

Uso de geometría y topología  
para ciencia de datos  
MA2007B.60

Dra. Lilia Alanís López

# **DETECTANDO ONDAS GRAVITACIONALES**

Pablo Monzon Terrazas (A01562619)  
Ángel David Ávila Pérez (A01562833)  
Luis Maximiliano López Ramírez (A00833321)  
Adrian Pineda Sanchez (A00834710)  
Kevin González Díaz (A01338316)



# Integrantes del equipo



ADRIÁN PINEDA



PABLO MONZÓN



LUIS LÓPEZ



KEVIN GONZÁLEZ



DAVID ÁVILA

# Calendario de Trabajo

## SEMANA 1

### **Artículo**

*Detection of gravitational waves using topological data analysis and convolutional neural network: An improved approach*

### **Video**

*Topological Methods for the Analysis of Data*

### **Ejercicio Fase 1**

*Reporte de la actividad  
Sliding Window*

## SEMANA 2

### **Artículo**

*Topological feature extraction*

### **Actividad Fase 3**

*Iniciar con el trabajo*

### **Terminar ejemplo de la librería**

*Gravitational wave detection*

## SEMANA 3

### **Actividad Fase 3**

*Continuación del trabajo*

### **Herramienta adicional**

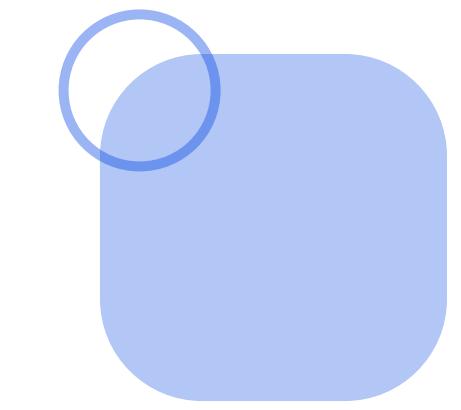
*Discusión de cómo mejorar el trabajo con aportes individuales*

### **Reporte**

*Realización de reporte y dudas con la maestra.*



# CONTENIDO



- 01** Generación de señales
  - 02** Reducción de dimensión
  - 03** Metodología TDA
  - 04** Evaluación Modelos de Machine Learning
  - 05** Validación cruzada de hiperparámetros
  - 06** Resultados y conclusiones
- 

