UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Classification of Cotopaxi Volcano Seismic Events using Semi-Supervised Learning

Proyecto de Titulación

Pavel Estrella Gordillo

Felipe Grijalva, Ph.D.

Director de Trabajo de Titulación

Trabajo de titulación de posgrado presentado como requisito para la obtención del título de Magíster en Ciencia de Datos

Quito, 02 de diciembre de 2024

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ
COLEGIO DE POSGRADOS

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Título Trabajo de Titulación

Pavel Estrella Gordillo

Nombre del Director del Programa:	Felipe Grijalva
Título académico:	Ph.D. en Ingeniería Eléctrica
Director del programa de:	Ciencia de Datos

Nombre del Decano del colegio Académico:	Eduardo Alba
Título académico:	Doctor en Ciencias Matemáticas
Decano del Colegio:	Ciencias e Ingenierías

Nombre del Decano del Colegio de Posgrados:	Dario Niebieskikwiat
Título académico:	Doctor en Física

Quito, diciembre 2024

© DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

Nombre del estudiante: Pavel Estrella Gordillo

Código de estudiante: 339246

C.I.: 0550010250

Lugar y fecha: Quito, 02 de diciembre de 2024.

ACLARACIÓN PARA PUBLICACIÓN

UNPUBLISHED DOCUMENT

DEDICATORIA

El presente trabajo va dedicado a las personas más importantes de mi vida, a mi familia. A Dacia, mi madre, quien con su comprensión siempre me ha brindado su apoyo en las decisiones que he tomado. A Hernan, mi padre, por su cariño, el cual me ha ayudado a moldear mi personalidad, caracter y pasión para enseñar. A Jadira, mi hermana, mi ejemplo a seguir tanto en su labor profesional como su capacidad para enfrentar retos y superar metas. A los demás mientros de mi familia quienes son la base de mi proyecto de vida.

Asimismo, deseo expresar mi sincero agradecimiento a mi amiga, Gisella, cuyo apoyo moral, sus grandiosas conversaciones y su perspectiva única, distinta a la mía, me ha brindado una visión más amplia y consciente de mi realidad. Su orientación contribuyó significativamente a la calidad final de este documento, ayudándome a lograr una presentación más clara, limpia y organizada, reflejo del crecimiento alcanzado gracias a su valiosa ayuda.

A la memoria de mi abuelita, Mariettita, cuyo legado siempre será la educación. Ella será mi eterna maestra, sus enseñanzas más valiosas fueron el trabajo duro, la perseverancia y la fortaleza. Se que estas impartiendo tu sabiduría en otro horizonte, y espero que nos volvamos a reencontrar.

AGRADECIMIENTOS

Los autores agradecen a la Escuela Politécnica Nacional por proporcionar el conjunto de datos utilizado en esta investigación. Su apoyo y contribución fueron esenciales para el desarrollo y éxito de este trabajo.

RESUMEN

La clasificación de microseísmos volcánicos es crucial para identificar el tipo de evento en situaciones potencialmente peligrosas. No obstante, el etiquetado de estos eventos es una tarea dificultosa y costosa en cuestion de tiempo, ya que requiere de expertos en el área. Para reducir la dependencia de grandes conjuntos de datos etiquetados, se propone un enfoque semi-supervisado que incorpora dos algoritmos: Self-Training con múltiples clasificadores base y Label Spreading. Adicionalmente, se implementa técnicas avanzadas de reducción y transformación de características para optimizar la representación de los datos de entrada. Todos los experimentos se realizaron utilizando la misma base de datos proporcionada. Los resultados indican que los modelos basados en Random Forest y SVM, empleando solo el 10% de los datos etiquetados, superan el rendimiento de los algoritmos supervisados tradicionales. Sin embargo, encontramos resultados menos satisfactorios con Naive Bayes, debido a la ausencia de ajuste de hiperparámetros, y con Label Spreading, atribuible a las limitaciones intrínsecas del propio algoritmo. Estos hallazgos destacan el significativo potencial de los enfoques semi-supervisados, especialmente cuando se seleccionan y optimizan adecuadamente los algoritmos base y las características utilizadas.

Palabras clave: Microseísmos,Clasificación semisupervisada,Análisis sísmico,Aprendizaje automático,Volcán Cotopaxi.

ABSTRACT

The classification of volcanic microseismic events is crucial to identify the type of event in potentially hazardous situations. However, labeling these events is a difficult and time-consuming task, as it requires experts in the field. To reduce the dependency on large labeled datasets, a semi-supervised approach is proposed that incorporates two algorithms: Self-Training with multiple base classifiers and Label Spreading. Additionally, advanced feature reduction and transformation techniques are implemented to optimize the representation of the input data. All experiments were performed using the same database provided. The results indicate that Random Forest and SVM based models, employing only 10% of the labeled data, outperform traditional supervised algorithms. However, we found less satisfactory results with Naive Bayes, due to the absence of hyperparameter fitting, and with Label Spreading, attributable to the intrinsic limitations of the algorithm itself. These findings highlight the significant potential of semi-supervised approaches, especially when the base algorithms and features used are properly selected and optimized.

Key words: Microseisms, Semi-supervised classification,Seismic analysis, Machine learning, Cotopaxi Volcano.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Classification of Cotopaxi Volcano Seismic Events using Semi-Supervised Learning

Pavel Estrella, Felipe Grijalva, *Senior Member, IEEE*

***Abstract*—The classification of volcanic microseismic events is crucial to identify the type of event in potentially hazardous situations. However, labeling these events is a difficult and time-consuming task, as it requires experts in the field. To reduce the dependency on large labeled datasets, a semi-supervised approach is proposed that incorporates two algorithms: Self-Training with multiple base classifiers and Label Spreading. In addition, advanced feature reduction and transformation techniques are implemented to optimize the representation of the input data. All experiments were performed using the same database provided. The results indicate that Random Forest and SVM based models, employing only 10% of the labeled data, outperform traditional supervised algorithms. However, we found less satisfactory results with Naive Bayes, due to the absence of hyperparameter fitting, and with Label Spreading, attributable to the intrinsic limitations of the algorithm itself. These findings highlight the significant potential of semi-supervised approaches, especially when the base algorithms and features used are properly selected and optimized.**

***Index Terms*—Microseisms, Semi-supervised classification,Seismic analysis, Machine learning, Cotopaxi Volcano.**

I. Introduction

VOLCANIC eruptions are one of the most destructive geological events on the planet, since they generate diverse consequences depending on the intensity with which they occur. These eruptions affect not only the populations near the volcano, but also more distant communities due to factors inherent to the eruptive activity. It is known that about 800 million people live within 100 km of an active volcano, 226 million live within 30 km and 29 million live within 10 km in 86 different countries [1].

