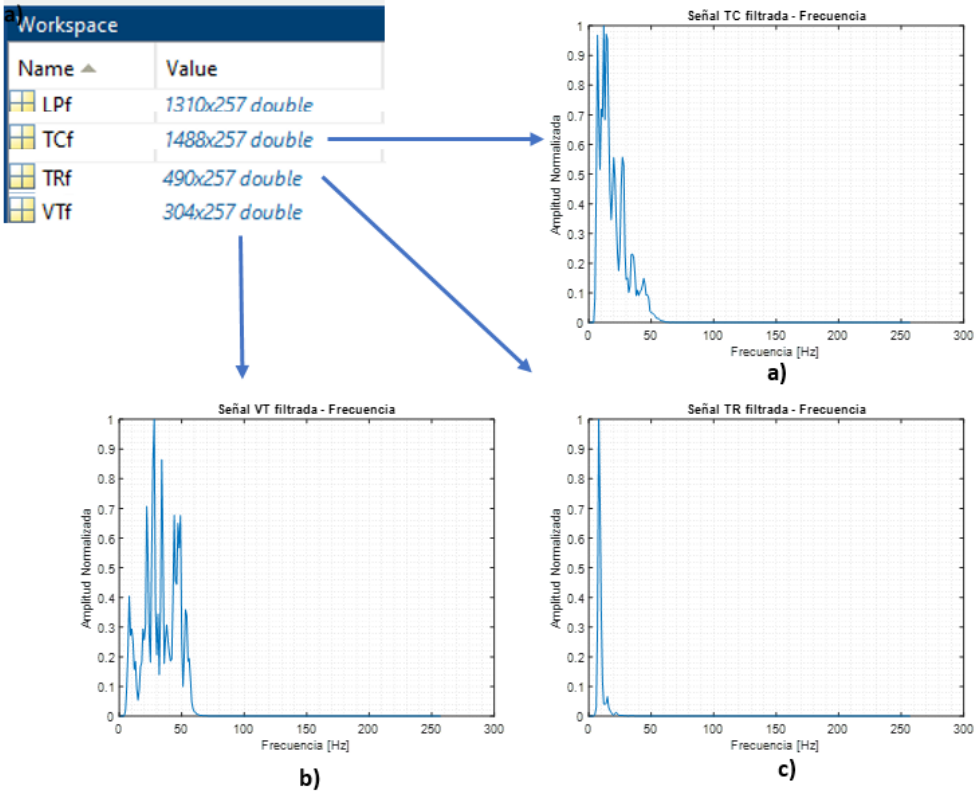
Procesamiento

Densidad espectral de potencia: Método Welch

En la Figura 19, se ilustra el proceso de obtención de la PSD Welch, se utiliza una ventana de *Hamming* con un tamaño de 512 puntos y un solapamiento del 50% como se explicó en el Capítulo II. Este proceso es aplicado a todos los microsismos contenidos en la base de datos. Se presenta la PSD Welch para diferentes microsismos como TC en la Figura 19a, VT en la Figura 19b y TR en la Figura 19c. Además estas señales se las almacena en una matriz.

Figura 19

Obtención PSD Welch de microsismo, con una resolución de 257 puntos



Extracción de 84 características

La Tabla 6 presenta un fragmento de las 84 características mencionadas en la Tabla 1. En esta instancia, se identifican en la columna f1, f2, f3, f4, …, f83 y f84 las distintas características tales como la

media, desviación estándar, varianza y la entropía, entre otras relevantes correspondientes a cada señal de microsismo.

Tabla 3

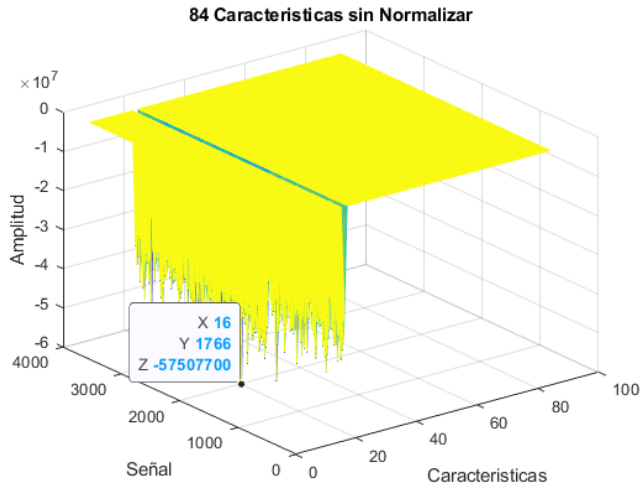
Obtención de 84 características

Señal	**84 características**					
	f1	***f2***	***f3***	***…***	***f83***	***f84***
1	-2.74e-05	0.26	0.07	…	4.49	0.07
2	-8.20e-05	0.19	0.03	…	5.42	0.03
3	3.31e-05	0.21	0.04	…	4.03	0.04
4	2.12e-05	0.20	0.04	…	6.33	0.03
5	5.45e-04	0.24	0.05	…	4.11	0.05
6	3.97e-04	0.17	0.03	…	4.93	0.03
7	5.81e-04	0.19	0.03	…	4.14	0.04
8	1.37e-04	0.20	0.04	…	4.64	0.04
9	2.15e-04	0.25	0.06	…	4.00	0.06
10	2.70e-04	0.22	0.04	…	4.29	0.04
11	4.22e-05	0.20	0.04	…	3.93	0.04
12	9.76e-05	0.26	0.06	…	4.78	0.06
…	…	…	…	…	…	…
3592	-6.28e-0.4	0.18	0.03	…	3.70	0.03

Obtenidas las 84 características se procede a graficarlas como se muestra en la Figura 20 se observa claramente como existe una amplia diferencia de rangos numéricos entre características.

Figura 20

Mesh de 84 características antes de normalizar

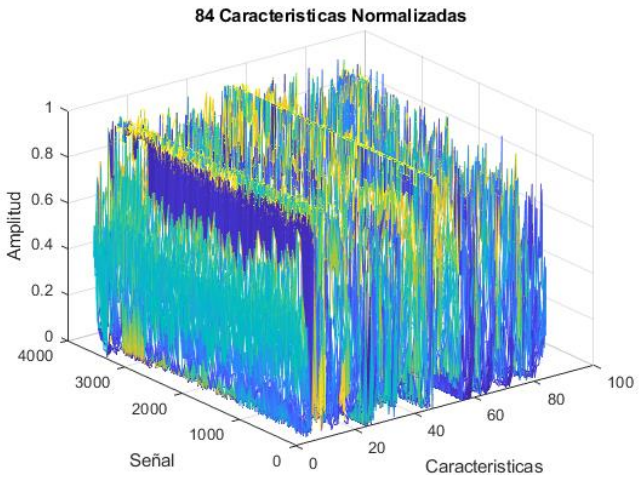


Nota. La diferencia de rangos numéricos es evidente en la característica f16.

Por tanto, las 84 características son normalizadas en el rango de 0 a 1, como se observa en la Figura 21. Se realiza esto para evitar las diferencias de rangos numéricos mencionados que afectan directamente al entrenar los algoritmos de ML.

Figura 21

Mesh de 84 características después de normalizar

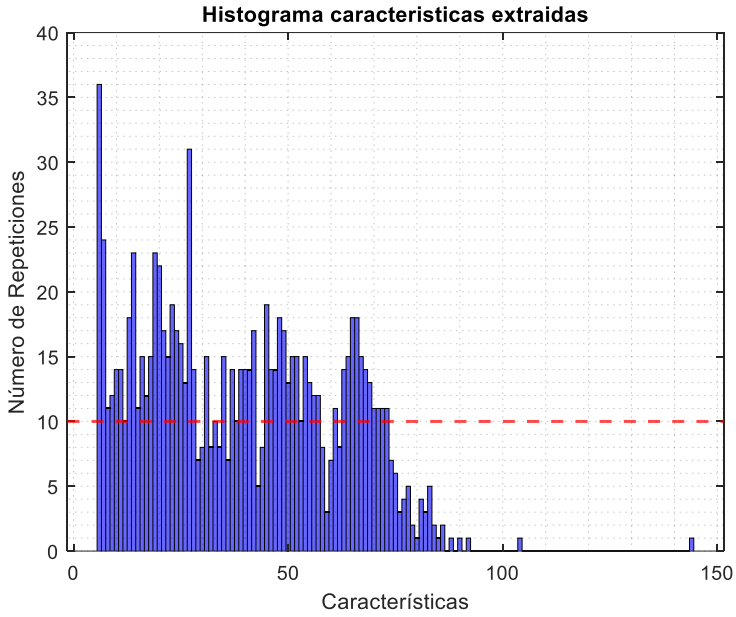


PSD Welch RFE por método de envoltura

El resultado de este análisis se muestra en el histograma de la Figura 22, donde se muestra la frecuencia de aparición de cada característica. Este histograma proporciona una visualización clara de las características más relevantes y su frecuencia relativa en el proceso de selección.

Figura 22

Histograma de características seleccionadas por método de envoltura



Dado que se obtuvieron el número de características que más se repiten. Se realiza un barrido y se selecciona varias de estas características, la selección de las mejores características se basa en los mejores resultados de pruebas obtenidas mas no en entrenamiento, es así que se presenta la Tabla 4, Tabla 5 y Tabla 6, se evalúan las métricas de desempeño de las pruebas de los diferentes modelos generados. Además, se presentan los parámetros de entrenamiento de cada algoritmo.

Selección de características DT.

Dado que en DT se varía el *MaxNumSplits* y sus restantes parámetros configurables se establecen por defecto. En la Tabla 4 presenta para cada selección de características el valor de *MaxNumSplits* que permite obtener el mejor BER.

Tabla 4

Evaluación del barrido PSD Welch RFE con DT

Selección de características	MaxNumSplits	A (%)	P (%)	S (%)	R (%)	BER
Nro. de características: 6	17	79	69	92	79	0.15
Nro. de características: 12	28	77	68	91	80	0.14
Nro. de características: 27	19	80	70	93	85	0.11
Nro. de características: 57	5	84	73	93	82	0.12
Nro. de características: 71	20	80	70	92	83	0.13

***Selección de características k*-NN.**

Dado que en *k*-NN se varía el *NumNeighbors* y sus restantes parámetros configurables se establecen por defecto. La Tabla 5 presenta para cada selección de características el valor de *NumNeighbors* que permite obtener el mejor BER.

Tabla 5

Evaluación del barrido PSD Welch RFE con k-NN

Selección de características	NumNeighbors	A (%)	P (%)	S (%)	R (%)	BER
Nro. de características: 6	6	77	64	92	81	0.14
Nro. de características: 12	6	79	66	93	83	0.12
Nro. de características: 27	9	81	67	93	85	0.11
Nro. de características: 57	9	83	68	94	86	0.10
Nro. de características: 71	5	84	69	94	84	0.11

Selección de características SVM.

Dado que en SVM se varía el *BoxConstraint* y sus restantes parámetros configurables se establecen por defecto. La Tabla 6 presenta para cada selección de características el valor de *BoxConstraint* que permite obtener el mejor BER.

Tabla 6

Evaluación del barrido PSD Welch RFE con SVM

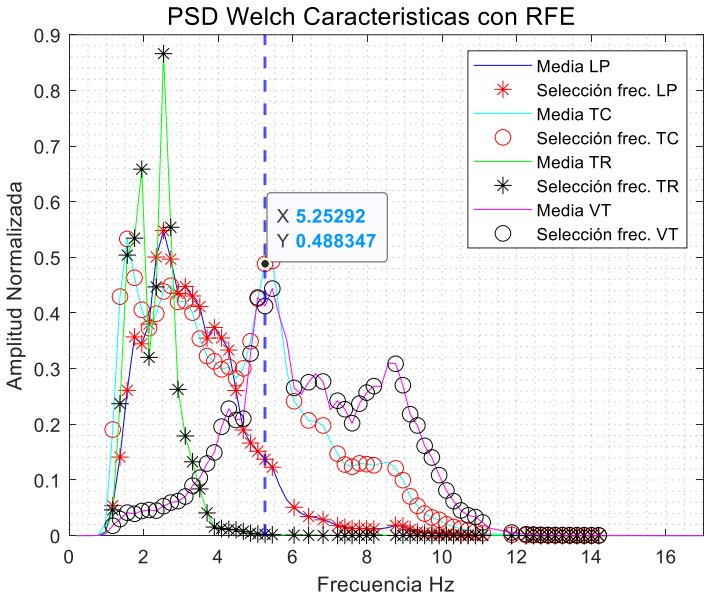
Selección de características	BoxConstraint	A (%)	P (%)	S (%)	R (%)	BER
Nro. de características: 6	0.7	78	65	92	82	0.13
Nro. de características: 12	3	82	69	94	85	0.11
Nro. de características: 27	4	83	70	94	86	0.1
Nro. de características: 57	2	85	72	95	88	0.08
Nro. de características: 71	0.2	79	68	93	85	0.11

Por tanto, dada la evaluación de las tablas anteriores, se establece que el número de características que maximizan los resultados de las métricas de desempeño en prueba corresponden a 57 características RFE, estas corresponden a los valores de: 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 31, 33, 35, 37, 38, 39, 40, 41, 42, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 61, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72 y 73.

