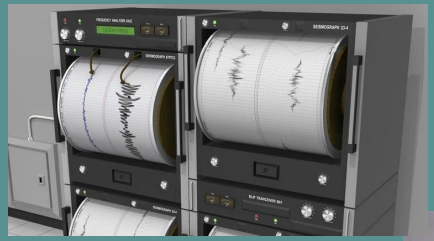
Aprendizaje Automático

Universidad Nacional de San Martín, **2021**



Grupo N°20
Proyecto VOLCANO

Iván Melchor
Estudiante de posgrado (UNRN - UGR)

Pablo Reynoso Peitsch
Estudiante de grado (CyT UNSAM)

disponible en: github.com/ifmelchor/volcano_ML-UNSAM

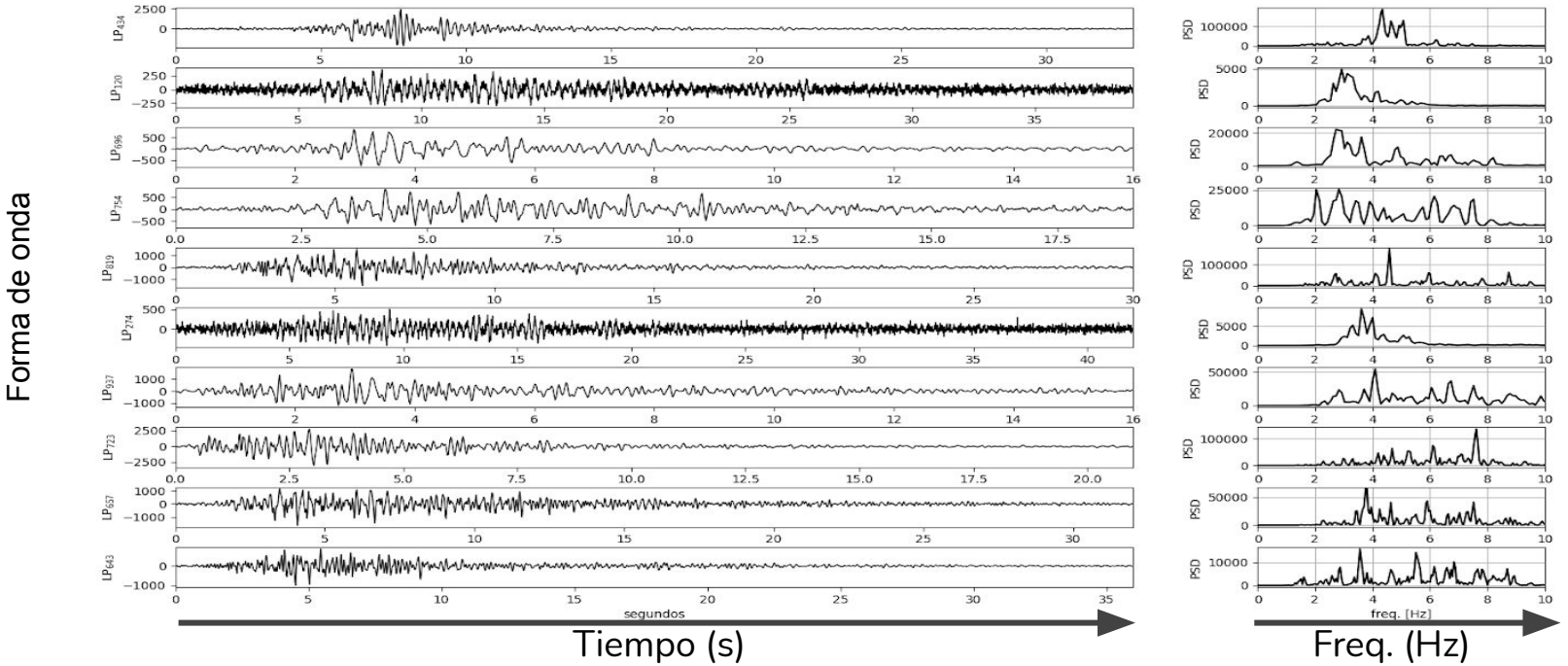
Contenido

1. Base de datos y problema a resolver
2. Parametrización de las señales/espectro
3. Algoritmos de clustering (no supervisado)
4. Mixtura y Regresor Logístico (semi supervisado)
5. Autoencoders

disponible en: igepn.edu.ec/senales-sismicas/fomulario-eseismic

Base de datos

1044 señales de Largo-Periodo (LP)



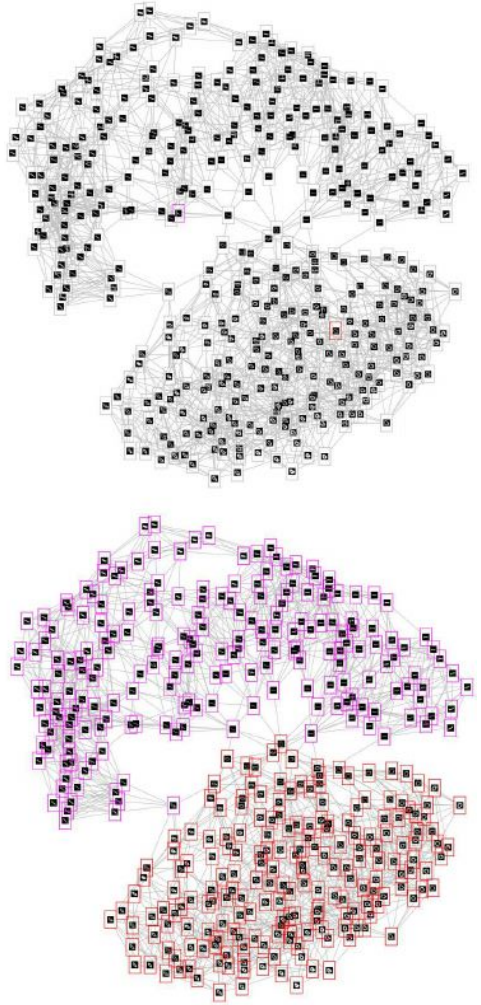
Problema:

Clasificar señales sismo-volcánicas (LP)