Inhabitants near the eruption may experience different risks, such as gas emissions, ash fall, lahars on the flanks of the volcano, lava flows, seismic activity, climate change, among others. According to [2], volcanic eruptions in the twentieth century have claimed the lives of about 80 thousand people around the world, this figure is limited not only to the eruptive event but also to the secondary risks that affected the population. Therefore, it is necessary to constantly monitor volcanoes with different sensors that capture their behavior, in order to prevent complex scenarios that could be triggered by an eruption.

In Ecuador, the entity in charge of the surveillance and monitoring of volcanic activity is Instituto Geofísico de la Escuela Politécnica Nacional (IG-EPN) since 1983, which observes volcanoes inside and outside the continent, and has classified them according to their last eruptive activity as: extinct, active, and erupting. According to [3] and data from Instituto Nacional de Estadística(2010), at least 35% of the Ecuadorian population lives in these regions and could be affected by volcanic activity.

In 2021, the IG-EPN had 266 stations installed in the 20 volcanoes of the continent and Galapagos [3]. It is important to note that it has strategically installed sensors to monitor the specific risks of each volcano. In continental territory, the Guagua Pichincha, Tungurahua, Reventador, Cotopaxi, and Sangay volcanoes are continuously monitored. In the Galapagos archipelago, monitoring covers the Sierra Negra, Fernandina, Cerro Azul and Wolf volcanoes [4].

The IG-EPN has short-period, broadband and lahars seismic sensors, geodetic sensors that include inclinometers, GPS, geochemical and remote sensors that are thermal cameras for the observation of national volcanology. One of the main interests are seismic sensors, since an increase in this activity is associated with possible internal changes of the volcano, which could provide the necessary information to timely alert risk management institutions and even the public.

Seismicity analysis is the most widely used method to determine the current state and future activity of a volcano, since it provides insight into magma and gas movements. Seismicity is evaluated by analyzing microseisms and other vibrations recorded in the ground by the seismometers of the monitoring network. The recovered seismic events are classified as Volcano-Tectonic (VT) type seismic events that can occur due to rock stress caused by the movement of magma and other fluids through preexisting cracks, Long-Period (LP) events caused by cracks that resonate as fluids move towards the surface, Tremors (TR), which is continuous seismicity [5], Hybrid (HYB) events are a mix between Long-Period (LP) and Volcano-Tectonic (VT) events [6], among others.

Microseism-related data is recorded on a daily basis, generating a significant volume of information, which cannot be processed efficiently since they require an expert to categorize them. Therefore, it is opts for the use of Machine Learning to classify behaviors and/or generate new structures from the data obtained. Currently, there are works that use supervised learning models [7], [8] to classify these events. However, this approach has significant disadvantages, one of which is the reliance on fully labeled databases, which ideally should be selected and validated by subject matter experts and require considerable evaluation time. Another significant disadvantage is the considerable increase in model training time when handling large volumes of data, potentially delaying the development process.

The proposed solution is to use semi-supervised learning with the objective of classifying seismic events through the knowledge of a fraction of labeled data, which contains structural information that is used to compare with the characteristics of unlabeled data, making possible the extension of labels [9]. The Semi-Supervised Learning approach to be used is particularly suitable for microseism data, since they are highly detailed, allowing one to take full advantage of the approach provided by these models. This is particularly valuable since semi-supervised learning does not rely on a large number of labels, thus facilitating the analysis of daily collected data, the manual labeling of which would be a complex and costly task.

The existence of research focused on the classification of microseisms by means of these algorithms is very low, for example, the one developed by [10]. In this study, the following signals of seismic systems of type "VT" and "LP" were analyzed using a single semi-supervised model. The present work contributes with a broader approach to the Self-Training Algorithm, since not only a base classifier will be used, but its implementation with different types of classifiers is explored. In addition, the Label Spreading Algorithm is integrated to extend the analysis and maximize the use of the available data.

Among the base classifiers used is one of each Machine Learning approach: Symbolists, Bayesians and Analogizers in combination with different proportions of labeled data.In addition, a Label Spreading approach is implemented, exploring its performance in data-limited scenarios. This paper analyzes the impact of labeled data proportions on model performance, using metrics such as Area Under the Curve (AUC), F1-Score, and Accuracy, in order to provide practical recommendations for the integration of unlabeled data in volcanic microseism classification problems.

The main contributions of this work include the development of a comparative methodology that evaluated semi-supervised techniques against traditional supervised methods, highlighting their advantages and limitations in the problem domain. A preprocessed and structured data set including labeled and unlabeled information is presented. In addition, semi-supervised learning approaches adapted to the specific characteristics of the problem are optimized. Finally, a detailed analysis of the effect of the percentage of labeled data on model performance is provided, providing key insights for applications in resource-constrained contexts with high uncertainty.

II. Methodology

This section presents the implementation that was carried out for the development of this project. First, in II-A, we describe the additional features of the database used for the model. Then, in II-B, a simple cross-validation technique is explained to obtain an optimality for the reduction of the number of features to be used. In the following section II-C, the Semi-Supervised Learning approach is explained and, finally, the experimental conditions II-D are detailed.

Fig. 1, shows the block diagram that explains the implementation process of the approach used. A database containing records of various micro-earthquake characteristics of the Cotopaxi Volcano is used. Then, a dimension reduction and transformation are applied on the same characteristics. Finally, the processed database, which contains records classified as VT and LP independently, is trained on the semi-supervised learning model to finally evaluate the metrics and compare the results.