Para una representación más completa de las características indicadas por el método de *envoltura*, en la Figura 23 se presenta la media de los microsismos junto con las características vistas en el histograma de la Figura 22.

Figura 23

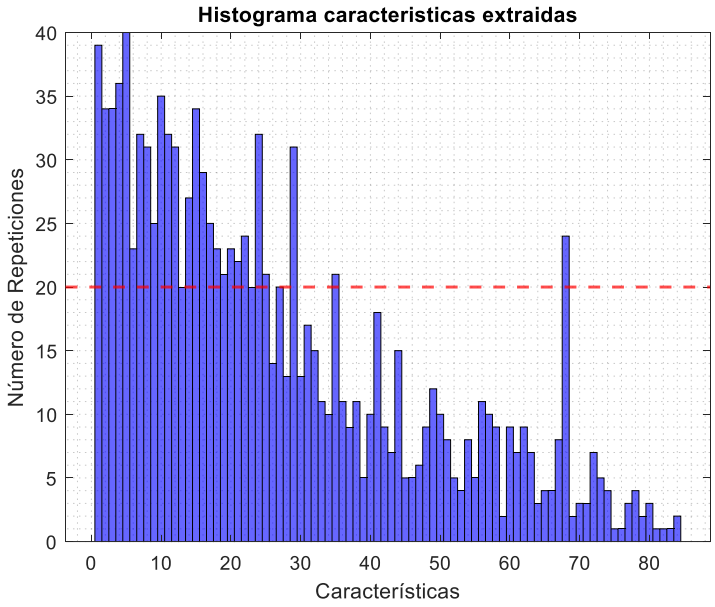
Representación de características sobre señales sismo volcánicas



84 características RFE por método de envoltura

Figura 24

Histograma de características seleccionadas por método de envoltura



De la misma manera que la sección anterior, el histograma de la Figura 24, se muestra la frecuencia de aparición de cada característica.

Dado que se obtuvieron el número de características que más se repiten. Se realiza un barrido y se selecciona varias de estas características, la selección de las mejores características se basa en los mejores resultados de pruebas obtenidas mas no en entrenamiento, es así que se presenta la Tabla 7, Tabla 8 y Tabla 9, se evalúan las métricas de desempeño de las pruebas de los diferentes modelos generados. Además, se presentan los parámetros de entrenamiento de cada algoritmo.

Selección de características DT

De la misma manera que en la sección anterior, la Tabla 7 presenta para cada selección de las 84 características el valor de *MaxNumSplits* que permite obtener el mejor BER.

Tabla 7

Evaluación del barrido 84 características RFE con DT

Selección de las 84 caract.	MaxNumSplits	A (%)	P (%)	S (%)	R (%)	BER
Nro. de características: 7	19	90	77	96	91	0.06
Nro. de características: 17	13	91	78	97	93	0.05
Nro. de características: 29	8	91	77	97	93	0.05
Nro. de características: 33	8	91	77	97	93	0.05
Nro. de características: 45	4	90	77	96	91	0.06

***Selección de características k*-NN.**

De la misma manera que en la sección anterior, la Tabla 8 presenta para cada selección de las 84 características el valor de *NumNeighbors* que permite obtener el mejor BER

Tabla 8

Evaluación del barrido 84 características RFE con k-NN

Selección de las 84 caract.	NumNeighbors	A (%)	P (%)	S (%)	R (%)	BER
Nro. de características: 7	1	89	75	96	91	0.06
Nro. de características: 17	5	92	78	97	93	0.05
Nro. de características: 29	3	94	82	98	94	0.04
Nro. de características: 33	5	94	82	98	95	0.04
Nro. de características: 45	3	94	81	98	94	0.04

Selección de características SVM.

De la misma manera que en la sección anterior, la Tabla 9 presenta para cada selección de las 84 características el valor de *BoxConstraint* que permite obtener el mejor BER.

Tabla 9

Evaluación del barrido 84 características RFE con SVM

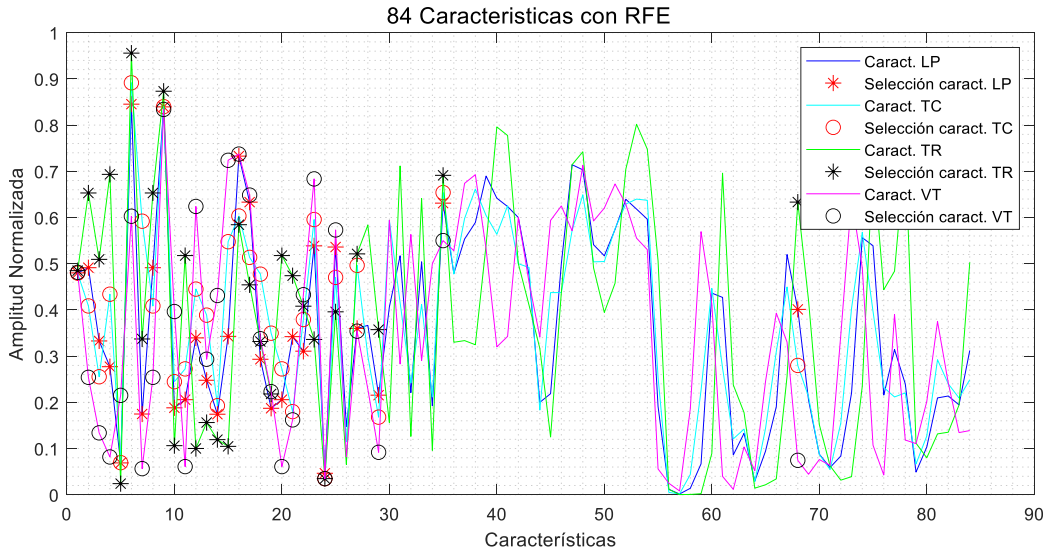
Selección de las 84 caract.	**BoxConstraint**	**A (%)**	**P (%)**	**S (%)**	**R (%)**	**BER**
Nro. de características: 7	2	92	80	97	93	0.05
Nro. de características: 17	7	94	82	98	95	0.03
Nro. de características: 29	6	96	85	99	96	0.03
Nro. de características: 33	2	95	84	99	96	0.03
Nro. de características: 45	2	95	83	98	95	0.03

Por tanto, dada la evaluación de las tablas anteriores, se establece que el número de característica que maximizan los resultados de las métricas de desempeño en prueba corresponden a 29 características RFE, estas corresponden a los valores de f1 al f13 en tiempo, f14 al f25, f27 y f29 en frecuencia. Por último, en escala se tiene el f35 y f68.

Para una representación más completa de las características indicadas por el método de *envoltura*, en la Figura 25 se presenta la media de cada característica separada por microsismo junto con la selección de características vistas en el histograma de la Figura 24.

Figura 25

Representación de características sobre señales sismo volcánicas



Bases de datos a entrenar

Dado los resultados de las anteriores secciones, la Tabla 10 se presenta las bases de datos finales a entrenar en los diferentes algoritmos de aprendizaje supervisado.

Tabla 10

Bases de datos y número de características utilizada para entrenamientos

Base de datos	**Nro. Características**		
	Tiempo	**Frecuencia**	**Escala**
PSD Welch	-	257	-
PSD Welch RFE	-	57	-
84 características	13	21	50
29 características RFE	13	14	2

Base de datos	Nro. Características		
	Tiempo	Frecuencia	Tiempo
Combinación PSD Welch + 84 Características	13	278	-
Combinación PSD Welch RFE + 29 Características RFE	13	71	2

Balanceo de datos

Para poder determinar el correcto balanceo de datos que maximice los resultados de pruebas y mas no en validación. Se empleo la base de datos de PSD Welch RFE ya que este requiere de un menor costo computacional para realizar el entrenamiento y pruebas con diferentes particiones. En la Tabla 11, Tabla 12 y Tabla 13, se muestran el barrido de particiones para entrenamiento y sus respectivas métricas de desempeño para cada algoritmo de aprendizaje implementado.

Tabla 11

Rendimiento de la variación de entrenamiento del modelo DT para PSD Welch RFE

División de datos	A (%)	P (%)	S (%)	R (%)	BER
Train: 50% - Test: 50%	82	83	94	82	0.12
Train: 60% - Test: 40%	83	84	94	83	0.11
Train: 70% - Test: 30%	82	83	94	82	0.12
Train: 80% - Test: 20%	86	86	95	86	0.09
Train: 90% - Test: 10%	86	86	95	86	0.09

Tabla 12

*Rendimiento de la variación de entrenamiento del modelo k-*NN *para PSD Welch RFE*

División de datos	A (%)	P (%)	S (%)	R (%)	BER
Train: 50% - Test: 50%	86	86	95	86	0.09
Train: 60% - Test: 40%	87	87	96	87	0.08
Train: 70% - Test: 30%	88	88	96	88	0.08
Train: 80% - Test: 20%	89	89	96	89	0.07
Train: 90% - Test: 10%	91	91	97	91	0.06

Tabla 13

Rendimiento de la variación de entrenamiento del modelo SVM para PSD Welch RFE

División de datos	A (%)	P (%)	S (%)	R (%)	BER
Train: 50% - Test: 50%	87	88	96	87	0.08
División de datos	**A (%)**	**P (%)**	**S (%)**	**R (%)**	**BER**
Train: 60% - Test: 40%	88	88	96	88	0.08
Train: 70% - Test: 30%	87	87	96	87	0.08
Train: 80% - Test: 20%	87	88	96	87	0.08
Train: 90% - Test: 10%	88	88	96	88	0.08

Puesto que se va a escoger la mejor partición en base a los resultados de las métricas de desempeño en las pruebas de cada modelo. En las Tabla 14, Tabla 15 y Tabla 16, se muestran mencionadas métricas para cada modelo de clasificación propuesto en base a su entrenamiento.

Tabla 14

Rendimiento de la variación de pruebas del modelo DT para PSD Welch RFE

División de datos	A (%)	P (%)	S (%)	R (%)	BER
Train: 50% - Test: 50%	83	73	93	82	0.12
Train: 60% - Test: 40%	81	71	93	84	0.11
Train: 70% - Test: 30%	84	73	93	82	0.12
Train: 80% - Test: 20%	84	73	94	83	0.12
Train: 90% - Test: 10%	84	73	93	82	0.12

Tabla 15

Rendimiento de la variación de pruebas del modelo k-NN *para PSD Welch RFE*

División de datos	A (%)	P (%)	S (%)	R (%)	BER
Train: 50% - Test: 50%	82	69	94	84	0.11
División de datos	**A (%)**	**P (%)**	**S (%)**	**R (%)**	**BER**
Train: 60% - Test: 40%	82	68	94	86	0.10
Train: 70% - Test: 30%	83	68	94	86	0.10
Train: 80% - Test: 20%	84	69	94	85	0.10
Train: 90% - Test: 10%	84	70	94	84	0.11

Tabla 16

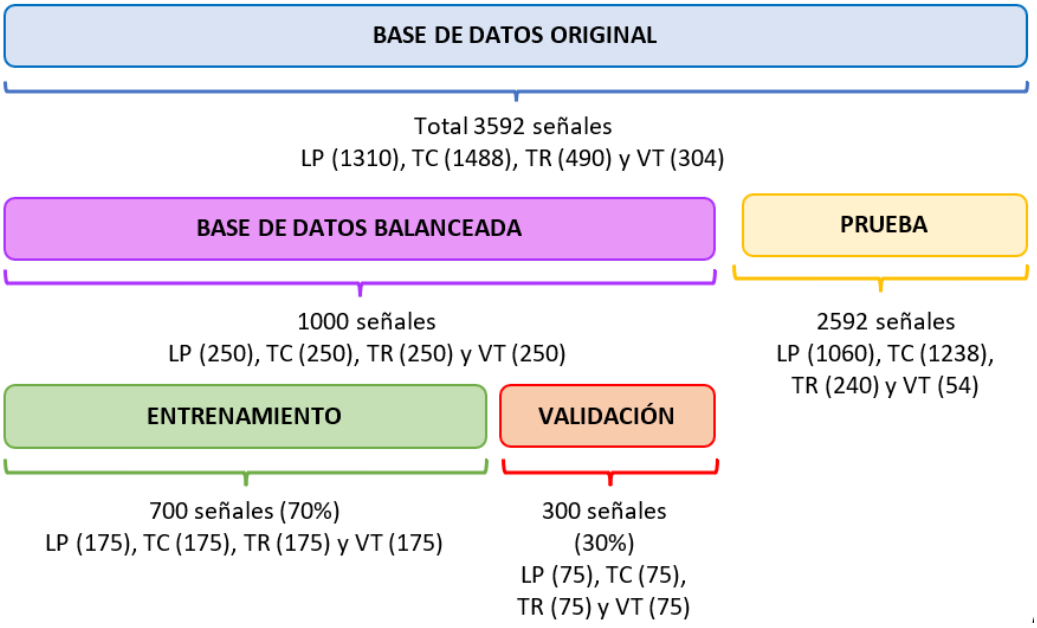
Rendimiento de la variación de pruebas del modelo SVM para PSD Welch RFE

División de datos	A (%)	P (%)	S (%)	R (%)	BER
Train: 50% - Test: 50%	83	70	94	87	0.09
Train: 60% - Test: 40%	81	68	94	86	0.10
Train: 70% - Test: 30%	85	72	95	88	0.08
Train: 80% - Test: 20%	83	70	94	87	0.09
Train: 90% - Test: 10%	85	72	95	88	0.09

Por tanto, se establece que para la división de datos óptima que maximiza los resultados de desempeño en pruebas es de 70-30. Debido a que en estos porcentajes se obtienen métricas de desempeño muy similares a 80-10 y 90-10 pero con la diferencia que en 70-30 se obtiene un número considerable de datos, tanto para entrenamiento del 70% (175 señales por microsismos) y validación del 30% (75 señales por microsismos) como se presenta en la Figura 26, lo cual lo hace óptimo para emplear en los entrenamientos al utilizar un distinto tamaño de la base de datos según el método aplicado.