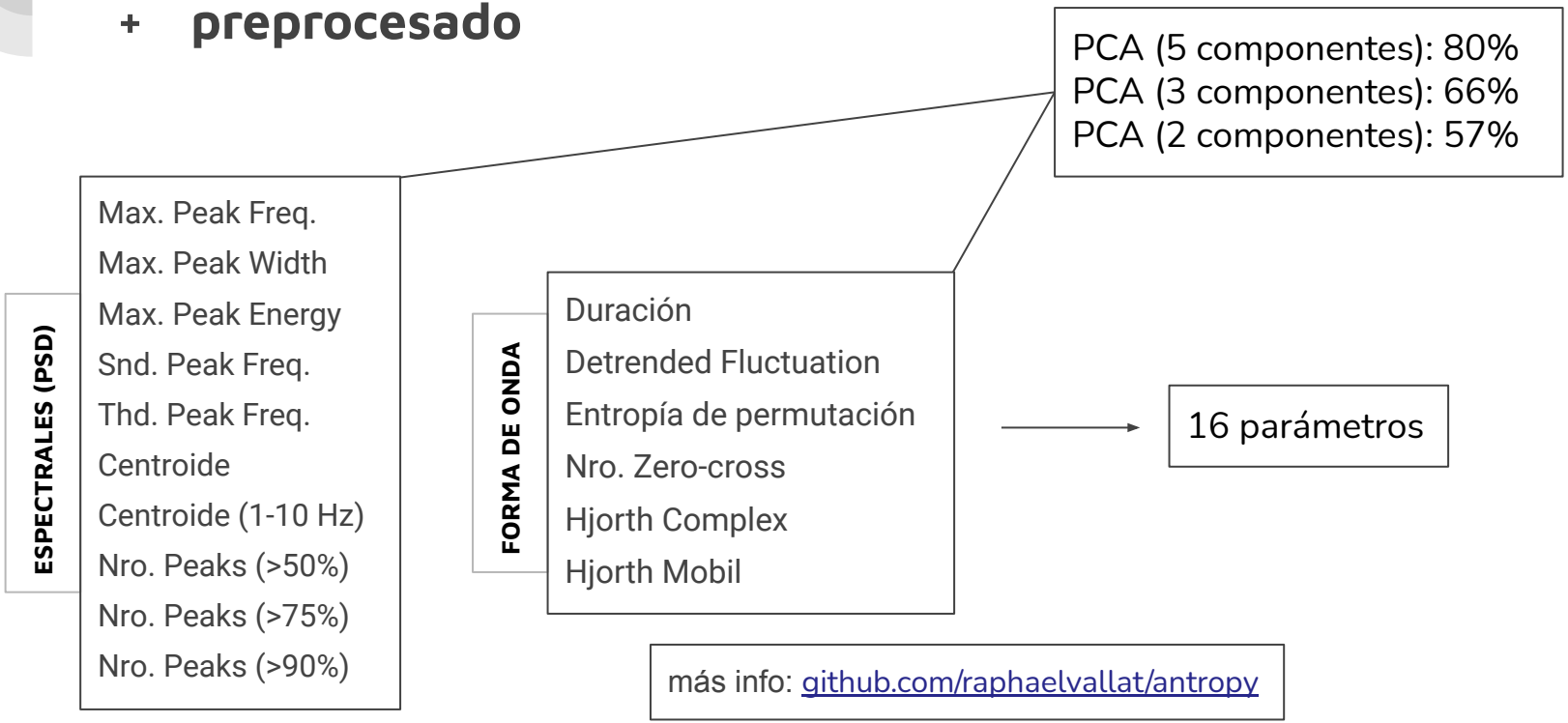
1: Extraemos parámetros (*features*) de los cuales no conocemos las etiquetas (*labels*) mediante técnicas de análisis de series temporales (y otros).

2: Entrenamos algoritmos de clustering (no-supervisados) para visualmente validar los resultados.

3: Seleccionamos el mejor modelo, fijamos etiquetas (cuantas más mejor) y entrenamos un clasificador para propagar las etiquetas (semi-supervisados).



Extracción de parámetros + preprocesado



Modelos no supervisados

Clustering

- Density-Based Spatial Clustering (DBSCAN)
- MeanShift
- Hierarchical
- K-Means
- Gaussian Mixture Model (GMM)