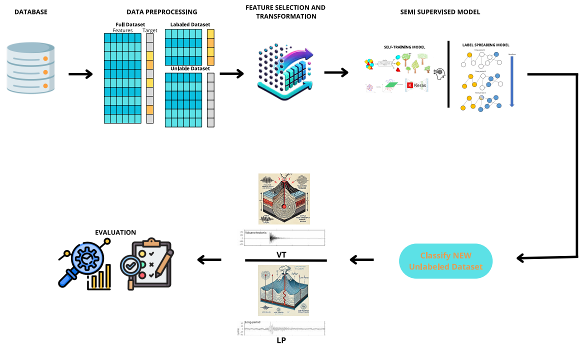

Figure 1. Block Diagram of the Methodological Stages

A. Database

It was proposed to use one of the databases that were collected from the BNAS, BREF, BTAM and BVC2 seismic sensors belonging to the Cotopaxi volcano, which we will call $\mathcal{D}$, containing 22,640 observations of different micro earthquakes such as VT, LP, TR, HYB, among others, as detailed in the following table I.

Table I
Frequency of Volcanic Microseisms by Type.

Type	Frequency
LP	11,553
VT	8,756
HB	533
TRE	252
OTHERS	1,546

This database contains 88 characteristics that are distributed in 1 type of earthquake, 3 related to data capture, and 84 related to the properties of the earthquake, as can be seen in table II. In the 84 characteristics, there are 13 in the time domain, 21 in the frequency domain, and 50 in the scale domain.

The extraction of Volcano-Tectonic (VT) and Long-Period (LP) type earthquakes was carried out, since these are the main ones that can inform of the beginning of an eruptive activity. As cited in [11] the document consulted, the "VT" earthquakes are in the frequency spectrum and can reach up to 15 Hz, and "LP" is of low frequency, compared to the "VT" events, it is between 2 to 5 Hz, reflecting a different dynamic behavior, possibly related to volcanic fluid interactions. This frequency distinction is crucial for classification and is a key tool for monitoring and minimizing volcanic hazards.

Table II
Seismic Event Features Table.

Feature Name	Description	Columns
Type	Type of microseism	1
Channel	Seismic event capture channel	1
Event Identifier	Unique identifier of the seismic event	1
Station	Station of origin of the seismic event	1
Time (t)	Time-related features, identified by “t”	13
Frequency (f)	Frequency-related features, identified by “f”	21
Scale (w)	Scale-related features, identified by “w”	50

Subsequent to the extraction of a subset of $\mathcal{D}$, composed only of two seismic events, this is divided into two sets: $\mathcal{D}_1$ and $\mathcal{D}_2$. The former was used for data training, while the latter was intended for the testing phase. Thus, $\mathcal{D}_1$ has $16,248$ records and $\mathcal{D}_2$ has $4,061$ records.

B. Preprocessing

For the development of this analysis, a Feature Reduction and Transformation Process was used using the Feature Selection process with mutual information. The data is scaled using the Standard Scaler tool to standardize the different scales of the characteristics, and, finally, a reduction in dimensionality is performed using PCA, preserving 95% of the variance. This allows the algorithm to simplify the processing of the data without losing significant information. Feature selection requires the specification of a k value, which it selects a maximum number of relevant features from the data set. In order to identify the optimal number, k, of the characteristics that will best contribute to the model, different values were taken. This made it possible to identify and select the most relevant characteristics for the performance of the model [12].

For optimization of the value of k in feature selection, an iterative procedure was implemented in which a base classifier is trained by varying the number of selected features [13].This process identifies the value of k that offers the most stable Area Under the Curve (AUC) metric, guaranteeing a balance between performance and consistency. Finally, we obtain the optimal value of k to be used in subsequent implementations, the evolution of this k is shown in Fig. 2.

We chose $k = 42$ as the optimal number of features out of the original 84, based on the observed performance of the ROC AUC, in this supervised classifier we used. This filtering helps in balancing model complexity and predictive reliability. In addition, it was considered to decrease the dimensionality by 50%, to optimize computational efficiency while retaining high predictive performance. This value was also estimated to be taken as k, since a balance between performance and simplicity is sought. If higher performance is desired, k values greater than 50 can be chosen.

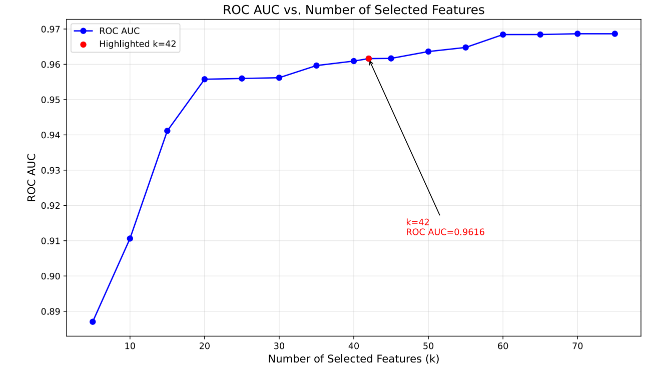

Figure 2. ROC AUC as a function of the number of selected features

C. Semi-Supervised Learning

In this section we detail the Semi-Supervised Learning models that were used, these models facilitate learning using labeled and unlabeled data sets. As mentioned above, the process of labeling earthquakes is very complex, so it is proposed to use this type of algorithm because it does not need a large amount of labeled data for training. Two approaches are used, Self-Training and Label Spreading.

1) Self-Training: For the Self-Training approach, it is necessary to have a base classifier to predict unlabeled data and select the most confident predictions and retrain the classifier [14]. Different classifiers are used to compare the results of their methods and obtain the classifier that best fits the data.

For the implementation of the Self-Training algorithm it is necessary to obtain a subset of the main database of labeled data L and unlabeled data U. From the labeled data we proceed to train a classifier f, which runs a hyperparameter optimizer using a cross-validation technique, and obtain the model with the best set of them. The optimized model serves as the basis for use in the Self-Training algorithm, where both labeled and unlabeled data are input. The whole procedure is shown in the flow of Fig. 3.