Figura 26

Balanceo de datos óptimo



Es importante resaltar que el proceso de equilibrio de datos mencionado se aplica de manera uniforme a todas las bases de datos que han sido previamente delineadas en la sección anterior.

Sistema de votación

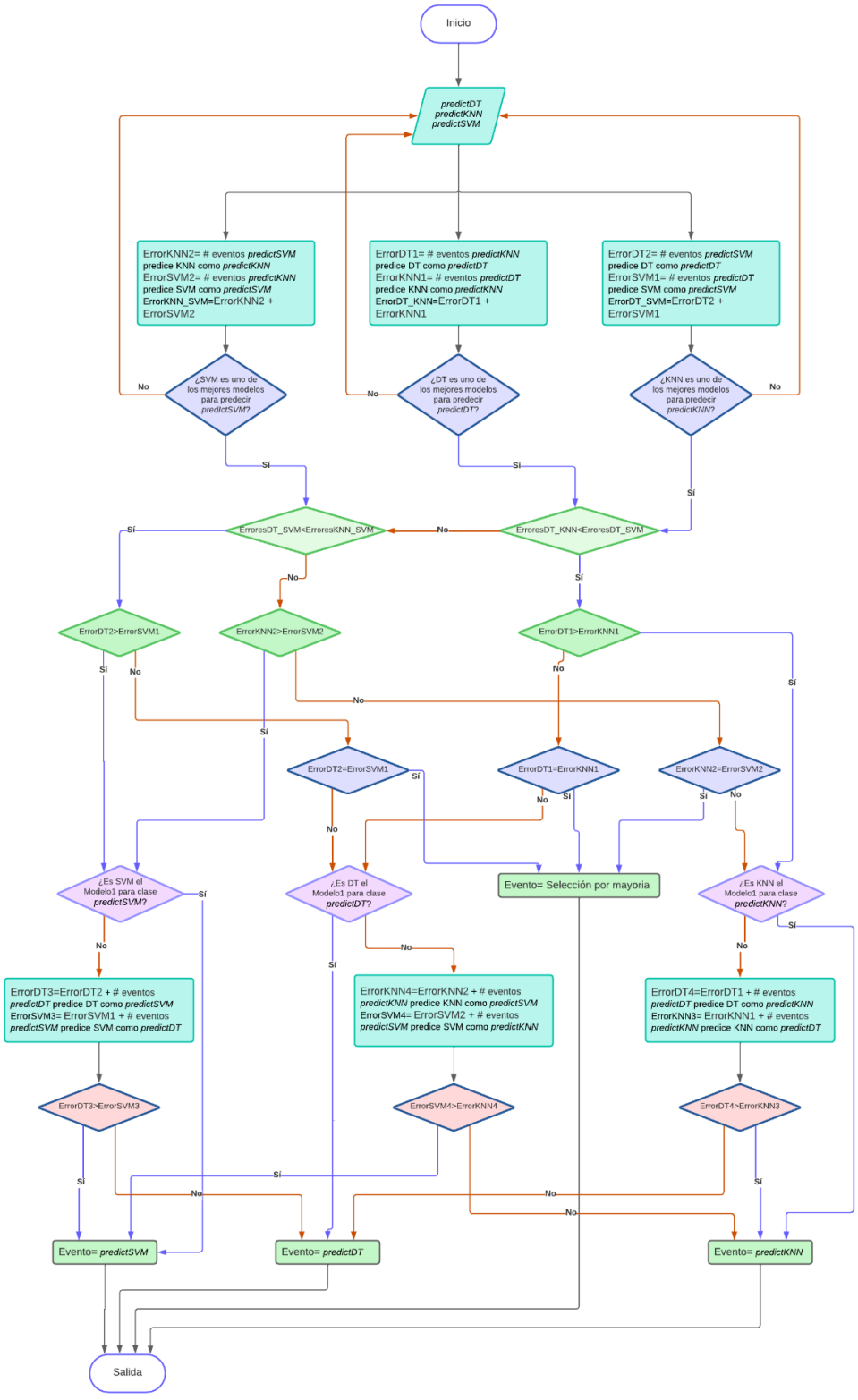
El sistema de votación descrito en el Capítulo II, sigue la secuencia de pasos indicados en el diagrama de flujo mostrado en la Figura 27.

Entrenamiento de algoritmos ML

El balanceo de datos establecido en 70-30 para 250 señales por microsismo en la sección anterior, se replica en las demás bases de datos, por tanto, los resultados de las métricas de desempeño se presentan en Tabla 17, Tabla 19 y Tabla 21 para entrenamiento y la Tabla 18, Tabla 20 y Tabla 22 para pruebas.

Figura 27

Votación entre modelos



Validación y pruebas de DT

Tabla 17

Evaluación de rendimiento en con datos de validación del modelo DT.

Base de Datos	A (%)	P (%)	S (%)	R (%)	BER
PSD Welch	85	85	95	85	0.10
PSD Welch RFE	82	83	94	82	0.12
84 características	91	91	97	91	0.06
29 características RFE	90	91	97	90	0.06
PSD WELCH + 84 características	92	92	97	92	0.05
PSD WELCH RFE + 29 características RFE	93	93	98	93	0.05

Tabla 18

Evaluación de rendimiento con datos de prueba del modelo DT

Base de Datos	A (%)	P (%)	S (%)	R (%)	BER
PSD Welch	80	70	92	83	0.12
PSD Welch RFE	84	73	93	82	0.12
84 características	92	79	97	92	0.05
29 características RFE	91	77	97	93	0.05
PSD WELCH + 84 características	92	79	97	93	0.05
PSD WELCH RFE + 29 características RFE	90	75	97	92	0.05

Validación y pruebas de k-NN

Tabla 19

Evaluación de rendimiento con datos de validación del modelo k-NN

Base de Datos	A (%)	P (%)	S (%)	R (%)	BER
PSD Welch	86	86	95	86	0.09
PSD Welch RFE	88	88	96	88	0.08
84 características	94	94	98	94	0.04
29 características RFE	93	93	98	93	0.05
PSD WELCH + 84 características	93	93	98	93	0.04
PSD WELCH RFE + 29 características RFE	92	93	97	92	0.05

Tabla 20

Evaluación de rendimiento con datos de prueba del modelo k-NN

Base de Datos	A (%)	P (%)	S (%)	R (%)	BER
PSD Welch	84	69	94	84	0.11
PSD Welch RFE	83	68	94	86	0.10
84 características	94	82	98	94	0.04
29 características RFE	94	82	98	94	0.04
PSD WELCH + 84 características	92	80	97	93	0.05
PSD WELCH RFE + 29 características RFE	93	81	97	91	0.06

Validación y prueba de SVM

Tabla 21

Evaluación de rendimiento con datos de validación del modelo SVM

Base de Datos	A (%)	P (%)	S (%)	R (%)	BER
PSD Welch	87	88	96	87	0.08
PSD Welch RFE	87	87	96	87	0.08
84 características	93	94	98	93	0.04
29 características RFE	96	96	99	96	0.02
PSD WELCH + 84 características	90	91	97	90	0.06
PSD WELCH RFE + 29 características RFE	91	91	97	91	0.06

Tabla 22

Evaluación de rendimiento con datos de prueba del modelo SVM

Base de Datos	A (%)	P (%)	S (%)	R (%)	BER
PSD Welch	82	69	94	87	0.10
PSD Welch RFE	85	72	95	88	0.08
84 características	95	84	98	95	0.03
29 características RFE	96	85	99	96	0.03
PSD WELCH + 84 características	86	76	96	89	0.07
PSD WELCH RFE + 29 características RFE	92	79	97	93	0.05

Generación de modelos y pruebas

Dada la sección anterior se seleccionó el mejor conjunto de datos para cada algoritmo de clasificación en pruebas como se presenta en la Tabla 23.

Tabla 23

Resultado de los modelos óptimos obtenidos

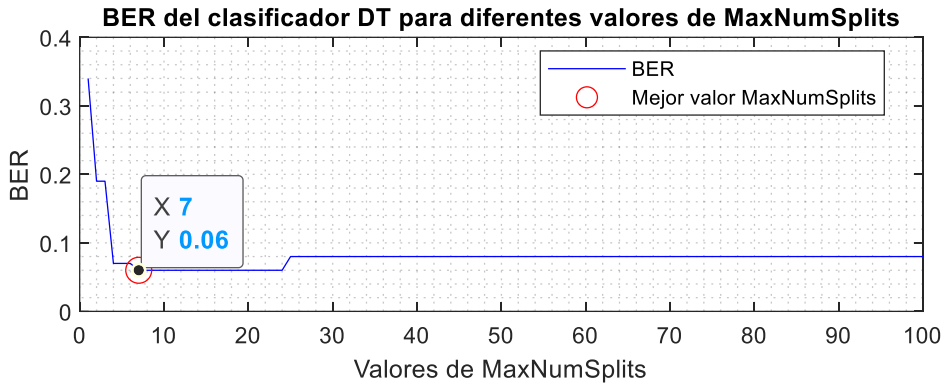
Modelo	**Base de Datos**	**Tiempo (s)**	**A (%)**	**P (%)**	**S (%)**	**R (%)**	**BER**
DT	84 características	0.79	92	79	97	92	0.05
k-NN	29 características RFE	0.88	94	82	98	94	0.04
SVM	29 características RFE	1.28	96	85	99	96	0.03

Cada entrenamiento del algoritmo de aprendizaje automático se la ejecuta en base a la metodología explicada en el Capítulo II. Por tanto, los resultados que se presentan a continuación ilustran la generación de las gráficas de desempeño en base al mejor BER a través de la variación de un parámetro de entrenamiento. Además de la presentación de los respectivos cuadros de confusión de entrenamiento y pruebas para la obtención de las métricas de desempeño que maximizan la clasificación microsismos de los tres mejores modelos obtenidos.

Modelo óptimo de DT

Figura 28

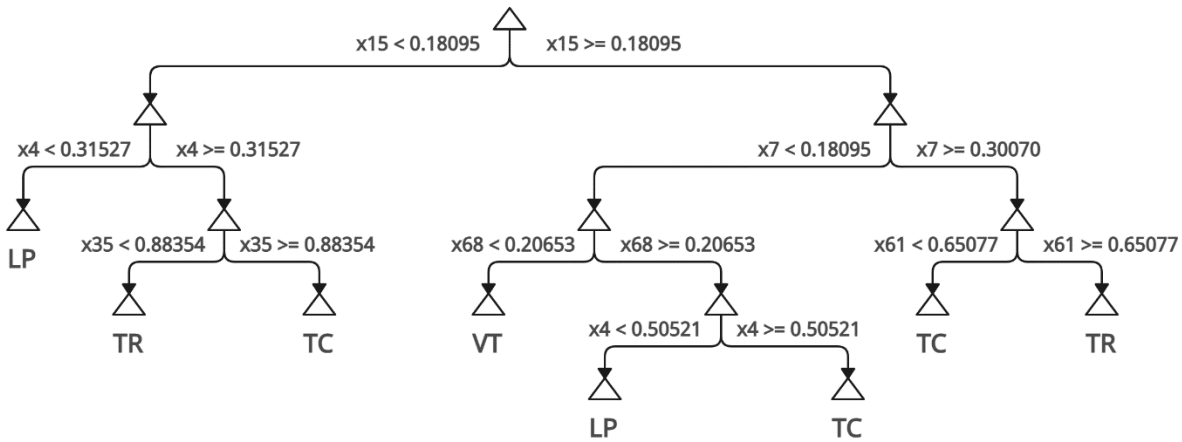
Mejor BER al entrenar el algoritmo DT



En la Figura 28 se presenta la variación del parámetro MaxNumSplits al realizar cada entrenamiento. El mejor BER obtenido de 0.06 corresponde a los parámetros: *MaxNumSplits* en 7, *MinLeafSize* en 1, *MinParentSize* en 10, *SplitCriterion* en *gdi*, *PruneCriterion* en *error*, *PruneAlpha* en 0.05 y *MergeLeaves* en *off*.