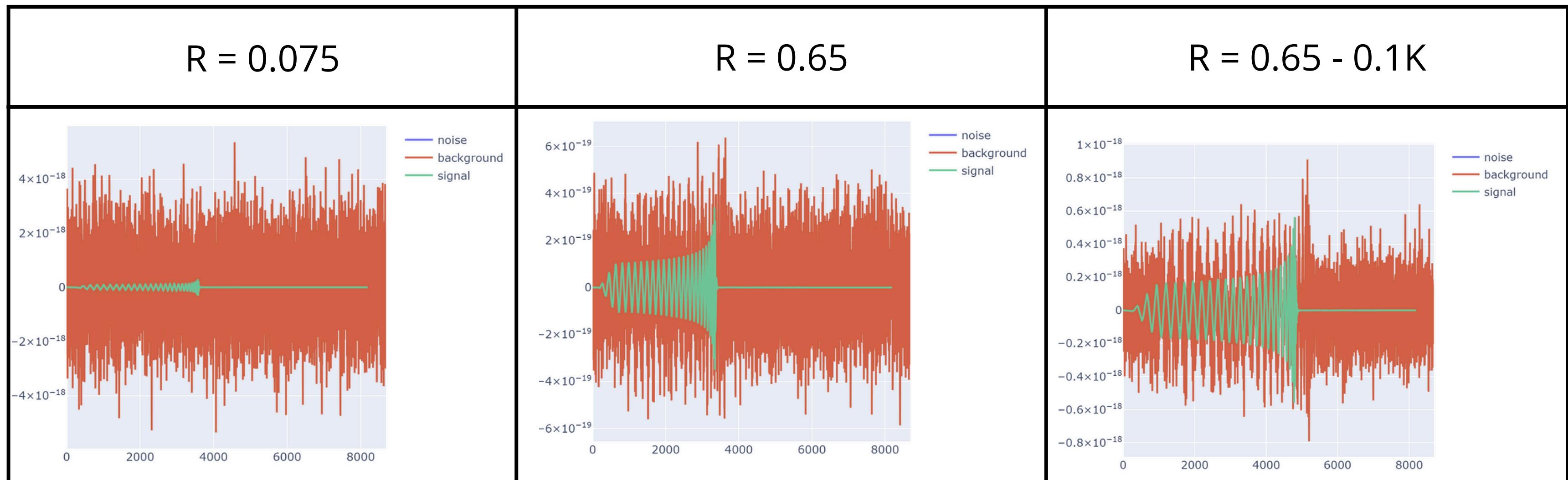
# Generación de señales

**make\_gravitational\_waves**

**Señal - Ruido (SNR)**

$$s = g + \epsilon \frac{1}{R} \xi$$

$$R \in (0.075, 0.65)$$



# Reducción de dimensión

	PCA	UMAP	t-SNE
Ventajas	<ul style="list-style-type: none"><li>• Simple y rápido.</li><li>• Bueno para preservar la varianza</li><li>• Fácil interpretación</li></ul>	<ul style="list-style-type: none"><li>• Preserva estructuras locales y globales.</li><li>• Escalable a grandes conjuntos de datos.</li><li>• Menos sensible a la dimensionalidad de los datos.</li></ul>	<ul style="list-style-type: none"><li>• Excelente en la preservación de estructuras locales.</li><li>• Muy efectivo para visualizar clusters en datos de alta dimensión.</li><li>• Se adapta bien a estructuras complejas.</li></ul>
Desventajas	<p>Lineal, puede no capturar relaciones no lineales. Puede perder estructuras locales importantes.</p>	<p>Los resultados pueden variar con diferentes hiperparámetros.</p>	<ul style="list-style-type: none"><li>• Computacionalmente intensivo, especialmente en grandes datasets.</li><li>• Sensible a los hiperparámetros como la perplexidad y la tasa de aprendizaje.</li></ul>