The input data consists of 84 characteristics, of which the data preprocessing mentioned in the previous section will be applied and where it was obtained that for the selection of features by mutual information a $k = 42$ characteristics, which are relevant, will be used. These relevant characteristics will go through a standardization process and finally the application of PCA to reduce dimensionality. These data will go through the Self-Training Algorithm procedure, referred to in the previous paragraph, and the output is a trained model that can predict the labels corresponding to VT and LP.

Regarding the f classifier, Random Forest, SVM and Naive Bayes are used. These algorithms are frequently used as binary classifiers, given their easy implementation, robustness, and efficiency, they are well adapted to high-

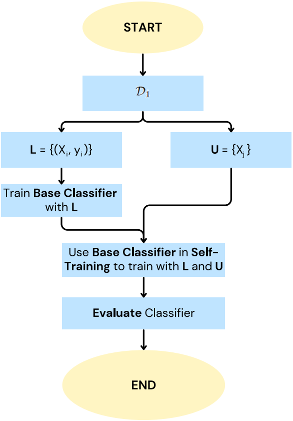
START
$\mathcal{D}_1$
L = {(X$_i$, y$_i$)}
U = {X$_j$}
Train Base Classifier with L
Use Base Classifier in Self-Training to train with L and U
Evaluate Classifier
END

Figure 3. Self-Training Process Flow

dimensionality numerical variables, which makes them ideal for this dataset. These classifiers are trained and optimized using k-fold cross-validation to evaluate combinations of hyperparameters, we specify five times, and upon completion, we retrain the model with the best values obtained from the search on the entire data set labeled L. The metric selected to optimize is the negative log loss, as in [15], and this function is particularly effective to evaluate the precision of probabilistic prediction algorithms. The above development was carried out in order to obtain a base classifier to be used in Self-Training. The configuration of hyperparameters is performed for Random Forest and for SVM, these values can be verified in the tables III and IV, for the case of Naive Bayes, the search of hyperparameters is not performed due to its simplicity.

Table III
Hyperparameter Grid for Random Forest.

Hyperparam.	**Description**	**Values**
n_estimators	Number of trees in the forest	10, 20, 50, 100
max_depth	Max. depth of the tree	1, 2, ..., 10
min_samples_split	Min. samples to split a node	2

Table IV
Hyperparameter Grid for Support Vector Machine(SVM).

Hyperparam.	**Description**	**Values**
C	Regularization parameter	0.1, 1, 10, 100
gamma	Kernel coefficient	1, 0.1, 0.01, 0.001
kernel	Kernel type	'rbf'

With the base algorithm already trained, it is implemented in Self-Training, where the only parameter that is fixed is the threshold, which in this case is 0.8. This threshold determines the confidence level required for a classifier prediction to be considered as a valid label for the unlabeled data set, U. As the algorithm progresses, high-confidence instances are automatically labeled and incorporated into the training set, allowing the classifier to iteratively improve its performance as it processes more data.

2) Label Spreading: This uses label propagation from a small set of labeled data to a much larger set of unlabeled data. The algorithm constructs a graph based on the similarity between the data points, generally using a base kernel, to model the relationships between the instances [16].

In the algorithm, labeled data is used as the source, while unlabeled records are initialized with uniform probabilities. Through an iterative process, Label Spreading propagates the labels to the unlabeled instances according to the similarity relations. During each iteration, the probability of the labels is adjusted according to the similarity of the points in the network, and the process is repeated until the labels converge.

The input data consists of 84 features from which the data preprocessing mentioned above will be applied. Unlike Self-Training, this algorithm does not use a base classifier, so it does not go through a hyperparameter optimization process, and therefore the total number of records (labeled L and unlabeled U) is input to the algorithm, Fig. 4, and performs the process described in the previous paragraph. For algorithm initialization, certain hyperparameters need to be adjusted to avoid under- and overfitting, thereby improving model performance.

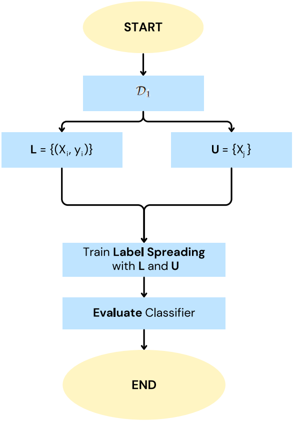

Figure 4. Label Spreading Process Flow

One of the most important hyperparameter to fix is the kernel, we used the radial basis function, "rbf", which is a Gaussian function that measures the similarity of the Euclidean distance between points, it is ideal for capturing nonlinear relationships in the data, avoiding linear assumptions about the data, and we allow the model to capture more complex patterns. Tie to the kernel, we have the value of gamma(γ) which is set as an inverse ratio to the square of the median of the distances between all pairs of points in the data, which reflects the actual structure of the data, avoiding being too wide (loss of local detail) or too narrow (lack of generalization). This value will avoid both overfitting and underfitting as it ensures the similarity of the data.

A value of alpha(α) is set at 0.30, which determines the weight given to the initial labels versus the propagated labels in each iteration of the algorithm, the value given gives as a rule that 30% importance is given to the original labels and 70% to the labels propagated in the training. This prevents the model from excessively depending on the propagated labels and also from depending too much on the initial labels, which would limit the propagation.

Likewise, we define the maximum number of iterations ($max\_iter = 100$) and the tolerance ($tol = 1e - 4$) values that allow the algorithm to converge correctly, ensuring that the model has enough time and accuracy to assign labels reliably.

D. Experiments

As described, for the Semi-Supervised Learning approach, both labeled and unlabeled data are needed, so, since the base $\mathcal{D}_1$ is fully labeled, a percentage was simulated to be unlabeled, and thus the database is obtained to have L and U (i.e., labeled and unlabeled data).