Figura 29

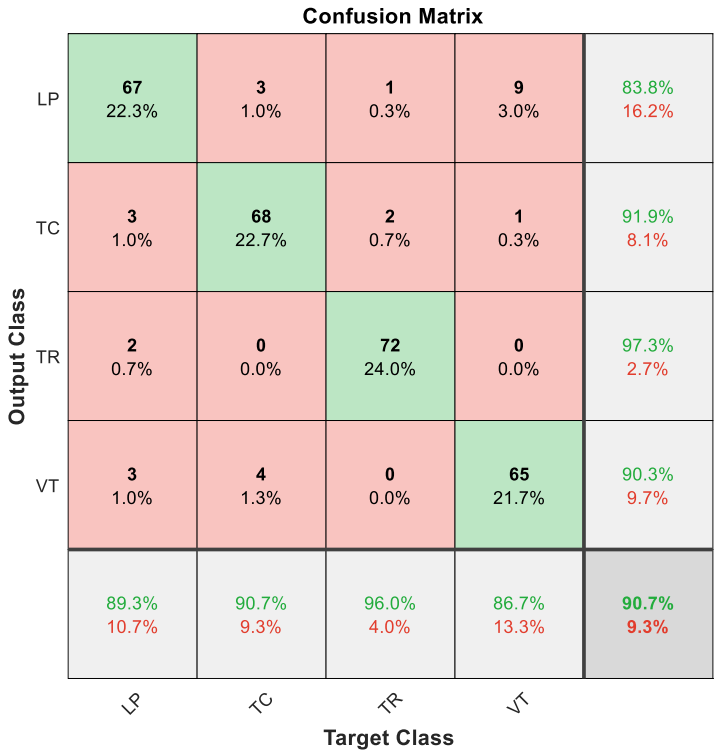
Árbol de decisión del Mejor BER al entrenar el algoritmo DT



Dado que se estableció el número máximo de división del árbol en 7, este valor es corroborado al presentar el árbol de decisión en la Figura 29, este árbol de decisión multiclase representa un modelo de clasificación que consta de 7 divisiones en sus nodos internos y distribuido en 4 niveles. Cada división en el árbol representa una decisión basada en un atributo específico del conjunto de datos de 84 características. Los datos se dividen en función de diferentes características al seguir el flujo de decisiones en los 4 niveles, lo que indica que el árbol no es complejo y no exige una alta capacidad para capturar patrones en los datos.

Figura 30

Cuadro de confusión por microsismo del mejor BER en validación del algoritmo DT



El resultado del entrenamiento del algoritmo DT se presenta en la matriz de confusión, reflejada en la Figura 30. A través de esta matriz, se puede observar que el mayor número de equivocaciones se presentan al clasificar microsismos VT como si fueran LP con un total de 9 señales, seguida por la

predicción de TC como si fuesen VT con un total de 4 señales. Esta evaluación pone de manifiesto que el modelo presenta inconvenientes en la predicción precisa de señales VT.

Tabla 24

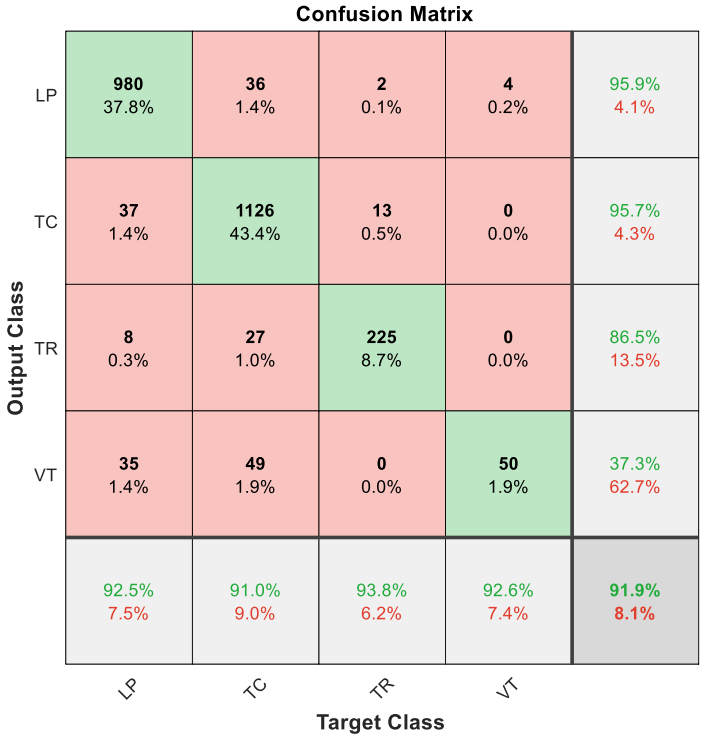
Evaluación de rendimiento en validación por microsismo del modelo DT

Microsismo	A (%)	P (%)	S (%)	R (%)	BER
LP	93	84	94	89	0.08
TC	96	92	97	91	0.06
TR	98	97	99	96	0.02
VT	94	90	97	87	0.08
General	91	91	97	91	0.06

En la Figura 31 se presenta la matriz de confusión correspondiente al mejor modelo DT con señales de prueba no utilizadas en el entrenamiento. En dicha matriz, se destaca que las predicciones erróneas están más acentuadas en la clasificación de microsismos TC, con un total de 49 señales incorrectamente predichas como si fueran VT, a esta le sigue la clasificación de LP como VT con un total de 35 señales clasificadas erróneamente.

Figura 31

Cuadro de confusión por microsismo del mejor BER en pruebas del algoritmo DT



En la Tabla 25 presenta en detalle las métricas de rendimiento del algoritmo DT entrenado, específicamente por microsismo. Al analizar esta tabla, se evidencia que la precisión en la clasificación del microsismo VT es notablemente inferior en comparación con los otros microsismos. Este aspecto constituye un factor significativo que debe ser considerado al considerar el modelo en la clasificación de microsismos de este tipo. Además, el tiempo de clasificación tomó 0.79 segundos.

Tabla 25

Evaluación de rendimiento en pruebas por microsismo del modelo DT

Microsismo	A (%)	P (%)	S (%)	R (%)	BER
LP	95	96	97	92	0.05
TC	94	96	96	91	0.06

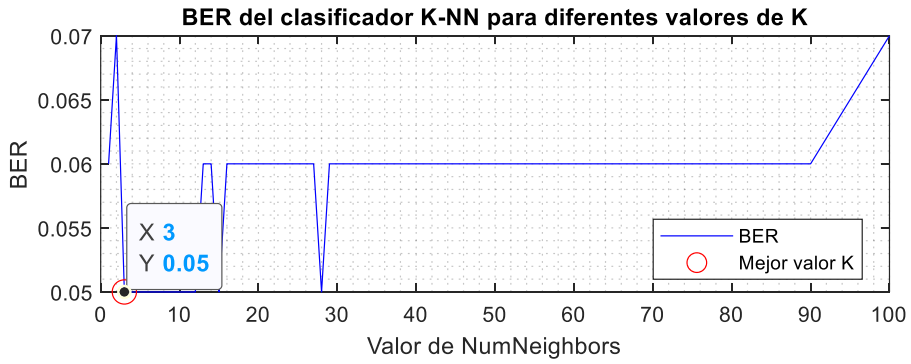
Microsismo	A (%)	P (%)	S (%)	R (%)	BER
TR	98	87	99	94	0.04
VT	97	37	97	93	0.05
General	92	79	97	92	0.05

Modelo óptimo de k-NN

En la Figura 32 se ilustra la variación del parámetro *k*, que representa el número de vecinos más cercanos, este se estableció en 3 con el objetivo de lograr un BER óptimo de 0.05. Los demás parámetros se mantuvieron con sus valores predeterminados. De esta manera, lo valores de los restantes parámetros son determinados por este proceso, los cuales son: *NumNeighbors* en 3*, Distance* en *Euclidean*, *NSMethod* en *Exhaustive*, *DistanceWeight* en *Equal* y *Standardize* en *false*.

Figura 32

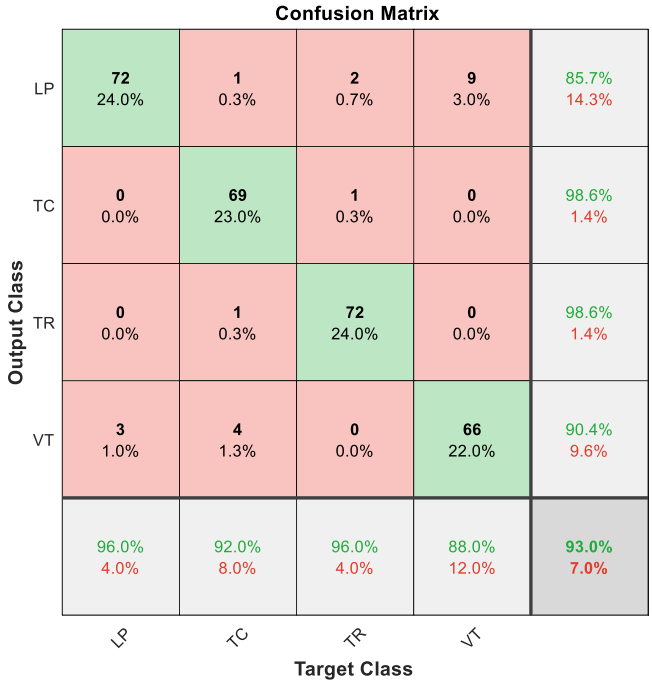
Mejor BER al entrenar el algoritmo k-NN



En la Figura 33, se puede observar la matriz de confusión en la que se destaca un número significativo de predicciones incorrectas, específicamente en el microsismo VT, donde se han clasificado erróneamente como LP un total de 9 señales. Este mismo patrón se repite en la clasificación de microsismos TC, donde algunas de ellas han sido confundidas y clasificadas como VT.

Figura 33

Cuadro de confusión por microsismo del mejor BER en validación del algoritmo k-NN



En la Tabla 26 se presenta en detalle las métricas de rendimiento del algoritmo *k*-NN con datos de validación y detallado por microsismo.

Tabla 26

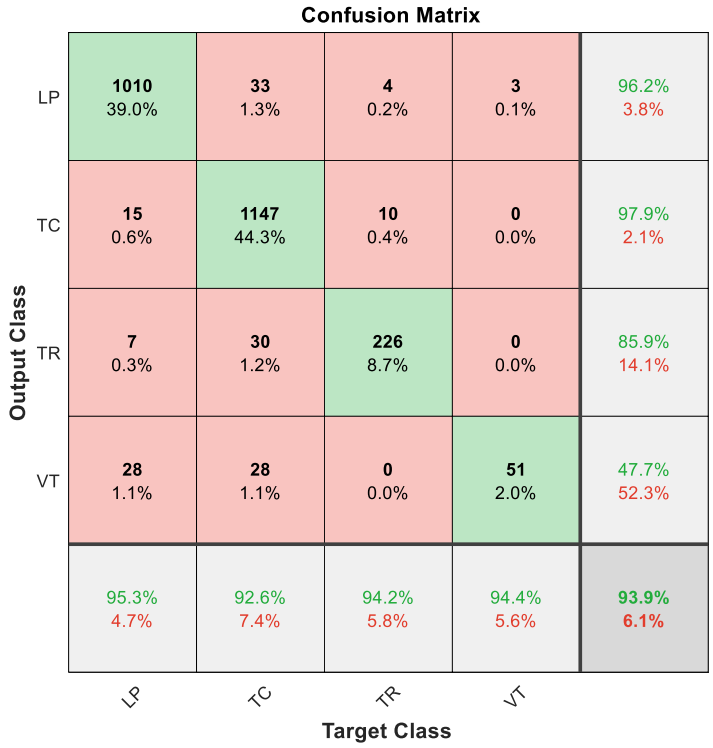
Evaluación de rendimiento en validación por microsismo del modelo k-NN

Microsismo	A (%)	P (%)	S (%)	R (%)	BER
LP	95	86	95	96	0.05
TC	98	99	100	92	0.04
TR	99	99	100	96	0.02
VT	95	90	97	88	0.08
General	93	93	98	93	0.05

Tal como se refleja en la matriz de confusión presentada en la Figura 34, al emplear los datos de prueba, el algoritmo *k*-NN evidencia dificultades en la tarea de clasificar microsismos TC al ser erróneamente clasificados como LP en un total de 33 señales. Además, se observa una tendencia similar en la clasificación de microsismos TC, confundiéndolos y clasificándolos como TR.

Figura 34

Cuadro de confusión por microsismo del mejor BER en pruebas del algoritmo k-NN



En la Tabla 27 se presenta lo resultados específicamente por microsismo. Al analizar esta tabla, se evidencia que la precisión en la clasificación del microsismo VT es mejor con respecto a algoritmo DT, este es un punto importante para considerar en la clasificación de este microsismo. Además, el tiempo de clasificación tomó 0.88 segundos.

