

Maestría en

Ciencia de Datos y Máquinas de Aprendizaje mención en Inteligencia Artificial

**Trabajo previo a la obtención de título de
Magister en**

AUTORES:

Cristian Fernando Tenecela Cabay

Eutimio Napoleón Arevalo Valarezo

José Miguel Ocaña Chiluisa

Yessenia Denisse Carrillo Toapanta

TUTORES

Alejandro Cortés

Karla Mora

TEMA

Comparativa de modelos basados en el análisis de datos sísmicos para la clasificación de erupciones en el volcán Cotopaxi.

CERTIFICACIÓN DE AUTORÍA

Nosotros, **Cristian Fernando Tenecela Cabay**, **Eutimio Napoleon Arevalo Valarezo**, **José Miguel Ocaña Chiluisa**, **Yessenia Denisse Carrillo Toapanta**, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

Firma
Cristian Fernando Tenecela Cabay

Firma
Eutimio Napoleón Arévalo Valarezo

Firma
José Miguel Ocaña Chiluisa

Firma
Yessenia Denisse Carrillo Toapanta

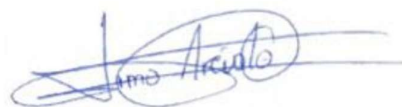
AUTORIZACIÓN DE DERECHOS DE PROPIEDAD INTELECTUAL

Nosotros, **Cristian Fernando Tenecela Cabay**, **Eutimio Napoleón Arévalo Valarezo**, **José Miguel Ocaña Chiluisa**, **Yessenia Denisse Carrillo Toapanta**, en calidad de autores del trabajo de investigación titulado ***Título del trabajo de investigación Comparativa de modelos de clasificación basados en el análisis de datos sísmicos para la predicción de erupciones en el volcán Cotopaxi***, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, enero 2026



Firma
Cristian Fernando Tenecela Cabay



Firma
Eutimio Napoleón Arévalo Valarezo



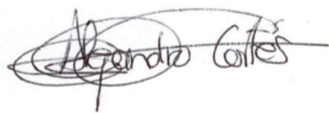
Firma
José Miguel Ocaña Chiluisa



Firma
Yessenia Denisse Carrillo Toapanta

APROBACIÓN DE DIRECCIÓN Y COORDINACIÓN DEL PROGRAMA

Nosotros, **Alejandro Cortés Director EIG** y **Karla Mora Coordinadora UIDE**, declaramos que: **Cristian Fernando Tenecela Cabay, Eutimio Napoleón Arévalo Valarezo, José Miguel Ocaña Chiluisa, Yessenia Denisse Carrillo Toapanta** son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés

Director de la Maestría en Ciencia de datos y máquinas aprendizaje con en mención Inteligencia Artificial



Karla Mora

Coordinadora de la Maestría en Ciencia de datos y máquinas aprendizaje con en mención Inteligencia Artificial

DEDICATORIA

A nuestros profesores y tutores, por su guía constante, su paciencia y sus valiosos comentarios y retroalimentaciones; su sabiduría y confianza han sido fundamentales en este proceso, y siempre llevaremos con nosotros las enseñanzas y el respaldo que nos brindaron.

A nuestras familias, por su apoyo incondicional y motivación diaria, este logro no es solo nuestro sino de todos. Y a nuestros compañeros, por los intercambios de ideas, por el conocimiento y la pasión con la que compartimos este camino y por haberlo hecho una experiencia enriquecedora.

AGRADECIMIENTOS

Expresamos nuestro sincero agradecimiento a todas las personas que hicieron posible la culminación de este camino académico.

A los docentes de la Maestría en Ciencia de Datos y Aprendizaje de Máquinas con mención en Inteligencia Artificial de la Universidad Internacional del Ecuador (UIDE), quienes, con su dedicación y compromiso, nos transmitieron valiosos conocimientos y se convirtieron en una fuente de inspiración.

A nuestros directores de tesis, por su orientación y apoyo constante en cada etapa de este proyecto de investigación.

A nuestros compañeros de equipo, por el esfuerzo compartido, la constancia y la colaboración que hicieron posible alcanzar este objetivo en conjunto.

De manera muy especial, a nuestras familias, por ser nuestro pilar de apoyo, brindarnos comprensión, cariño y aliento en los momentos de mayor exigencia, motivándonos siempre a perseverar hasta la culminación de este reto.

RESUMEN

La presente investigación examina y contrasta los marcos de clasificación empleados en los datos sísmicos con el objetivo de pronosticar las erupciones en el volcán Cotopaxi, uno de los volcanes más activos y peligrosos de Ecuador. La importancia de la investigación reside en la necesidad de contar con sistemas predictivos más precisos que las metodologías convencionales, capaces de mitigar los riesgos sociales, económicos y ambientales asociados a los fenómenos volcánicos. El objetivo principal era evaluar la viabilidad de los algoritmos de aprendizaje automático para predecir escenarios eruptivos con una precisión superior al 70%. La metodología se basó en el marco CRISP-ML (Q), que consistía en comprender el problema, preparar datos sísmicos históricos y en tiempo real del IG-EPN y llevar a cabo la capacitación y la validación del modelo. Algoritmos como Random Forest, SVM, XGBoost y MLP se yuxtapusieron y evaluaron utilizando parámetros de precisión, recuperación, puntuación de F1 y AUC-ROC. Los hallazgos indican que los modelos de aprendizaje profundo, en particular las redes neuronales recurrentes, muestran un rendimiento superior a la hora de discernir los patrones precursores. Se concluye que la incorporación de la inteligencia artificial en la vulcanología constituye una contribución científica y social fundamental para mejorar los sistemas de alerta temprana en Ecuador.

Palabras clave: aprendizaje automático, clasificación, predicción sísmica, Cotopaxi, CRISP-ML(Q).

ABSTRACT

This research examines and compares the classification frameworks used in seismic data with the aim of predicting eruptions at Cotopaxi volcano, one of Ecuador's most active and dangerous volcanoes. The importance of the research lies in the need for predictive systems that are more accurate than conventional methodologies and capable of mitigating the social, economic, and environmental risks associated with volcanic phenomena. The main objective was to evaluate the feasibility of machine learning algorithms to predict eruptive scenarios with an accuracy greater than 70%. The methodology was based on the CRISP-ML (Q) framework, which consisted of understanding the problem, preparing historical and real-time seismic data from the IG-EPN, and carrying out model training and validation. Algorithms such as Random Forest, SVM, XGBoost, and MLP were juxtaposed and evaluated using precision, recall, F1 score, and AUC-ROC parameters. The findings indicate that deep learning models, particularly recurrent neural networks, show superior performance in discerning precursor patterns. It is concluded that the incorporation of artificial intelligence in volcanology constitutes a fundamental scientific and social contribution to improving early warning systems in Ecuador.

Keywords: machine learning, classification, seismic prediction, Cotopaxi, CRISP-ML(Q).

TABLA DE CONTENIDOS

CERTIFICACIÓN DE AUTORÍA	i
AUTORIZACIÓN DE DERECHOS DE PROPIEDAD INTELECTUAL	ii
ACUERDO DE CONFIDENCIALIDAD	¡Error! Marcador no definido.
APROBACIÓN DE DIRECCIÓN Y COORDINACIÓN DEL PROGRAMA.....	iii
DEDICATORIA	iv
AGRADECIMIENTOS	v
RESUMEN	vi
ABSTRACT	vii
TABLA DE CONTENIDOS	viii
LISTA DE TABLAS	x
LISTA DE FIGURAS	xi
CAPITULO 1	1
1. INTRODUCCIÓN	1
Justificación e importancia del trabajo de investigación	3
1.1 Alcance	5
Componentes del alcance.....	6
1.2 Idea a defender	8
1.3 Objetivos	8
1.4 Objetivo general	8
1.5 Objetivos específicos.....	8
2. REVISIÓN DE LITERATURA.....	10
2.1 Estado del Arte.....	10
2.2 Contexto geodinámico y volcanismo en Ecuador	14
2.3 Volcanes activos en Ecuador	14
2.4 Riesgo volcánico y sus implicaciones sociales	16
2.5 Riesgo social y humano de la actividad volcánica	16
2.6 Señales precursoras de erupciones volcánicas	16
2.7 Modelos automáticos de detección y clasificación	17
2.8 Integración de ciencia de datos e inteligencia artificial.....	18
2.9 Metodología CRISP-ML(Q).....	18
2.10 Comprensión del problema.....	19
2.11 Comprensión y preparación de datos	19

2.12	Selección e ingeniería de modelos	20
2.13	Validación y aseguramiento de calidad	20
2.14	Implementación y escalabilidad	21
CAPITULO 3.....		22
3.	DESARROLLO DEL TRABAJO	22
3.1	Proceso ETL	22
3.2	Clasificación Peligrosidad.....	24
3.3	Balanceo de Clases.....	25
3.4	Entrenamiento.....	27
3.4.1	Random Forest.....	30
3.5.2	XGBoost.....	30
3.5.3	MLP.....	31
3.5.4	SVM	32
CAPITULO 4.....		34
4.	ANÁLISIS DE RESULTADOS.....	34
4.1	Resultados del Pre procesamiento	34
4.2	Resultados del entrenamiento de los modelos	41
CAPITULO 5.....		70
5.1	CONCLUSIONES.....	70
5.2	RECOMENDACIONES	73
APÉNDICES		78

LISTA DE TABLAS

Tabla 1	<i>Listado de volcanes Potencialmente activos en el Ecuador</i>	15
Tabla 2	<i>Criterios para clasificación de categoría</i>	24
Tabla 3	<i>Componentes de Hiperparámetros Random Forest</i>	30
Tabla 4	<i>Componentes de Hiperparámetros XGBoost</i>	31
Tabla 5	<i>Resumen del reporte de clasificación de todos los modelos</i>	62

LISTA DE FIGURAS

Figura 1 <i>Volcanes del Ecuador</i>	15
Figura 2 <i>Reporte interactivo con la funcion ydata_profiling</i>	23
Figura 3 <i>Validación de criterios de peligro</i>	25
Figura 4 <i>Distribución de tipos de magnitud</i>	34
Figura 5 <i>Distribución de las magnitudes Mw</i>	35
Figura 6 <i>Frecuencia de magnitudes</i>	36
Figura 7 <i>Resumen proceso ETL</i>	37
Figura 8 <i>Resultado clasificación binaria con eventos reales</i>	37
Figura 9 <i>Clasificación de sismos considerados de Potencial Peligro</i>	38
Figura 10 <i>Resultado del balanceo de clases</i>	39
Figura 11 <i>Sismos sintéticos generados como Potencial Peligro</i>	40
Figura 12 <i>Mejores hiperparametros para random forest</i>	42
Figura 13 <i>Reporte de Clasificación y Matriz de Confusión de Random Forest</i>	43
Figura 14 <i>Curva ROC- Random Forest</i>	44
Figura 15 <i>Curva Precision-Recall Random Forest</i>	45
Figura 16 <i>Curva de Aprendizaje-Random Forest</i>	46
Figura 17 <i>Mejores hiperparametros de xgboost</i>	47
Figura 18 <i>Reporte de Clasificación y Matriz de Confusión de XGBoost</i>	48
Figura 19 <i>Curva ROC-XGBoost</i>	49
Figura 20 <i>Curva Precision-Recall XGBoost</i>	50
Figura 21 <i>Curva de Pérdida (Loss)-XGBoost</i>	51
Figura 22 <i>Curva de Accuracy- XGBoost</i>	52
Figura 23 <i>Mejores hiperparámetros de red neuronal-MLP</i>	53
Figura 24 <i>Reporte de Clasificación y Matriz de Confusión MLP Classifier</i>	54
Figura 25 <i>Curva Precision-Recall MLP</i>	55
Figura 26 <i>Curva ROC- MLP Classifier</i>	56
Figura 27 <i>Curva de Pérdida MLP Classifier</i>	57
Figura 28 <i>Validación de Exactitud MLP Classifier</i>	58
Figura 29 <i>Mejores hiperparámetros para SVM</i>	59
Figura 30 <i>Reporte de Clasificación y Matriz de Confusión SVM</i>	60
Figura 31 <i>Curva de Aprendizaje SVM</i>	61
Figura 32 <i>Comparativa mejor modelo</i>	63
Figura 33 <i>Interfaz del aplicativo</i>	64
Figura 34 <i>Interfaz carga de archivo</i>	65
Figura 35 <i>Interfaz con clasificación de eventos sísmicos</i>	66
Figura 36 <i>Interfaz con gráfica de distribución</i>	67
Figura 37 <i>Interfaz con histograma de probabilidad</i>	68
Figura 38 <i>Interfaz con tabla de eventos críticos</i>	68

CAPÍTULO 1

1. INTRODUCCIÓN

El avance tecnológico en microelectrónica, informática, computación y telecomunicaciones ha llevado a un crecimiento exponencial en la capacidad de generar y almacenar datos. Esta gran cantidad de datos ha impulsado la necesidad de técnicas que transformen grandes volúmenes de información en conocimiento útil, siendo el Descubrimiento de Conocimiento en Bases de Datos (KDD) y el Data Mining técnicas emblemáticas en este sentido (Ramos, 2019). En el contexto de la gestión de desastres naturales, específicamente la predicción de erupciones volcánicas, la sismología volcánica se ha consolidado como una de las herramientas más cruciales para el monitoreo y la predicción a corto plazo. Los procesos internos de los volcanes generan señales sísmicas variadas que son fundamentales para entender la dinámica eruptiva (Perales Palacios et al., 2021).

La integración de técnicas avanzadas de inteligencia artificial y aprendizaje automático ha revolucionado el campo de la volcanología predictiva, permitiendo el procesamiento de grandes volúmenes de datos sísmicos en tiempo real (Carniel & Guzmán, 2020; Carniel & Guzmán, 2025).

Este proyecto de investigación se orienta al análisis de datos sísmicos mediante técnicas estadísticas y de aprendizaje automático, con el fin de detectar y caracterizar precursores de erupciones volcánicas en Ecuador. La propuesta se fundamenta en la creciente necesidad de contar con modelos predictivos que integren señales sismo-volcánicas, tales como eventos volcano-tectónicos (VT), de largo periodo (LP), tremores y sismos híbridos, en un sistema capaz de reconocer patrones anómalos y anticipar escenarios eruptivos.

Estudios recientes han demostrado que los modelos de aprendizaje automático pueden detectar hasta 7.5 veces más eventos sísmicos que los métodos manuales tradicionales,

mejorando significativamente la capacidad de identificar precursores eruptivos (Yukutake et al., 2023).

Este proyecto se define como una investigación aplicada que utilizará técnicas de análisis de datos, incluyendo métodos estadísticos y de aprendizaje automático (machine learning), para examinar los datos sísmicos históricos y en tiempo real de los volcanes más activos de Ecuador. El objetivo es identificar y caracterizar patrones precursores de actividad eruptiva. Por ejemplo, la aplicación de redes neuronales y algoritmos de aprendizaje automático ha permitido reconocer patrones de energía sísmica y caídas de entropía que anteceden a erupciones en volcanes como Etna, Kilauea y Augustine (Rey-Devesa, Benítez, et al., 2023). Además, enfoques innovadores como el aprendizaje por transferencia han demostrado ser efectivos para mejorar el pronóstico en volcanes con datos escasos, generalizando conocimiento adquirido de múltiples sistemas volcánicos (Ardid et al., 2025).

Se analizarán distintos tipos de señales sismo-volcánicas registradas por la red de monitoreo del Instituto Geofísico de la Escuela Politécnica Nacional (IG-EPN), institución que ha liderado el monitoreo volcánico en Ecuador durante cuatro décadas, enfrentando desafíos instrumentales y desarrollando capacidades técnicas para la vigilancia de volcanes de alto riesgo como Cotopaxi (Hidalgo et al., 2023).

De esta manera, se pretende disponer de un modelo que permita alcanzar un nivel de exactitud superior al 70% en la predicción de escenarios eruptivos, contribuyendo al desarrollo de nuevas aplicaciones tecnológicas orientadas a la gestión de riesgos volcánicos en el país. La implementación operativa de sistemas de monitoreo en tiempo real basados en machine learning, como los desarrollados para el volcán Axial Seamount, demuestra la viabilidad de estos enfoques para sistemas de alerta temprana (Wang et al., 2024).

Justificación e importancia del trabajo de investigación

El Ecuador alberga 35 volcanes potencialmente activos, cuya dinámica representa una amenaza latente para la población, la infraestructura y la economía del país (Vaca Molina, 2024). Estudios recientes han permitido establecer un Índice de Peligrosidad Volcánica (IPV) preliminar que clasifica a los principales volcanes en diferentes categorías de riesgo. Sin embargo, este tipo de indicadores se fundamenta principalmente en parámetros históricos y geológicos, lo que limita su capacidad predictiva frente a la evolución en tiempo real de la actividad sísmica volcánica.

Bajo este contexto, se considera necesario complementar el enfoque tradicional con técnicas modernas de aprendizaje automático y modelos de clasificación, capaces de identificar patrones complejos en los registros sísmicos tectónico-volcánicos y de aportar mayor precisión en la detección temprana de escenarios eruptivos. Modelos generalizados entrenados con datos de múltiples volcanes han demostrado capacidad para detectar precursores sísmicos en sistemas volcánicos no incluidos en el conjunto de entrenamiento, superando métodos basados únicamente en amplitud promedio (Ardid et al., 2023). La comparación sistemática de diferentes algoritmos de clasificación permitirá reconocer cuáles ofrecen mejores resultados en términos de exactitud, robustez y reducción de sesgos, aportando bases científicas sólidas para el desarrollo de sistemas predictivos confiables.

Desarrollar este proyecto como aplicación en volcanes de alta peligrosidad permitirá generar conocimiento extrapolable a otros escenarios volcánicos, consolidando un aporte tanto científico como social en la prevención de desastres naturales en Ecuador.

El presente proyecto adquiere relevancia por tres razones principales:

- *Riesgo social y humano:* La cercanía de los principales volcanes a zonas densamente pobladas convierte la predicción temprana en un factor determinante para salvar vidas. La experiencia de erupciones repentinas, como la de Whakaari en 2019,

evidencia la necesidad de modelos de alerta basados en datos y no únicamente en juicios expertos (Dempsey et al., 2020). En Ecuador, volcanes como el Cotopaxi, Tungurahua y Sangay han demostrado su potencial destructivo a lo largo de la historia. Un pronóstico más preciso y oportuno de una erupción puede salvar innumerables vidas al permitir evacuaciones preventivas y bien organizadas. Sistemas integrados de alerta temprana que combinan umbrales estadísticos de precursores con difusión automatizada de alarmas han demostrado su utilidad en contextos operativos (Spina et al., 2020).

- *Impacto económico:* Las erupciones volcánicas tienen repercusiones directas en la agricultura, transporte aéreo y turismo. La detección temprana de señales precursoras, como la variación de la entropía de Shannon en registros sísmicos, posibilita reducir pérdidas al activar protocolos preventivos con suficiente antelación (Rey-Devesa et al., 2023). Una erupción puede paralizar sectores económicos clave como la agricultura, la ganadería, el turismo y la aviación, debido a la caída de ceniza y los flujos laháricos. Este proyecto busca mitigar el impacto económico al proporcionar información que permita a las industrias y al gobierno prepararse y proteger sus activos.
- *Avance científico y tecnológico:* La evidencia sugiere que muchos procesos eruptivos presentan señales precursoras sutiles y de largo plazo que no son detectadas por métodos tradicionales. El análisis del nivel sísmico de fondo en Shinmoe-dake, Japón, reveló que la preparación eruptiva puede extenderse por meses y ser registrada solo mediante técnicas avanzadas de procesamiento (Ichihara et al., 2023b). La implementación de modelos predictivos avanzados basados en datos sísmicos posicionaría a Ecuador a la vanguardia de la vulcanología en la región. Además, este proyecto contribuirá a un entendimiento más profundo de la dinámica interna de los volcanes andinos, generando conocimiento valioso para la comunidad científica

internacional, al tiempo de fortalecer las capacidades del IG-EPN, la institución encargada del monitoreo, fundamental para la seguridad nacional. Técnicas innovadoras como el emparejamiento de plantillas multiescala han permitido descubrir precursores eruptivos recurrentes en escalas temporales de 1 a 60 días en diversos contextos volcánicos (Ardid et al., 2024).