To train each algorithm, we performed variations on the initial size of L ranging from 10% to 90%, note that U is the remaining data in the database. On the one hand, in the Self-Training model, each classifier is trained using the set L, adjusting their respective hyperparameters each time we increase the labeled set. The size of L was selected and the experiment was repeated 10 times employing different subsets of the same size. To ensure a fair comparison between the base classifiers, both the size and the content of the data were kept constant by setting a random seed. On the other hand, in the Label Spreading model, the same procedure is performed in the same way, obviating the training of the base classifier, in this case passing the data set L and U, with their respective sizes directly to the model training. As mentioned, the inputs to the trainings were the features obtained by preprocessing dimensionality reduction and transformation.

The trained models were subjected to a testing phase that consists of using the $\mathcal{D}_2$ data set to evaluate various performance metrics. For this purpose, the Self-Training and Label Spreading models were selected in their different variants, which have been trained using between 10% and 90% of the set L, performing 10 repetitions for each percentage. The corresponding metrics were then calculated and a confidence interval was applied. Finally, the results obtained were compared between the different classifiers and configurations to determine their performance. This process is based on a variant of Monte Carlo-based cross-validation, as described [17], [18], having different data randomization of L and U, and repeating the process 10 times to average evaluation metrics.

For self-training-focused models, the semi-supervised approach was compared in terms of the area under the curve (AUC) and F1-Score metrics against a Supervised Learning scheme. On the one hand, AUC allows us to evaluate the ability of the model to discriminate between classes, while F1-Score is especially useful in unbalanced data sets, which is the case studied. In addition, the ROC curve and its corresponding AUC are used in the plots for clear and accurate visualization and comparison. In [14] it is argued that these metrics provide robust evaluations for these methods. For supervised training, the same hyperparameter search setup with k-fold cross-validation with k = 5 is used to evaluate performance. The same preprocessing setup was performed to achieve fair comparisons. For the Label-Spreading approach, only comparisons are made within experiments of the different experienced ratios of L.

In addition, the best results obtained by each classifier were compared to determine which classifier was best suited to the data set. The classifiers that achieved optimal performance using the least amount of labeled data are considered the most important, highlighting their ability to adapt and train efficiently with a minimum amount of labeled data and to the same data structure.

To obtain a more organized work, a preprocessing, training and evaluation flow tracking and orchestration tool, Prefect, was used. It allowed detailed tracking of each stage of the pipeline, managing dependencies between tasks and recording key performance metrics and failures, as well as enabling multitasking, optimizing time. This was especially relevant to ensure the reproducibility of the experiments and the analysis of intermediate results.

III. Results

In this section, the results of the Semi-Supervised Learning approach used for the binary classification of Cotopaxi Volcano microseisms are presented. The performance obtained with the limited use of data, in comparison with the Supervised Learning that uses the whole base, is highlighted, thus demonstrating the effectiveness of these approaches. Additionally, it is appreciated that the search for hyperparameters for the base model in the Self-Training approach improves the performance metrics.

For performance evaluation, we focused on 4 metrics, AUC as a global metric, F1-Score and Accuracy, including the ROC AUC curve for visual interpretation. The ROC AUC curve is a graphical representation that evaluates the ability of a model to discriminate between positive and negative classes. The curve plots the true positive rate (TPR) versus the false positive rate (FPR) for different decision thresholds. An AUC closer to 1 indicates better performance, reflecting a superior ability to distinguish between classes. While the AUC provides a numerical value summarizing its overall performance, a higher value indicates that the model is more effective in terms of sensitivity and specificity simultaneously.

Furthermore, the F1-Score allows us to assess the balance between accuracy and sensitivity, which is especially relevant in scenarios with unbalanced classes. However,

accuracy provides an overview of the percentage of correct predictions, although its effectiveness may be limited in unbalanced datasets. For this reason, in this study, precision is complemented with more robust metrics, such as AUC and F1-Score, to provide a more complete and accurate assessment of the performance of the models analyzed.
Therefore, for each model, a graph of the ROC-AUC curve is developed for a clear interpretation, and additionally different tables are made to summarize the results obtained by performing 10 training repetitions of the algorithm with different database percentages. For each percentage, we have a mean, a minimum, and a maximum value of the AUC for confidence intervals of 95%. Finally, the AUC result of the supervised learning algorithm, obtained under the same conditions as those used during training, is included in the table.

A. *Results of the Naive Bayes Self-Training Model*

It can be seen in Fig. 5, that since the algorithm was trained with higher percentages of data it has a higher accuracy, but it does not outperform a Supervised model. Therefore, Naive Bayes Self-Training does not obtain good results with small proportions of labeled data.

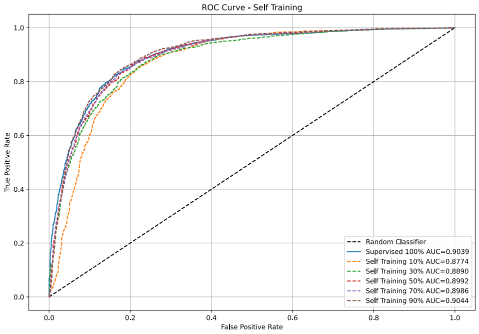

Figure 5. ROC Curves of Naive Bayes Self-Training and Supervised Learning Algorithms using LP events as Positive Class

The results of the self-training model, with the Naive-Bayes base classifier, in which no hyperparameter optimization was performed, are shown in Table V. It can be seen that the AUC value when trained using 90% of labeled data is close to the AUC of the Supervised model, suggesting that Self-Training performs acceptably well with less labeled data. From 50% of the labeled data, the confidence intervals become narrower, indicating stability and less sensitivity to the initialization data.

B. *Results of the Random Forest Self-Training Model*

It can be observed in Fig. 6, that since the algorithm was trained with higher percentages of data, it has a higher accuracy, and from a minimum amount of labeled data, which is 10, it manages to outperform the supervised model. Therefore, by using Random Forest in the Self-Training it manages to take advantage of the unlabeled data in a high capacity. The curves are stable, suggesting the robustness of the model.