Tabla 27

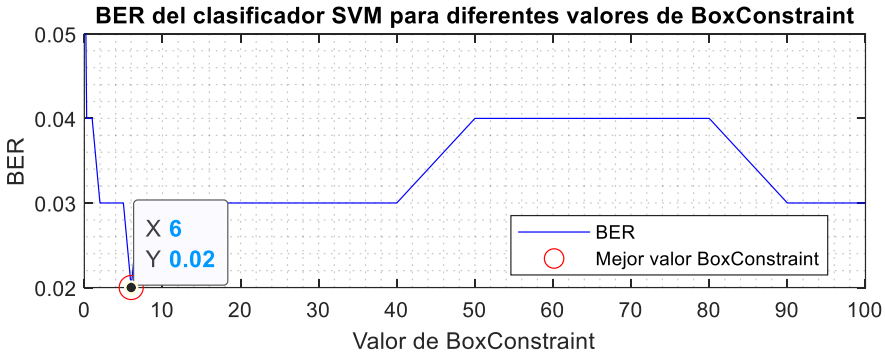
Evaluación de rendimiento en pruebas por microsismo del modelo k-NN

Microsismo	A (%)	P (%)	S (%)	R (%)	BER
LP	97	96	97	95	0.04
TC	96	98	98	93	0.05
TR	98	86	98	94	0.04
VT	98	48	98	94	0.04
General	94	82	98	94	0.04

Modelo óptimo de SVM

Figura 35

Mejor BER al entrenar el algoritmo SVM

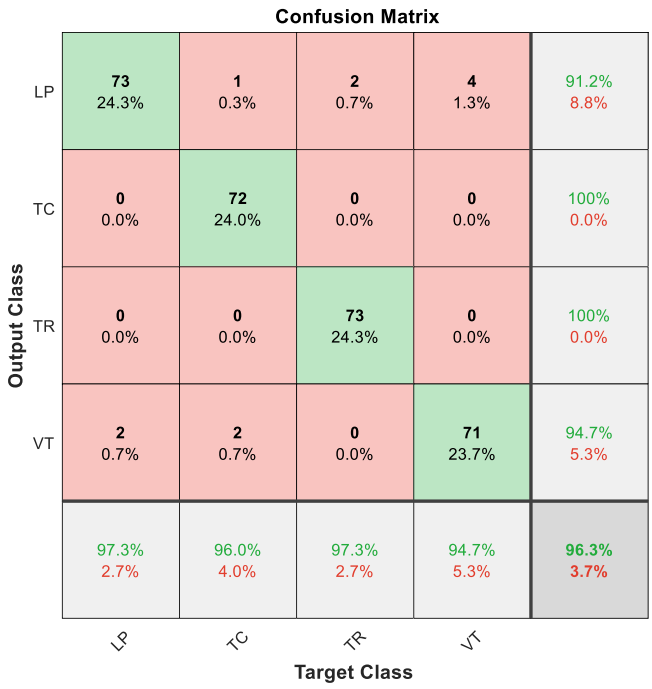


En este algoritmo se estableció el BER en 0.02 para el valor del BoxConstraint en 6 como se presenta en la Figura 35. Esto indica que el modelo generado penalizara más los errores en la clasificación durante el proceso de entrenamiento, lo que puede resultar en un ajuste más restrictivo y una menor tolerancia a los errores en el conjunto de entrenamiento. Sus demás parámetros se los dejo por defecto como: *Coding* en *onevsone*, *Cost* en *vacío, FitPosterior* en *false*.

En la Figura 36 se muestra la matriz de confusión del algoritmo SVM entrenado. En el microsismo VT, se observa una clasificación errónea como microsismo LP con un total de 4 señales. Esto representa una cifra inferior a la obtenida en los análisis anteriores con las matrices de los algoritmos de DT y *k*-NN.

Figura 36

Cuadro de confusión por microsismo del mejor BER en validación del algoritmo SVM



La Tabla 28 presenta los resultados a detalle por microsismo del algoritmo SVM en validación.

Tabla 28

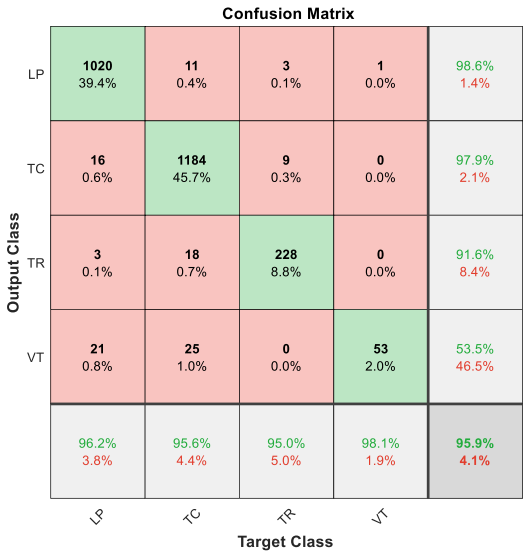
Evaluación de rendimiento en validación por microsismo del modelo SVM

Microsismo	A (%)	P (%)	S (%)	R (%)	BER
LP	97	91	97	97	0.03
TC	99	100	100	96	0.02
TR	99	100	100	97	0.01
VT	97	95	98	95	0.04
General	96	96	99	96	0.02

La Figura 37 se muestra la matriz de confusión de algoritmo SVM evaluado con datos de prueba. Se observa dificultades en la clasificación de los microsismos VT. Se pueden notar conflictos similares en la clasificación de microsismos VT y TC en este algoritmo, al igual que los dos algoritmos anteriores.

Figura 37

Cuadro de confusión por microsismo del mejor BER en pruebas del algoritmo SVM



La Tabla 29 se detallan las métricas de rendimiento en la clasificación de cada microsismo. En este algoritmo, se destaca su mayor precisión en la clasificación del microsismo VT, al alcanzar un 54%. Esta cifra supera la precisión vista en los algoritmos anteriores. Además, el tiempo de clasificación tomó 1.28 segundos.

Tabla 29

Evaluación de rendimiento en pruebas por microsismo del modelo SVM

Microsismo	A (%)	P (%)	S (%)	R (%)	BER
LP	98	99	99	96	0.02
TC	97	98	98	96	0.03
TR	99	92	99	95	0.03
VT	98	54	98	98	0.02
General	96	85	99	96	0.03

Algoritmo de Votación

Basado en los mejores modelos de clasificación entrenados presentados en la Tabla 30, se lleva cabo la prueba del algoritmo de votación basado en prioridades mostrado en el Capítulo II.

Tabla 30

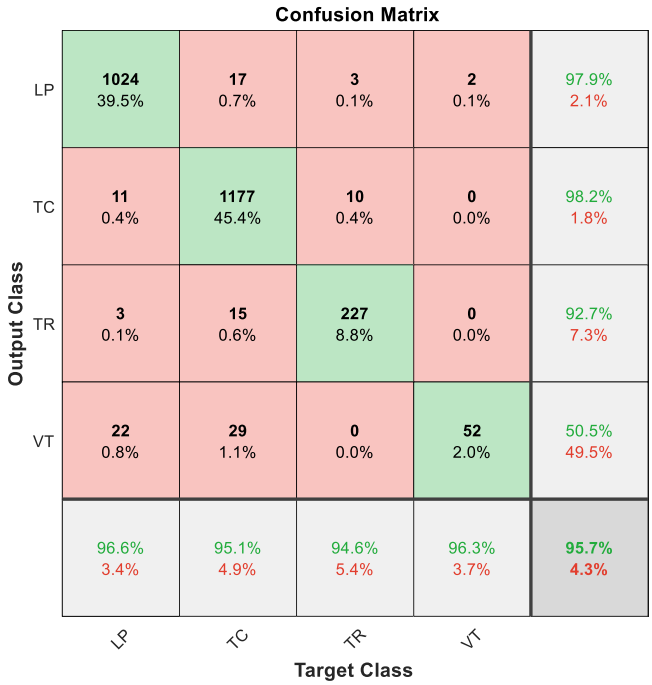
Evaluación de rendimiento en pruebas del algoritmo de votación

Modelo	Base de Datos	Tiempo (s)	A (%)	P (%)	S (%)	R (%)	BER
DT	84 características	0.79	92	79	97	92	0.05
k-NN	29 características RFE	0.88	94	82	98	94	0.04
SVM	29 características RFE	1.28	96	85	99	96	0.03
	Votación	17.84	96	85	99	96	0.03

El cuadro de confusión por microsismos del algoritmo de votación corresponde a la Figura 38.

Figura 38

Cuadro de confusión por microsismo del mejor BER en pruebas del algoritmo votación



Y sus métricas de desempeño por microsismo, se representan en la Tabla 31. Además, el tiempo de clasificación tomó 17.84 segundos.

Tabla 31

Evaluación de rendimiento en pruebas por microsismo del algoritmo de votación

Microsismo	A (%)	P (%)	S (%)	R (%)	BER
LP	98	98	99	97	0.02
TC	97	98	98	95	0.03
TR	99	93	99	95	0.03
VT	98	50	98	96	0.03
General	96	85	99	96	0.03

Pruebas con señales sintéticas

Una vez que se determinaron los algoritmos de clasificación óptimos, se procedió a llevar a cabo pruebas con señales sintéticas. Estas señales sintéticas se generaron mediante los métodos de Bootstrap y CGAN simultáneamente con los grupos de trabajo de investigación. Ambas bases de datos conforman 10 000 señales y se dividen en 2 500 por microsismos LP, TC TR y VT.

Microsismos por Bootstrap

Estas señales sintéticas de microsismos son generadas a partir de la base de datos del volcán Llaima mediante la técnica del Bootstrap. Los resultados obtenidos se presentan en la Tabla 32. Cabe recalcar que su tiempo de preprocesamiento y procesamiento tomó 31 min debido a la cantidad de las señales, sin embargo, los tiempos de clasificación no son elevados.

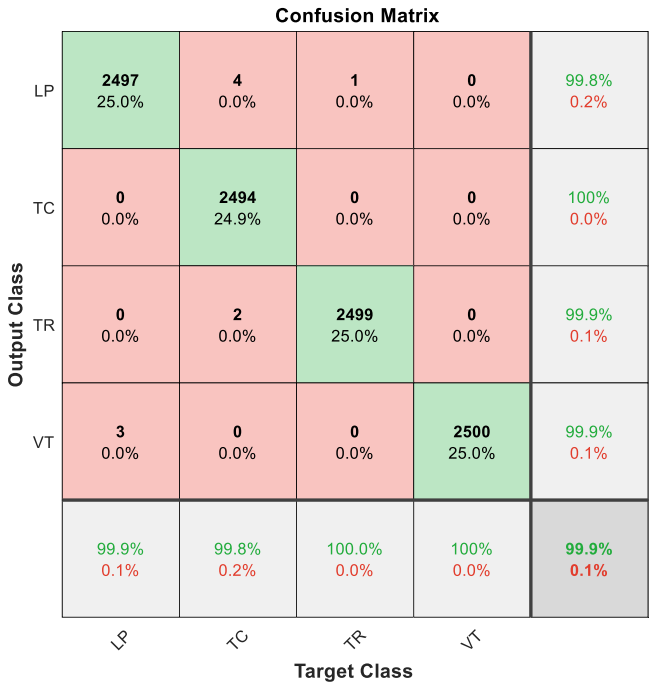
Tabla 32

Evaluación de rendimiento en pruebas Bootstrap sintéticas del algoritmo de votación

Modelo	Base de Datos	Tiempo (s)	A (%)	P (%)	S (%)	R (%)	BER
DT	84 características	0.96	88	91	96	88	0.08
k-NN	29 características RFE	0.84	100	100	100	100	0
SVM	29 características RFE	1.72	100	100	100	100	0
	Votación	45.22	96	100	100	100	0

Figura 39

Cuadro de confusión por microsismo sintético Bootstrap del algoritmo de votación



Microsismos por CGAN

Las señales sintéticas de microsismos son derivadas de la base de datos del volcán Llaima por medio del método CGAN. Los resultados obtenidos se exhiben en la Tabla 33. Además, los tiempos

empleados para preprocesamiento y procesamiento tuvo una duración de 33 min debido a la cantidad de las señales, sin embargo, los tiempos de clasificación no son elevados.