Problema de investigación

El pronóstico temprano de las erupciones volcánicas en Ecuador sigue planteando un problema, a pesar de la accesibilidad de la información sísmica histórica y en tiempo real. Las técnicas de monitoreo convencionales basadas en parámetros geológicos e históricos proporcionan información valiosa; sin embargo, presentan dificultades para pronosticar con precisión la incidencia de una erupción. Esta deficiencia complica la ejecución de medidas preventivas oportunas, lo que amplifica los riesgos sociales, económicos y ambientales en el contexto de fenómenos volcánicos importantes, como los relacionados con el Cotopaxi. En consecuencia, existe la necesidad de investigar la utilización de modelos de clasificación basados en el aprendizaje automático para discernir los patrones sísmicos precursores y evaluar su viabilidad como instrumento predictivo a corto plazo.

1.1 Alcance

El estudio se centrará en el volcán Cotopaxi como caso principal, dada su alta peligrosidad y la disponibilidad de registros sísmicos históricos y recientes provistos por el Instituto Geofísico de la Escuela Politécnica Nacional (IG-EPN). Se analizarán tanto periodos de calma como fases de reactivación, en especial la crisis eruptiva de 2015, para entrenar y validar modelos de clasificación.

Aunque el énfasis inicial será en Cotopaxi, los resultados serán extrapolables a otros volcanes activos del país, como Tungurahua y Reventador, permitiendo diseñar sistemas predictivos replicables a nivel regional.

Aunque el énfasis inicial será en Cotopaxi, los resultados serán extrapolables a otros volcanes activos del país, como Tungurahua y Reventador, permitiendo diseñar sistemas predictivos replicables a nivel regional. La experiencia exitosa del modelo de pronóstico desarrollado para el volcán Copahue en los Andes del Sur, que logró emitir alertas pseudo-prospectivas entre 5 y 75 horas antes de erupciones pequeñas, demuestra la viabilidad de este enfoque en contextos andinos similares (Cabrera et al., 2024).

Componentes del alcance

1. *Volcán de Estudio Principal:* Se propone iniciar el análisis con el volcán Cotopaxi, debido a su historial eruptivo y al elevado riesgo que representa para áreas estratégicas del país, incluyendo el Distrito Metropolitano de Quito y la ciudad de Latacunga.
2. *Datos para Utilizar:* Se utilizarán las bases de datos sísmicos históricos y los datos en tiempo real proporcionados y curados por el IG-EPN. Esto incluye catálogos de eventos sísmicos clasificados y formas de onda continuas de las estaciones de monitoreo del Cotopaxi.
3. *Período de Análisis:* El estudio abarcará los datos sísmicos recopilados durante los periodos de calma y de reactivación del volcán, con especial énfasis en el proceso eruptivo de 2015 para entrenar y validar los modelos.
4. *Resultados Esperados:* El proyecto se limitará a la creación y validación de un modelo predictivo. La implementación de este modelo en un sistema de alerta temprana en tiempo real quedará fuera del alcance de esta fase, pero se sentarán las bases para su futuro desarrollo.

Se ha identificado la relevancia de los volcanes potencialmente activos y el impacto de contar con modelos predictivos confiables en el contexto del problema volcánico en

Ecuador, Paralelamente, se revisan las fuentes de datos disponibles y se determina su utilidad para el entrenamiento de modelos de clasificación.

Ingeniería de datos (preparación de datos): Se realizará el preprocesamiento necesario para transformar los datos crudos en insumos utilizables por los modelos. Esto incluye limpieza de registros, segmentación de eventos sísmicos, extracción de características, balanceo de clases y construcción del dataset final con las etiquetas de clasificación.

Ingeniería de modelos de aprendizaje automático: Se diseñan y entrenan diferentes algoritmos de clasificación (Random Forest, SVM, XGBoost, LightGBM, CNN, LSTM/GRU), ajustando hiperparámetros y evaluando su rendimiento. Esta etapa tiene un carácter experimental y comparativo, ya que se busca determinar qué modelo ofrece mejores métricas de precisión, recall, F1-score y AUC-ROC en la predicción de actividad eruptiva.

Estudios comparativos han demostrado que diferentes arquitecturas (RF, ANN, CNN, LSTM) presentan sensibilidades variables al ruido y desempeños relativos distintos según las características del conjunto de datos (Mustafa et al., 2022).

Garantía de calidad para aplicaciones de aprendizaje automático: Se validarán los modelos obtenidos bajo un esquema de pruebas, verificando su robustez, ausencia de sobreajuste y capacidad de generalización. Se emplearán técnicas de validación mismas que permitirán seleccionar el mejor modelo de clasificación para el escenario planteado.

Implementación: Se integrará el modelo con mejor clasificación en una aplicación piloto o prototipo, para que el público objetivo pueda verificar el funcionamiento.

Monitoreo y Mantenimiento: Después del despliegue se realizarán pruebas necesarias para que el aplicativo prototipo funcione correctamente, donde en caso de presentarse fallos puedan realizarse los ajustes necesarios.

1.2 Idea a defender

Con la evaluación de la información sísmica utilizando metodologías de categorización y aprendizaje computacional puede discernir los motivos precursores de las erupciones y lograr grados de precisión superiores al 70% en la predicción a corto plazo del comportamiento eruptivo del volcán Cotopaxi, lo que constituye una alternativa factible para mejorar los sistemas de vigilancia volcánica y contribuir a la reducción del riesgo en Ecuador.

1.3 Objetivos

1.4 Objetivo general

Evaluar la factibilidad de modelos de clasificación basado en datos sísmicos para la identificación de escenarios de amenazas volcánicas en el volcán Cotopaxi, como apoyo a la mitigación del riesgo volcánico en Ecuador.

1.5 Objetivos específicos

- Caracterizar estadísticamente la sismicidad de eventos eruptivos históricos del volcán Cotopaxi mediante el análisis de sismicidad, y energía liberada.
- Comparar diferentes modelos de clasificación aplicados a datos sísmicos, con el fin de identificar el modelo más adecuado para la obtención de una alerta temprana de escenarios de amenazas volcánicas
- Comprobar el desempeño del modelo de clasificación seleccionado utilizando un conjunto de datos de prueba independiente, considerando métricas como precisión, sensibilidad, especificidad y tasa de falsas alarmas.

El presente documento se organiza en cinco capítulos. El *Capítulo 1* introduce la temática, justifica la investigación y establece los objetivos. El *Capítulo 2* realiza una revisión

de la literatura relevante, abarcando el estado del arte y el marco teórico sobre sismología volcánica, procesamiento de señales, Machine Learning y gestión de riesgos. El *Capítulo 3* detalla la metodología propuesta, siguiendo un enfoque estructurado. El *Capítulo 4* presenta el análisis y discusión de los resultados obtenidos. Finalmente, el *Capítulo 5* expone las conclusiones del trabajo y sugiere líneas de investigación futuras.

CAPÍTULO 2

2. REVISIÓN DE LITERATURA

2.1 Estado del Arte

La predicción de erupciones volcánicas ha evolucionado gracias a la integración de métodos sismológicos avanzados y técnicas de inteligencia artificial. Tradicionalmente, la identificación de precursores se basaba en el análisis manual de catálogos sísmicos y en la experiencia de los observatorios locales. Sin embargo, la complejidad y diversidad de los procesos volcánicos han impulsado la adopción de enfoques automatizados y basados en datos. Las revisiones recientes del campo destacan la transición desde métodos heurísticos hacia sistemas predictivos basados en evidencia cuantitativa y aprendizaje de patrones (Carniel & Guzmán, 2020).

Estudios recientes han demostrado la utilidad de aplicar métricas de información como la entropía de Shannon, la cual tiende a decrecer horas antes de una erupción. Esta metodología fue validada en volcanes como Etna, Kilauea y Mount St. Helens, mostrando que los cambios en la entropía, combinados con indicadores de energía y frecuencia, pueden anticipar escenarios eruptivos con varias horas de antelación (Rey-Devesa et al., 2023).

Investigaciones específicas en el Volcán de Colima han confirmado que la disminución de la entropía de Shannon en señales sísmicas continuas puede preceder explosiones con 2 a 6 días de anticipación, consolidando esta métrica como herramienta complementaria para el monitoreo operativo (Rey-Devesa et al., 2023). Herramientas computacionales desarrolladas recientemente en Python permiten estimar en tiempo real la entropía de Shannon, curtosis, índice de frecuencia y energía sobre registros sísmicos, facilitando el análisis de eventos VT, LP y tremores (Gutiérrez et al., 2025).

Asimismo, para Ichihara et al. (2023) la detección de señales precursoras de baja amplitud ha cobrado relevancia. Investigaciones en el volcán Shinmoe-dake (Japón) evidenciaron que el análisis del nivel sísmico de fondo (SBL) permite identificar procesos de preparación eruptiva que se desarrollan a lo largo de meses, incluso cuando no hay señales visibles en la superficie. Este hallazgo sugiere que parte de los “falsos negativos” en la vigilancia volcánica se deben a la falta de metodologías sensibles a estas variaciones.

Enfoques basados en ruido sísmico ambiental, incluyendo interferometría y análisis de variaciones de velocidad (dv/v), están siendo explorados para detectar cambios precursores a erupciones freáticas, con propuestas de integrar machine learning para mejorar la interpretación (Lecocq et al., 2023).

Otro avance significativo proviene del uso de aprendizaje automático supervisado y no supervisado. En Whakaari (Nueva Zelanda), un modelo de machine learning logró detectar secuencias energéticas que ocurrían entre horas y días antes de erupciones, emitiendo alertas tempranas en tiempo real. Este estudio constituye un caso pionero de implementación práctica de algoritmos predictivos en sistemas de alerta volcánica (Dempsey et al., 2020).

Para Ardid et al. (2023) los enfoques de aprendizaje por transferencia han demostrado ser particularmente prometedores para volcanes con datos escasos. Modelos entrenados con registros sísmicos de 24 volcanes diferentes han logrado detectar precursores generales y predecir erupciones en volcanes no incluidos en el conjunto de entrenamiento, superando métodos basados únicamente en amplitud promedio.

El uso de estadística ergódica combinada con transferencia de aprendizaje ha mejorado significativamente el pronóstico a corto plazo mediante generalización entre volcanes similares (Ardid et al., 2025).

El análisis de clústeres de tremores ha permitido descubrir patrones que los métodos clásicos pasaban por alto. Durante la erupción del Geldingadalir en 2021, el uso de técnicas de deep learning sobre datos sísmicos continuos permitió identificar secuencias precursoras tres días antes del inicio de la erupción, así como las fases de transición entre efusión continua de lava y pulsos de fountaining (Mousavi et al., 2023). Redes híbridas que combinan scattering y ConvLSTM han demostrado capacidad para detectar comportamiento metastable en señales sísmicas polifónicas, con aplicabilidad para transferencia entre diferentes sistemas volcánicos (Bueno et al., 2021).

Finalmente, los avances en la clasificación y localización de señales sismo-volcánicas también han fortalecido la predicción. El análisis multicomponente en el Piton de la Fournaise (La Reunión) mostró que los enjambres sísmicos migran desde niveles profundos hacia zonas someras antes de una erupción, información crucial para estimar la dinámica internacional (Journeau et al., 2020).

La evidencia internacional también incluye el uso de redes sísmicas locales en la vigilancia volcánica. En Bali, (Syahbana et al., 2019) analizaron más de 2.700 eventos volcano-tectónicos previos a la erupción del Agung, diferenciando entre fracturamiento de roca y migración de magma, lo que permitió reconstruir la secuencia eruptiva con mayor detalle.

Asimismo, Yukutake et al. (2023) mostraron que los modelos de machine learning entrenados con catálogos sísmicos pueden detectar hasta 7,5 veces más eventos volcánicos que los métodos manuales. En el Kirishima (Japón), esta técnica permitió identificar hipocentros de enjambres y variaciones de b-values previas a erupciones magmáticas y freáticas. Para, Fenner et al. (2022) el desarrollo de herramientas de detección automática también ha beneficiado a volcanes con actividad estromboliana el módulo AWESAM, capaz de detectar cientos de miles de eventos en Stromboli, Italia, sin necesidad de grandes catálogos de entrenamiento, alcanzando una precisión del 95% en comparación con registros ma-

nuales. Métodos de análisis espectral continuo operando en tiempo casi real han sido desarrollados para la red sísmica del Etna, permitiendo detección de anomalías y estimación indicativa de intensidad eruptiva potencial (Ollino, 2022).

Finalmente, los avances en modelos de riesgo sísmico y volcánico en contextos urbanos complementan estos desarrollos. En Quito, Córdoba et al. (2025) elaboraron un modelo de riesgo probabilista que integra escenarios volcánicos y sísmicos, aportando insumos clave para la planificación preventiva en ciudades altamente vulnerables.

Aplicaciones específicas en la región andina han demostrado la viabilidad de estos enfoques. Un pipeline de machine learning aplicado al volcán Copahue integró extracción de precursores sísmicos y decisión de alerta, reportando alertas pseudo-prospectivas entre 5 y 75 horas antes de erupciones pequeñas (Cabrera et al., 2024).

Diseños híbridos que combinan series temporales y machine learning para predicción eruptiva en volcanes como Merapi ilustran flujos de preprocesamiento y validación adaptables a conjuntos de datos ecuatorianos (Mandita et al., 2024). Estudios comparativos de diferentes arquitecturas de redes neuronales han evaluado el desempeño de RBFNN, ANN, CNN y LSTM para predicción de erupciones, mostrando sensibilidades variables al ruido y desempeños relativos distintos según las características del conjunto de datos (Shyamala et al., 2022; Mustafa et al., 2022). La aplicación de Random Forests para predicción eruptiva en el Anillo de Fuego del Pacífico ha sido evaluada en grandes conjuntos regionales, discutiendo su eficacia frente a otras técnicas de clasificación (Kaza et al., 2025).

Sistemas operativos de monitoreo en tiempo real

La implementación de sistemas operativos de monitoreo volcánico ha avanzado significativamente. Sistemas integrados de alerta temprana que combinan umbrales estadísticos de precursores con difusión automatizada de alarmas a dispositivos móviles ofrecen referencias de arquitectura operativa (Spina et al., 2020). La construcción de catálogos sísmicos

etiquetados con machine learning y el despliegue de monitoreo en tiempo real, integrando aprendizaje no supervisado para discriminación de señales, han demostrado su utilidad en volcanes submarinos como Axial Seamount (Wang et al., 2024).

Herramientas de inteligencia artificial y machine learning para mejorar sistemas de alerta temprana han sido aplicadas exitosamente en Stromboli, incluyendo predicción de intensidad de eventos explosivos basándose en deformación y series temporales locales (Longo et al., 2024). Enfoques bayesianos de aprendizaje profundo para detección y clasificación de transientes en flujo continuo han demostrado capacidad para detectar cambios previos a erupciones, mostrando utilidad para monitoreo probabilístico (IEEE, 2022).

Marco teórico

2.2 Contexto geodinámico y volcanismo en Ecuador

El Ecuador se encuentra en un entorno tectónico complejo, dominado por la subducción de la placa de Nazca bajo la Sudamericana. Este proceso genera una zona de Wadati-Benioff caracterizada por sismicidad intermedia y superficial, así como por un notable volcanismo en la región andina (Hanu & Vanek, 1987). La variabilidad en la inclinación de la losa subducida y la presencia de zonas asísmicas se relacionan directamente con el emplazamiento de volcanes activos como Cotopaxi, Tungurahua y Reventador, cuya proximidad a zonas densamente pobladas incrementa la vulnerabilidad social y económica.

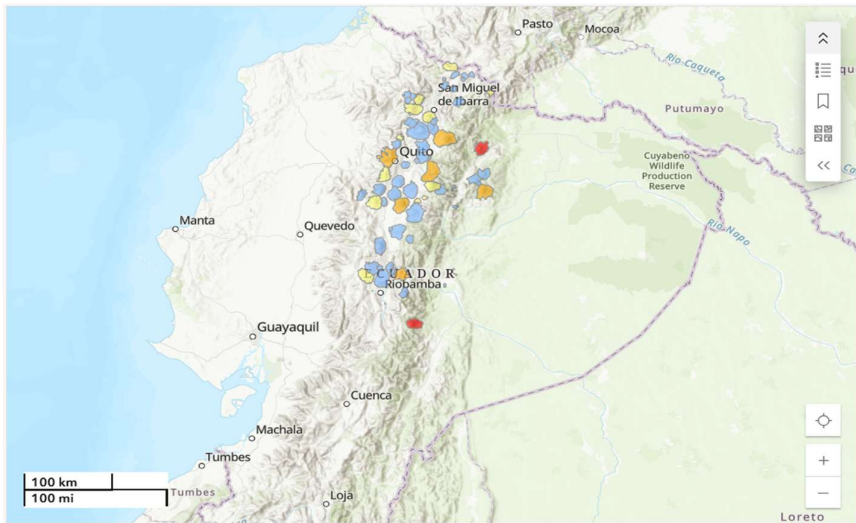
2.3 Volcanes activos en Ecuador

Ecuador se ubica en un contexto geodinámico complejo, influenciado por la subducción de la placa de Nazca bajo la Sudamericana. Este proceso da lugar a la formación de una zona de Wadati-Benioff y a la presencia de más de 30 volcanes potencialmente activos,

entre los que destacan Cotopaxi, Tungurahua, Sangay y Reventador (Chlieh, 2021). La proximidad de estos sistemas volcánicos a áreas densamente pobladas incrementa la vulnerabilidad social y económica del país.

El monitoreo sistemático de estos volcanes ha sido liderado por el Instituto Geofísico de la Escuela Politécnica Nacional (IG-EPN) durante cuatro décadas, enfrentando desafíos instrumentales significativos mientras desarrolla capacidades técnicas para la vigilancia de volcanes de alto riesgo (Hidalgo et al., 2023).

Figura 1
Volcanes del Ecuador



Nota. Adaptado de Instituto Geográfico Militar (2025).

Tabla 1
Listado de volcanes Potencialmente activos en el Ecuador

Ubicación	Cantidad	Punto A
Volcanes continentales	7	Cayambe, Reventador, Guagua Pichincha, Cotopaxi, Tungurahua, Sangay y Potrerillos-Chacana
Volcanes de Galápagos	7	Marchena, Cerro Azul, Fernandina, Santo Tomás Chico, Alcedo, Darwin y Wolf

Nota. Adaptado de Instituto Geográfico Militar (2025).

2.4 Riesgo volcánico y sus implicaciones sociales

La actividad volcánica representa una amenaza significativa para la seguridad de las comunidades y la infraestructura crítica. Estudios de riesgo en ciudades como Quito han integrado escenarios sísmicos y volcánicos, demostrando que la exposición urbana y la concentración poblacional amplifican los posibles impactos (Córdoba et al., 2025). Casos internacionales, como la erupción súbita de Whakaari en 2019, han puesto en evidencia que los sistemas de alerta basados solo en criterios heurísticos son insuficientes, dado que muchas erupciones ocurren con poca o ninguna señal perceptible en superficie (Dempsey et al., 2020)

2.5 Riesgo social y humano de la actividad volcánica

La historia eruptiva reciente del Ecuador muestra que fenómenos como la caída de ceniza, flujos piroclásticos y lahares generan impactos directos sobre la población, la infraestructura crítica y la economía nacional. En este sentido, contar con sistemas predictivos confiables permite mitigar pérdidas humanas y materiales. Investigaciones internacionales resaltan que las erupciones súbitas, como la de Whakaari en 2019, subrayan la urgencia de modelos de alerta más dinámicos y automáticos (Dempsey et al., 2020).

2.6 Señales precursoras de erupciones volcánicas

Los precursores volcánicos se manifiestan como anomalías sísmicas, geoquímicas o geodésicas que anteceden a una erupción. Entre las más estudiadas se encuentran los eventos volcano-tectónicos (VT), los de largo periodo (LP) y los tremores armónicos. La literatura reciente destaca la utilidad de nuevas métricas en su caracterización.