Reducción de dimensionalidad para graficar

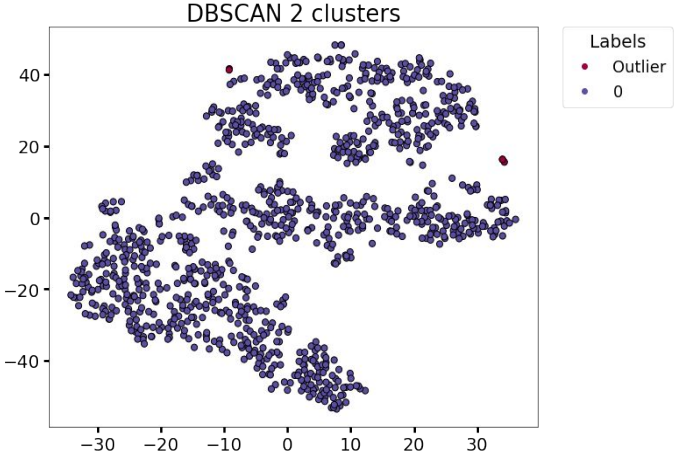
- tSNE

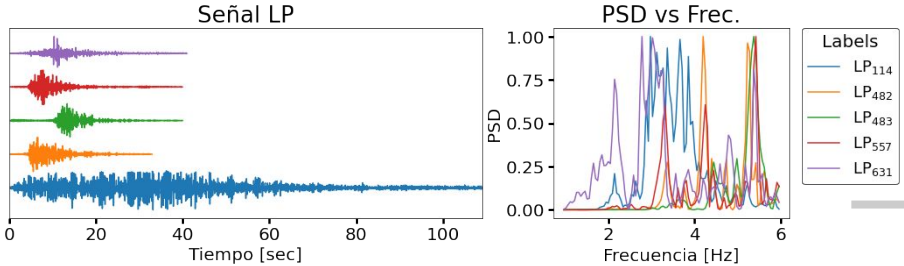
Aprendizaje profundo

- AutoEncoder

TSNE

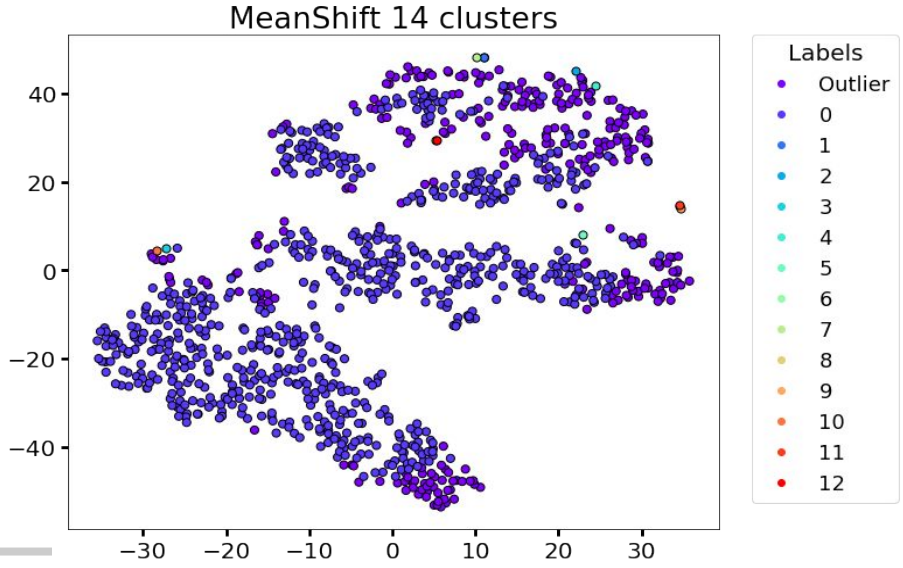
DBSCAN



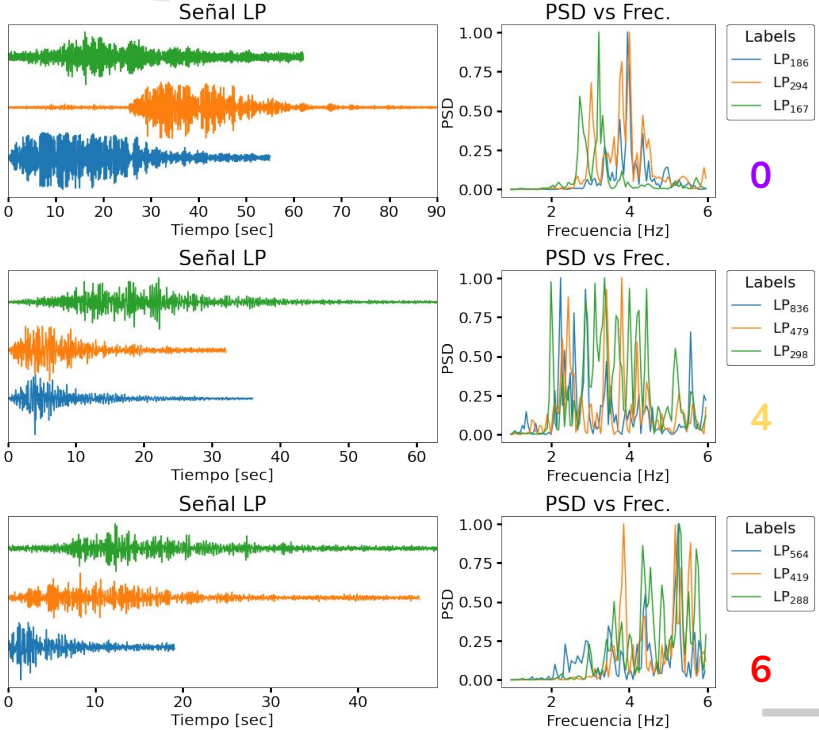


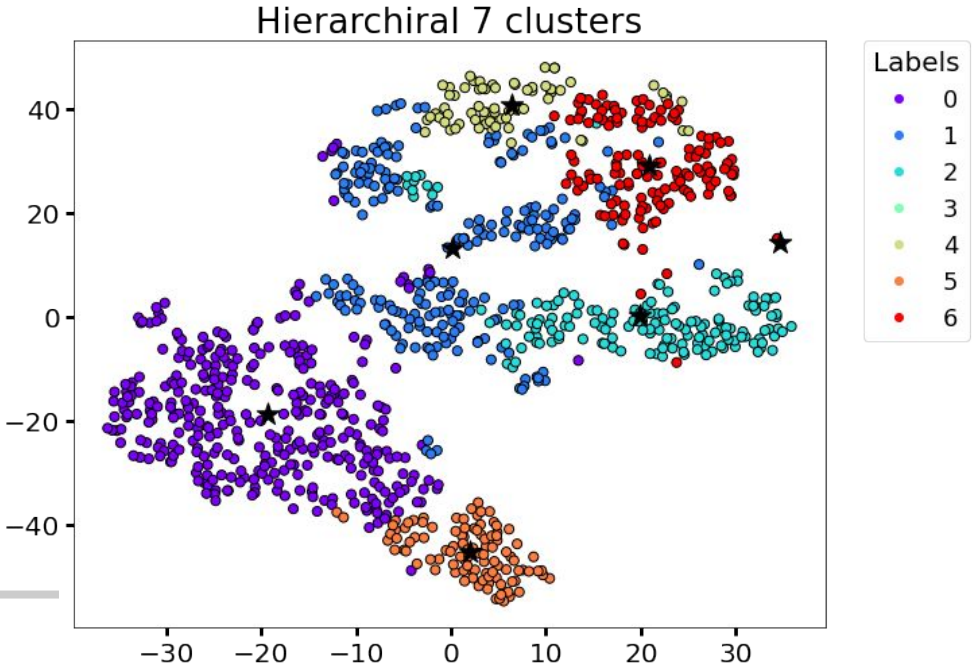
Mean Shift

Num. labels: 14 Score: 0.21932905612489612
Counter({0: 729, -1: 302, 1: 2, 12: 1, 10: 1, 4: 1, 7: 1, 3: 1, 11: 1, 9: 1, 8: 1, 5: 1, 6: 1, 2: 1})