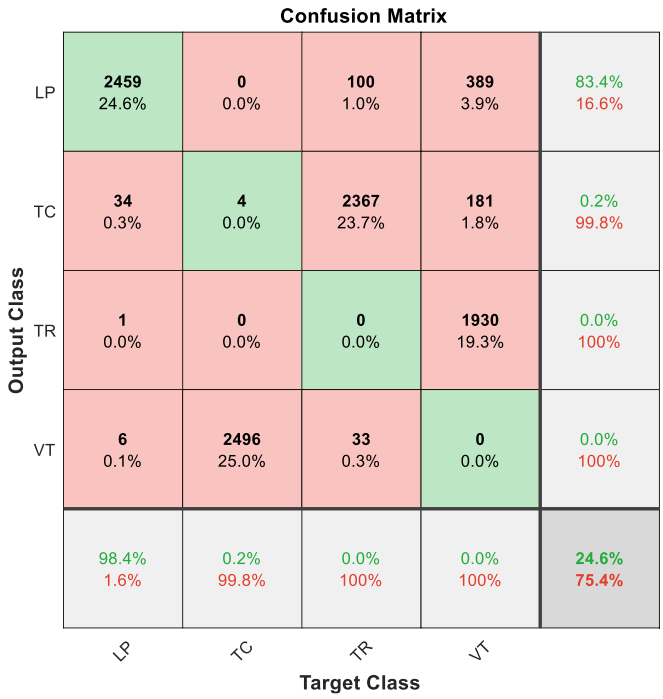
Tabla 33

Evaluación de rendimiento en pruebas CGAN del algoritmo de votación

Modelo	Base de Datos	Tiempo (s)	A (%)	P (%)	S (%)	R (%)	BER
DT	84 características	0.75	24	24	75	24	0.50
k-NN	84 características RFE	0.60	22	22	74	22	0.52
SVM	84 características RFE	1.28	20	26	73	20	0.54
	Votación	45.44	25	21	75	25	0.52

Figura 40

Cuadro de confusión por microsismo sintético CGAN del algoritmo de votación



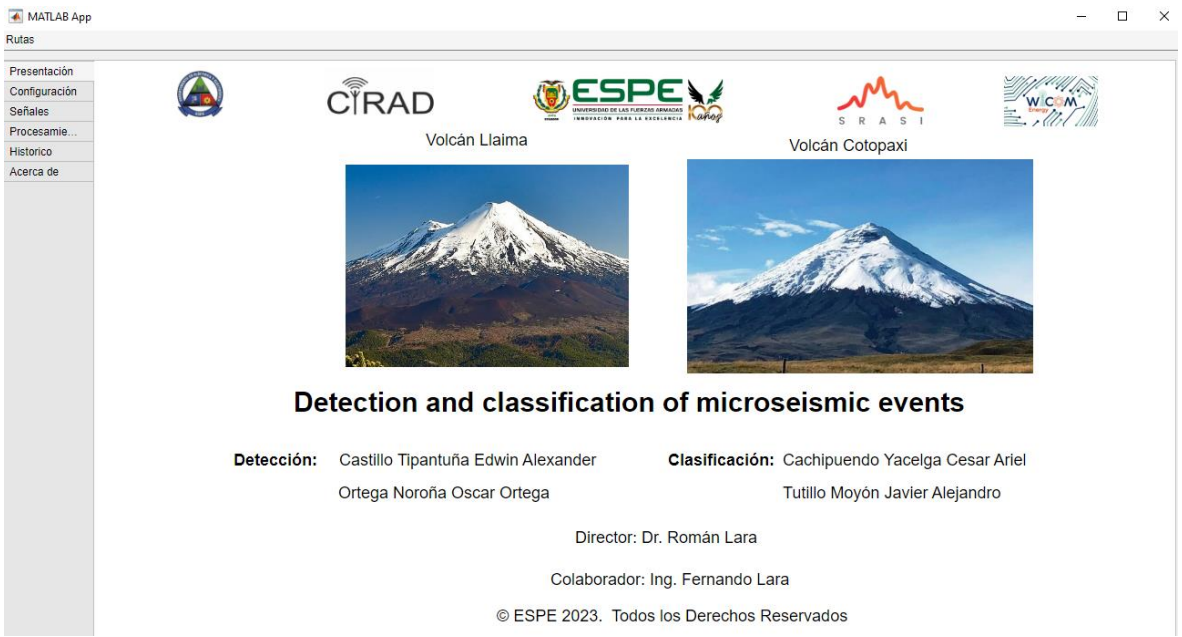
Interfaz Gráfica

La presentación del clasificador desarrollado mediante técnicas de ML ha sido incorporada en el trabajo de detección y clasificación de microsismos del volcán Cotopaxi, llevado a cabo por Fernando Lara. Para la creación de esta interfaz, se utilizó la herramienta de diseño de aplicaciones Matlab App. El objetivo de esta interfaz es proporcionar a geofísicos una forma sencilla de determinar la clase de microsismo a partir de una señal ingresada.

La interfaz presenta diversas pestañas que permiten seleccionar el volcán y los métodos de detección y clasificación. En la Figura 41 se muestra la pestaña principal de la interfaz, que proporciona información sobre la institución, los autores que han participado en la creación de los detectores y clasificadores del volcán Llaima, así como el nombre del director del proyecto de investigación y el colaborador responsable de los detectores y clasificadores del volcán Cotopaxi.

Figura 41

Interfaz gráfica: presentación de autores



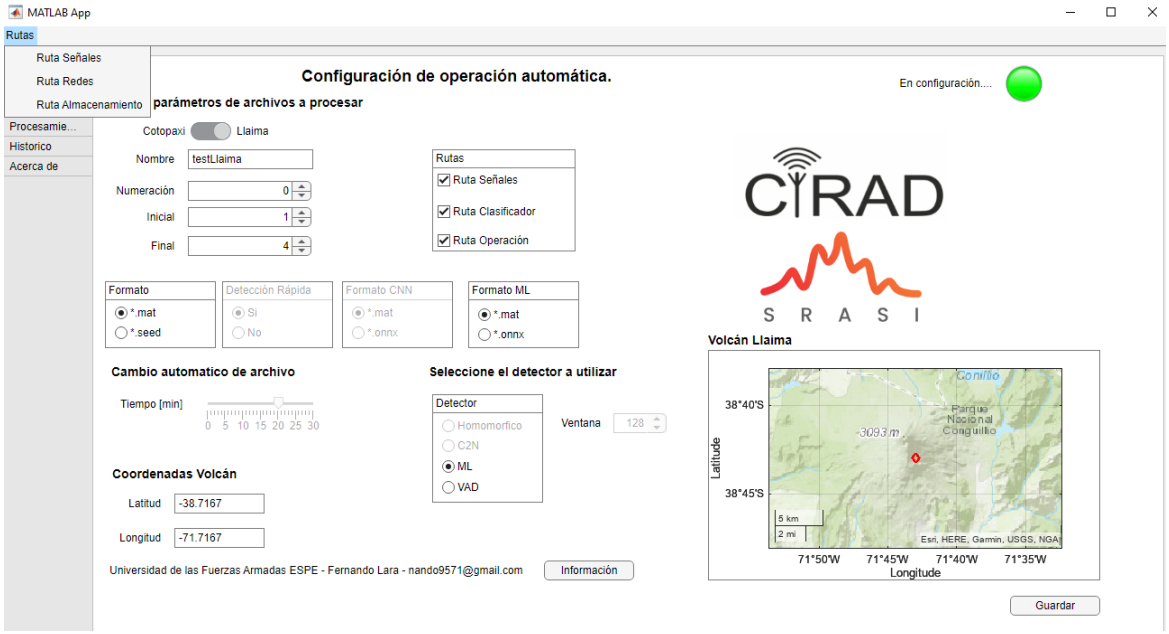
En la Figura 42 se muestra la pestaña de configuración de la interfaz. En esta selección, se permite al usuario seleccionar el volcán de interés, su ubicación geográfica mediante un mapa ilustrativo con las coordenadas de longitud y latitud correspondientes. Además, se presenta los métodos de detección específicamente diseñados para cada volcán.

En la esquina superior izquierda de esta pestaña, se encuentran las opciones para agregar las rutas de las diferentes carpetas necesarias para el funcionamiento de la interfaz. Esto incluye la dirección de la carpeta que contiene las señales a ser probadas, los modelos entrenados utilizados en el proceso y las ubicaciones donde se almacenarán los resultados de la clasificación. Una vez que el usuario ha realizado la configuración correspondiente, puede guardarla al presionar el botón “Guardar”. Una vez que se haya completado el proceso de guardar la configuración, la lámpara ubicada en la parte superior de la interfaz cambiará de color, al pasar de rojo a verde.

Esta pestaña de configuración brinda a los geofísicos la flexibilidad de personalizar la aplicación según sus necesidades, permitiéndoles trabajar con diferentes volcanes y métodos de detección de manera intuitiva y eficiente.

Figura 42

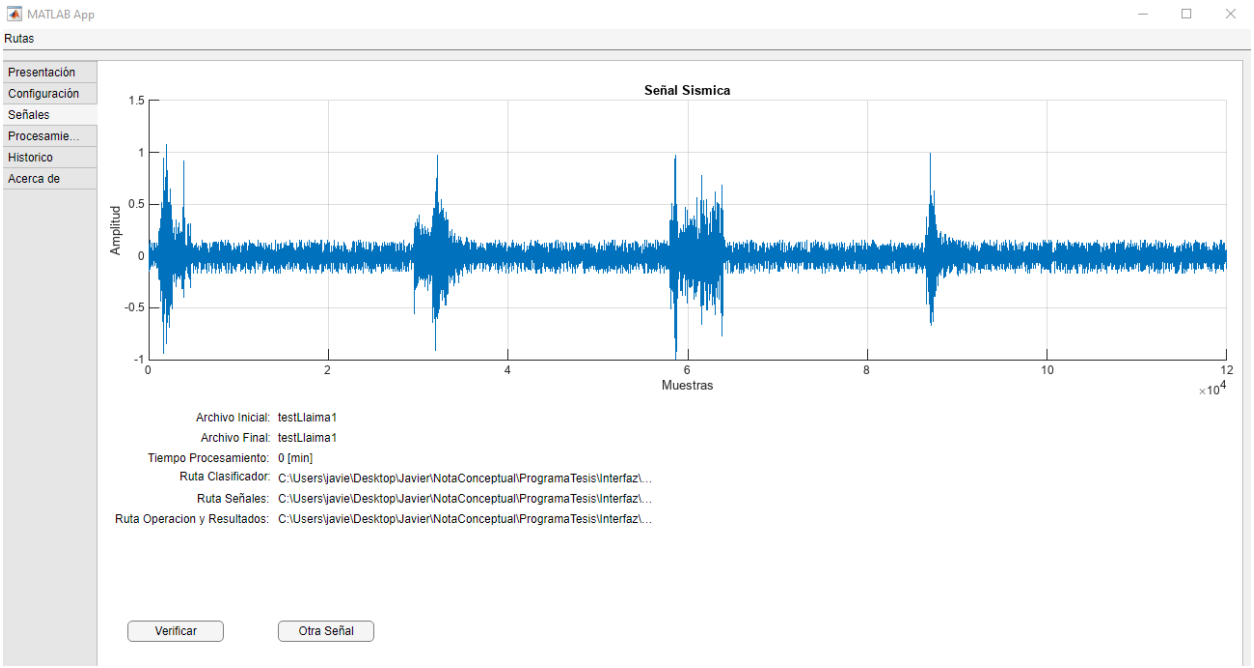
Interfaz gráfica: configuración de operación automática



Después de haber guardado la configuración del volcán seleccionado en la Figura 43, en la pestaña de “Señales” se muestran las señales correspondientes a ese volcán. En esta pestaña, cada una de estas señales se presenta junto con la ubicación de las carpetas donde se encuentran las señales a ser procesadas, los modelos clasificadores utilizados y el lugar donde se almacenará los resultados de la clasificación. Esta representación permite a los usuarios visualizar de manera organizada y clara las señales específicas del volcán elegido.

Figura 43

Interfaz gráfica: presentación de señal ingresada



La Figura 44 ilustra la pestaña de “Procesamiento”, la cual presenta varias secciones que guían a través del procesamiento de la señal ingresada. Estas secciones incluyen:

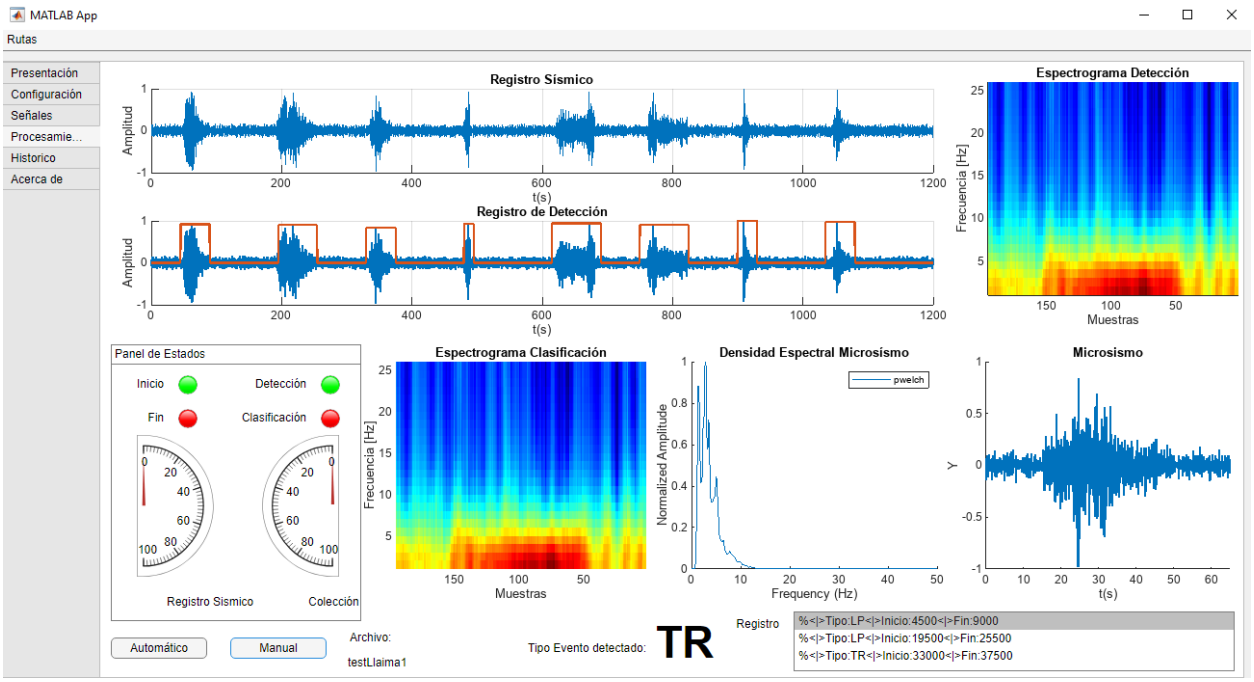
- *Registro sísmico*. - muestra el registro visual de la señal ingresada, permite una vista completa de la información sísmica a lo largo del tiempo.
- *Registro y espectrograma de detección.* - presenta el registro sísmico junto con el espectrograma correspondiente, destaca las áreas donde se han detectado posibles microsismos.
- *Espectrograma de clasificación.* - exhibe el espectrograma de la señal procesada, se resaltan las características relevantes para la clasificación de los microsismos.
- *Señal detectada en tiempo y PSD.* - muestra la señal detectada en el dominio de tiempo y su PSD, proporciona una comprensión detallada de las propiedades de la señal.