En primer lugar, la entropía de Shannon, que decrece de manera sistemática previo a erupciones, permitiendo anticipar la transición hacia estados críticos; aplicaciones específicas han demostrado pronósticos exitosos con 2 a 6 días de anticipación en volcanes como Colima (Rey-Devesa et al., 2023).

Como segunda característica, el nivel sísmico de fondo (SBL), que ha revelado procesos eruptivos de lenta evolución no detectables con métodos convencionales, como en el caso del Shinmoe-dake en Japón (Ichihara et al., 2023b).

En tercer lugar, la identificación de enjambres sísmicos migratorios, observados en La Reunión, que marcan la intrusión de magma hacia zonas someras antes del inicio eruptivo (Journeau et al., 2020).

Y finalmente, el registro de secuencias prolongadas de sismicidad, como en el Agung (Indonesia), donde miles de eventos VT precedieron la erupción de 2017, distinguiendo entre fracturamiento de rocas y migración de magma (Syahbana et al., 2019).

El procesamiento efectivo de señales sísmicas volcánicas requiere una comprensión profunda de los pasos clave: filtrado, representación tiempo-frecuencia (TFR), detección de eventos y extracción de características. La integración de machine learning en estas etapas mejora significativamente la discriminación de señales volcánicas y la vigilancia en tiempo real (Khan et al., 2019).

Enfoques basados en ruido sísmico ambiental, incluyendo interferometría y análisis de variaciones de velocidad, están siendo explorados para detectar cambios precursores, con propuestas de integrar modelos numéricos y machine learning para interpretación avanzada (Lecocq et al., 2023).

2.7 Modelos automáticos de detección y clasificación

La necesidad de procesar grandes volúmenes de datos ha impulsado el desarrollo de métodos automáticos de detección:

- Fenner et al. (2022) desarrollaron el módulo AWESAM, que automatiza la identificación de eventos en volcanes estrombolianos como Stromboli, alcanzando un 95% de precisión sin depender de catálogos extensos.

- Yukutake et al. (2023) aplicaron machine learning en el Kirishima (Japón), logrando detectar hasta 7,5 veces más eventos que los métodos tradicionales y revelando clústeres hipocentrales vinculados a procesos eruptivos.
- Zali et al. (2024) mostraron que el clustering de tremores con técnicas de deep learning permite distinguir fases eruptivas y reconocer secuencias precursoras con varios días de antelación.

Estos avances evidencian que los algoritmos supervisados y no supervisados (Random Forest, SVM, CNN, LSTM/GRU) se están consolidando como herramientas robustas para la predicción de erupciones, capaces de superar las limitaciones del análisis manual.

2.8 Integración de ciencia de datos e inteligencia artificial

La predicción de erupciones requiere no solo de la detección de señales anómalas, sino también de su integración en modelos predictivos reproducibles y escalables. El uso de metodologías como CRISP-ML(Q) ofrece un marco estructurado para proyectos de machine learning aplicados a volcanología, asegurando la calidad en las fases de preparación de datos, entrenamiento, validación y despliegue. En este enfoque, la detección temprana se convierte en una tarea de clasificación y predicción probabilística, reduciendo la incertidumbre y mejorando la gestión del riesgo volcánico.

Revisiones comprehensivas de métodos de machine learning aplicables a volcanología discuten vectores de características y enfoques para combinación de múltiples tipos de datos en observatorios, cubriendo clasificación, reducción de dimensión y aplicaciones a geofísica, geodesia y teledetección (Carniel & Guzmán, 2020, 2025).

2.9 Metodología CRISP-ML(Q)

El desarrollo de modelos predictivos en vulcanología requiere procesos estandarizados que aseguren calidad, reproducibilidad y escalabilidad. Para este fin, se adopta CRISP-

ML(Q), una adaptación del modelo clásico CRISP-DM a proyectos de *machine learning*, recomendado en el diseño metodológico del trabajo base.

Este marco metodológico comprende fases iterativas: comprensión del negocio y de los datos, preparación, modelado, validación, implementación y mantenimiento. Su valor agregado reside en incorporar criterios de calidad (Q) en todas las fases, garantizando que los modelos predictivos no solo alcancen métricas elevadas de precisión, sino que también puedan ser auditados y aplicados en escenarios críticos como la predicción de erupciones volcánicas.

2.10 Comprensión del problema

El objetivo principal es anticipar escenarios eruptivos en volcanes de alto riesgo, como Cotopaxi, a partir del análisis de datos sísmicos. Se establecen criterios de éxito medibles, como precisión superior al 70%, baja tasa de falsas alarmas e interpretabilidad de resultados. La relevancia de estas condiciones radica en que eventos súbitos, como la erupción de Whakaari en 2019, mostraron la insuficiencia de los sistemas tradicionales de alerta (Dempsey et al., 2020).

2.11 Comprensión y preparación de datos

Las fuentes de información incluyen catálogos sísmicos clasificados y series continuas de formas de onda del Cotopaxi. Entre las variables derivadas destacan la entropía de Shannon, validada como indicador precursor en Etna, Kilauea y Mount St. Helens (Rey Delvesa et al., 2023), y el nivel sísmico de fondo (SBL), empleado con éxito en el Shinmoe-dake (Ichihara et al., 2023b). También se consideran parámetros de migración hipocentral y secuencias sísmicas, como las identificadas en Piton de la Fournaise (Journeau et al., 2020) y en el Agung (Syahbana et al., 2021).

El preprocesamiento incluye limpieza de registros, segmentación de señales, balanceo de clases dado que los eventos eruptivos son poco frecuentes y normalización de variables. Este enfoque responde a la problemática identificada en estudios japoneses, donde los algoritmos detectaron hasta 7,5 veces más eventos que los catálogos manuales (Yukutake et al., 2023).

2.12 Selección e ingeniería de modelos

Se aplican y comparan diferentes algoritmos de aprendizaje automático:

- *Random Forest* y *SVM*, eficaces en clasificación de eventos VT y LP (Dempsey et al., 2020).
- *CNN*, apropiadas para la clasificación de formas de onda complejas (mousavi et al., 2023).
- *LSTM/GRU*, útiles en el análisis de series temporales prolongadas (Ichihara et al., 2023b).
- *XGBoost* y *LightGBM*, por su capacidad de optimizar rendimiento en bases de datos de gran escala.

El sistema AWESAM de Stromboli, que detectó con un 95 % de precisión más de 290.000 eventos, refuerza la pertinencia de los enfoques automatizados (Fenner et al., 2022).

2.13 Validación y aseguramiento de calidad

La evaluación se realizará mediante validación cruzada y conjuntos de prueba independientes, empleando métricas como *accuracy*, *recall*, F1-score y AUC-ROC. Para aplica-

ciones críticas, se dará prioridad a la reducción de falsos positivos, ya que una alerta innecesaria puede generar desconfianza en la población y en las autoridades de riesgo (Córdova et al., 2025).

2.14 Implementación y escalabilidad

El modelo seleccionado se integrará en un prototipo de sistema de alerta temprana, capaz de procesar datos en tiempo real. Su escalabilidad permitirá extrapolarlo a otros volcanes activos de la región andina, aprovechando arquitecturas cloud y técnicas de transferencia de aprendizaje. La experiencia exitosa de implementación operativa en volcanes como Axial Seamount demuestra la viabilidad de estos sistemas (Wang et al., 2024). En este sentido, la reproducibilidad se garantiza con documentación abierta, control de versiones y almacenamiento transparente de datos y código, siguiendo principios de calidad establecidos en CRISP-ML(Q).

CAPÍTULO 3

3. DESARROLLO DEL TRABAJO

3.1 Proceso ETL

Se inicia con un proceso de ETL en donde el objetivo es unificar, estandarizar y limpiar varios catálogos de sismicidad histórica para generar un proxy homogéneo de magnitudes Mw. Las librerías utilizadas durante esta fase son pandas, numpy, matplotlib, seaborn, sys y os.

Se inicia con la carga de los 3 catálogos distintos para transformarlos en un formato común. En el catálogo 1901-2009 se seleccionaron las columnas relevantes y se les asignó un nombre estándar, se realiza además una conversión de fecha a formato datetime. En el catálogo 2010_2011 se define el tipo_magnitud, se renombra las variables preincipales y se añade una etiqueta de origen. Finalmente, el catálogo 2012-2025 selecciona solo las columnas existentes, se convierte la fecha y se estandarizan los nombres.

Como siguiente paso se genera una magnitud homogénea Mx_Proxy en escala en donde se vectoriza todas las magnitudes para convertirlas en aproximaciones a Mw. La conversión utilizada fue $Mw = [(0.85 \times mb) + 1.03]$

Esta conversión se caracteriza por ser empírica, no depende de otras características como región, profundidad o tipo de evento; y además se puede introducir incertidumbre para magnitudes grandes o pequeñas.

Posteriormente, se inicia un proceso de limpieza en donde se elimina filas donde falten datos críticos (Fecha.UTC, Latitud, Longitud), así como también se elimina duplicados de sismos. Para ello, utiliza un enfoque de redondeo de coordenadas (4 decimales) antes de aplicar drop_duplicates basado en la terna (Fecha, Latitud, Longitud) para identificar y retener solo el primer registro del evento.

Carga Final: El catálogo final limpio y consolidado se exporta a un archivo CSV (1_Sismos_Consolidados.csv).

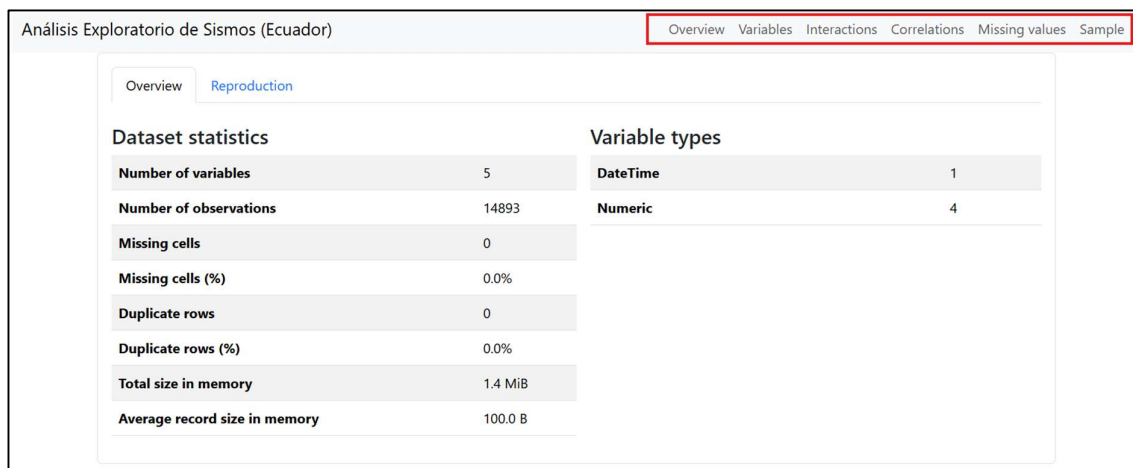
ANÁLISIS EDA

Esta fase es de vital importancia ya que el análisis exploratorio de los datos permite realizar una visión general de la data además de corregir particularidades como valores nulos, duplicados, datos redundantes, sesgos, entre otros. Un análisis completo incluye estadística descriptiva, distribuciones y correlaciones entre variables. En este proyecto se ha realizado un análisis completo con la función `ydata_profiling`

Visualización en Notebook: El método `perfil.to_notebook_iframe()` despliega el reporte interactivo dentro del entorno del notebook (Colab o Jupyter), facilitando la validación inmediata del dataset post-ETL.

Figura 2

Reporte interactivo con la funcion `ydata_profiling`



Nota: El reporte muestra varias pestañas en la esquina superior derecha con el objetivo de que el usuario interactúe con la data, se recomienda visualizarlo desde el Notebook.

Esta etapa es crítica para validar que el proceso ETL previo no introdujo artefactos y que el dataset está listo para el Feature Engineering y el modelado.

3.2 Clasificación Peligrosidad

Se ha definido la clasificación de la actividad sísmica en dos categorías: “Potencial Peligro” y “Sin Peligro”. De acuerdo con la literatura un sismo representa una amenaza cuando se encuentran cerca al volcán y cuando son de magnitud superior a 6. (Volcano Foundation, 2021)

Por lo tanto, es importante definir cuáles son los sismos que se encuentran mayormente relacionados con la actividad volcánica del Cotopaxi. Para ello se define las coordenadas geográficas del volcán (Ubicación focal) y se utiliza una función de distancia con la fórmula *Haversine Vectorizada* para determinar la distancia geodésica entre cada epicentro del evento sísmico y las coordenadas del volcán, esta distancia es calculada en kilómetros.

Un sismo se marca como "Potencial Peligro" si cumple alguno de los siguientes criterios:

1. *Amenaza Local Crítica*: Sismos cercanos de radio corto que pueden ocurrir en el edificio volcánico o fallas aledañas
2. *Amenaza Regional Fuerte*: Sismos de radio medio, son sismos fuertes que sacuden la estructura.
3. *Gran Evento Lejano*: Sismos importantes de subducción de radio amplio.

Tabla 2

Criterios para clasificación de categoría

Tipo	Distancia Geodésica	Magnitud
<i>Amenaza Local Crítica</i>	Menor a 30 km	Mayor a 4.0
<i>Amenaza Regional Fuerte</i>	Menor a 100 km	Mayor a 5.5
<i>Gran Evento Lejano 1</i>	Menor a 250 km	Mayor a 7.0
<i>Gran Evento Lejano 2</i>	Cualquiera	Mayor a 8

Basados en la lógica descrita anteriormente, se procedió con la generación de una columna en donde se aplica la función de clasificación sobre el DataFrame (`df.apply(..., axis=1)`) a la cual se la llamó: `Clasificacion_Amenaza`.

Para validar el proceso anterior se imprimió el conteo de valores y un ejemplo de los eventos clasificados como "Potencial Peligro"

Figura 3

Validación de criterios de peligro

=== Ejemplo de Alerta ===				
		Fecha.UTC	Mw.Proxy	Distancia_Cotopaxi_km
1	1906-01-31	15:36:00	8.35	208.592911
2	1906-09-28	15:24:54	7.50	159.206997
8	1924-07-22	04:04:16	6.50	71.788127
40	1942-05-14	02:13:28	7.80	230.404166
48	1949-08-05	19:08:47	6.80	60.926928

Nota: Se muestran aleatoriamente 5 casos de ejemplo.

El DataFrame se guarda en un archivo llamado `2_Sismos_Clasificados.csv` incluyendo a la nueva columna de riesgo.

Para la visualización geográfica se utilizó la librería *FOLIUM* el cual permite generar un mapa interactivo en donde se graficaron los puntos sísmicos que representan una amenaza para la actividad volcánica del Cotopaxi.

3.3 Balanceo de Clases

Se vuelve necesario realizar un balanceo de las clases ya que se identifica que existe un sesgo marcado en la distribución total de los eventos pues son mucho menor los sismos de categoría "Potencial Peligro" en relación a los eventos clasificados como "Sin Peligro", por lo tanto, se generan datos sintéticos que son geográficamente plausibles y se

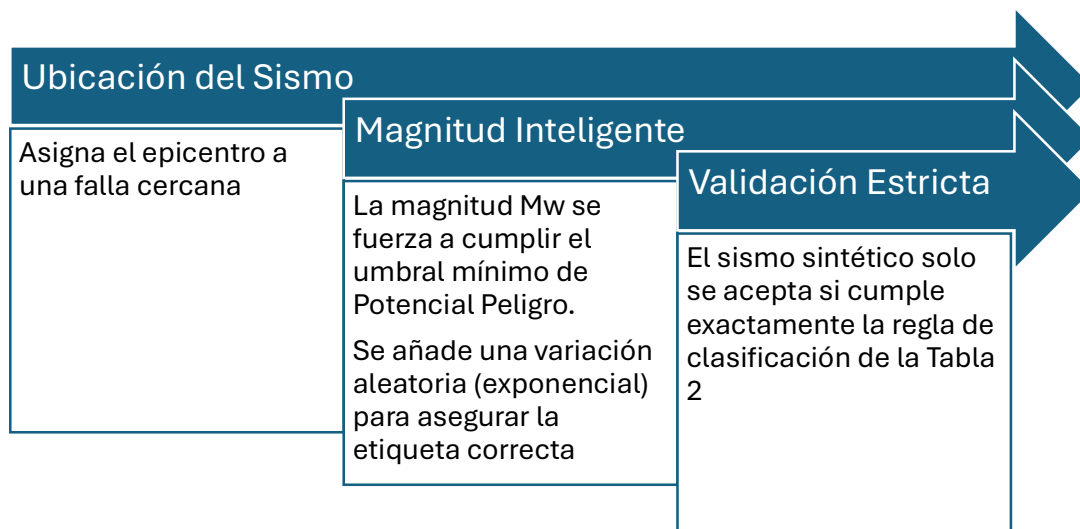
añade además ruido a las características. Este paso es necesario ya que permite evitar el sobreajuste y soluciona el problema de desbalance de clases.

Para la generación de datos sintéticos se utilizan las librerías `geopandas` y `shapely`, los cuales permiten crear una función para extraer las fallas activas del archivo (`gem_active_faults.geojson`). Posteriormente, se proyecta las geometrías a un sistema de coordenadas planas (UTM 17S) para calcular distancias precisas en metros y finalmente se filtra las fallas que se encuentran en un radio de 100km alrededor del Cotopaxi, asegurando que los sismos sintéticos se generen en ubicaciones geológicamente relevantes.

El sismo sintético se ubica aleatoriamente sobre una línea de falla (usando `linea_falla.interpolate`). Se añade ruido gaussiano ($\sigma=0.015$) a las coordenadas para simular la dispersión real de los epicentros alrededor del plano de falla (zona de ruptura).

```
Fallas encontradas en el radio de 100km: 55
Detectado desbalance. Generando 963 eventos de 'Potencial Peligro'...
Generados 963 registros de Potencial Peligro.
```

A pesar de que sea datos genéricos, es importante que estos sean de alto valor y buena calidad para el estudio, por lo tanto, se esquematiza el siguiente flujo a seguir:



Una vez que ya se han generado los datos sintéticos, es necesario consolidarlos con los datos reales, para esto se llama a la función de generación solo para la clase minoritaria y se concatenan los datos sintéticos con los datos históricos (`df_final_balanceado`).

Como resultado se obtiene un dataset balanceado y se guarda en `3_Sismos_Balanceado.csv`.

Como fase final de este proceso se identifica la necesidad de introducir ruido para evitar el sobreajuste, esta sección es crítica para preparar el dataset para Machine Learning.

Ruido en Características (`Mw_Noise`, `Dist_Noise`, `Prof_Noise`): Se añade ruido gaussiano a las variables predictoras (magnitud, distancia, profundidad) para simular la incertidumbre de la medición instrumental (error de localización, error de magnitud), haciendo que el modelo aprenda la regla subyacente y no memorice fronteras perfectas.

Ruido en Etiquetas (`Label Noise`): Se introduce un error intencional (5% de etiquetas invertidas) en la columna `Clasificacion_Amenaza`, esto obliga al modelo de ML a ser robusto y tolerante al error, impidiendo que el *Accuracy* llegue al 100% (lo cual indicaría un fallo en la validación).

El dataset resultante (`df_model`) es la versión final y más robusta, lista para el entrenamiento de un clasificador que debe aprender la frontera de decisión ruidosa entre "Potencial Peligro" y "Sin Peligro".

3.4 Entrenamiento

El dataset se encuentra listo para ser entrenado por lo que se preparó el ambiente de Python importando todas librerías necesarias para los 4 modelos a ser estudiados: SVM, MLP, RANDOM FOREST, XGBOOST.

Es necesario que se unifique las mismas condiciones para evaluar todos los modelos con el objetivo de poder compararlos de manera objetiva, por lo tanto, se realizó:

Selección de Features

Se define explícitamente el vector de características (**X**) utilizando las variables ruidosas y ubicación Prof_Noise, Mw_Noise, Dist_Noise, Latitud, y Longitud.