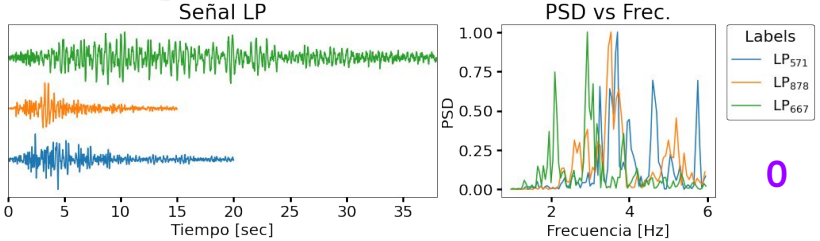


Hierarchical

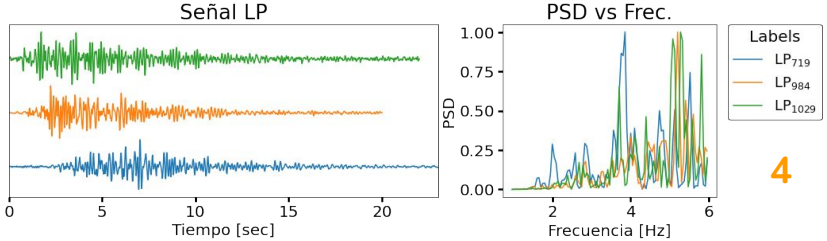




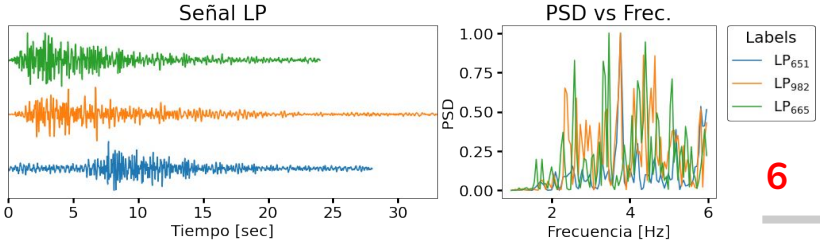
K-Means


Señal LP
Tiempo [sec]
PSD vs Frec.
PSD
Frecuencia [Hz]
Labels
LP571
LP878
LP667

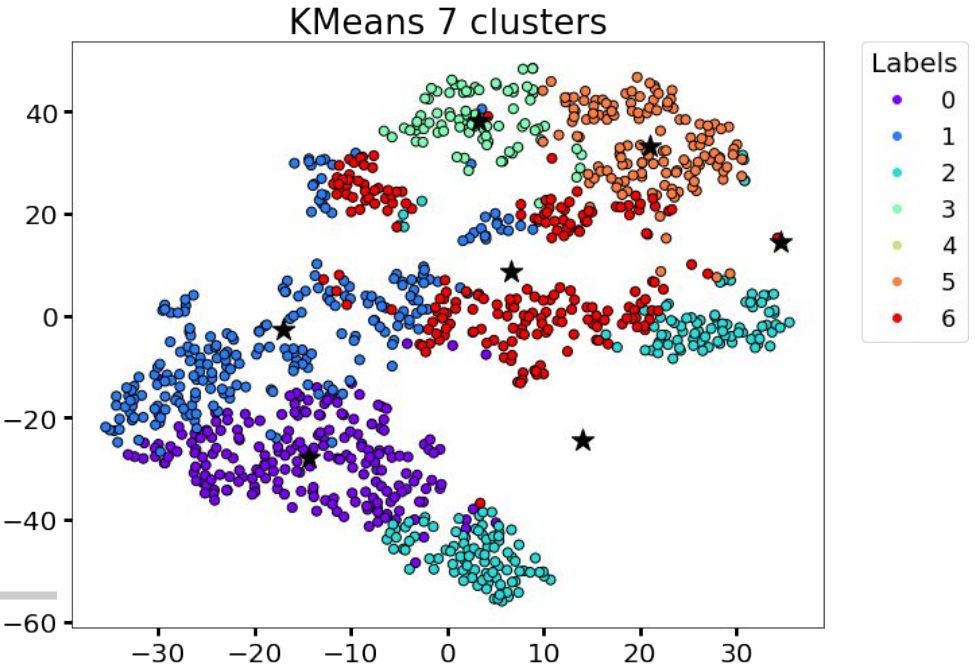
0


Señal LP
Tiempo [sec]
PSD vs Frec.
PSD
Frecuencia [Hz]
Labels
LP719
LP984
LP1029

4


Señal LP
Tiempo [sec]
PSD vs Frec.
PSD
Frecuencia [Hz]
Labels
LP651
LP982
LP665

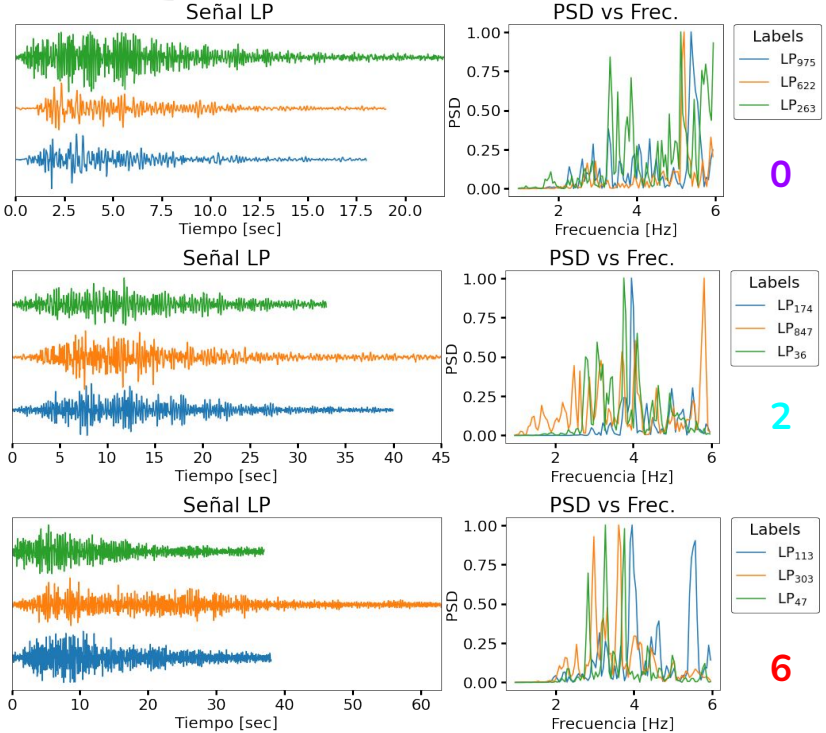
6

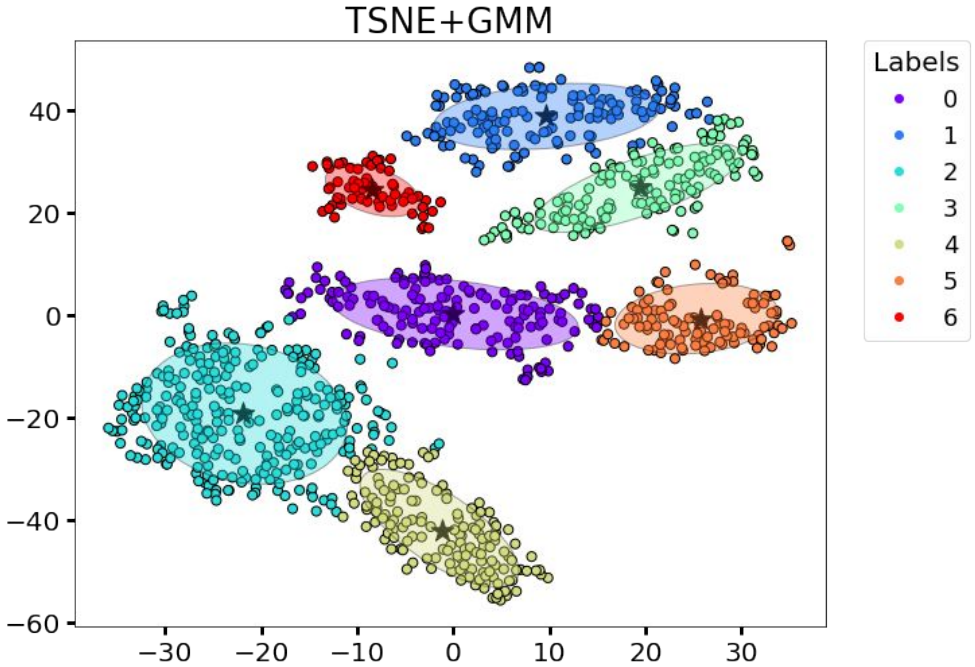

KMeans 7 clusters
Labels
0
1
2
3
4
5
6

CONCLUSIONES PARCIALES

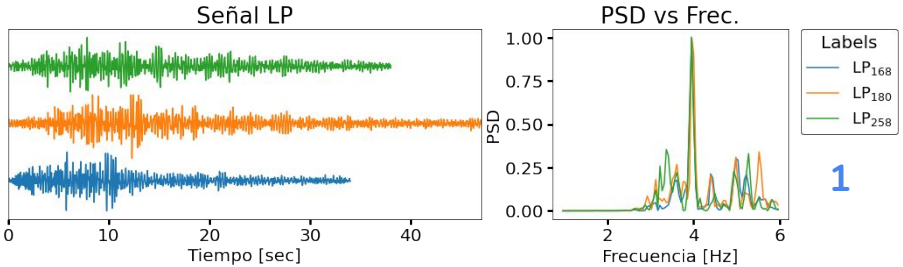
- La métrica Silhouette no es conveniente para seleccionar el mejor modelo en nuestro caso. Hay que aplicar *validación visual*.
- El dataset es separable en siete clusters.
- El modelo Hierarchical es el que da mejor resultado.