Table V
Performance Evaluation: AUC Metrics Obtained from Naive Bayes Self-Training Across Dataset Splits (95% CI)

Training Dataset Utilization (%)	**Mean**	**Confidence Intervals**
10%	0.86554	[0.85873, 0.87235]
30%	0.88249	[0.87901, 0.88597]
50%	0.88793	[0.88232, 0.89353]
70%	0.89143	[0.88752, 0.89535]
90%	0.90287	[0.90172, 0.90401]
AUC SUPERVISED	**0.90389**	

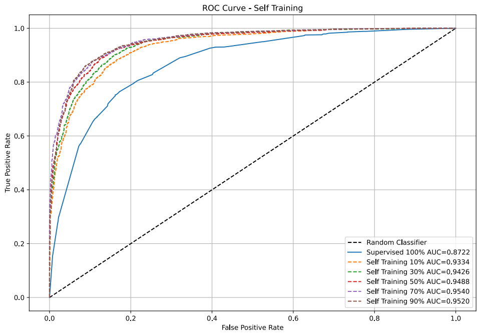

Figure 6. ROC Curves of Random Forest Self-Training and Supervised Learning Algorithms using LP events as Positive Class

In table VI, the results for the Self-Training model, with Random Forest base classifier, are shown, in which a search is carried out for the best hyperparameters to be the base model of Self-Training, it is clearly distinguished that the AUC value when trained only with 10% of labeled data exceeds the AUC of the Supervised model, suggesting that the base model is effective in scenarios with different amounts of labeled data, particularly only with 10% of labeling. Because of the tightness of the confidence intervals in all cases, it fits the data set well, achieving excellent stability.

Table VI
Performance Evaluation: AUC Metrics Obtained from Random Forest Self-Training Across Dataset Splits (95% CI)

Training Dataset Utilization (%)	**Mean**	**Confidence Intervals**
10%	0.92100	[0.91603, 0.92597]
30%	0.93732	[0.93520, 0.93944]
50%	0.94594	[0.94466, 0.94723]
70%	0.94939	[0.94778, 0.95099]
90%	0.95049	[0.94977, 0.95120]
AUC SUPERVISED	**0.87215**	

C. Results of the SVM Self-Training Model

As can be seen in Fig. 7, as with Random Forest, training with higher percentages of data yields higher accuracy and outperforms Supervised with only training 10% of labeled data. The curves are stable and much closer to the perfect value.

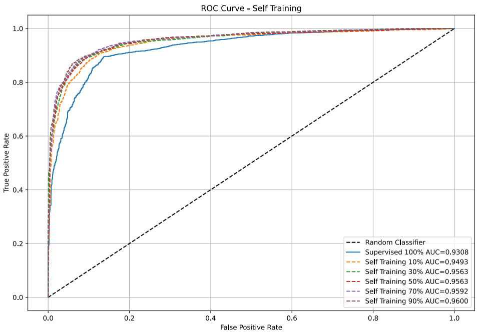

Figure 7. ROC Curves of SVM Self-Training and Supervised Learning Algorithms using LP events as Positive Class

The table VII shows the results for the Self-Training model, with SVM base classifier, as the previous model, a search is performed for the best hyperparameters to be the base model of self-training; likewise the AUC value when trained only with 10% of labeled data exceeds the AUC of the supervised model. The remarkable thing about this base model is the confidence intervals that are even narrower than when using Random Forest, describing a greater stability.

Table VII
Performance Evaluation: AUC Metrics Obtained from SVM Self-Training Across Dataset Splits (95% CI)

Training Dataset Utilization (%)	Mean	Confidence Intervals
10%	0.94442	[0.94196, 0.94687]
30%	0.94948	[0.94743, 0.95153]
50%	0.95271	[0.95127, 0.95416]
70%	0.95582	[0.95416, 0.95748]
90%	0.95748	[0.95630, 0.95867]
AUC SUPERVISED	**0.93079**	

D. Results of the Label Spreading model

It can be observed in Fig. 8, that the curves are closer to each other, indicating that despite training the model with larger amounts of labeled data, no improvement in prediction is achieved. Additionally, the curves are closer to the Random Classifier straight line, which shows that the algorithm has low performance and fails to identify significant patterns in the data.

In table VIII, the results for the Label Spreading model are shown, in this approach the AUC value does not exceed the value of 0.89, and there are no significant changes when increasing the amount of data with which the models were trained. The confidence intervals are narrow, suggesting good stability. Since the approach relies on label propagation using a graph, it may have limitations, especially for our type of data.

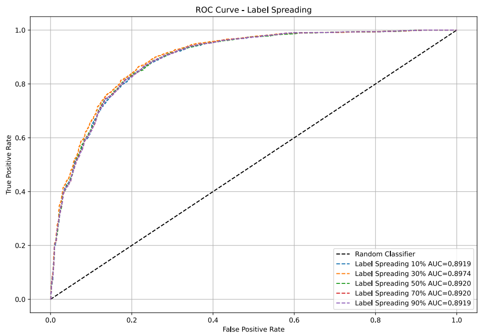

Figure 8. ROC Curves of Label Spreading using LP events as Positive Class

Table VIII
Performance Evaluation: AUC Metrics Obtained from Label Spreading Across Dataset Splits (95% CI)

Training Dataset Utilization (%)	Mean	Confidence Intervals
10%	0.88040	[0.87381, 0.88698]
30%	0.88781	[0.88407, 0.89155]
50%	0.88638	[0.88410, 0.88866]
70%	0.88668	[0.88450, 0.88886]
90%	0.88730	[0.88522, 0.88935]

E. Comparative Analysis

In this subsection, to refer to self-training models with their different base models, only the name of the base model is mentioned to simplify the wording and avoid unnecessary repetitions.

In general, it can be seen that the Random Forest and SVM base models obtain outstanding results with a minimum amount of labeled data; in particular, SVM showed consistency in its ROC curve and narrow confidence intervals, which is reflected, on the one hand, in the consistency of its ROC curve and, on the other hand, in its narrow confidence intervals, which translates into greater robustness and reliability in the face of possible variations in the data. In contrast, when Naive Bayes is used as a basis, it does not present a substantial improvement when compared to its supervised version. Finally, Label Spreading has even more restrictions, since in spite of increasing the training data, the AUC metric remains in similar ranges, but it does not outperform the Supervised and Self-Training approaches.