Codificación de la Variable Objetivo

Se utiliza *LabelEncoder* para convertir las etiquetas de texto (Clasificación_Amenaza) a valores numéricos (0 y 1), un requisito para la mayoría de los modelos de ML. Se almacenan los nombres de las clases (class_names).

División Estratificada

Función: `train_test_split`. Se utiliza `stratify=y_encoded` para garantizar que el conjunto de entrenamiento y el de prueba contengan la misma proporción de eventos de "Potencial Peligro" [Image of Stratified sampling illustration] (Clase 1), lo cual es crucial para una evaluación imparcial del riesgo.

Escalado de Características

Función: `StandardScaler`. Se aplica escalado a todos los conjuntos (ajustado en `X_train_scaled`), ya que es un requisito de rendimiento obligatorio para modelos basados en la distancia (SVM) y en gradientes (MLP). Los modelos basados en árboles (RF, XGBoost) también se benefician de la uniformidad.

Finalmente, se consideran funciones adicionales que permitirán optimizar los entrenamientos de los modelos:

Función `mostrar_top5_hiperparametros`

Procesa los resultados de `GridSearchCV` (`grid_search.cv_results_`), muestra las cinco mejores combinaciones de hiperparámetros ordenadas por

el `mean_test_score` (promedio del *accuracy* en la validación cruzada), esto permite al analista comprender la configuración de los hiperparámetros que ofrece el mejor rendimiento y la variabilidad de ese rendimiento (`std_test_score`).

Función `graficar_metricas_estandar`

Genera una figura con tres subtramas para la evaluación estándar del clasificador:

Matriz de Confusión: Visualiza los aciertos y errores (Falsos Negativos y Falsos Positivos) en formato *heatmap* de Seaborn. [Image of Confusion Matrix heatmap example]

Curva ROC: Muestra el *trade-off* entre TPR (Verdaderos Positivos) y FPR (Falsos Positivos). El valor AUC (Area Under the Curve) se muestra como una métrica de la capacidad de discriminación.

Curva Precision-Recall: Esencial en problemas con desbalance de clases (como este de riesgo sísmico). Evalúa la capacidad del modelo para predecir la clase positiva con alta Precision a diferentes umbrales de Recall.

Función `graficar_historia_entrenamiento`

Diagnostica el comportamiento del modelo durante el entrenamiento, diferenciando entre modelos basados en *Épocas* y modelos estáticos:

Modelos Iterativos (MLP, XGBoost): Grafica las curvas de Pérdida (Loss) y/o Accuracy en los conjuntos de entrenamiento y validación a través de las *Épocas* (o iteraciones de *boosting*). Esto es vital para detectar el punto exacto de sobreajuste.

Modelos Estáticos (RF, SVM): Genera una Curva de Aprendizaje utilizando `learning_curve`. Esta curva muestra si el modelo está sufriendo de bajo sesgo/alta varianza (sobreajuste) o si la adición de más datos mejoraría significativamente el rendimiento.

3.4.1 Random Forest

Se inicia este modelo con la búsqueda de hiperparámetros mediante la función *GridSearch*, para ello se define el espacio de búsqueda que está compuesto por los componentes detallados en la tabla a continuación:

Tabla 3

Componentes de Hiperparámetros Random Forest

Componente	Descripción
<i>n_estimators</i>	Número de árboles
<i>max_depth</i>	Profundidad máxima de los árboles
<i>min_samples_split</i>	Nro mínimo de muestras para dividir un nodo
<i>class_weight</i>	Opción para equilibrar los pesos de las clases

La función *GridSearch* utilizó una validación cruzada *cv=3*, el objetivo es encontrar la mejor combinación que maximice el *accuracy*. Una vez que se ha encontrado la mejor combinación de hiperparámetros se entrena el modelo para la clasificación final.

Se procede a medir el rendimiento del modelo generando un Reporte de Clasificación en donde constan las métricas *Precision*, *Recall* y *F-1 Score*. Adicionalmente, se generó la Curva ROC y la Curva de Aprendizaje, esta última es necesaria ya que permite determinar visualmente si existe *overfitting* o *underfitting*.

Finalmente, el mejor modelo es almacenado y posteriormente se graba en un diccionario para la comparativa final.

3.5.2 XGBoost

Se inicializa este modelo con la optimización de los hiperparámetros mediante *Grid Search* y se realiza un diagnóstico detallado del entrenamiento utilizando las curvas de historia (pérdida/error por epoch).

Se define el parámetro de búsqueda con los componentes detallados en la tabla a continuación:

Tabla 4

Componentes de Hiperparámetros XGBoost

Componente	Descripción
<i>n_estimators</i>	Número de iteraciones (boosting)
<i>max_depth</i>	Profundidad del árbol
<i>learning_rate</i>	Velocidad de aprendizaje del algoritmo
<i>subsample</i>	Fracción de datos a muestrear para cada árbol

En este algoritmo es necesario realizar el re-entrenamiento con Curvas de Historia para poder graficar la evolución del rendimiento (curva de pérdida), el mejor modelo (*best_xgb*) se re-entrena utilizando el parámetro *eval_set*.

Eval_set: Permite a XGBoost calcular la logloss (pérdida) y el error ($1 - accuracy$) en tiempo real tanto en el conjunto de entrenamiento como en el de prueba en cada *epoch* de *boosting*.

Con este modelo se puede medir el rendimiento con el reporte de Clasificación, matriz de confusión, Curva ROC, Curva de pérdida, Curva de Accuracy, por lo que se procede a generar estas gráficas. El mejor modelo es almacenado para la comparativa final.

3.5.3 MLP

Al igual que los dos modelos anteriores se inicia con la optimización de la arquitectura y los hiperparámetros mediante Grid Search, con esto se busca la configuración que mejor capture las fronteras de decisión no lineales del riesgo sísmico.

Los parámetros que se deben considerar en este modelo son *hidden_layer_sizes* el cual permite la exploración de arquitecturas de una capa (50, 100) y dos capas (50, 50); *activation* que son funciones de activación, entre las más comunes se tiene *relu* y *tanh*, por

otro lado, está *Alpha* que se definió con un término de regularización L2 y finalmente *learning_rate_init* que corresponde a la tasa de aprendizaje inicial.

El entrenamiento se realiza con `early_stopping=True`, lo que utiliza una fracción del conjunto de entrenamiento como validación interna y detiene el proceso si el rendimiento deja de mejorar, previniendo el sobreajuste y ahorrando tiempo.

A diferencia de XGBoost o Random Forest, el diagnóstico de la MLP se centra en la evolución del entrenamiento, por un lado se tiene la *Curva de Pérdida* (`best_mlp.loss_curve_`) aquí se muestra cómo disminuye el error (*loss*) en el conjunto de entrenamiento a lo largo de las épocas. Una caída constante es señal de un buen proceso de optimización. Por otro lado, se tiene la *Curva de Validación* (`best_mlp.validation_scores_`) en este caso se muestra el *score* (precisión interna) del modelo en el conjunto de validación interno utilizado para el *early stopping*. Es clave para verificar que el modelo no se haya sobreajustado antes de finalizar el entrenamiento.

3.5.4 SVM

Se optimizó el clasificador SVM mediante Grid Search para encontrar el hiperplano óptimo que separe las clases de riesgo en el espacio de características escalado. Además, se consideran también los siguientes parámetros: C (Parámetro de Regularización) el cual controla el *trade-off* entre la penalización por error de clasificación y la simplicidad de la frontera; la gamma por su parte define la influencia de un solo ejemplo de entrenamiento (esencial para el kernel RBF) y finalmente el kernel en donde se realizan pruebas con los kernels RBF (Radial Basis Function), que maneja fronteras no lineales, y poly (Polinomial).

Se usa el modelo base SVC(`probability=True`) para asegurar que el modelo pueda devolver las probabilidades, lo cual es necesario para trazar la Curva ROC/AUC. El resultado del modelo son las predicciones binarias (`y_pred_svm`) y las probabilidades (`y_proba_svm`).

Finalmente, se realizó una comparativa de todos los modelos para identificar el clasificador de riesgo sísmico con el mejor *Accuracy* y guardarlo junto con los artefactos de pre-procesamiento necesarios para su despliegue.

Se creó un diccionario (artifact) que empaqueta todos los componentes necesarios para usar el modelo en un entorno nuevo, asegurando la coherencia entre el entrenamiento y la inferencia. El artefacto completo se guardó en un archivo final (best_model_[Nombre del Modelo].pkl).

El sistema desarrollado corresponde a una aplicación de escritorio orientada al análisis y clasificación del riesgo sísmico en el territorio ecuatoriano, con énfasis en la evaluación de eventos que podrían representar un potencial peligro para la actividad volcánica, específicamente en el entorno del volcán Cotopaxi. La herramienta integra técnicas de aprendizaje automático con una interfaz gráfica intuitiva, facilitando la interacción del usuario con modelos predictivos avanzados sin requerir conocimientos técnicos especializados.

CAPITULO 4

4. ANÁLISIS DE RESULTADOS

4.1 Resultados del Pre procesamiento

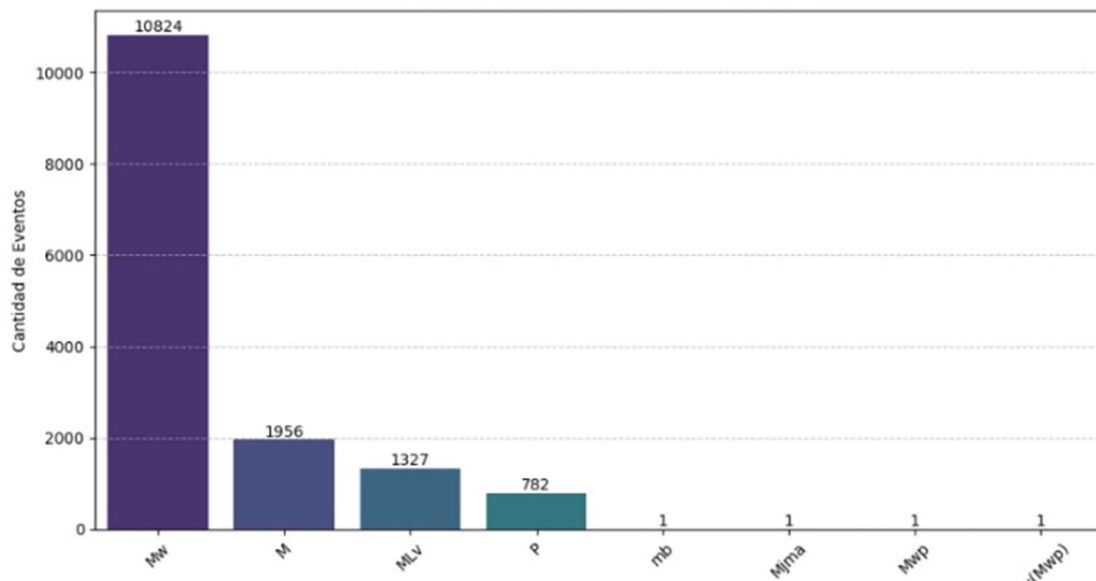
A continuación, se muestra los resultados por cada etapa del trabajo desarrollado.

En la etapa ETL fue posible la unificación de los tres catálogos sísmicos de diferentes periodos de tiempo, se considera una integración exitosa ya que a pesar de que cada catálogo tenía diferente formato, columna distintas y tipos de magnitudes diferentes, se logró estandarizar a una sola fuente de datos sólida y robusta.

Por otro lado, considerando que la variable objetivo debía ser homogenizada debido al diferente formato y origen de esta, este cálculo a Mw permitió comparar eventos sísmicos de diferentes épocas y evitar el sesgo por el uso de escalas incompatibles.

Figura 4

Distribución de tipos de magnitud



Nota: Se muestran todas las magnitudes originales de todos los catálogos.

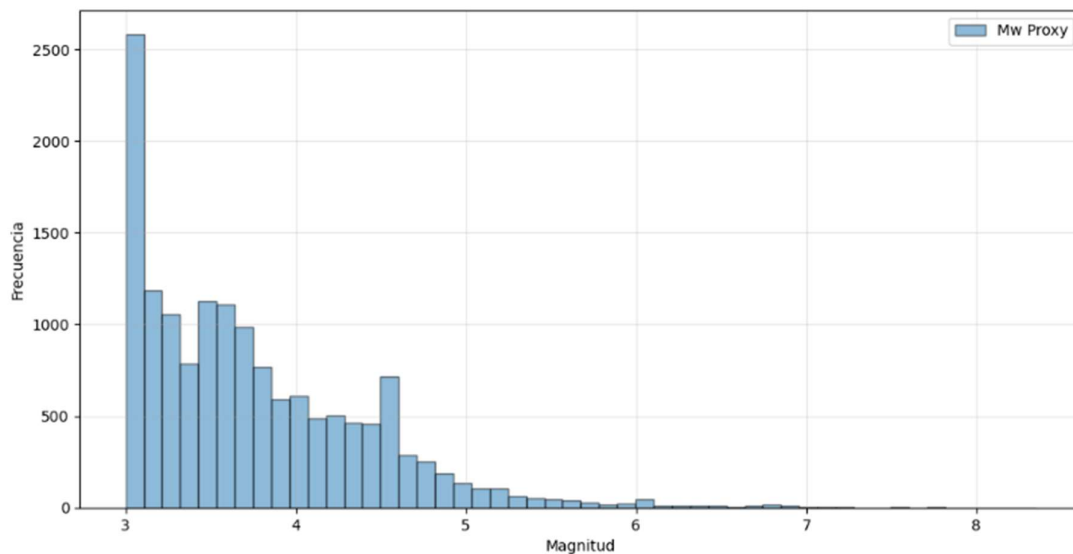
Se debe considerar además que M_w no es una magnitud real como tal sino mas bien una aproximación estadística válida para análisis globales

Como se puede ver en la gráfica, la conversión no distorsiona significativamente la distribución de las magnitudes, ya que existe una alta similitud entre la variable original y la ajustada. Así mismo se observa cómo se conservan los patrones como por ejemplo la forma de la curva, la asimetría y la cola. No se introducen artefactos visibles.

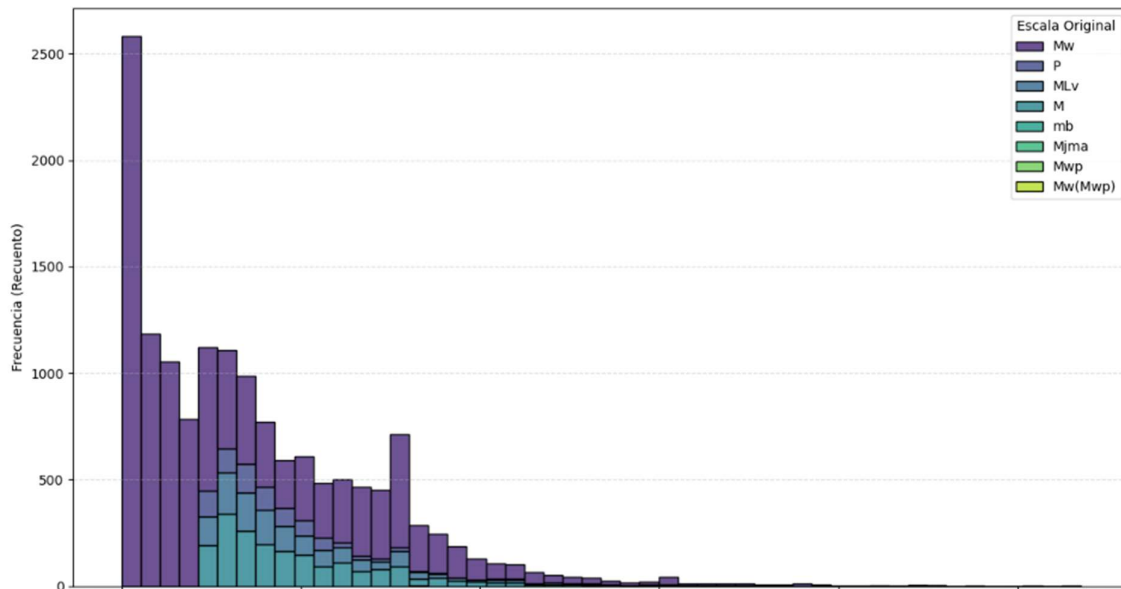
La comparación de las gráficas con las magnitudes originales y homogenizadas permite entender que la aproximación M_w es válida para este tipo de estudios temporales.

Figura 5

Distribución de las magnitudes M_w



Nota: Distribución de las magnitudes una vez que se ha aplicado la conversión

Figura 6*Frecuencia de magnitudes*

Nota: Se muestra la distribución de las magnitudes originales.

Por último, aquí en el ETL se realizó el primer proceso de limpieza de los datos en donde se realizó una limpieza controlada de los duplicados. El riesgo que se generó en este proceso es que se eliminan por error dos eventos sísmicos que se encuentren cercanos en tiempo y espacio; sin embargo, la literatura nos indica que este proceso es correcto debido a que las mediciones son históricas, para el caso de enjambres sísmicos se debe realizar otra validación pues esta no es la más adecuada.

El resultado final del ETL fue un catálogo sísmico confiable para continuar con el análisis exploratorio.

Figura 7*Resumen proceso ETL*

Etapa	Filas	Columnas	Datos_Nulos_Totales	Descripción
Consolidación	14893	7	0	Unión de archivos
Estandarización	14893	8	0	Mw_Proxy generado
Limpieza Final	14893	8	0	Nulos y duplicados eliminados

Archivo guardado: 1_Sismos_Consolidados.csv

Nota: Imagen del resultado obtenido en el Notebook.

Durante el proceso de análisis de la calidad de los datos, se realizó la clasificación de los eventos en dos categorías: “Potencial Peligro” y “Sin peligro”, esta clasificación se basó en criterios físicos de la data recolectada como distancia, magnitud y profundidad; es decir que no es un modelo de aprendizaje automático sino mas bien está basado en reglas geofísicas.

El volcán Cotopaxi fue modelado como un punto de coordenadas a partir del cual se generaron las distancias geodésicas y en base a la Tabla 2 los eventos fueron clasificados. El resultado fue el esperado:

Figura 8*Resultado clasificación binaria con eventos reales*

```

=== Resumen de Clasificación (Binaria) ===
Clasificacion_Amenaza
Sin Peligro      14856
Potencial Peligro    37
Name: count, dtype: int64

```

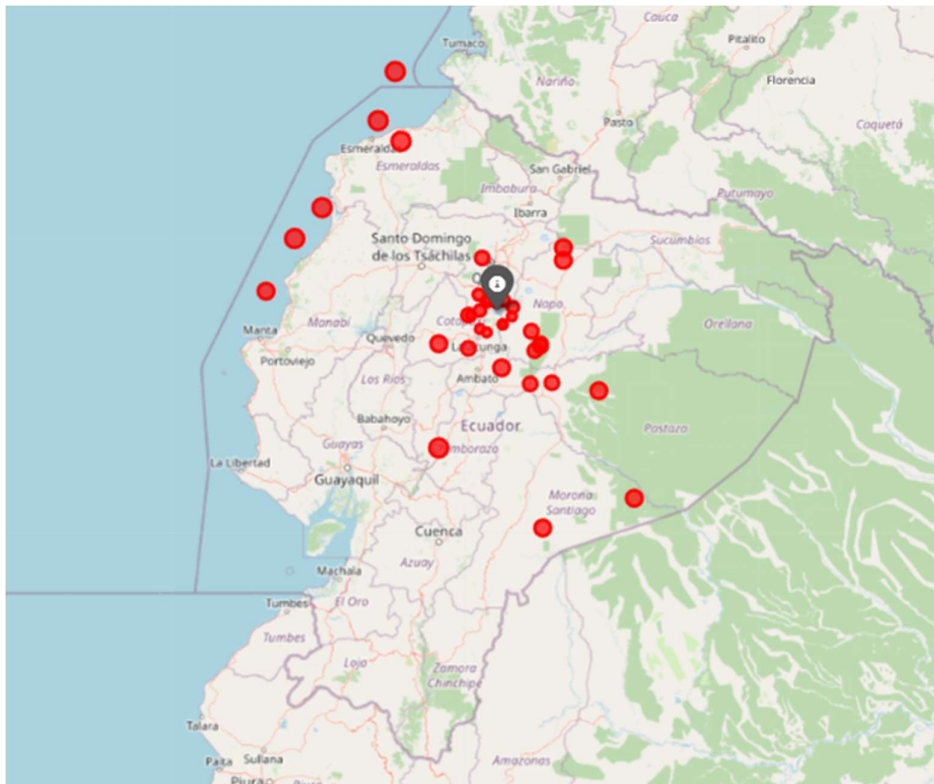
En condiciones geofísicas este resultado tiene sentido ya que siendo conscientes con la realidad sísmica la gran mayoría de los sismos no afectan la actividad del sistema volcánico. Sin embargo, con este resultado se pudo observar que para el entrenamiento con Machine Learning si representa un problema pues evidencia un desbalance de clases.