T-SNE + Gaussian Mixture Model





Etiquetado Manual



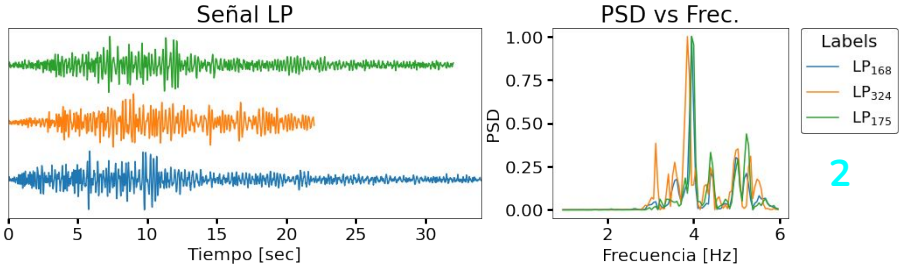
1

supervised_label: 1
LP_index, GMM_label

206	0
315	2
340	2
162	2
116	2
778	1
959	2
115	2
112	2
156	2

Supervised labels: 97
Sup_label, nro_LPs

1	28
2	10
3	4
4	30
5	3
6	3
7	2
8	3
9	2
10	2
11	2
12	2
13	2
14	2
15	2



2

No hay compatibilidad entre nuestros resultados y los resultados del TSNE+GMM

Como no podemos entrenar un clasificador sin etiquetas. Consideraremos las etiquetas arrojadas por el modelo TSNE+GMM como correctas

REGRESIÓN LOGÍSTICA

tSNE + GMM (prob. > 99.6%): 538 etiquetas

Entrenamiento: 403

Validación: 135

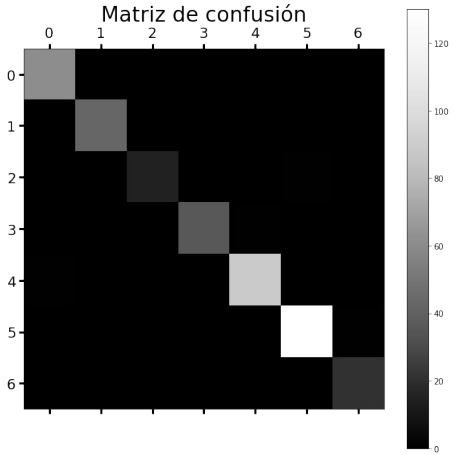
Regularización Ridge (C=1.0)