The gradual increase of the labeled data leads both SVM and Random Forest to have constant improvement until they reach maximum with the data labeled at 90%, where SVM reaches a maximum AUC of 0.9600 and Random Forest 0.9520. However, Naive Bayes does not exceed a limit, which is approximately 0.9044, and shows no noticeable improvement with different percentages of labeled data. Label Spreading remains virtually unchanged around 0.8919, indicating a possible lack of significant relationship between labeled and unlabeled samples.

The table III-E was constructed to group relevant information and add additional metrics, providing a comparative summary of the four analyzed models. The results show that SVM and Random Forest consistently outperform the supervised model, while Naive Bayes and Label Spreading present significant limitations. In terms of AUC, SVM obtains the best overall performance, followed by Random Forest. On the other hand, both Naive Bayes and Label Spreading fail to achieve competitive results.

In addition, metrics such as F1-Score were included in the table III-E , which confirms the results. Similarly, SVM and Random Forest stand out with outstanding values, particularly SVM, reflecting an excellent balance between accuracy and sensitivity. In contrast, Naive Bayes and Label Spreading do not exceed the 0.85 threshold, evidencing a low performance in this metric. Accuracy yields similar results, where SVM and Random Forest continue to lead. In the case of a supervised model, Naive Bayes stands out with acceptable metrics, although it continues to be outperformed by Self-Training models. When comparing the robustness of the models, both SVM and Random Forest demonstrate narrow confidence intervals and high stability to variations in the data. In contrast, Naive Bayes shows lower robustness, with moderately wide intervals, while Label Spreading offers moderate robustness but does not reach the levels of the other two main models.

Table IX
Comparison of Semi-Supervised Models

Feature	**Naive B.**	**R. Forest**	**SVM**	**Label Spreading**
AUC (%) (90% L)	90.29	95.05	**95.75**	0.88730
AUC (%) (10% L)	86.55	92.10	**94.44**	0.88040
Supervised AUC(%)	90.39	87.22	**93.08**	Not reported
F1 (%) (90% L)	85.61	89.83	**90.93**	0.84927
F1 (%) (10% L)	80.01	86.44	**89.67**	0.77548
Supervised F1 (%)	**85.42**	80.67	61.40	Not reported
Acc.(%) (90% L)	83.56	88.39	**89.90**	0,80823
Acc.(%) (10% L)	78.53	84.51	**88.42**	0.67057
Supervised Acc.(%)	**83.21**	79.09	67.87	Not reported
Exceeds Supervised?	NO	YES	YES	NO
Impact more L	Moderate	High	High	Low
Confidence Intervals	Moderately wide	Narrow	Narrow	Moderately wide
Robustness	Low	High	High	Moderate

IV. Discussion

In this work, we were able to evaluate and compare four Semi-Supervised Learning models, three of them based on Self-Training but based on different algorithms Naive Bayes, Random Forest, and SVM, and one based on graphs which is Label Spreading. The results obtained allow analyzing their capabilities and limitations in contexts of different percentages of labeled data, providing a broader view of performance, since key metrics such as AUC, F1-Score and Accuracy were compared, and confidence intervals are considered to estimate aspects such as robustness and sensitivity to the amount of labeled data.

The SVM and Random Forest-based models demonstrated outstanding performance in terms of AUC and F1-Score. SVM led in AUC, reaching maximum values of **95.75%** at 90% of the labeled data and maintaining outstanding performance with only 10% of the labeled data (**94.44%**). This implies that it has a high ability to generalize adequately, even with a minimal amount of labeled data. On the other hand, Random Forest obtained an AUC close to that of SVM (**95.05%** at 90% and **92.10%** at 10%), also showing robustness against scenarios with little labeled data. It should also be noted that the confidence intervals for the metric are extremely narrow, which translates into greater robustness and reliability in the face of possible variations in the data. These characteristics reinforce the notion that both models are highly adaptive and effective within this data set.

In contrast, Naive Bayes performs less well, achieving a maximum AUC of 90.29% on 90% labeled data, which barely matches its supervised performance (90.39%). To a large extent, it may be because the hyperparameters were fixed and the model was not optimized to enhance Self-Training. Furthermore, since the experiment was performed for the different percentages with the hyperparameters fixed, it is highly likely that despite increasing labeled data it fails to capture more information, resulting in relatively low AUC values for any proportion of the training set.

With Label Spreading, poor performance is also evident. Label Spreading was the lowest performing model, with AUC values remaining nearly constant regardless of the percentage of labeled data, indicating a lack of significant improvement when labeled data were incorporated. This behavior can be attributed to the dependence of Label Spreading on the structure of the similarity graph and the quality of the relationships between labeled and unlabeled samples.

In table III-E, additional metrics were added to verify the performance of the models. Thus, similarly both F1-Score and Accuracy confirm that Self-Training based on SVM and Random Forest has exceptional performance and superior accuracy. Since there will be an imbalance in the "VT" and "LP" classes, as can be seen in the table I, the "LP" earthquakes have more records than the "VT", therefore validating the F1-Score is extremely important and in these models it has promising results. In terms of Accuracy, the trend is maintained, so we can affirm that both Self-Training based on SVM and Random Forest stand out as an approach with high effectiveness and accuracy for Semi-Supervised Learning scenarios. Their ability to maintain high levels of accuracy with limited labeled data is ideal for our data set since manual labeling is very costly. These models are not only robust but also efficient, showing a balance between performance and resource usage.

From the metrics obtained and their comparisons, it is important to note that when performing the hyperparameter search for the Self-Training base algorithm, better results are obtained. This confirms what was pointed out in [10] regarding the performance of the base classifier when optimized. In addition, the present work contributes to the study of more base classifiers for Self-Training, highlighting the SVM and Random Forest classifiers as the best options for the classification of microseisms. Likewise, different from [10] work, a dataset with a larger number of records was used, which allowed extending the approach to larger amounts of data, and another different approach was added with respect to the Semi-Supervised Algorithm, which is the use of Label Spreading.