La limitación de esta clasificación binaria es que únicamente los divide en dos grupos, pero no los diferencia por niveles de peligro ya que dentro de las clasificaciones iniciales también pudiera haber subclasificaciones.

A continuación, se graficaron los eventos considerados como “Potencial Peligro” ya que esto permite visualizar la amenaza real que podrían afectar la estabilidad de la actividad volcánica del Cotopaxi y esto a su vez nos deja ver porque este volcán es el más vigilado a nivel nacional y de gran interés a nivel internacional. Este gráfico sería ideal como un filtro inicial de eventos relevantes para estudios futuros relacionados con la sismología del volcán.

Figura 9

Clasificación de sismos considerados de Potencial Peligro



Nota: El radio del marcador corresponde a la magnitud del sismo, el color rojo es para los eventos clasificados como potencial peligro y el pin plomo marca la ubicación del volcán Cotopaxi.

Este análisis permitió clasificar sismos que, por su ubicación y energía liberada, representan una amenaza directa a la estabilidad del sistema volcánico a Cotopaxi.

Posteriormente, al haber evidenciado que existe un desbalance de clases se generó un proceso *data augmentation* para generar datos sintéticos que sean físicamente coherentes y apegados a las condiciones de los datos reales. Con este método se resolvió el desbalance de clases para poder obtener un dataset listo para el entrenamiento y que no sobreajuste a un accuracy de 1, lo cual es común cuando no existen suficientes datos para entrenar en una de las clases.

Figura 10

Resultado del balanceo de clases

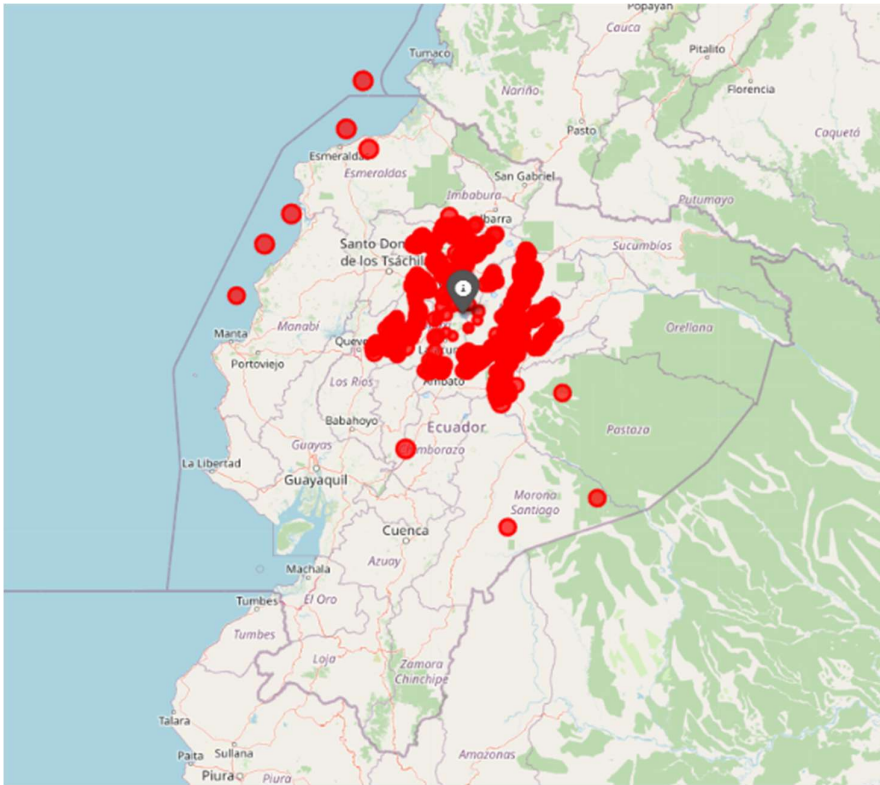
```
=== Resultado Final del Balanceo ===  
Clasificacion_Amenaza  
Sin Peligro          14856  
Potencial Peligro    1000  
Name: count, dtype: int64
```

Se valida este proceso ya que el objetivo principal no es el análisis sísmico como tal sino es construir un dataset que sea entrenable y a su vez sea científicamente defendible.

Para validar geográficamente que son datos sintéticos plausibles y de calidad, se procedió con la ilustración de los sismos para hacer una comparativa con los datos reales que se muestran en la *FIGURA 9*

Figura 11

Sismos sintéticos generados como Potencial Peligro



Nota: Los sismos sintéticos se generaron cerca de fallas geológicas en un perímetro de 100km alrededor del Volcán Cotopaxi.

Se utilizaron los mismos criterios de validación de la categoría “Potencial Peligro”, garantizando así la plausibilidad física de los nuevos datos.

Se puede evidenciar que se mantiene la misma lógica que la clasificación de eventos reales puesto que la ubicación geográfica de los puntos sintéticos encierra áreas muy similares a los sismos reales y debido a que la clase desbalanceada era la de “Potencial Peligro”, se evidencia que la *FIGURA 11* es mucho más poblada pues claramente ese es el objetivo en este proceso.

Y de igual manera que los datos reales tienen ruido y errores, es necesario también simular los mismos “errores” en los datos sintéticos por lo que es agregado el ruido en las

variables continuas, este paso introduce zonas grises de igual manera que existe en los eventos reales. La generación de ruido es clave y obligatorio hasta cierto punto en la generación de datos sintéticos pues de esta manera se evita un overfitting y un accuracy artificial de 1. Por lo tanto, con esto se garantiza que la data sintética respete la tectónica la real, incorpora incertidumbre y evita el aprendizaje trivial.

La limitación de esta técnica es que no hay un enfoque sobre el estado dinámico del volcán, el ruido sigue distribuciones simples y no existe un mecanismo focal, sin embargo, como el objetivo de este trabajo es la aplicación de algoritmos de machine learning, el tratamiento es el adecuado, en otros casos de estudio se puede mitigar la debilidad de la actividad sísmica teórica que al momento es mínima.

Finalmente, se prepararon los datos para el entrenamiento. El dataset preprocesado se encuentra listo para su división en entrenamiento 80% y prueba 20%; así como también se define las variables características y la variable objetivo, es importante considerar que se utilizan las variables “ruidosas”, para evitar que el modelo aprenda reglas “perfectas”.

```
Datos preparados. Train shape: (12684, 5), Test shape: (3172, 5)
```

4.2 Resultados del entrenamiento de los modelos

El primer modelo de entrenamiento a ser comparado es Random Forest cuyo desarrollo y evaluación se estructura de tal forma que sea rigurosa y sistemática. En primer lugar, se realiza una búsqueda profunda de hiperparámetros, el cual tiene por objetivo identificar la configuración óptima del modelo para maximizar su desempeño y al mismo tiempo evitar el sobreajuste; este proceso se complementa con la aplicación de validación cruzada, lo que a su vez permite evaluar la capacidad de generalización del modelo sobre diferentes subconjuntos de datos y de esta manera garantizar la estabilidad de los resultados. De igual manera, se cuantifica el desempeño del modelo a través de métricas objetivas, permitiendo así una base para el análisis y toma de decisiones. Para finalizar, el proceso fue reforzado

con la generación de visualizaciones que facilitan la interpretación y el almacenamiento del mejor modelo obtenido para su posterior uso.

La búsqueda de los mejores hiperparámetros se filtró con la métrica de optimización accuracy y con una validación cruzada de 3-fold. El resultado es la impresión del top 5 de mejores combinaciones de hiperparámetros, esto es útil para analizar la sensibilidad y además de otorgar transparencia metodológica.

Figura 12

Mejores hiperparametros para random forest

=====				
MODELO 1: RANDOM FOREST				
=====				
--- TOP 5 MEJORES HIPERPARÁMETROS ---				
	params	mean_test_score	std_test_score	rank_test_score
{'class_weight': None, 'max_depth': None, 'min_samples_split': 5, 'n_estimators': 100}		0.9480	0.003017	1
{'class_weight': None, 'max_depth': None, 'min_samples_split': 5, 'n_estimators': 200}		0.9480	0.003191	1
{'class_weight': None, 'max_depth': None, 'min_samples_split': 5, 'n_estimators': 300}		0.9480	0.003191	1
{'class_weight': None, 'max_depth': None, 'min_samples_split': 2, 'n_estimators': 100}		0.9480	0.003017	1
{'class_weight': None, 'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 100}		0.9479	0.002956	5

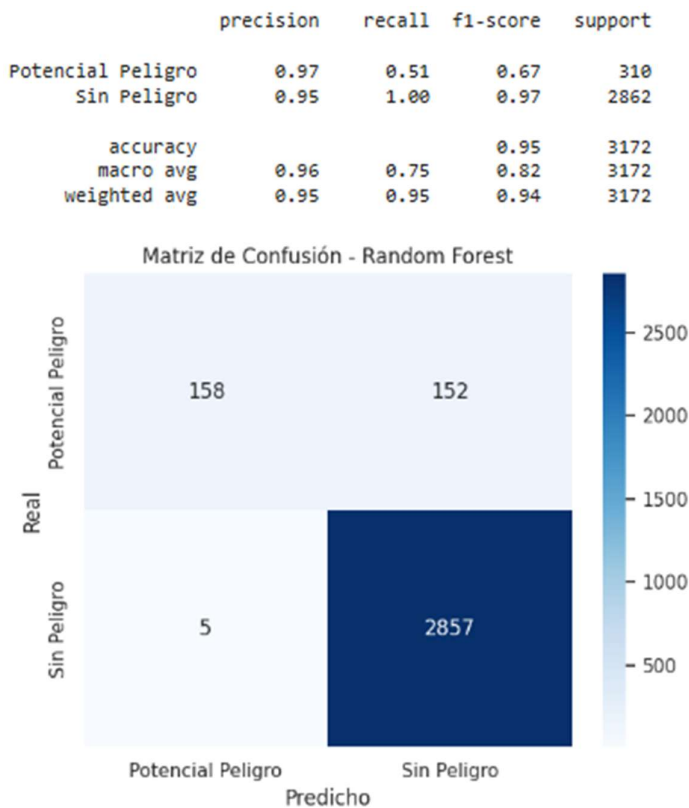
Nota: Se muestran únicamente las 5 combinaciones de hiperparámetros que obtuvieron mejor promedio en CV.

Se puede ver que existe una clara tendencia en cada uno de los hiperparámetros: 'n_estimators': 100, 'max_depth': None, 'min_samples_split': 5, 'class_weight': None

El modelo fue entrenado y se obtuvieron las siguientes métricas de rendimiento:

Figura 13

Reporte de Clasificación y Matriz de Confusión de Random Forest



Se observa que el accuracy del modelo alcanzó un valor de 0.95 lo cual, dentro del contexto de este trabajo, se interpreta como una métrica global aceptable del modelo en general.

Por otro lado, la matriz de confusión nos indica que los eventos considerados en la categoría “Sin Peligro” se predicen bastante bien ya que únicamente 5 muestras son falsos positivos; sin embargo la categoría “Potencial Peligro” muestra bastante ambigüedad ya que entre los casos predichos y los reales hay una tendencia el 50/50, esto se puede ver en el Recall de “Potencial Peligro” donde se obtuvo un puntaje de 0.51 indicando que aproximadamente la mitad de los eventos críticos no fueron detectados correctamente, esto puede deberse a una frontera de decisión conservadora, posiblemente por la incorporación de

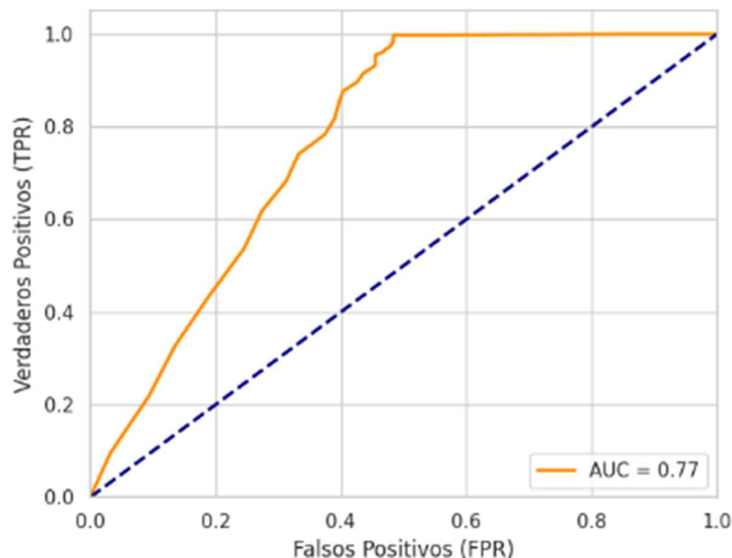
ruido instrumental de etiquetado, lo cual fue diseñado para evitar el sobreajuste y simular condiciones reales de incertidumbre.

Por otro lado, si bien el desempeño global del modelo es aceptable con un 0.95, la baja sensibilidad con relación a los eventos considerados potencialmente peligrosos, limita a que pueda ser un modelo utilizado como sistema de alerta temprano ya que en la mitad de los casos los riesgos reales no van a ser detectados ni alertados. Sin embargo, a pesar de llegar a esta conclusión, estos resultados llegan a tener sentido hasta cierto punto ya que la amenaza de la actividad volcánica no depende ni única ni directamente de la sismología, sino también de condiciones geológicas y meteorológicas adicionales.

Otra métrica de rendimiento del modelo es la curva ROC, esta curva permite medir la capacidad de la discriminación del modelo, es decir que tan bien separa los eventos en promedio “Potencial Peligro” de los “Sin Peligro” para todos los umbrales.

Figura 14

Curva ROC- Random Forest



Nota: AUC siglas para Area Under Curve

Para este modelo se obtuvo un AUC de 0.77 este valor se puede interpretar como una probabilidad en donde si se selecciona un evento al azar de “Sin Peligro” y “Potencial

Peligro”, el modelo le asignará una probabilidad mayor al evento peligroso el 77% de las veces.

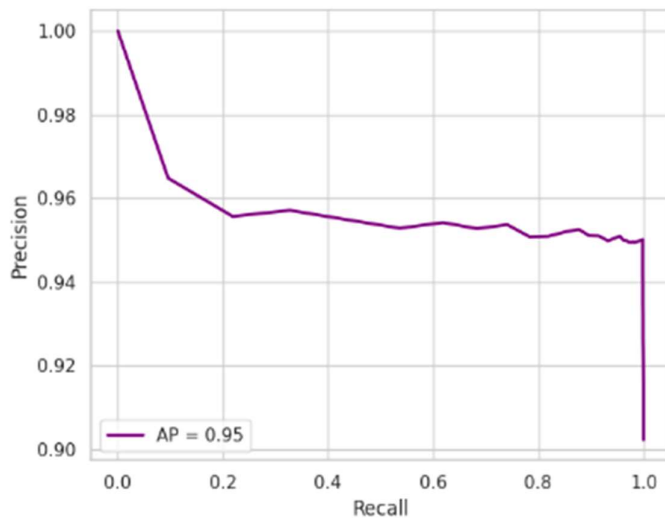
En una escala de interpretación global, un AUC en una escala entre 0.7 y 0.8 es bueno-aceptable, el AUC bueno-ideal habría sido entre 0.8 y 0.9. Este resultado nos indica que el modelo tiene una buena capacidad de discriminación de clases y captura patrones relevantes en los datos, sin embargo, la selección final limita la sensibilidad frente a la clase crítica.

La discrepancia entre un AUC bueno y un recall reducido, se puede interpretar como una frontera difusa de decisión y claramente existe una superposición entre clases, condiciones que se ajustan a escenarios realistas sísmicos.

La curva a continuación, dentro de este contexto, responde a la pregunta ¿Qué tan confiables son las alertas a lo largo de todos los umbrales, cuando el modelo decide emitir una alerta?

Figura 15

Curva Precision-Recall Random Forest



Nota: AP siglas para Average Precision

El valor obtenido de 0.95 se puede interpretar como el acierto en un 95% de las veces en promedio cuando el modelo se clasifica como “Potencial Peligro”.

Esta interpretación podría parecer contradictoria con el Recall de 0.51 discutido anteriormente, sin embargo, no lo es ya que esta métrica se enfoca únicamente en las pocas alertas que emite el modelo, pero de estas pocas alertas se acierta casi siempre, ósea un 95%.

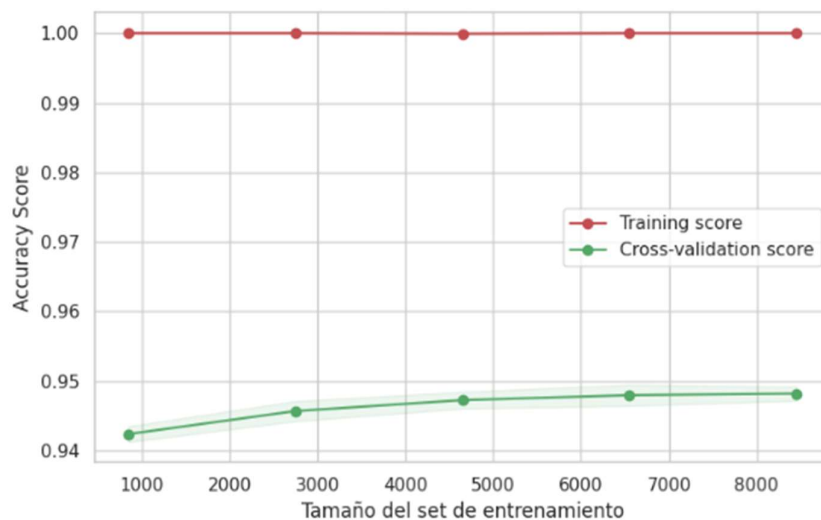
A nivel de la interpretación en el contexto sísmico quiere decir que, si el modelo marca un sismo como potencialmente peligroso, tiene una probabilidad casi certera que si lo es; sin embargo, en el contexto global hay muchos eventos peligrosos que no están siendo detectados.

En un contexto geopolítico, es preferible tener un AP alto con un recall ajustable, a tener un recall bajo con muchas falsas alarmas, ya que en este contexto implicaría una desestabilización a nivel económico y social el generar una alarma innecesaria.

Finalmente, para este modelo se generó una gráfica de curva de aprendizaje

Figura 16

Curva de Aprendizaje-Random Forest



Nota: Este algoritmo no entrena por épocas.

La curva de entrenamiento aproximadamente igual a 1 implica que el modelo aprende muy bien los datos de entrenamiento, esto muestra que puede memorizar patrones complejos. Por otro lado, la curva de validación cruzada que se mantiene en 0.95 aproximadamente, indica una buena generalización. La brecha entre ambas curvas es pequeña lo cual es un resultado positivo.

En base a todos los resultados obtenidos de Random Forest, los valores no son contradictorios sino complementarios, el modelo aprende bien, pero decide con cautela.

El segundo modelo de entrenamiento a ser comparado es XGBoost, el resultado de este modelo se puede resumir a los siguientes puntos:

- Búsqueda sistemática de hiperparámetros,
- Re entrenamiento registrando métricas por épocas
- Métricas cuantitativas: Precision, Recall, AUC, etc.
- Almacenamiento del mejor modelo.