Validación cruzada 5 folds

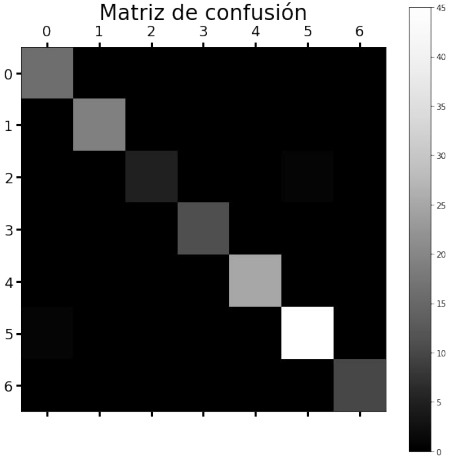
Recall 1.0 / Precisión 1.0

Recall 0.82 / Precisión 0.83

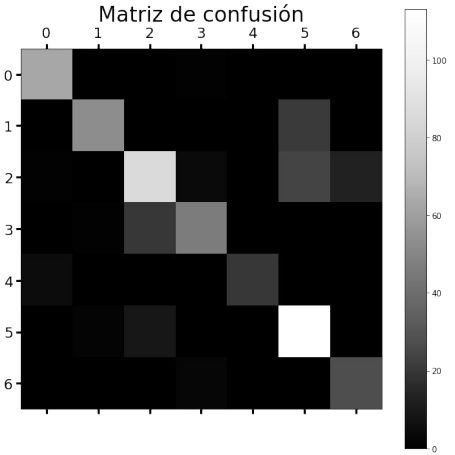
Entrenamiento



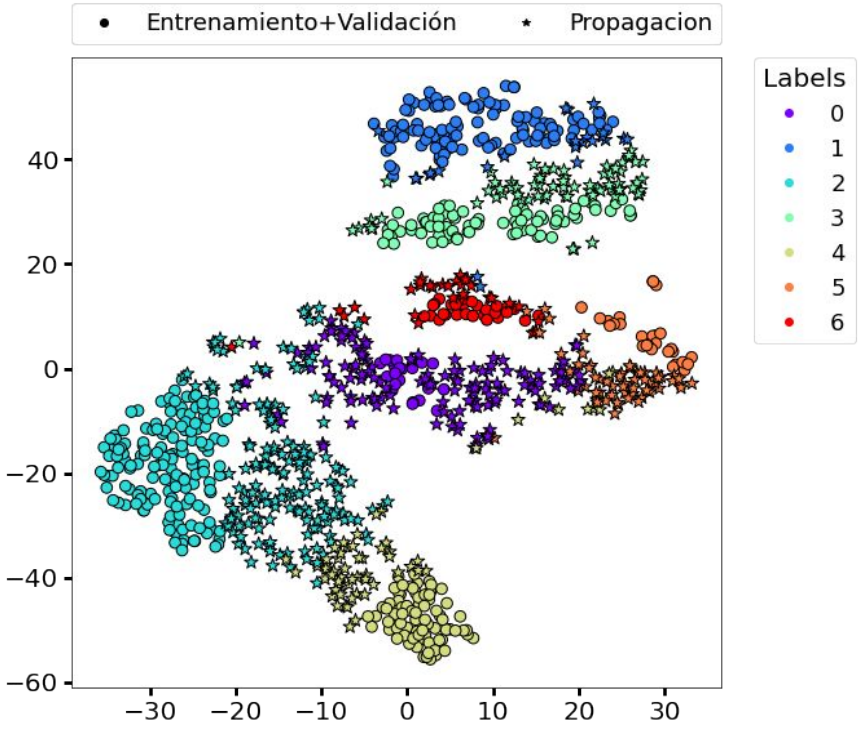
Validación



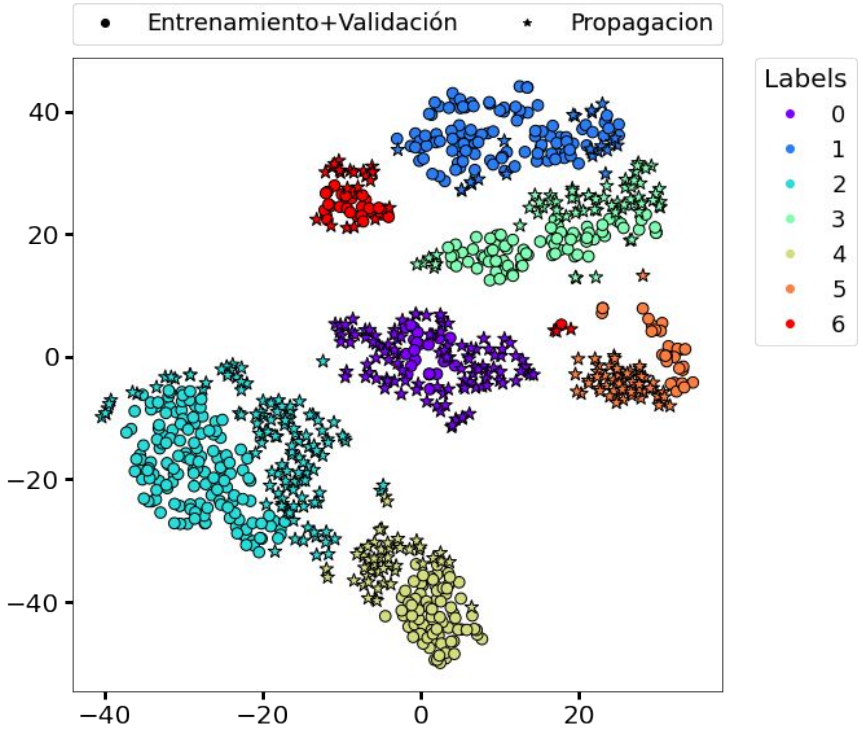
GMM



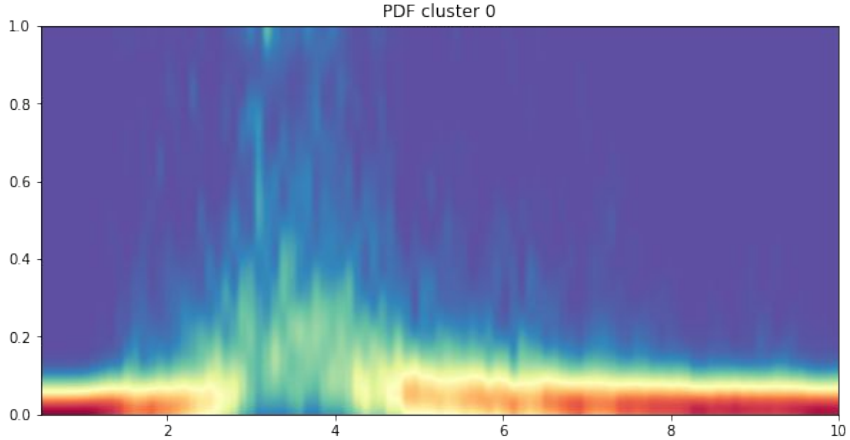
Propagación de etiquetas

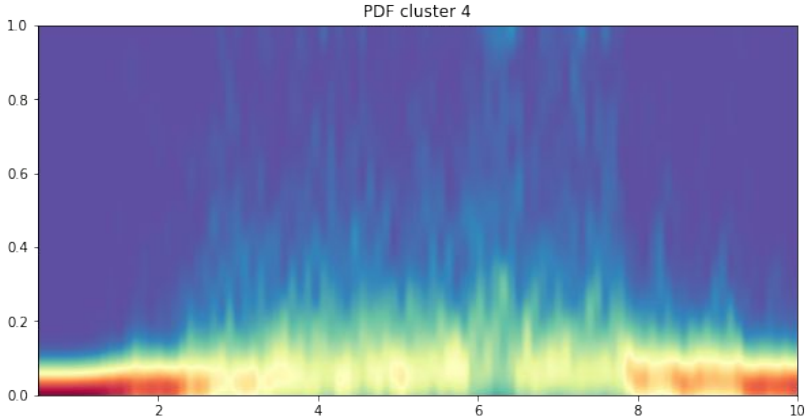


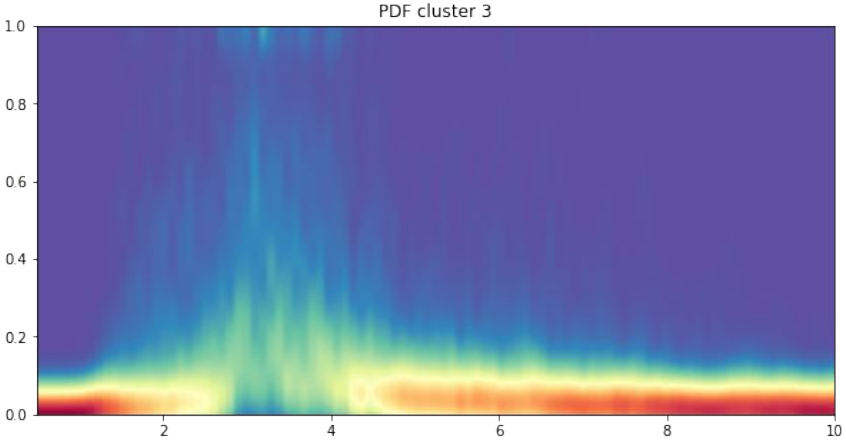
129 etiquetas (27%) no coinciden

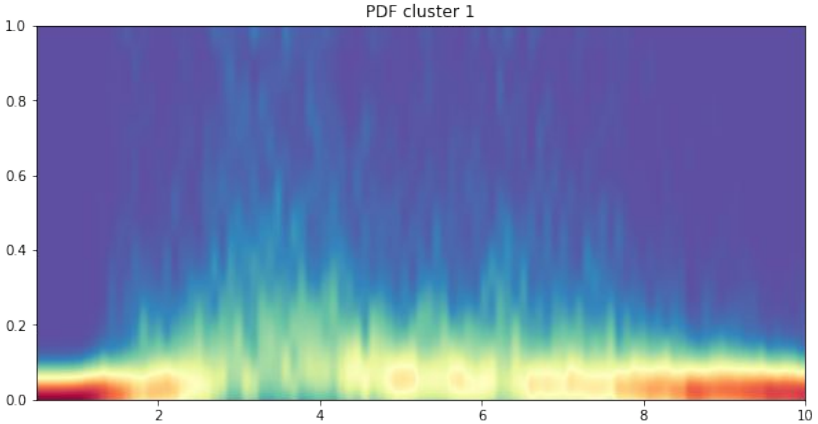


PDF por clase









¿Problema resuelto?

Un problema complejo. Pocas señales clasificadas manualmente e incapacidad de establecer etiquetas.

tSNE+GMM provee buenos resultados que permiten entrenar un regresor capaz de diferenciar entre las diferentes etiquetas.

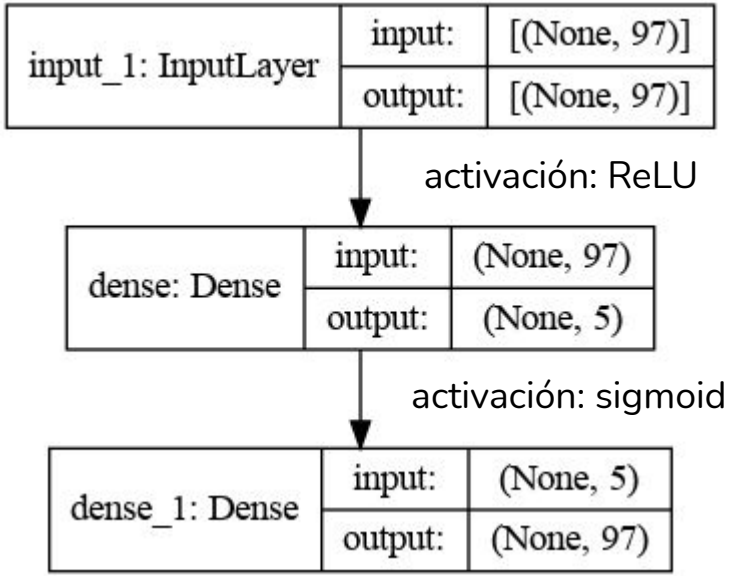
Sin embargo, las PDFs no muestran claras diferencias (generales) que permitan zanjar el problema.