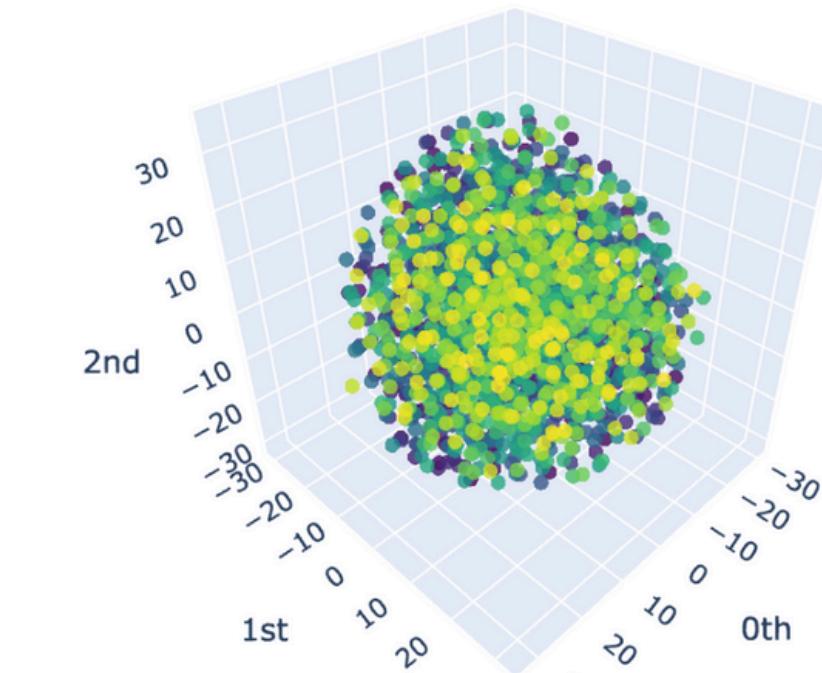
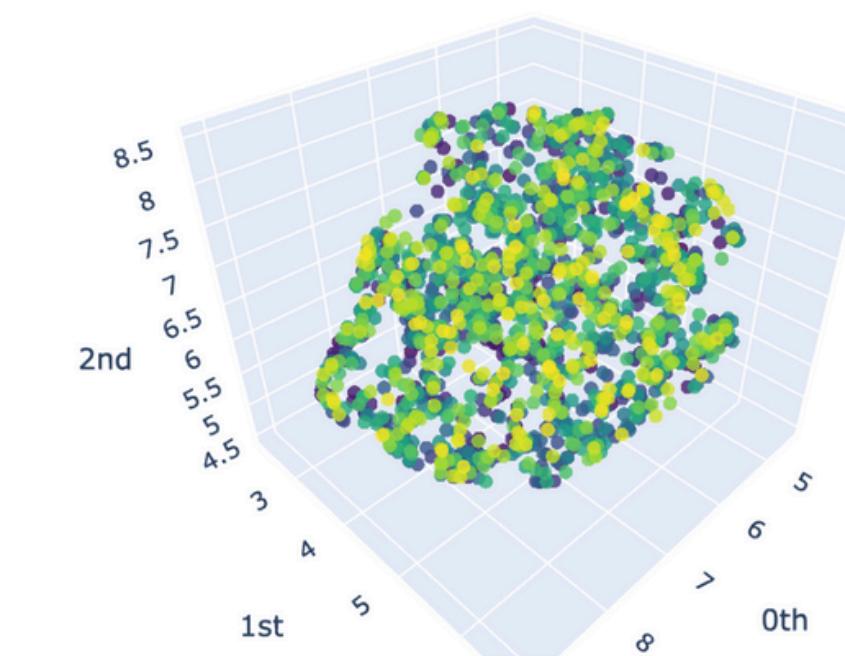
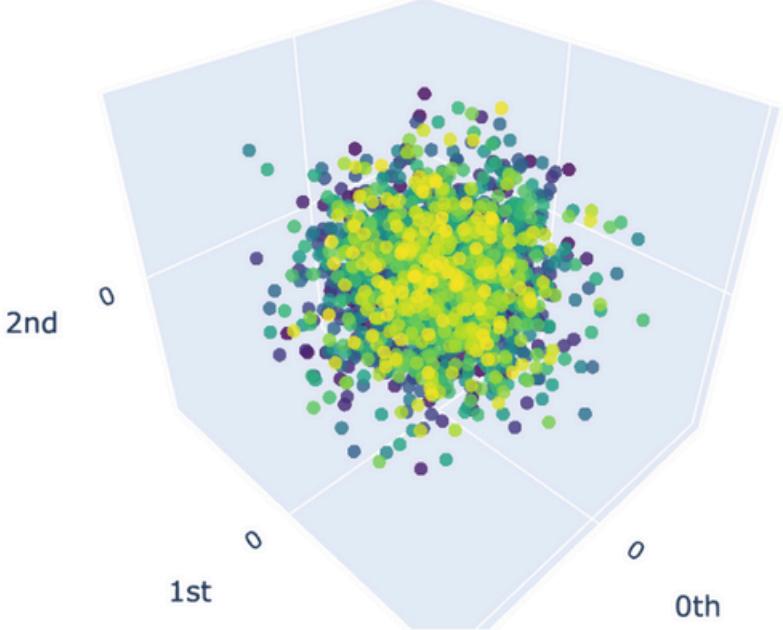
# RUIDO