Se obtuvo los siguientes resultados para la combinación de mejores hiperparámetros:

Figura 17

Mejores hiperparametros de xgboost

```
=====
MODELO 2: XGBOOST
=====
```

--- TOP 5 MEJORES HIPERPARÁMETROS ---

params	mean_test_score	std_test_score	rank_test_score
{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 200, 'subsample': 1.0}	0.9469	0.003256	1
{'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 100, 'subsample': 1.0}	0.9468	0.003367	2
{'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 100, 'subsample': 1.0}	0.9468	0.002858	2
{'learning_rate': 0.01, 'max_depth': 7, 'n_estimators': 200, 'subsample': 1.0}	0.9466	0.002899	4
{'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 200, 'subsample': 1.0}	0.9466	0.002925	4

Se puede observar que se marca una clara tendencia:

n_estimators: 200, max_depth: 3, learning_rate: 0.1, subsample: 1

Se almacena el mejor conjunto de hiperparámetros para entrenar el modelo, de igual manera se lo filtra por el mejor promedio obtenido.

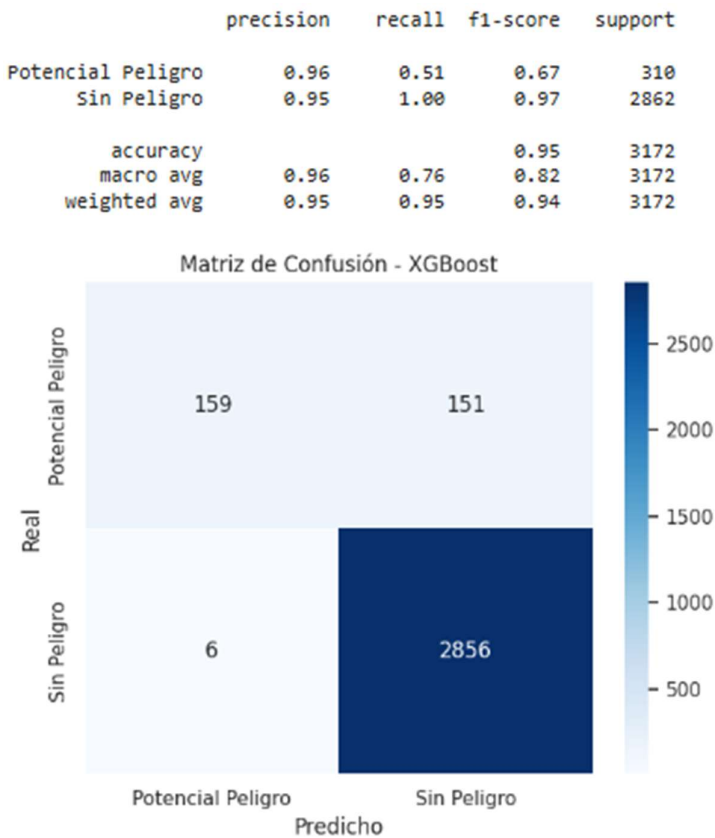
La exploración del Grid en XGBoost se destaca por árboles simples con aprendizaje lento versus árboles complejos con aprendizaje rápido.

En este algoritmo se realiza un re entrenamiento lo cual permite ver el sobreajuste y analizar la estabilidad del modelo.

Se muestra a continuación las métricas del rendimiento general del modelo:

Figura 18

Reporte de Clasificación y Matriz de Confusión de XGBoost



Del reporte de clasificación se puede observar que se obtuvo un puntaje global del rendimiento del modelo con un accuracy de 0.95

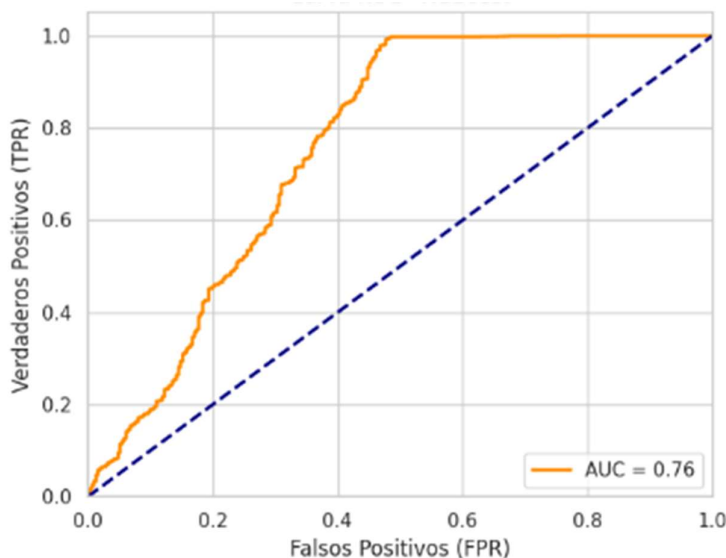
Por otro lado, la matriz de confusión nos indica que los eventos considerados en la categoría “Sin Peligro” se predicen bastante bien ya que únicamente 6 muestras son falsos positivos; sin embargo, la categoría “Potencial Peligro”, al igual que el modelo anterior, muestra bastante ambigüedad ya que 151 casos que se predijeron como “Sin Peligro” resultaron si siendo una amenaza, y de forma paralela casi el mismo número de eventos resultaron siendo eventos reales potencialmente peligrosos. La tendencia es que 1 de cada 2 eventos peligrosos no son discriminados correctamente esto se puede ver en el Recall de “Potencial Peligro” donde se obtuvo un puntaje de 0.51 esto puede deberse a una frontera de decisión conservadora, posiblemente por la incorporación de ruido.

Se puede observar además una ligera mejora que Random Forest (RF) en relación a los eventos peligrosos predichos correctamente, ya que en RF fueron 158 eventos y con XGBoost son 159 eventos.

Por otro lado, se mide también al modelo con la Curva ROC

Figura 19

Curva ROC-XGBoost



Nota: AUC siglas para Area Under Curve

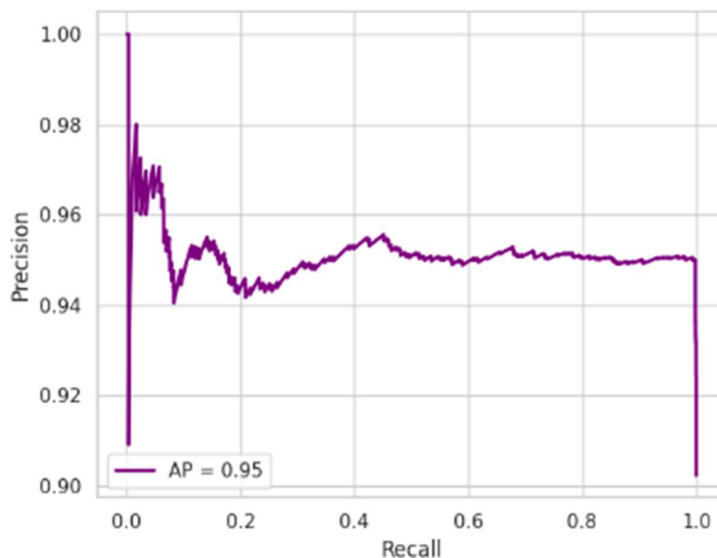
Se obtuvo un resultado similar a RF con un AUC de 0.76 que se interpreta como una probabilidad del 76% de que el modelo asigne mayormente una probabilidad a un sismo potencialmente peligroso que a uno que no sea peligroso, eligiendo de manera aleatoria cualquier evento sísmico. El 0.76 es una métrica buena sin embargo no sobresaliente y es una oportunidad de mejora.

En el contexto de las geociencias es un valor esperado un AUC que se encuentre entre 0.7 y 0.8; sin embargo, con datos sintéticos se esperaría que esta métrica sea al menos 0.8, el ruido añadido intencionalmente está cumpliendo su función y el problema ya no es trivial.

A continuación, se muestra la gráfica Precision-Recall

Figura 20

Curva Precision-Recall XGBoost



Nota: AP siglas para Average Precision

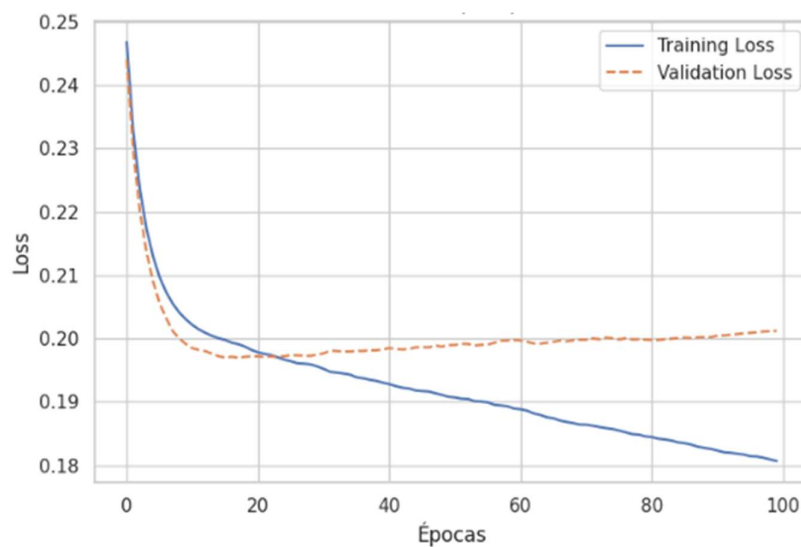
Se obtiene el mismo resultado de AP que en Random Forest, la interpretación es la misma ya que esta métrica hace referencia únicamente a los eventos peligrosos y sus aciertos, es

decir que cuando se generen alarmas de sismos potencialmente peligrosos, con casi toda seguridad los eventos sí serán peligrosos.

A continuación, se muestra la gráfica de pérdida, esta gráfica es una de las más importantes en el algoritmo de XGBoost ya que mide que tan buenas son las probabilidades más que la clase final, sin embargo, penaliza fuertemente las predicciones muy seguras pero que terminan siendo incorrectas.

Figura 21

Curva de Pérdida (Loss)-XGBoost



Se puede observar que la curva de pérdida muestra una disminución progresiva de ambas curvas al inicio, en la fase intermedia se marca una distancia entre la curva de entrenamiento y la curva de validación, y en la etapa final la curva de validación se aplana mientras que la curva de entrenamiento sigue bajando. Esta fase final es la deseable o esperada ya que significa que el límite de la información útil ha sido alcanzado.

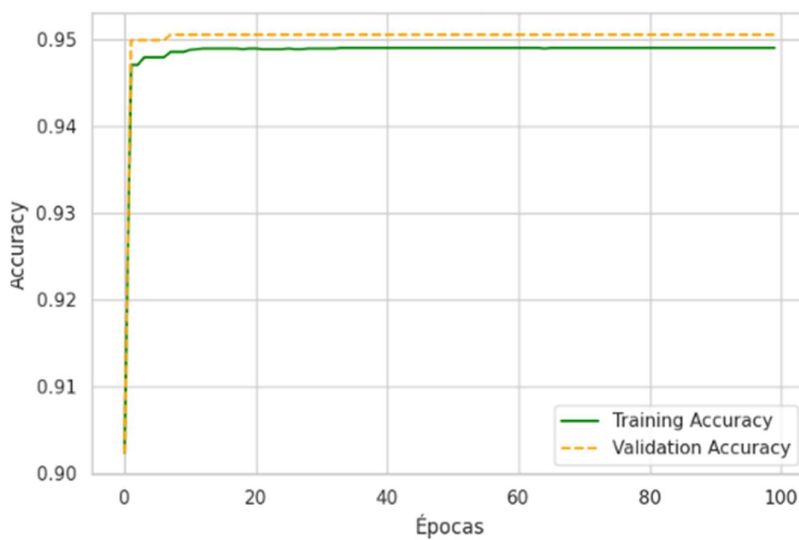
En este contexto, la pérdida no llega a cero ya que en un sistema volcánica no se encuentra determinada una frontera perfecta y clara entre el peligro y no peligro, por lo tanto, el modelo está mostrando la realidad.

El modelo está evitando el sobreajuste y tratando de generalizar adecuadamente a las clases.

A continuación, se muestra la siguiente métrica que es la Curva de Accuracy, esta curva tiene una interpretación de mucho cuidado ya que el Accuracy no es la métrica principal del algoritmo, sin embargo, si aporta información complementaria.

Figura 22

Curva de Accuracy- XGBoost



El resultado esperado de esta gráfica es que ambas curvas suban rápidamente y se estabilicen ambas horizontalmente y mantengan una brecha cercana entre las dos curvas.

Si ambas curvas se encuentran en un accuracy entre 0.95 y 0.99 se puede interpretar como que el modelo aprende bien, pero no memoriza completamente los datos.

Adicionalmente, se observa que las curvas se mantienen altas y estables sin mayores fluctuaciones, lo que indica que es un resultado esperado y correcto.

Es importante mencionar que un accuracy alto, en este contexto, no significa que sea un buen detector de amenaza de eventos potencialmente peligrosos, mas bien la interpretación correcta sería que el modelo es bueno detectando casos “fáciles” es decir los

eventos “Sin Peligro”, esto ayuda a que el accuracy suba, sin embargo, en los eventos críticos tiene baja sensibilidad para discriminar este tipo de eventos.

Por lo tanto, se debe considerar que el accuracy sirve para medir la estabilidad global del modelo, pero no debe ser utilizado como referencia de un umbral en específico, especialmente en este caso no debe interpretarse como métrica de capacidad de detección de eventos sísmicos peligrosos.

El tercer modelo que ha sido entrenado es MLP- Red Neuronal. Este modelo no utiliza reglas explícitas y aprende fronteras no lineales continuas.

Al Igual que en los dos modelos anteriores, se utilizó GridSearch para definir la mejor combinación de hiperparámetros:

Figura 23

Mejores hiperparámetros de red neuronal-MLP

=====				
MODELO 3: MLP (RED NEURONAL)				
=====				
--- TOP 5 MEJORES HIPERPARÁMETROS ---				
	params	mean_test_score	std_test_score	rank_test_score
{'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (50, 50), 'learning_rate_init': 0.01}		0.9443	0.003031	1
{'activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (50,), 'learning_rate_init': 0.01}		0.9438	0.002733	2
{'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (50,), 'learning_rate_init': 0.01}		0.9436	0.003068	3
{'activation': 'tanh', 'alpha': 0.0001, 'hidden_layer_sizes': (100,), 'learning_rate_init': 0.01}		0.9435	0.002373	4
{'activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (50,), 'learning_rate_init': 0.001}		0.9434	0.002520	5

En este proceso de GridSearch se utilizó EarlyStopping, lo cual es una validación interna que detiene el entrenamiento cuando ya el desempeño deja de mejorar, esto a su vez redujo el sobreajuste y el entrenamiento innecesario.

La arquitectura de la red se entrenó con los mejores hiperparámetros:1 capa (50 o 100 neuronas) y 2 capas (50–50)

Función de activación

Relu: rápida y robusta

Regularización: alpha

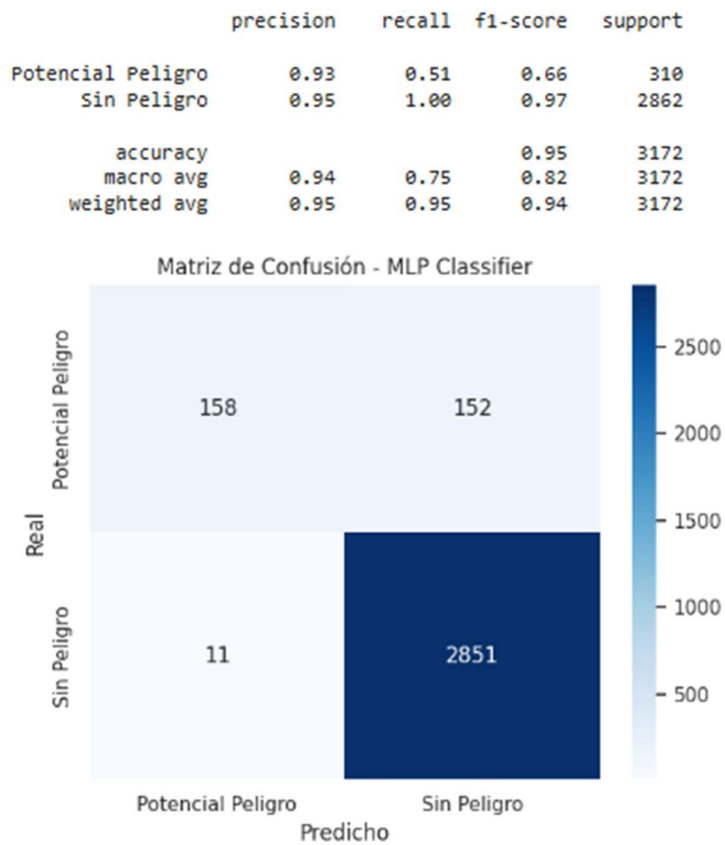
Penaliza pesos grandes evitando así el sobreajuste.

Learning Rate

Controla estabilidad contra rapidez. Una vez que se realizó el entrenamiento, se obtuvieron los siguientes resultados que miden la capacidad de discriminación de clases del modelo:

Figura 24

Reporte de Clasificación y Matriz de Confusión MLP Classifier



Se puede observar que se mantiene la tendencia con un accuracy del modelo global como tal de 0.95, los eventos clasificados “Sin Peligro” parecen ser de algún modo, bastante fáciles de clasificar pues además de ser la clase mayoritaria, los eventos peligrosos

históricamente se presentan en intervalos largo de tiempo, por lo tanto, es lo esperado que entre la gran mayoría de eventos a los clasificados como “Sin Peligro”

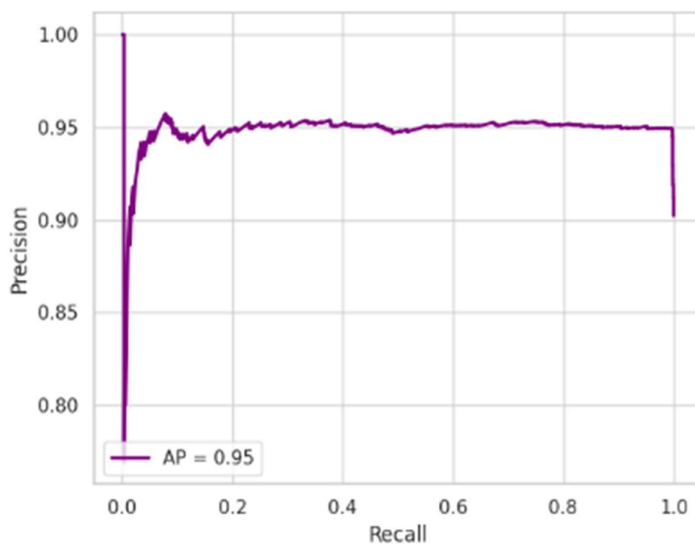
Por otro lado, se obtuvo un F1-Score de 0.66 para el evento de “Potencial Peligro” lo cual es de esperarse un valor bajo al obtener un 0.51 en el Recall ya que estas métricas se encuentran directamente relacionadas, el F1-Score se calcula a partir del Recall y Precision, y este cálculo castiga fuertemente si uno de estos dos valores es bajo.

El F1-Score entre 0.6 y 0.7 se considera aceptable, y dentro del contexto volcánico es realista, ya que no mide perfección sino compromiso operacional. Un F1-Score aceptable es consistente con un sistema de alerta aceptable, mas no un clasificador que sea determinante.

Posteriormente, se obtuvo la Curva de Precision-Recall donde se observa que se mantiene la tendencia.

Figura 25

Curva Precision-Recall MLP



Nota: AP siglas para Average Precision

La gráfica mantiene la misma interpretación en donde se mide la calidad de las alertas peligrosas y que en este caso casi siempre acierta.

De este resultado se puede interpretar a nivel del contexto sísmico en donde el modelo es muy conservado, es muy confiable cuando alerta eventos altamente peligrosos, sin embargo, es poco sensible ya que no detecta casi la mitad de los eventos que deberían marca una alerta.