En la parte inferior de la interfaz, se indica el tipo de microsismo detectado en la señal procesada. Además, se presenta un panel de estados que informa en que etapa del programa se encuentra en ejecución, este brinda una guía clara sobre el progreso del procesamiento.

Esta pestaña de “Procesamiento” proporciona una vista exhaustiva y estructurada de los diversos aspectos del procesamiento de la señal, permite a los geofísicos analizar y comprender en profundidad los resultados obtenidos durante la detección y clasificación de los microsismos.

Figura 44

Interfaz gráfica de procesamiento de señal ingresada



Conforme avanza la detección y clasificación de los microsismos en la pestaña “Históricos”de la Figura 45, se almacenan los datos de los microsismos reconocidos. Este registro consta de las siguientes columnas:

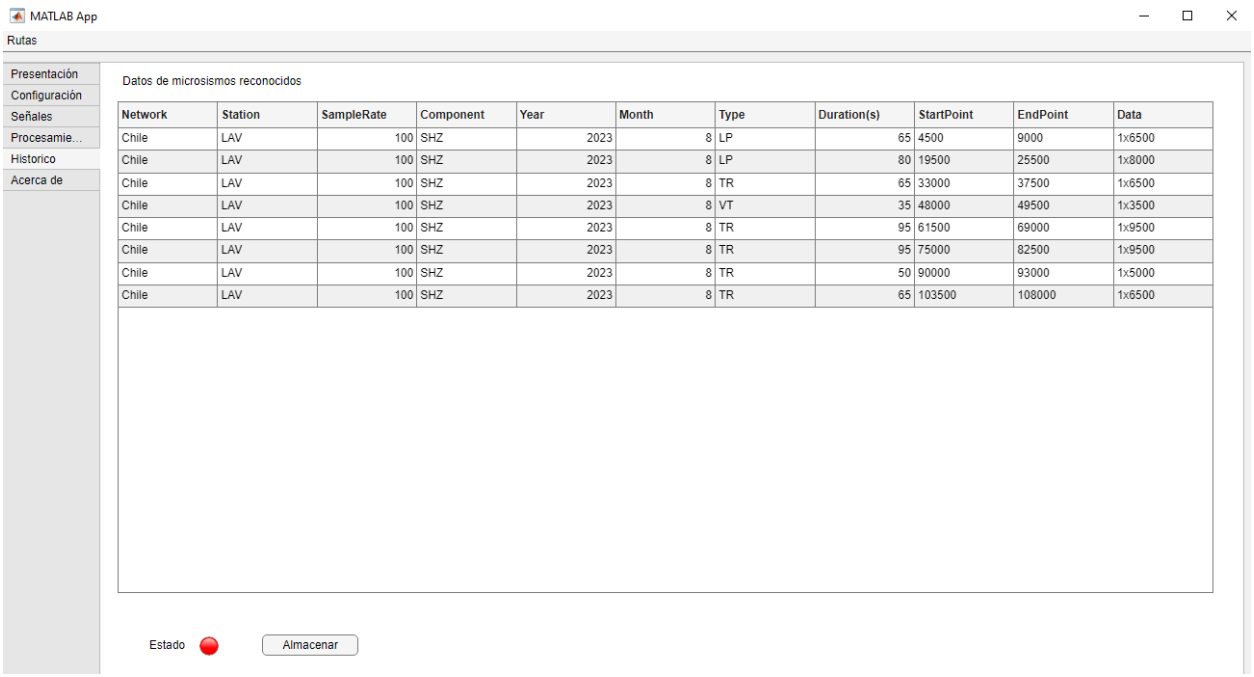
- *Network*. - indica la ubicación de volcán.
- *Station.* - representa la estación que ha capturado los microsismos.

- *Sample Rate*. - indica la tasa de muestreo utilizada.
- *Component*. - hace referencia a la dirección de la componente, como SHZ.
- *Date*. - muestra la fecha de detección y clasificación del microsismo.
- *Type.* - describe el tipo de microsismo detectado.
- *Duration*. - indica la duración de los microsismos detectados en segundos.
- *StartPoint y EndPoint*. - representan el tiempo en que comenzó y finalizó la detección del microsismo.

Todos estos datos pueden ser almacenados en un archivo en formato .dat o xls. Esta elección de formato proporciona flexibilidad para guardar y compartir los registros históricos de los microsismos.

Figura 45

Interfaz gráfica: registro de señal detectada y clasificada



Network	Station	SampleRate	Component	Year	Month	Type	Duration(s)	StartPoint	EndPoint	Data
Chile	LAV	100	SHZ	2023	8	LP	65	4500	9000	1x6500
Chile	LAV	100	SHZ	2023	8	LP	80	19500	25500	1x8000
Chile	LAV	100	SHZ	2023	8	TR	65	33000	37500	1x6500
Chile	LAV	100	SHZ	2023	8	VT	35	48000	49500	1x3500
Chile	LAV	100	SHZ	2023	8	TR	95	61500	69000	1x9500
Chile	LAV	100	SHZ	2023	8	TR	95	75000	82500	1x9500
Chile	LAV	100	SHZ	2023	8	TR	50	90000	93000	1x5000
Chile	LAV	100	SHZ	2023	8	TR	65	103500	108000	1x6500

Finalmente, en la pestaña “Acerca de”, se proporciona información detallada sobre el formato requerido para el archivo del microsismo que se ingresará para el procesamiento de detección y

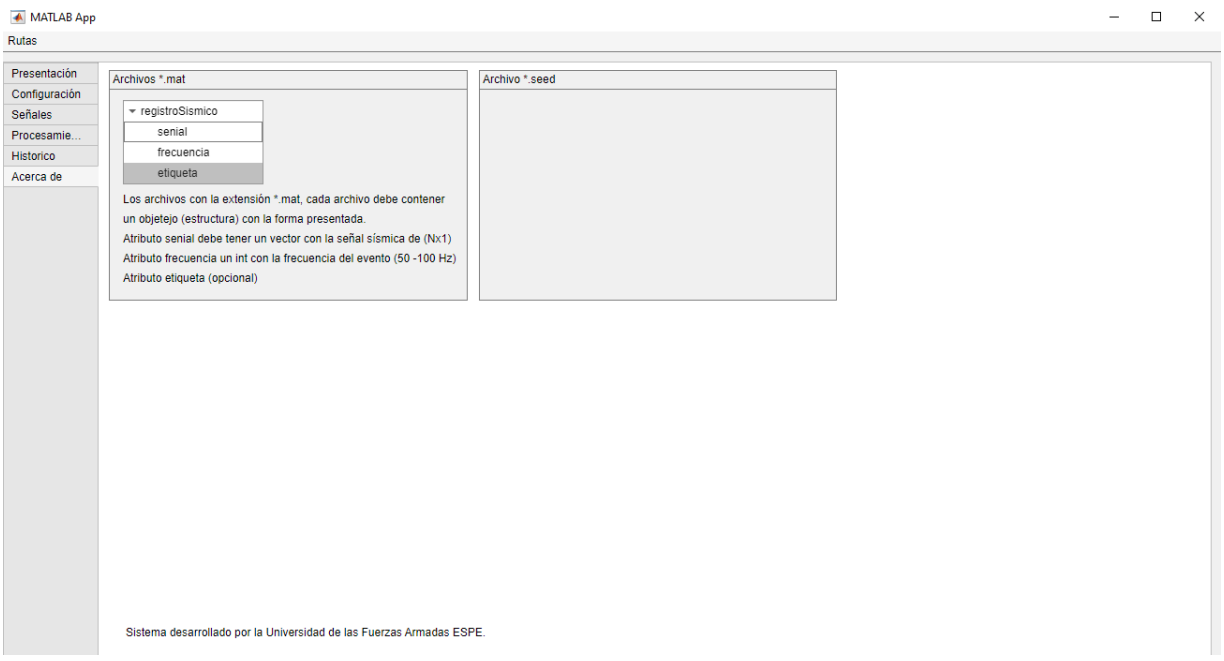
clasificación, tal como se muestra en la Figura 46. Para que la señal sea procesada adecuadamente, debe cumplir con los siguientes requisitos:

- *Señal*. - la señal del microsismo debe estar en formato de un vector NX1, donde N representa la longitud de la señal en puntos de muestreo.
- *Frecuencia de muestreo*. - 50 - 100 Hz
- *Etiqueta*. - opcionalmente, se pude incluir una etiqueta que describa la señal, proporciona información adicional sobre su origen, ubicación o cualquier otro detalle relevante.

Estos requisitos aseguran que la señal del microsismo pueda ser procesada de manera efectiva por el sistema de detección y clasificación.

Figura 46

Interfaz gráfica: Acerca de



Capítulo IV

Conclusiones

Se ha logrado crear un sistema de clasificación para microsismos del volcán Llaima, al utilizar los modelos más destacados de tres algoritmos de clasificación: Árboles de Decisión (DT), k-Nearest Neighbors (*k*-NN) y Máquinas de Vectores de Soporte (SVM). Adicionalmente, se aplicó un algoritmo de votación para mejorar aún más los resultados obtenidos.

A través de la revisión exhaustiva de investigaciones anteriores y proyectos afines, se establecieron criterios esenciales para el análisis de microsismos. Estos criterios abarcan aspectos clave como la extracción y selección de características mediante el uso de Recursive Feature Elimination (RFE), el empleo de algoritmos de aprendizaje supervisado y la evaluación a través de métricas de rendimiento. Mediante esta metodología, se logró evitar la adopción de técnicas que pudieran tener un impacto negativo en las métricas de desempeño esperadas.

La normalización de la base de datos emerge como un paso crítico e indispensable en el proceso de desarrollo de algoritmos de aprendizaje automático para la clasificación de microsismos en el volcán Llaima. Los resultados evidencian que la normalización no solo mejora de manera significativa las métricas de desempeño, sino que también agiliza los procesos de entrenamiento y extracción de características. Esta práctica se establece como un componente esencial para maximizar la efectividad y la eficiencia del proceso de clasificación, por tanto, se permite que los modelos generados sean más precisos, adaptables y rápidos en su respuesta ante nuevas señales sísmicas.

La extracción y selección de 84 características relacionadas con los microsismos en los dominios del tiempo, frecuencia y escala demostraron ser elementos esenciales para la exitosa implementación de las técnicas de aprendizaje supervisado y la obtención de los modelos de clasificación más efectivos.

A pesar de que las alteraciones en los resultados pueden no ser de gran magnitud al emplear la selección de características en contraste con el conjunto de datos completo, esta técnica presenta notables ventajas en cuanto a la eficiencia computacional y la mitigación del sobreajuste. Constituye una estrategia recomendada para realzar la efectividad y la viabilidad de los modelos de aprendizaje automático, especialmente en contextos donde la sencillez y la interpretación juegan un papel fundamental.

Los mejores resultados de clasificación obtenidos corresponden en primera instancia al algoritmo SVM con el método RFE, el cual redujo de las 84 características en múltiples dominios a solo 29 y se obtuvo un BER del 0.03. En segundo lugar, se presenta el algoritmo k-NN con el método RFE antes mencionado con un BER del 0.04. Por último, el algoritmo DT empleado las 84 características en múltiples dominios con un BER del 0.05. A pesar de obtener valores bajos en la métrica del BER, no se logró cumplir con la solicitud planteada por el Instituto Geofísico de la Escuela Politécnica Nacional, que requería un BER de 0.01.

El uso del algoritmo de votación reveló un efecto neutro en las métricas de rendimiento, con un ligero incremento del 0.5% tanto en la Exactitud como en la Precisión. A pesar de este aumento, dichas métricas se mantuvieron en niveles comparables a las obtenidas por el modelo SVM con el porcentaje más alto en las demás áreas evaluadas.

En lo que concierne a los tiempos de clasificación se utiliza el conjunto de pruebas, se logró un tiempo total de 25.49 segundos. Esto se distribuye entre los algoritmos de la siguiente manera: 0.79 segundos para DT, 0.88 segundos para *k*-NN, 1.28 segundos para SVM y 17.84 segundos para el algoritmo de votación. Los restantes intervalos de tiempo corresponden a ejecuciones entre líneas de código adicionales.

Trabajos Futuros

A medida que el campo de la clasificación de microsismos en entornos volcánicos continua en constante evolución, se abren múltiples oportunidades para explorar y mejorar los enfoques existentes. Este estudio ha sentado las bases para la creación de algoritmos inteligentes basados en la teoría de ML tradicional, específicamente para la clasificación de microsismos en el volcán Llaima en un escenario multiclase. Sin embargo, existen diversos caminos a seguir para ampliar y perfeccionar esta investigación en el futuro.