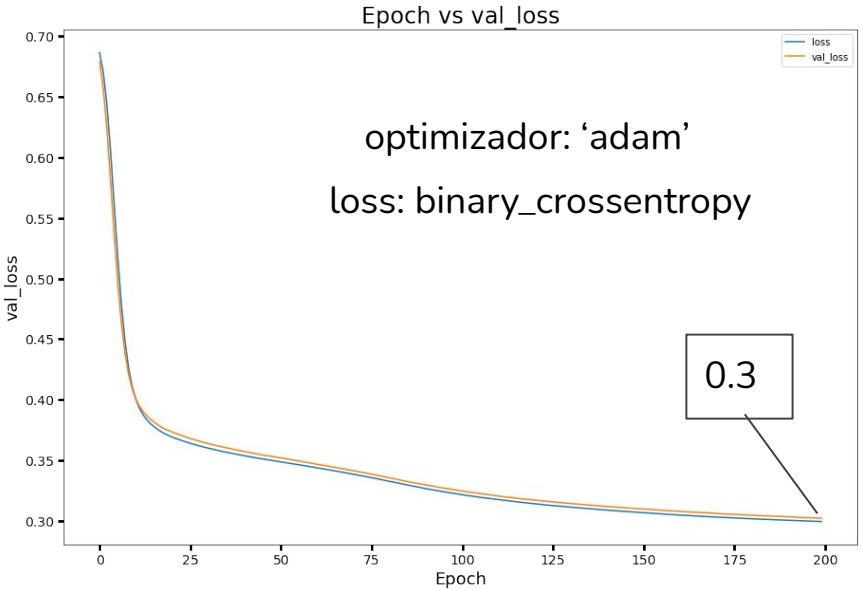
A continuación, dos opciones:

1. Continuar la búsqueda de señales similares y el etiquetado manual. Se tiene mayor control sobre el proceso, pero requiere muchas horas de trabajo que no siempre van a dar resultado.