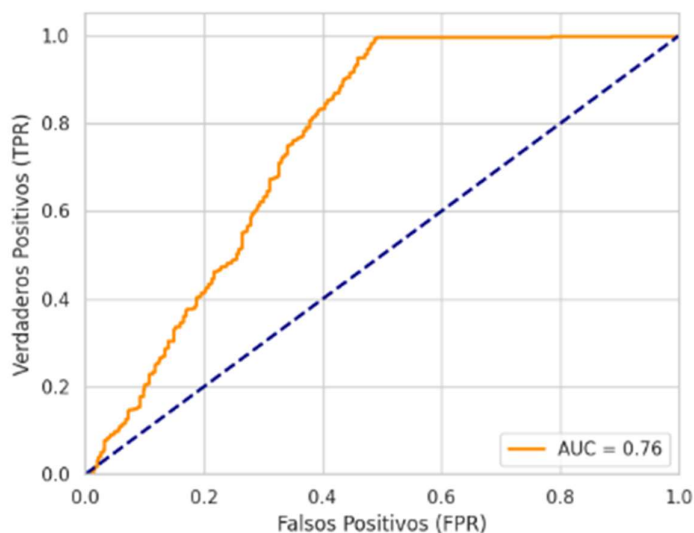
Un AP alto se puede interpretar también como un ranking interno dentro del modelo muy bueno para la clase de “Potencial Peligro”, aunque la frontera de decisión fina sea muy conservadora.

Dentro de un contexto aplicativo, el modelo no aporta mayormente para generar una alerta temprana automática ya que, no es suficiente porque pierde muchos eventos críticos. Sin embargo, sería excelente para un sistema de apoyo de decisión ya que da prioridad a los eventos con una alta probabilidad real.

En la curva ROC se observaron los siguientes resultados:

Figura 26

Curva ROC- MLP Classifier



Nota: AUC siglas para Area Under Curve

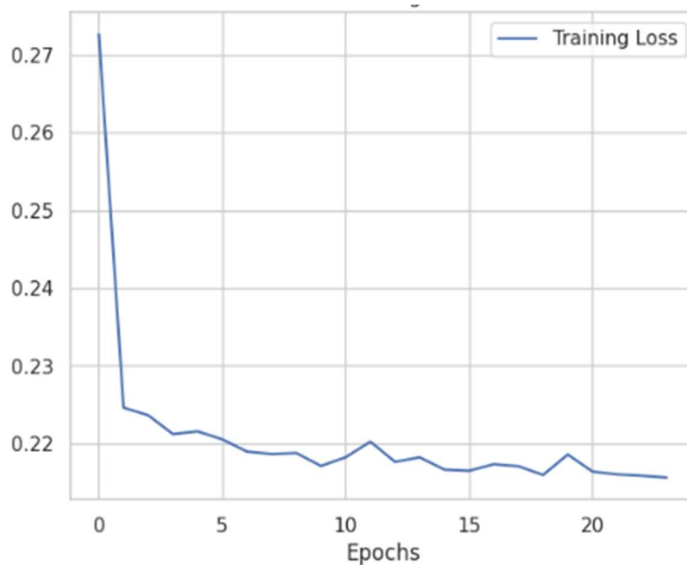
Al igual que las otras métricas, el AUC es consistente, marcando así un valor de 0.76, en este contexto es un valor bueno y aceptable ya que de haber obtenido un AUC >0.9 habría significado algún tipo de problema como fuga de información o aprendizajes triviales; el valor de 0.77 es el esperado ya que el modelo está trabajando a propósito con datos ruidosos e intenta imitar un escenario real.

Con este resultado se puede decir que, el modelo clasifica mejor que al azar y de manera consistente, pero le hace falta la optimización para maximizar los eventos altamente peligrosos.

Se muestra a continuación la curva de pérdida para este modelo.

Figura 27

Curva de Pérdida MLP Classifier



Se puede observar que la curva cae rápidamente en las primeras épocas, esto significa que el modelo ha encontrado rápidamente patrones informativos en el conjunto de datos, este es el resultado esperado debido a que las variables tienen un significado físico y se realizó correctamente el escalado.

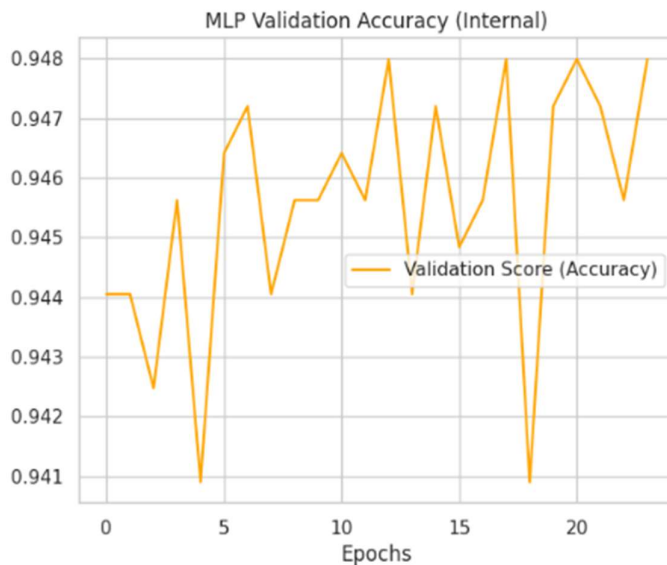
Luego, se observa una estabilización en la curva a medida que se van aumentando las épocas, sin embargo, no implica que más épocas mejoraría en modelo, esto quiere decir que el modelo ha extraído toda la información útil.

Desde el punto vista volcánico, se podría decir que no todos los sismos cercanos tienen la capacidad de activar al volcán y de igual manera que no todos los sismos grandes producen un efecto en la actividad sísmica y esto se debe principalmente a que se deben considerar además factores externos como el estado del magma, la presión, las fallas geológicas, entre otros muchos más para que se conjugue todo el escenario para un evento potencialmente peligroso.

La gráfica a continuación muestra una validación interna del accuracy en donde una fracción del conjunto de entrenamiento pasa a ser entrenada y calcula el accuracy en cada época.

Figura 28

Validación de Exactitud MLP Classifier



En esta gráfica se espera que las oscilaciones no son fuertes y no existan caídas abruptas, sin embargo, en esta gráfica se puede ver que existen dos picos negativos muy marcados, esto puede deberse al ruido gaussiano y el ruido intencional en las etiquetas.

También puede deberse al tamaño reducido del set de validación en donde si unos eventos mal clasificados pueden provocar las caídas en la curva.

De todas maneras, las caídas abruptas no son preocupantes ya que se puede ver que la gráfica se recupera rápidamente y vuelve a elevar la curva, aquí se puede ver que el early stopping no se detiene por una caída, sino que continúa mejorando. En términos generales, la curva se estabiliza alrededor de los valores medios de la figura y a pesar de las oscilaciones con caídas abruptas debido a la exploración bajo incertidumbre, no necesariamente es una mala generalización, sería preocupante en el caso de que mantuviera una tendencia descendente sostenida.

Finalmente, se muestran los resultados para el último modelo Support Vector Machine (SVM), al igual que los otros modelos se realizó la búsqueda de los mejores parámetros con Grid Search, el resultado obtenido fue el siguiente:

Figura 29

Mejores hiperparámetros para SVM

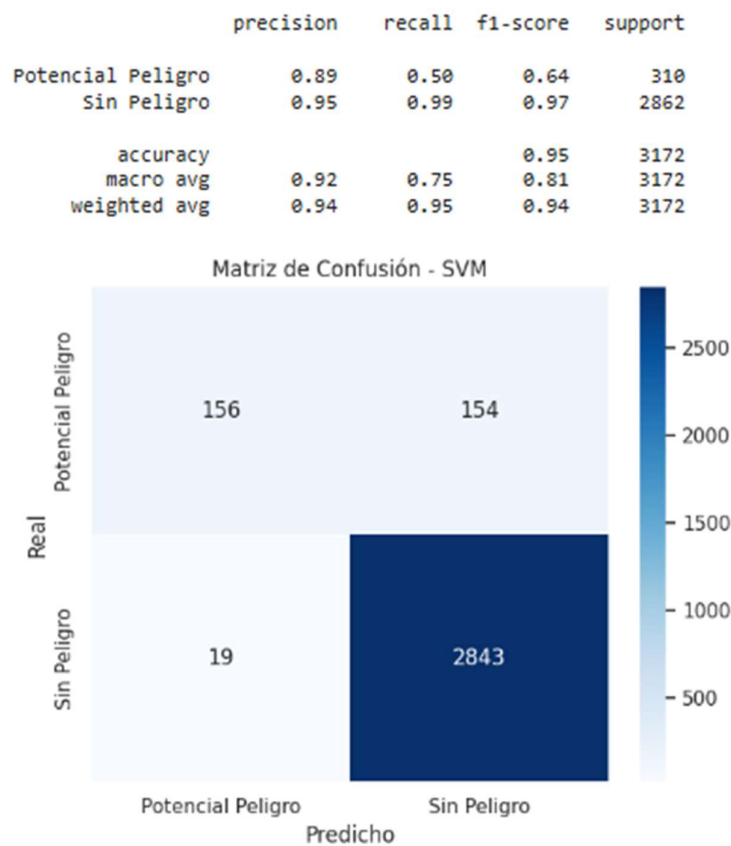
--- TOP 5 MEJORES HIPERPARÁMETROS ---				
	params	mean_test_score	std_test_score	rank_test_score
	{'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}	0.9441	0.002408	1
	{'C': 1, 'gamma': 'scale', 'kernel': 'rbf'}	0.9430	0.001769	2
	{'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}	0.9424	0.001658	3
	{'C': 0.1, 'gamma': 0.1, 'kernel': 'poly'}	0.9421	0.002117	4
	{'C': 1, 'gamma': 0.01, 'kernel': 'rbf'}	0.9419	0.002178	5

Los hiperparámetros evaluados solo fueron tres, esto indica que es un Grid razonable y controlado evitando así las combinaciones extremas.

Se muestra a continuación las métricas obtenidas en este modelo:

Figura 30

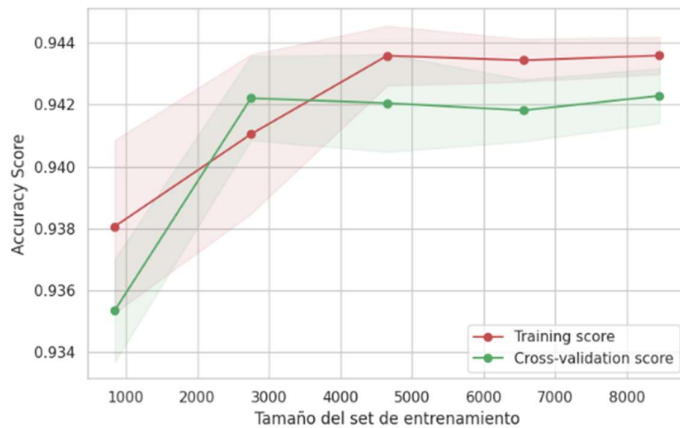
Reporte de Clasificación y Matriz de Confusión SVM



En este caso se observa que la predicción de los eventos predichos y acertados como potencialmente peligrosos ha disminuido, lo cual hace que el Recall quede en un 50/50 en la capacidad de discriminar esta clase. De igual manera aumenta las falsas alarmas, en donde se catalogan 19 eventos como potencialmente peligrosos cuando en realidad no lo eran.

Se observa también que el accuracy global se mantiene en el 0.95, lo cual es un rendimiento bueno del modelo en general.

Se realizó la curva de aprendizaje para evaluar al modelo:

Figura 31*Curva de Aprendizaje SVM*

Nota: Este Algoritmo no entrena por Epocas

Esta gráfica muestra una curva saludable ya que espera que el training score se encuentre por encima de la curva de cross-validation, con una ligera brecha entre ellas. Además, se observa que las curvas no se acercan al 1 y esto valida que no existe sobreajuste, así como también las curvas no se juntan demasiado indicando que no existe underfitting; finalmente se muestra una tendencia a la convergencia de las curvas a medida que el número de datos va en aumento.

El modelo SVM mostró un comportamiento competitivo, aunque ligeramente inferior a los anteriores modelos, especialmente en la capacidad de detección de eventos peligrosos, esto podría sugerir que a pesar de que el margen es efectivo, el ruido y el solapamiento entre las clases limita la capacidad de discriminación.

Los resultados obtenidos en los cuatros modelos pueden ser comparados ya que fueron pre procesados para que la competitividad entre estos, se encuentren en las mismas condiciones. A continuación, se muestra una tabla resumen de los resultados más relevantes:

Tabla 5*Resumen del reporte de clasificación de todos los modelos*

Clasificación	Precision	Recall	F1-Score	Accuracy
Random Forest				0.95
Potencial Peligro	0.97	0.51	0.67	
Sin Peligro	0.95	1	0.97	
XGBoost				0.95
Potencial Peligro	0.96	0.51	0.67	
Sin Peligro	0.95	1	0.97	
Red Neuronal MLP				0.95
Potencial Peligro	0.93	0.51	0.66	
Sin Peligro	0.95	1	0.97	
SVM				0.95
Potencial Peligro	0.89	0.5	0.64	
Sin Peligro	0.95	0.99	0.97	

Se puede observar que en todos los modelos el Accuracy fue de 0.95 lo cual indicaría que es un excelente rendimiento, sin embargo, dentro del contexto de este estudio esta métrica es importante pero no determinante. Este resultado puede deberse a que clasificó bien la mayoría de eventos, de los cuales en su mayoría caían en una sola clase, por ello el accuracy es alto de forma global, sin embargo, en el resultado por clases muestra métricas más bajas.

Los resultados en general son casi iguales en todos los modelos con ligeras variaciones en algunas métricas, a simple vista se puede descartar a SVM debido a que muestra una precisión y un recall menor en la clase de potencial peligro.

Por lo tanto, ahora se determinó cual es el mejor modelo por accuracy ya que es el que mide el desempeño global en el conjunto de prueba.

Figura 32*Comparativa mejor modelo*

```
=====
RESUMEN FINAL Y GUARDADO
=====

Ranking de Modelos por Accuracy:
      Model Accuracy
Random Forest 0.950504
      XGBoost 0.950504
      MLP    0.948613
      SVM    0.945460

El mejor modelo fue Random Forest.
Archivo guardado exitosamente como: best_model_Random_Forest.pkl
```

Nota: El mejor modelo se graba en un archivo tipo pkl para posteriormente ser consumido por un aplicativo Tkinter y analizar nueva data.

El modelo final fue llevado con su pipeline de pre procesamiento, asegurando así su reproducibilidad. A pesar de que la selección final fue por *Accuracy* , otras métricas como Recall y F1-score fueron analizadas y discutidas durante el análisis.

A continuación, se muestran los resultados obtenidos en el aplicativo desarrollado con Tkinter, la interfaz p del sistema presenta un encabezado informativo que identifica el propósito de la aplicación, la institución académica responsable y los integrantes del equipo de desarrollo. Este diseño permite el uso del sistema desde el primer contacto del usuario, reforzando su carácter analítico.

Figura 33*Interfaz del aplicativo*

La interfaz gráfica muestra un despliegue amigable para el usuario ya que muestra pasos simples a seguir, además de una barra de colores donde existen botones para:

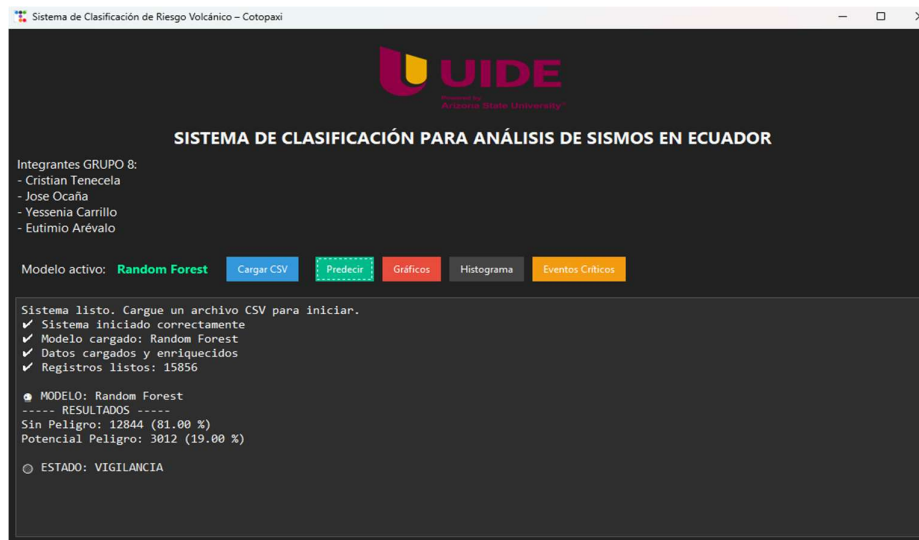
- Carga de archivo CSV (Botón celeste)
- Predecir (Botón verde)
- Gráficos (Botón rojo)
- Histograma (Botón negro)
- Eventos críticos (Botón naranja)

Cada botón muestra su funcionalidad y permite al usuario elegir la opción que requiera analizar.

Figura 34*Interfaz carga de archivo*

Nota: Se debe cargar un archivo en formato CSV

El sistema incorpora un modelo de clasificación supervisada basado en Random Forest, previamente entrenado y validado, el cual utiliza como variables de entrada la latitud, longitud, profundidad del evento sísmico, magnitud y la distancia respecto al volcán Cotopaxi. A través de la interfaz, el usuario puede cargar archivos en formato CSV que contienen registros sísmicos, los cuales son procesados automáticamente para garantizar la consistencia y correcta estructuración de los datos antes de ser analizados por el modelo.

Figura 35*Interfaz con clasificación de eventos sísmicos*

Una vez cargada la información, la aplicación permite ejecutar el proceso de predicción, clasificando cada evento en dos categorías: *Sin Peligro* y *Potencial Peligro*. El sistema utiliza probabilidades generadas por el modelo y un umbral de decisión configurable para determinar la clase final, lo que proporciona mayor control y flexibilidad en la interpretación de los resultados. Los resultados de la clasificación se presentan tanto en valores absolutos como en porcentajes, permitiendo una comprensión clara de la distribución del riesgo en el conjunto de datos analizado.

Adicionalmente, la interfaz incluye módulos gráficos que permiten visualizar la distribución de las predicciones mediante gráficos de barras e histogramas de probabilidad, incorporando una línea de referencia correspondiente al umbral de decisión. Estas visualizaciones facilitan el análisis exploratorio y la interpretación del comportamiento del modelo frente a distintos escenarios sísmicos.

Figura 36*Interfaz con gráfica de distribución*

El modelo se encuentra entrenado para dar una clasificación Binaria: Sin Peligro & Potencial Peligro; es de esperarse que exista una mayor cantidad de eventos sísmicos considerados como “Sin Peligro” ya que las grandes erupciones volcánicas del Cotopaxi se han dado cada 100 años y en intervalos menores, han sido erupciones moderadas.

Figura 37

Interfaz con histograma de probabilidad



Nota: La línea roja marca el límite entre Potencial Peligro (derecha) y Sin Peligro (izquierda)

Figura 38

Interfaz con tabla de eventos críticos



El sistema también dispone de una tabla de eventos críticos, en la cual se destacan aquellos registros clasificados como *Potencial Peligro*, proporcionando información clave

como fecha, magnitud, distancia y probabilidad estimada. Este componente resulta fundamental para el análisis focalizado de eventos relevantes y para la toma de decisiones basada en evidencia.

Finalmente, la aplicación incluye un indicador de estado del sistema que informa al usuario sobre la correcta carga del modelo, el procesamiento de los datos y la disponibilidad de resultados, asegurando transparencia y confiabilidad durante el uso de la herramienta. En conjunto, el sistema constituye una solución integral que combina modelado predictivo, visualización interactiva y una interfaz amigable, aportando una herramienta de apoyo al análisis sísmico y a la evaluación del riesgo volcánico en Ecuador.