Un primer paso fundamental consiste en validar la precisión del clasificador desarrollado mediante la evaluación de nuevas señales emitidas por el volcán Llaima. A medida que la base de datos se actualiza con datos en tiempo real, será crucial comprobar cómo el clasificador responde a estos nuevos microsismos y si mantiene su capacidad de predicción precisa. Este proceso permitirá ajustar y afinar el modelo en función de los datos más recientes, mejorar su capacidad de adaptación a condiciones cambiantes.

Además, una expansión natural de este trabajo es evaluar el rendimiento del clasificador al utilizar bases de datos de otros volcanes. Se podría considerar la inclusión de bases de datos que contengan microsismos similares a los investigados en este estudio, como microsismos LP y VT del volcán Cotopaxi. Esto permitiría verificar si el modelo desarrollado es transferible y puede generalizar patrones en diferentes contextos volcánicos, lo que ampliaría su aplicabilidad y utilidad en un espectro más amplio de situaciones.

Además, para futuros trabajos, resulta crucial tener en cuenta la incorporación de otras técnicas de selección de características, como el método de filtros o el método de embebidos. La exploración de estas alternativas podría brindar un enfoque más sólido en la identificación de las características fundamentales. Al llevar a cabo una comparativa de los resultados obtenidos mediante estos métodos,

se podría lograr una mejor comprensión de cuales características son realmente relevantes para optimizar la clasificación de los microsismos.

Otro ámbito de investigación interesante implica la exploración de técnicas de *Deep Learning* u otras técnicas de ML alternativas para generar bases de datos sintéticas. Al utilizar estas técnicas, es posible crear señales sintéticas que capturen patrones más complejos y sutiles en los microsismos. Investigar cómo estas señales sintéticas afectan el rendimiento del clasificador permitiría una comparación con los enfoques tradicionales empleados al brinda una perspectiva más completa de las capacidades de cada método.

Además, es fundamental analizar en profundidad la eficacia del algoritmo de votación. Si bien se ha demostrado su impacto neutral en los resultados, sería interesante explorar cómo se combina con otros modelos generados a partir de técnicas de aprendizaje de máquina, como Naive Bayes (NB) o Extreme Gradient Boosting (XG). Asimismo, se podría considerar la combinación con modelos desarrollados mediante enfoques de Deep Learning. Esta evaluación de la sinergia entre distintos algoritmos de clasificación contribuiría a definir estrategias más sólidas y robustas para abordar la clasificación de microsismos.

Otro aspecto valioso por considerar es la exploración más profunda de los parámetros configurables en cada algoritmo de Aprendizaje de Máquina utilizado. En este estudio, se varió un único parámetro en cada algoritmo mientras se mantuvieron los demás en sus valores por defecto. Sin embargo, existe una amplia gama de parámetros que pueden ajustarse para optimizar aún más el rendimiento y la adaptabilidad de estos algoritmos. La variación sistemática de estos parámetros puede ofrecer una visión más completa de cómo cada algoritmo responde a diferentes configuraciones y cómo se comporta en un rango más amplio de escenarios. Este enfoque permitiría identificar configuraciones

óptimas para cada algoritmo en el contexto específico de la clasificación de microsismos en el volcán Llaima.

Por todo lo anterior, la creación de algoritmos inteligentes para la clasificación de microsismos en el volcán Llaima es prometedor y lleno de posibilidades. La validación continua, la adaptabilidad a nuevos datos, la exploración de diferentes bases de datos y técnicas, así como la combinación de modelos, se perfilan como áreas clave para una mayor mejora y desarrollo en esta investigación. El resultado final será un enfoque más sólido y confiable para la clasificación de microsismos en entornos volcánicos, con importantes implicaciones para la detección temprana y la toma de decisiones en la mitigación de riesgos volcánicos.

Bibliografía

Alessio, S. M. (2016). *Digital Signal Processing and Spectral Analysis for Scientists: Concepts and Applications.* Switzerland: Springer Cham. doi:https://doi.org/10.1007/978-3-319-25468-5

Altamirano, R. B. (2021). *Sistema de reconocimiento de microterremotos en tiempo real del volcán Cotopaxi aplicando aprendizaje supervisado.* Universidad de las Fuerzas Armadas ESPE, Eléctrica, Electrónica y Telecomunicaciones, Quito. Retrieved from http://repositorio.espe.edu.ec/xmlui/bitstream/handle/21000/23743/T-ESPE-044263.pdf?sequence=1&isAllowed=y

aprendeIA. (2019). *Métodos de seleccion de características machine learning*. Retrieved from https://aprendeia.com/metodos-de-seleccion-de-caracteristicas-machine-learning/

Arana, C. (2021). *MODELOS DE APRENDIZAJE AUTOMÁTICO MEDIANTE ÁRBOLES DE DECISIÓN.* Universidad del CEMA, Negocios, Buenos Aires. Retrieved from https://ucema.edu.ar/publicaciones/download/documentos/778.pdf

Battaglia, J., & Aki, K. (2003). *Location of seismic events and eruptive fissures on the Piton de la.* doi:10.1029/2002JB002193.

Brownlee , J. (2020, Junio 30). *Rescaling Data for Machine Learning in Python with Scikit-Learn*. Retrieved from Machine Learning Mastery: Making Developers Awesome at Machine Learning: https://machinelearningmastery.com/rescaling-data-for-machine-learning-in-python-with-scikit-learn/

Canário, P., Fernandes, R., Curilem, M., Huenupan, F., & Araujo, R. (2020). Llaima volcano dataset: In-depth comparison of deep artificial neural network architectures on seismic events classification. *Journal of Volcanology and Geothermal Research, 401*, 106881.

CENAPRED. (2018). *Sistema de reconocimiento automático de señales sísmicas.* Mexico D.F. Retrieved from http://www1.cenapred.unam.mx/DIR_INVESTIGACION/2109/FRACCION_XLI/RV/12_%20HTK_2018.pdf

Franco Marín, L. (2019). *Comportamiento eruptivo del volcán Llaima (2007-2010) e incidencia del terremoto del Maule MW 8.8 en la actividad volcánica y tectónica local.* Universidad de Concepción , Facultad de Ciencias Químicas, Concepción. Retrieved from http://repositorio.udec.cl/xmlui/bitstream/handle/11594/1139/Tesis%20comportamiento%20eruptivo%20del%20volcan%20Llaima.pdf?sequence=1&isAllowed=y

Galarza, C. C., & Vega, V. J. (2022). *Generación de señales sintéticas de eventos sismo-volcánicos del volcán Cotopaxi a través de un modelo de red neuronal adversario generativo condicional.* Universidad de las Fuerzas Armadas ESPE, Eléctrica, Electrónica y Telecomunicaciones. Retrieved from https://repositorio.espe.edu.ec/bitstream/21000/31542/1/T-ESPE-052391.pdf

Ginez, O. D. (2019). *SISTEMA PARA DETERMINACIÓN DE EVENTOS SÍSMICOS EN UNA SECUENCIA DE TIEMPO.* PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR, ESCUELA DE SISTEMAS, QUITO. Retrieved from http://repositorio.puce.edu.ec/bitstream/handle/22000/17058/Disertaci%C3%B3n%20Ginez%20Daniel.pdf?sequence=1&isAllowed=y

IGEPN. (2023). *Cotopaxi*. Retrieved from Instituto Geofísico de la Escuela Politécnica Nacional: https://www.igepn.edu.ec/cotopaxi

Iglesias, G. I., & Rosero, A. A. (2023, 07 10). *Implementación de un sistema clasificador de microterremotos del volcán Cotopaxi basado en técnicas de Deep Learning.* Universidad de las

Fuerzas Armadas ESPE, Eléctrica, Electrónica y Telecomunicaciones , Sangolqui. Retrieved from https://repositorio.espe.edu.ec/jspui/bitstream/21000/36707/1/T-ESPE-058092.pdf

InteractiveChaos. (2023, 07 11). *Índice de Gini*. Retrieved from https://interactivechaos.com/es/manual/tutorial-de-machine-learning/indice-gini

Lara, F., Lara-Cueva, R., Larco, J., Carrera, E. V., & León, R. (2021). A deep learning approach for automatic recognition of seismo-volcanic events at the Cotopaxi volcano. *Journal of Volcanology and Geothermal Research, 409*(107142). doi:https://doi.org/10.1016/j.jvolgeores.2020.107142

Lara-Cueva, R., Carrera, E. V., Morejon, J. F., & Benitez, D. (2016). Comparative analysis of automated classifiers applied to volcano event identification. *IEEE Colombian Conference on Communications and Computing (COLCOM)*, 1-6. doi:10.1109/ColComCon.2016.7516377

Lara-Cueva, R., D. B., Carrera, E., Ruiz, M., & Rojo-Álvarez, J. (2016). Feature selection of seismic waveforms for long period event detection at Cotopaxi Volcano. *Journal of Volcanology and Geothermal Research*, 34-49. doi:https://doi.org/10.1016/j.jvolgeores.2016.02.022

Lara-Cueva, R., Larco, J. C., Benítez, D. S., Pérez, N., Grijalva, F., & Ruiz, M. (2020). On finding possible frequencies for recognizing microearthquakes at Cotopaxi volcano: A machine learning based approach. *Journal of Volcanology and Geothermal Research*(407).

Lee, W., & Valdes, C. (1985). *HYPO71PC; a personal computer version of the HYPO71 earthquake location program.* doi:10.3133/ofr85749

Maisueche, A. C. (2019). *UTILIZACIÓN DEL MACHINE LEARNING EN LA INDUSTRIA 4.0.* Universidad de Valladolid, Eescuela de Ingenierias Industriales, Valladolid. Retrieved from https://core.ac.uk/download/pdf/228074134.pdf

Minango, G. M. (2022). *Clasificación de eventos sismo volcánicos usando características psicoacústicas mediante técnicas de aprendizaje automático supervisado y no supervisado.* Universidad de las Fuerzas Armadas ESPE, Quito. Retrieved from http://repositorio.espe.edu.ec/bitstream/21000/31541/1/T-ESPE-052390.pdf

Moreno, H. N. (2009). *Estilos eruptivos 2007-2008 del volcán Llaima, Andes del sur.* Santiago: XII Congreso Geológico Chileno. Retrieved from https://www.researchgate.net/publication/326449956_Estilos_eruptivos_2007-2008_del_volcan_Llaima_Andes_del_sur

MySQL. (2003). *Normalizacion de bases de datos*. Retrieved from Mallorca.

OVSICORI. (2023). *Vulcanología*. Retrieved from Universidad Nacional Costa Rica: http://www.ovsicori.una.ac.cr/index.php/faqs/vulcanologia/acerca-de-sismos-asociados-con-actividad-volcanica

Pérez Quisaguano, A. S. (2018). *Detección automática de eventos sísmicos en el volcán Cotopaxi mediante técnicas de Aprendizaje de Máquinas:.* Retrieved from https://repositorio.espe.edu.ec/bitstream/21000/13812/1/T-ESPE-057527.pdf

Pérez, N., Benítez, D., Grijalva, F., Lara, R., Ruiz, M., & Aguilar, J. (2020). ESeismic: Towards an Ecuadorian volcano seismic repository. *Journal of Volcanology and Geothermal Research*, ELSEVIER. doi:https://doi.org/10.1016/j.jvolgeores.2020.106855

SERNAGEOMIN. (2023). *Volcán Llaima*. Retrieved from https://rnvv.sernageomin.cl/volcan-llaima/

TIBCO. (2023). *¿Qué es el aprendizaje supervisado?* Retrieved from https://www.tibco.com/es/reference-center/what-is-supervised-learning

Trnkoczy, A. (1999). *Understanding and parameter setting of STA/LTA trigger algorithm.* Retrieved from https://gfzpublic.gfz-potsdam.de/rest/items/item_4097/component/file_4098/content#:~:text=An%20initial%20setting%20for%20the,significant%20man%2Dmade%20seismic%20noise.

Vásconez Gómez, F. S. (2020, Junio 30). *Desarrollo de un sistema de clasificación de eventos sismo volcánicos usando librerías de Machine Learning en Python.* Retrieved from http://repositorio.espe.edu.ec/jspui/bitstream/21000/22408/1/T-ESPE-043768.pdf

Yugsi, J. P. (2022). *GENERACIÓN DE SEÑALES VOLCÁNICAS ARTIFICIALES DE TIPO LP (LONG-PERIOD) Y VT (VOLCANO-TECTONIC) A PARTIR DE UNA BASE DE DATOS DEL VOLCÁN COTOPAXI USANDO LA TÉCNICA DE BOOTSTRAPPING.* Escuela Politécnica Nacional, Quito. Retrieved from https://bibdigital.epn.edu.ec/bitstream/15000/23276/1/CD%2012691.pdf

Zapata, J. Y. (2022). *GENERACIÓN DE SEÑALES VOLCÁNICAS ARTIFICIALES DE TIPO LP (LONG-PERIOD) Y VT (VOLCANO-TECTONIC) A PARTIR DE UNA BASE DE DATOS DEL VOLCÁN COTOPAXI USANDO LA TÉCNICA DE BOOTSTRAPPING.* ESCUELA POLITÉCNICA NACIONAL, ELÉCTRICA Y ELECTRÓNICA, Quito. Retrieved from https://bibdigital.epn.edu.ec/bitstream/15000/23276/1/CD%2012691.pdf

Apéndices