2. Buscar más parámetros que caractericen mejor las señales (*feature engineering*).

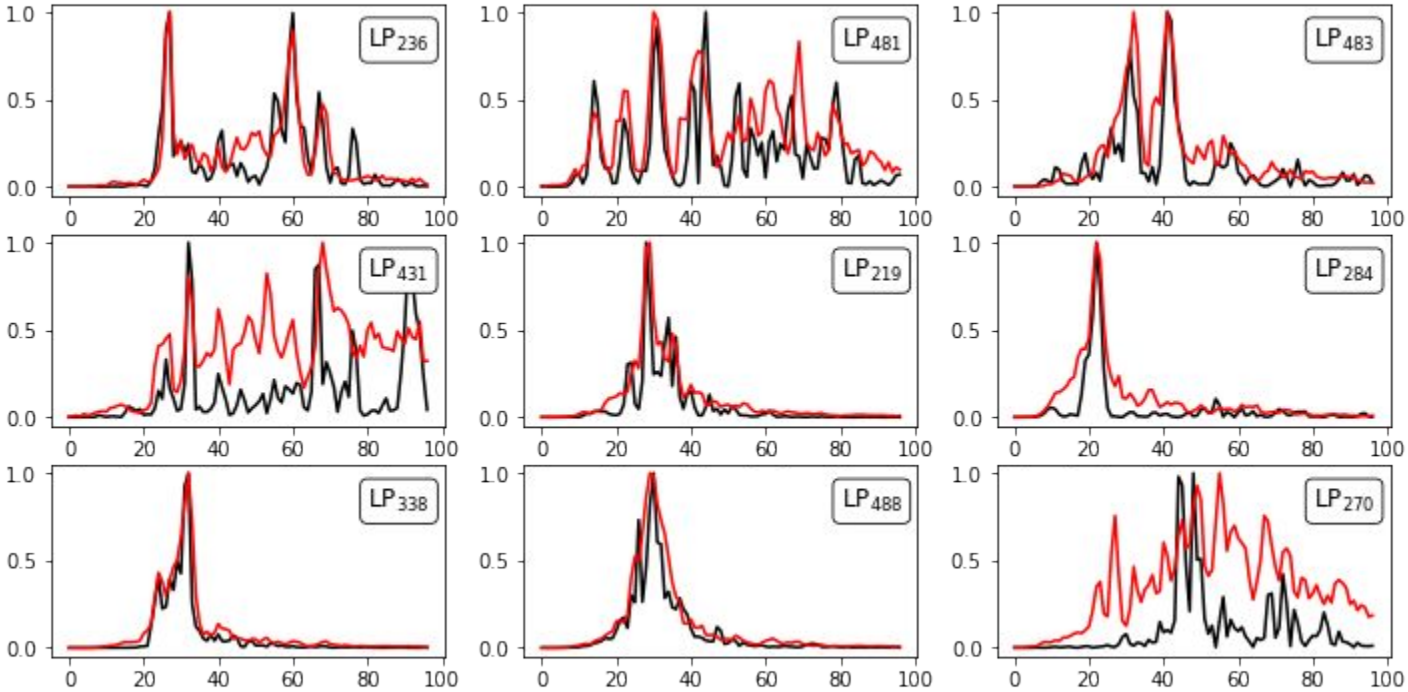
AUTOENCODERS

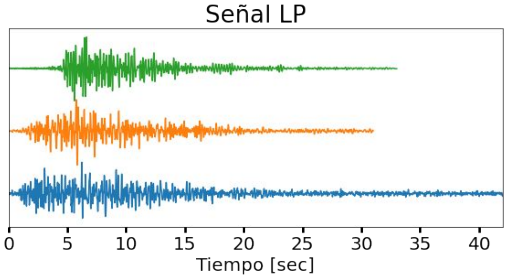


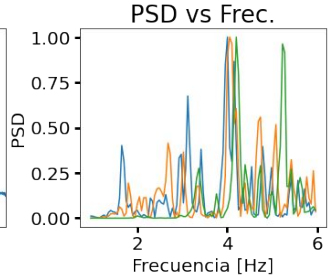


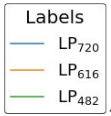
Trainable params: 1,072

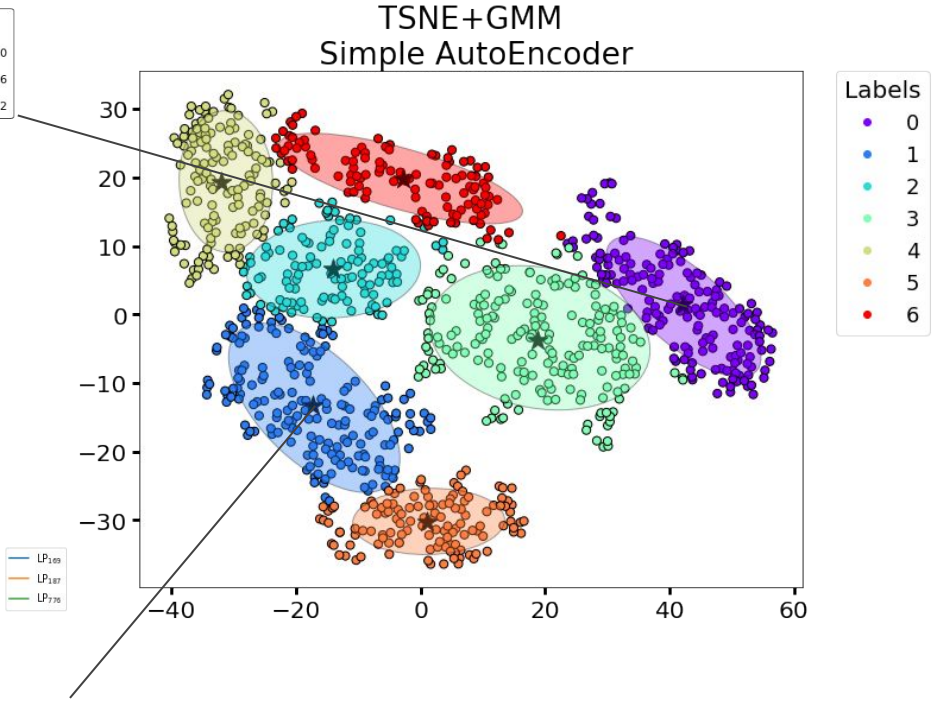
AUTOENCODERS

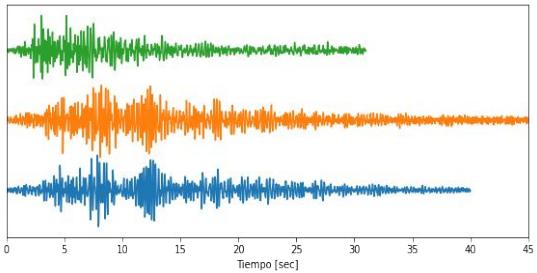


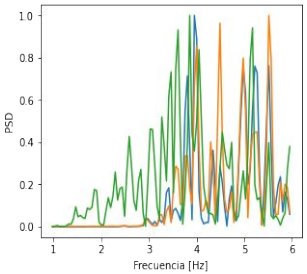






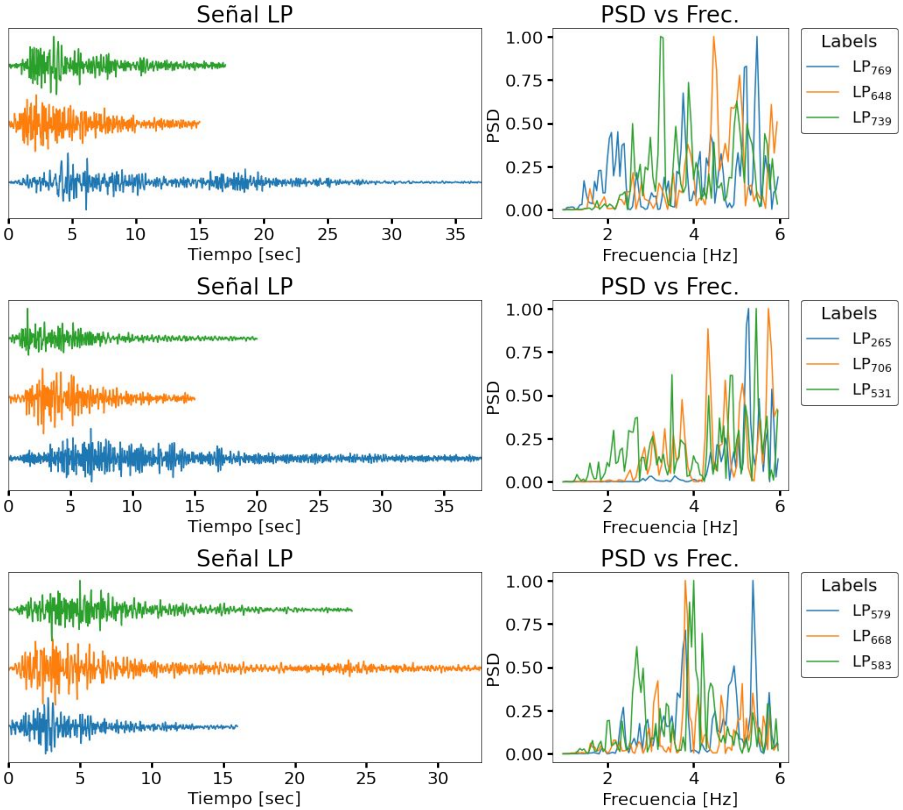




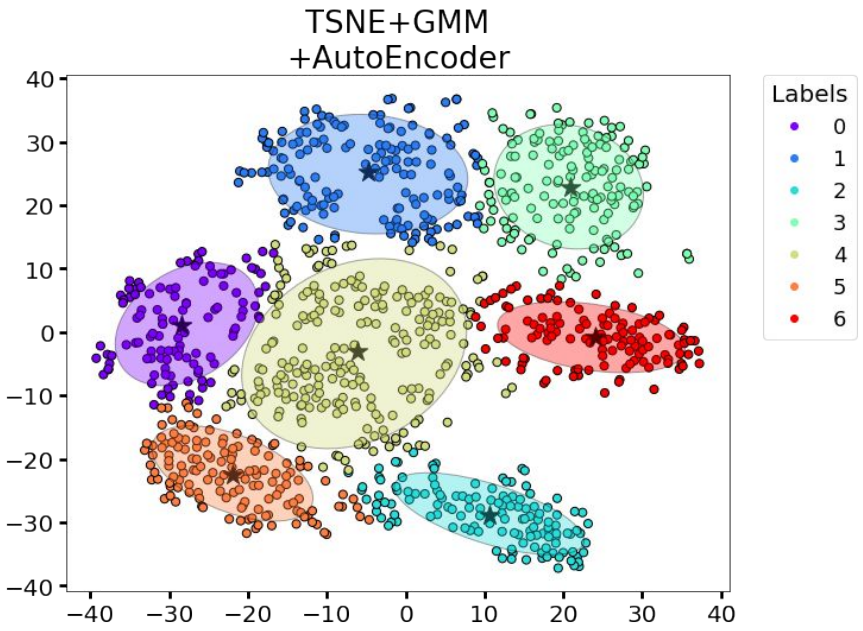


> valores AIC y BIC

< Nro. de etiquetas con prob. > 99.6%
(46% respecto del original)



16 + 5 (StandarScaler) parámetros



Reducción de métricas AIC y BIC

< Nro. de etiquetas con prob. > 99.6%
(84% respecto del original)

Conclusiones

Los mejores resultados se obtuvieron al aplicar tSNE+GMM, filtrar etiquetas por probabilidad, propagar con Logistic Regression y remover aquellas etiquetas que no eran consistentes.

El AE Simple logra reproducir el espectro de las señales, pero los parámetros sí permiten diferenciar entre clases, pero no mejor a como veníamos haciendo.

Combinar los parámetros obtenidos con AE Simple con los parámetros iniciales tampoco mejoró el procedimiento original.

Trabajo futuro. Aplicar Self-Organizing Map (SOM), y Deep Embedding Clustering (DEC). Probar con AEs más sofisticados. Evaluar el peso de cada parámetro.

Estamos en github!

Todas las funciones y procedimientos pueden encontrarse en:

github.com/ifmelchor/volcano_ML-UNSAM