Although Random Forest was effective in the scenarios we ran, it has significant challenges in terms of computational complexity, as the increase in the number of trees and features results in long training times. Similarly, SVMs, when applied to very large datasets, have a high computational cost and the effectiveness of SVMs is highly dependent on proper kernel selection. However, Naive Bayes shows limited applicability in complex tasks due to its assumption of independence between features, which degrades its predictive power. On the other hand, Label Spreading requires meticulous adjustments in the construction of the similarity graph to be competitive, which implies additional complexity and the need for deep domain-specific knowledge.

To extend the scope of the present study, we propose exploring the use of Self-Training with neural networks as a base classifier. Incorporating more complex models, such as neural networks, could potentially improve generalization ability and performance in microseism classification tasks. One could also opt for research with hybrid approaches that combine SVM or Random Forest with semi-supervised methods to take advantage of their combined strengths. In addition, it is essential to optimize the hyperparameters of both Label Spreading and Naive Bayes to maximize their performance. Optimization of these hyperparameters will allow for better adaptation of the models to the specific characteristics of the data, which could result in significant improvements in the accuracy and efficiency of the classification process.

Furthermore, it is recommended to evaluate the proposed models on different datasets to validate the robustness and generalization of the techniques employed. Diversification of data sets will facilitate the identification of the strengths and weaknesses of each model in different contexts and conditions. However, with respect to the dimensionality reduction techniques used, it is suggested to investigate and apply other methodologies to determine their impact on the performance of the classifiers. The combination of various dimensionality reduction techniques with varied base models and hyperparameter optimization may lead to a more robust and versatile approach to microseism classification.

V. Conclusion

By selecting SVM and Random Forest as base models and also through hyperparameter optimization, both models consistently outperform the supervised model in all metrics evaluated. SVM shows the best overall performance, especially in AUC and F1-Score, highlighting its ability to balance accuracy and sensitivity. Random Forest, although slightly behind in performance, demonstrated exceptional robustness and stability, making it suitable for scenarios with noise or unbalanced data.

While using Naive Bayes, and even more so when using fixed hyperparameters, it fails to capitalize on labeled data effectively, barely matching its supervised performance. Label Spreading, on the other hand, shows stagnant performance, with no significant improvement with more labeled data, indicating a critical dependence on the quality of the similarity graph.

Both SVM and Random Forest are scalable, achieving excellent performance even with 10% labeled data. This makes them ideal for this data set, which is expensive to label manually. In contrast, Naive Bayes and Label Spreading quickly reach a performance limit, limiting their practical usefulness.

When analyzing the confidence intervals of SVM and Random Forest, they are narrow and decrease as the number of labeled data scales, reflecting high reliability and lower sensitivity to variations in the data. These attributes position them as preferred choices in real-world environments.

In summary, this study confirms that SVM and Random Forest are the best choices in semi-supervised learning, while Naive Bayes and Label Spreading require significant adjustments to be competitive in more complex problems.

Acknowledgment

The authors thank the Escuela Politécnica Nacional for providing the data set utilized in this investigation. Their support and contribution were essential for the development and success of this work.

References

[1] M. Auker *et al.*, "Global volcanic hazards and risk," in *Global Volcanic Hazards and Risk.* Cambridge University Press, 2015, pp. 81–173.

[2] V. Kirianov, "Environmental impacts of volcanic eruptions," in *Natural and Human Induced Hazards.* UNESCO-EOLSS, 2004, vol. I.

[3] P. Ramon, S. Vallejo, P. Mothes, D. Andrade, F. Vásconez, H. Yepes, S. Hidalgo, and S. Santamaría, "Instituto Geofísico – Escuela Politécnica Nacional, the Ecuadorian Seismology and Volcanology Service," *Volcanica*, vol. 4, no. S1, pp. 93–112, 2021.

[4] S. Hidalgo, A. Robles, D. Andrade, B. Bernard, P. Ramón, P. Mothes, J. Ordoñez, and G. Ruiz, *Los volcanes activos y potencialmente activos del Ecuador continental y sus redes de monitoreo.* Quito, Ecuador: Instituto Geofísico – Escuela Politécnica Nacional, 2014. [Online]. Available: https://www.igepn.edu.ec

[5] United States Geological Survey (USGS), "Monitoring volcano seismicity provides insight to volcanic structure," https://www.usgs.gov/programs/VHP/monitoring-volcano-seismicity-provides-insight-volcanic-structure, 2024, [Accessed: Oct. 12, 2024].

[6] S. Petrosino and P. Cusano, "Low frequency seismic source investigation in volcanic environment: the mt. vesuvius atypical case," *Advances in Geosciences*, vol. 52, pp. 29–39, 2020.

[7] R. A. Lara-Cueva, D. Benitez, E. Carrera, M. Ruiz, and J. Rojo-Alvarez, "Automatic recognition of long period events from volcano tectonic earthquakes at cotopaxi volcano," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 9, pp. 5247–5257, 2016.

[8] R. A. Lara-Cueva, A. S. Moreno, J. C. Larco, and D. S. Benítez, "Real-time seismic event detection using voice activity detection techniques," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 12, pp. 5533–5542, 2016.

[9] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning.* Cambridge, MA: MIT Press, 2006.

[10] C. Brusil, F. Grijalva, R. Lara-Cueva, M. Ruiz, and B. Acuña, "A semi-supervised approach for microseisms classification from cotopaxi volcano," in *2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI).* Guayaquil, Ecuador: IEEE, November 11–15 2019, pp. 1–6.

[11] D. A. R. Carrillo, "Implementación de un método de agrupación de señales sísmicas generadas por el volcán cotopaxi basado en aprendizaje automático no supervisado utilizando el modelo de mezcla gaussiana," Master's thesis, Escuela Politécnica Nacional, Quito, Ecuador, febrero 2022.

[12] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[13] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[14] X. Zhu and A. B. Goldberg, *Introduction to Semi-Supervised Learning*, ser. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2009.

[15] V. Vovk, "The fundamental nature of the log loss function," *arXiv preprint arXiv:1502.06254*, 2015.

[16] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*, ser. Adaptive Computation and Machine Learning. MIT Press, 2006.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[18] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R.* New York: Springer, 2013.