CAPÍTULO 5

5.1 CONCLUSIONES

El alcance de este trabajo comprende un análisis profundo de la actividad sísmica relacionada al volcán Cotopaxi. En el Ecuador existen 34 volcanes activos y el que representa mayor amenaza es el Cotopaxi debido al impacto ambiental, social y económico que puede desencadenar la erupción de este. El estudio de la actividad volcánica es muy complejo ya que, los científicos deben analizar características como la deformación del suelo, los gases, la temperatura, observaciones visuales y acústicas y la actividad sísmica para la caracterización de una erupción volcánica. Este estudio se basa en una de estas características que es la sismicidad, para lo cual se analizó patrones en características como latitud, longitud, profundidad, entre otros.

El objetivo del proyecto fue entrenar modelos de clasificación que permitan discriminar la variable objetivo en una clasificación binaria: “Potencial Peligro” y “Sin Peligro” en un conjunto de datos de eventos sísmicos históricos, los cuales fueron recolectados por el Instituto Geofísico de la Escuela Politécnica Nacional y son de dominio público, el principal reto con la recolección de data fue que por ser mediciones de muchos años atrás, estos catálogos se obtuvieron en formatos y características diferentes, por lo tanto fue necesario un proceso de ETL, este permitió la consolidación de varios catálogos sísmicos heterogéneos en un solo catálogo homogéneo, fue necesaria la estandarización de las magnitudes mediante un proxy Mw con el objetivo de poder comparar temporalmente los eventos, manteniendo la coherencia de la distribución sísmica.

Una vez que se obtuvo un único catálogo con la data procesada, fue necesario trabajar en la variable objetivo y estructurar el problema de este estudio, para lo cual se realizó la clasificación de la peligrosidad de un sismo en función de la actividad volcánica del Cotopaxi. Por lo tanto, se definió los criterios de aceptación, basados en la literatura, de la clase

“Potencial Peligro” en la que los eventos sísmicos deben tener características físicas (distancia y magnitud) específicas para que puedan alterar o influenciar la estabilidad sísmica del volcán Cotopaxi. Durante este proceso de clasificación de la peligrosidad, se identificó que existe un claro desbalance de clases, pues casi la totalidad de eventos se clasificaron como “Sin Peligro”, este resultado tuvo mucha lógica ya que, en el contexto volcánico, el Cotopaxi ha tenido 5 grandes periodos de erupciones desde la época colonial, por lo tanto, casi toda la data debe representar los periodos en los que no hubo actividad sísmica considerable. En base a este resultado, fue necesario hacer un balanceo de clases ya que, para los entrenamientos de los algoritmos, la data real no aportaba datos de entrenamiento para la clase minoritaria y hubiera caído en el sobreajuste.

El balanceo de clases consistió en la generación de datos sintéticos que sean considerados como Potencial Peligro, los datos no fueron generados al azar, sino que cumplieron parámetros técnicos apegados a la teoría sísmica, estadística y matemática que fueron lo más reales posibles y al ser introducidos en la data con los datos reales, se encuentren en las mismas condiciones y los entrenamientos no aprendan patrones triviales equivocados. Esta generación de datos sintéticos permitió construir una base de datos balanceada en donde se introdujo también ruido para evitar el sobreajuste. La preparación de los datos fue un proceso fundamental, el cual incluyó la incorporación de ruido y una estandarización total de las variables, de esta manera se garantizó que los modelos no graben ni reproduzcan reglas predefinidas, esto permitió evaluar la capacidad real de la generalización de los clasificadores en datos inciertos.

El modelo de clasificación que se estructuró, permitió identificar eventos sísmicos con potencial capacidad de alterar la actividad volcánica del Cotopaxi, basándose en características físicas como la distancia y la magnitud. Se puede decir que el enfoque es conservador y los resultados son coherentes con la teoría y la realidad sísmo-volcánica, sin embargo, se limita su capacidad clasificatoria debido a la ausencia de información interna del volcán y el análisis de otras magnitudes de interés como la presión, la dinámica de del

magma y gases internos, entre otros. Por lo tanto, los resultados deben ser interpretados como indicadores de atención más como predicciones eruptivas.

Por otro lado, se pudo observar los resultados (Tabla 3) de rendimiento de los cuatro modelos entrenados: Random Forest, XGBoost, Red Neuronal MLP y Support Vector Machine; en donde las métricas son constantes para todos los modelos con ligeras variaciones. Se puede concluir que, a manera global, los modelos presentan un buen rendimiento ya que todos obtuvieron un accuracy de 0.95, esto debido a que de los 3172 eventos sísmicos que fueron parte del conjunto de prueba, un promedio de 2851 sismos fue clasificados correctamente, es decir un 90% de los datos de prueba son bien clasificados, por ello se obtiene un accuracy alto.

Sin embargo, al analizar individualmente las categorías, se pudo observar que la categoría crítica de “Potencial Peligro” tiene un bajo nivel predictivo pues al obtener un recall en promedio de 0.51 esto nos indica que cerca de la mitad de eventos sísmicos no son detectados como amenaza real cuando sí lo son, por lo tanto el modelo se limita a ser aplicado como una herramienta de prevención de amenaza temprana, a pesar de ello, se pudo destacar que los modelos obtuvieron en promedio un AP de 0.95 lo cual se puede interpretar como una veracidad cuando los modelos detectan la amenaza, es decir que cuando un evento es clasificado como potencialmente peligro, tiene un 95% de probabilidades que realmente lo sea, por tanto el modelo puede ser aplicado como un modelo de apoyo de decisión sobre otro estudio a futuro que tenga mejor capacidad predictiva.

Así mismo, se pudo analizar el AUC que en promedio se obtuvo un valor de 0.77 lo cual se puede considerar como un valor bueno-aceptable y esto nos indica que posee en general una capacidad de clasificación adecuada en condiciones reales de incertidumbre, aun que su desempeño operativo va a depender del umbral de decisión que decida.

En términos generales podemos concluir que los modelos con mejores rendimientos fueron Random Forest y XGBoost, ya que sus métricas fueron ligeramente mejor en comparación a MLP y SVM, sin embargo, los 4 modelos tuvieron buenos resultados ya que ninguno mostró comportamiento errático o aprendizaje colapsado. En tanto MLP es competitivo, pero menos interpretable y estable en este contexto, que a diferencia de RF. Por su parte XGBoost tuvo las métricas más bajas en esta comparativa, por lo que se puede decir que, es válido pero el menos óptimo en este contexto.

Finalmente, se puede concluir que los resultados de los algoritmos mostrados, pueden capturar patrones complejos de la actividad sismo-volcánica. De ellos, se determinó que Random Forest, fue el que obtuvo mejor equilibrio entre desempeño global, estabilidad y capacidad de generalización, por ello se seleccionó como modelo final para la reproducibilidad en el aplicativo desarrollado en Tkinter. Sin embargo, al haber obtenido un recall moderado en la clase crítica de mayor interés, se genera una necesidad de realizar en el futuro ajustes necesarios para maximizar la detección de eventos potencialmente peligrosos.

5.2 RECOMENDACIONES

Se recomienda enfocarse en mejorar el recall de la clase crítica, ya que el objetivo es la detección de eventos sísmicos que representen una amenaza real, se propone optimizar los modelos enfocándose en Recall o F1-score y ya no en accuracy, así como también ajustar el umbral de decisión y una implementación de cost-sensitive learning ya que de esta manera penaliza más fuerte a los falsos negativos. Con estas técnicas se esperaría reducir el riesgo de no detectar eventos críticos.

Por otro, al nivel de la calidad de los datos se recomienda enriquecer las características mediante la implementación de variables que midan la energía sísmica liberada, el tipo de evento sísmico ya sea volcánico-tectónico (VT), long-period (LP) o tremor, esto permitiría

hacer agrupaciones de donde se puedan generar patrones específicos por cada grupo y hacer un estudio personalizado para cada uno, en donde se esperaría mejorar la frontera de decisión.

También se puede fortalecer la generalización del modelo validando los datos con los de otros volcanes activos, evaluando la robustez del modelo en distintos periodos eruptivos y realizando un cross validation para evitar la fuga de información geográfica; de esta manera se puede medir la transferibilidad y reproducibilidad del modelo.

Se recomienda también la incorporación de intervalos de confianza para predicciones y un análisis de sensibilidad a ruido, esto es un paso fundamental para sistemas de alerta temprana.

Se recomienda también una integración de dimensión temporal ya este estudio analizó los sismos de manera independiente, sin embargo, el comportamiento volcánico sugiere una dependencia temporal, por lo tanto, se recomienda explorar modelos que sean capaces de capturar secuencias temporales tales como LSTM o GRU.

Se sugiera finalmente, contrastar los resultados con métodos tradicionales empíricos y estadísticos utilizados ya en la vigilancia volcánica, esto con el objetivo de cuantificar objetivamente las ventajas, desventajas y limitaciones del enfoque basado en Machine Learning versus a metodologías clásicas.

Este trabajo constituye un aporte inicial en el análisis de riesgo sismo-volcánico con técnicas de aplicación de aprendizaje automático en un ambiente real en el volcán Cotopaxi. Se espera que la implementación de las recomendaciones planteadas anteriormente, permita la optimización de la precisión, la discriminación de clases, la interpretabilidad y la aplicabilidad del modelo a otras fuentes de datos y de esta manera se consolide como una herramienta de apoyo para las alertas tempranas y la toma de decisiones en sistemas de vigilancia y análisis de riesgo de la actividad volcánica.

REFERENCIAS BIBLIOGRÁFICAS

- Ardid, A., Dempsey, D., Caudron, C., Cronin, S., & Kennedy, B. (2023). *Generalized eruption forecasting models using machine learning trained on seismic data from 24 volcanoes*. *Research Square*. <https://doi.org/10.21203/rs.3.rs-3483573/v1>
- Ardid, A., Dempsey, D., Caudron, C., Cronin, S., & Kennedy, B. M. (2025). *Ergodic seismic precursors and transfer learning for short term eruption forecasting at data scarce volcanoes*. *Nature Communications*, 16, Article 56689. <https://doi.org/10.1038/s41467-025-56689-x>
- Ardid, A., Dempsey, D., Corry, J., & Lamb, O. D. (2024). *Multitimescale template matching: Discovering eruption precursors across diverse volcanic settings*. *Seismological Research Letters*. <https://doi.org/10.1785/0220240012>
- Perales Palacios, F. J., Carrillo Rosúa, F. J., García Yeguas, M. A., & Vázquez Vilchez, M. M. (2021). Los volcanes: Algunas perspectivas para un conocimiento científico y didáctico. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias*, 18(3), Article 3105. https://doi.org/10.25267/Rev_Eureka_ensen_divulg_cienc.2021.v18.i3.3105
- Bueno, A., Balestrieri, R., De Angelis, S., Benitez, C., & Zuccarello, L. (2021). *Recurrent scattering network detects metastable behavior in polyphonic seismo-volcanic signals for volcano eruption forecasting*. *IEEE Transactions on Geoscience and Remote Sensing*, 60, Article 3134198. <https://doi.org/10.1109/TGRS.2021.3134198>
- Cabrera, L., Ardid, A., Melchor, I., Ruiz, S., & Symmes-Lopetegui, B. (2024). *Eruption forecasting model for Copahue volcano using seismic data and machine learning*. *Seismological Research Letters*, 95(6), 3186–3199. <https://doi.org/10.1785/0220240022>
- Córdoba, G. A., Costa, A., Rodríguez, D. M., Thouret, J. C., Criollo, R. A., & Guerrero, D. A. (2025). *Assessing probabilistic hazard and risk for building, road network, and ecosystems: the case study of La Florida municipality, Galeras Volcano, Colombia*. *Frontiers in Earth Science*, 13, 1632282. <https://doi.org/10.3389/feart.2025.1632282>
- Carniel, R., & Guzmán, S. (2020). Machine learning in volcanology: A review. In *Volcanoes - Updates in Volcanology*. *IntechOpen*. <https://doi.org/10.5772/INTECHOPEN.94217>
- Carniel, R., & Guzmán, S. (2025). *Machine learning for volcanology and volcano monitoring*. En D. M. Pyle & T. A. Mather (Eds.), *Advances in volcanology*. Springer. https://doi.org/10.1007/978-3-031-86841-2_14
- Dempsey, D. E., Cronin, S. J., Mei, S., & Kempa-Liehr, A. W. (2020). *Automatic precursor recognition and real-time forecasting of sudden explosive volcanic eruptions at Whakaari (White Island), New Zealand*. *Nature Communications*, 11, Article 3562. <https://doi.org/10.1038/s41467-020-17375-2>
- Di Stasio, P., Sebastianelli, A., Meoni, G., & Ullo, S. L. (2022). *Early detection of volcanic eruption through artificial intelligence on board*. En *2022 IEEE International Workshop on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroX-RAINE)* (pp. 616–621). *IEEE*. <https://doi.org/10.1109/MetroXRAINE54828.2022.9967616>
- Fenner, D., Ruz Ginouves, C., Koulakov, I., Schurr, B., Gottsmann, J., & GEOFON Data Centre. (2022). AWESAM: A tool for the automatic detection, location, and analysis of volcanic

- seismicity based on waveform stacking. *Journal of Volcanology and Geothermal Research*, 432, Article 107689. <https://doi.org/10.1016/j.jvolgeores.2022.107689>
- Ramos, L., Schab, E. A., Rivera, R., Richard, C., Cristaldo, P., Núñez, J. P., ... & Herrera, N. E. (2019). *Descubrimiento de conocimiento en bases de datos*. <https://doi.org/11336/16126>
- Gutiérrez, L., Rey Devesa, P., Ibáñez, J., & Benítez, C. (2025). *Shannon entropy estimator for the characterization of seismic-volcanic signals using Python*. Preprint. <https://doi.org/10.48550/arXiv.2511.01530>
- Chlieh, M., Beauval, C., Yepes, H., Marinière, J., Saillard, M., & Audin, L. (2021). *Seismic and aseismic cycle of the Ecuador–Colombia Subduction Zone*. *Frontiers in Earth Science*, 9, Article 701720. <https://doi.org/10.3389/feart.2021.701720>
- Hidalgo, S., Bernard, B., Mothes, P., Ramos, C., & Aguilar, J. (2023). *Hazard assessment and monitoring of Ecuadorian volcanoes: Challenges and progress during four decades since IG-EPN foundation*. *Bulletin of Volcanology*, 85, Article 71. <https://doi.org/10.1007/s00445-023-01685-6>
- Ichihara, M., Takeo, M., Ohminato, T., Abe, Y., & Maeda, Y. (2023). *Long-term monitoring reveals unique features of volcanic seismicity at Shinmoe-dake, Kirishima Volcano, Japan*. *Earth, Planets and Space*, 75, 119. <https://doi.org/10.1186/s40623-023-01883-8>
- IEEE. (2022). Bayesian monitoring of seismo-volcanic dynamics. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-17. <https://doi.org/10.1109/tgrs.2021.3076012>
- Journeau, C., Shapiro, N. M., Seydoux, L., Soubestre, J., Ferrazzini, V., & Peltier, A. (2020). *Detection, classification, and location of seismo-volcanic signals with multicomponent seismic data: Example from the Piton de la Fournaise volcano (La Réunion, France)*. *Journal of Geophysical Research: Solid Earth*, 125(8), e2019JB019333. <https://doi.org/10.1029/2019JB019333>
- Kaza, V. S., Reddy, C. K. K., Reddy, K. N., Sreelatha, G., & Lippert, K. (2025). *Random forests to predict volcanic eruptions in the Pacific Ring of Fire*. In *Lecture Notes in Electrical Engineering* (pp. 49-60). Springer. https://doi.org/10.1007/978-981-96-6297-5_5
- Khan, M. S., Curilem, M., & Huenupan, F. (2019). *A signal processing perspective of monitoring active volcanoes*. In *2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA)* (pp. 1-5). IEEE. <https://ieeexplore.ieee.org/document/8887573>
- Lecocq, T., Caudron, C., & Toyon, M. A. S. (2023). Towards monitoring phreatic eruptions using seismic noise. *EGU sphere*. <https://doi.org/10.5194/egusphere-egu23-7166>
- Longo, R., Lacanna, G., Innocenti, L., & Ripepe, M. (2024). *Artificial intelligence and machine learning tools for improving early warning systems of volcanic eruptions: The case of Stromboli*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. <https://doi.org/10.1109/TPAMI.2024.3399689>
- Mandita, F., Ashari, A., Wibowo, M. E., & Suryanto, W. (2024). *Propose design of Merapi volcano eruption prediction using hybrid time series and machine learning*. *2024 International Conference on ICT for Smart Society (ICISS)*, 1-6. <https://doi.org/10.1109/icic64337.2024.10957023>

- Mousavi, S. M., Beroza, G. C., Somala, S. N., Baraldi, F., & Zhu, W. (2023). *Deep learning for fast and reliable seismic monitoring*. *Science*, 382(6675), eadh9859. <https://doi.org/10.1146/annurev-earth-071822-100323>
- Mustafa, A. H., AbdelMoneim, F. M., Matta, M. G., Barghash, T. O., & Gomaa, W. (2022). *Automatic forecasting of volcanoes eruption time*. *2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, 648-653. <https://doi.org/10.1109/IMCOM53663.2022.9721804>
- Ollino, A. (2022). A machine learning method for seismic signal monitoring: *A contribution to the detection of the potential volcanic hazard on Etna, Italy*. *EGU General Assembly 2022*. <https://doi.org/10.5194/egusphere-egu22-1151>
- Rey-Devesa, P., Prudencio, J., Martí, J., Benítez, C., Ibáñez, J. M., García-Yeguas, A., & Almendros, J. (2023). Using *Shannon entropy of seismic signals for volcanic eruption forecasting*. *Scientific Reports*, 13(1), 14756. <https://doi.org/10.1029/2023JB026684>
- Rey-Devesa, P., Ibáñez, J. M., Benítez, C., Prudencio, J., Martí, J., García-Yeguas, A., & Almendros, J. (2023). Forecasting and tracking volcanic explosions using Shannon entropy at Volcán de Colima. *Research Square*. <https://doi.org/10.21203/rs.3.rs-2718630/v1>
- Syabhana, D. K., Kasbani, K., Suantika, G., Prambada, O., Andreas, A. S., Saing, U. B., Lowenstern, J. B. (2019). The 2017–19 activity at Mount Agung in Bali (Indonesia): Intense unrest, monitoring, crisis response, evacuation, and eruption. *Scientific Reports*, 9, Article 8848. <https://doi.org/10.1038/s41598-019-45295-9>
- Shyamala, K., David Solomon Raju, Y., Swarnalatha, C., Saidulu, D., & Premala Tha, M. (2022). *Leveraging machine learning to predict volcanic eruptions by function neural networks*. *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1310-1315. <https://doi.org/10.1109/ICIRCA54612.2022.9985483>
- Spina, R., Fornaia, A., & Tramontana, E. (2020). *VSEW: An early warning system for volcanic and seismic events*. *2020 IEEE International Conference on Smart Computing (SMART-COMP)*, 316-321. <https://doi.org/10.1109/SMARTCOMP50058.2020.00084>
- Vaca Molina, V. P. (2024). Estimación del índice de peligrosidad de los volcanes potencialmente activos del Ecuador. *ESPOL. FICT*. <http://www.dspace.espol.edu.ec/handle/123456789/63114>
- Wang, K., Waldhauser, F., Schaff, D. P., Tolstoy, M., & Wilcock, W. S. D. (2024). Real-time detection of volcanic unrest and eruption at Axial Seamount using machine learning. *ESSOAr*. <https://doi.org/10.22541/essoar.171415883.36620444/v1>
- Yukutake, Y., Kim, A., & Ohminato, T. (2023). Reappraisal of volcanic seismicity at the Kiri-shima volcano using machine learning. *Earth, Planets and Space*, 75(1), 183. <https://doi.org/10.1186/s40623-023-01939-9>
- Zali, M., Gudmundsson, M. T., & Einarsson, P. (2024). *Deep learning for seismic tremor clustering and eruption forecasting at Geldingadalir, Iceland*. *Journal of Volcanology and Geothermal Research*, 445, 107951. <https://doi.org/10.1038/s43247-023-01166-w>

APÉNDICES

Link del Notebook

<https://colab.research.google.com/drive/1b5guuVq4IHZMHjU8si66HV-kGPpclCgn?usp=sharing#scrollTo=p408HBXtzUcw>