PCA

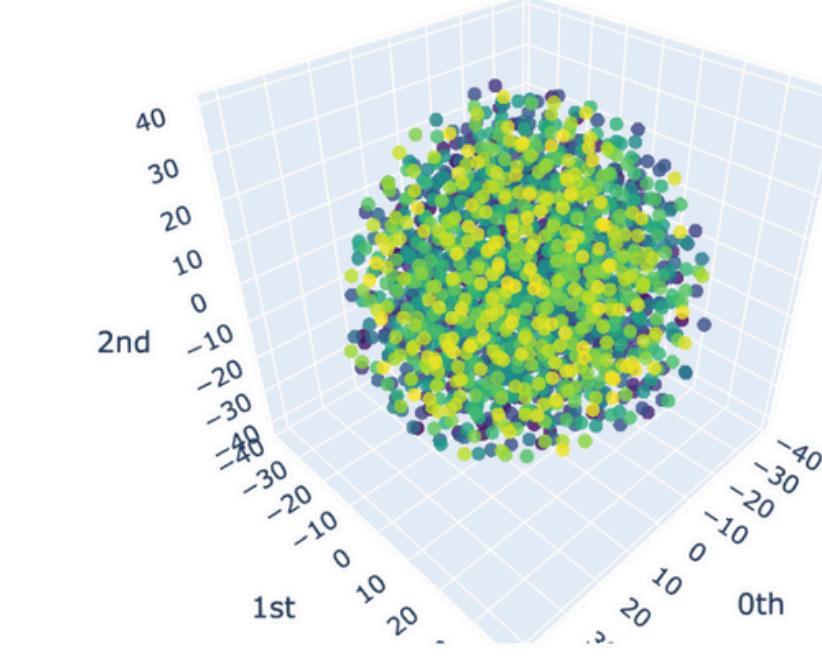
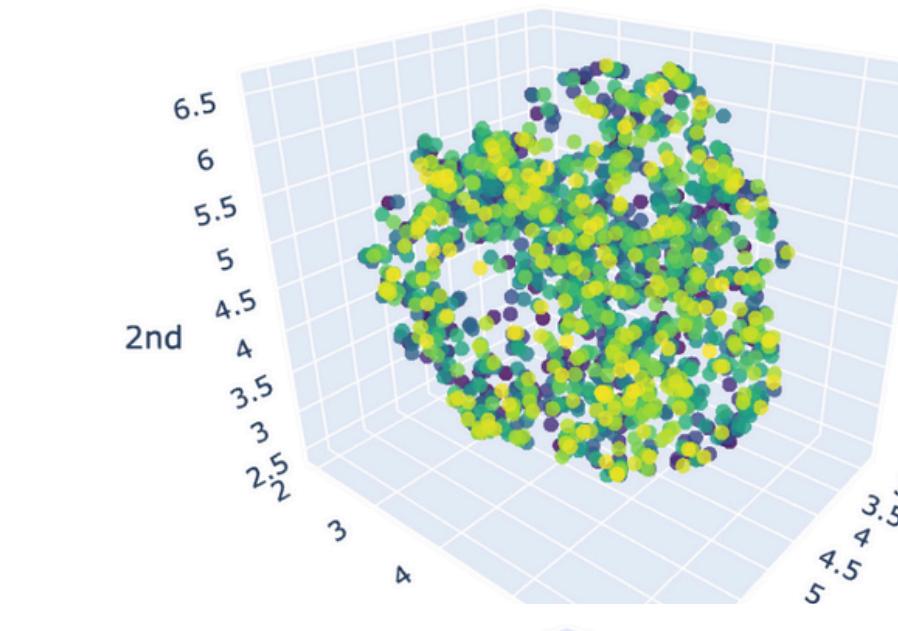
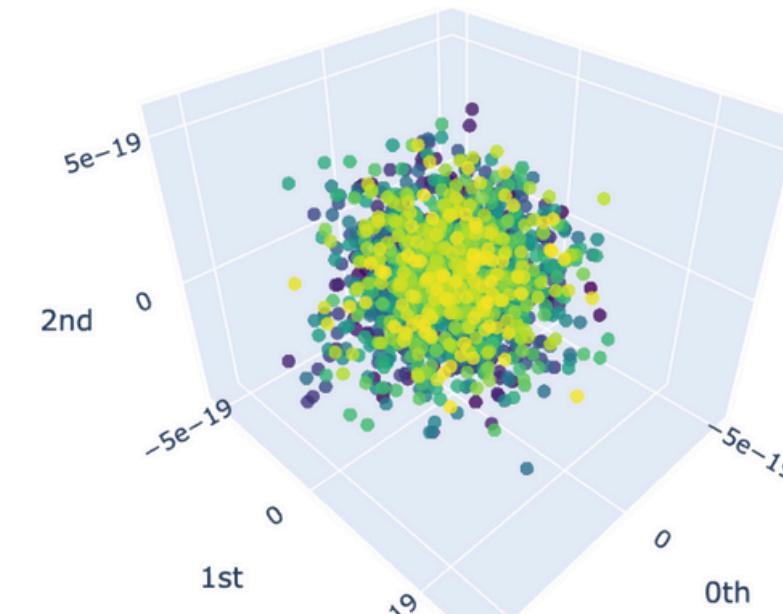
UMAP

t-SNE

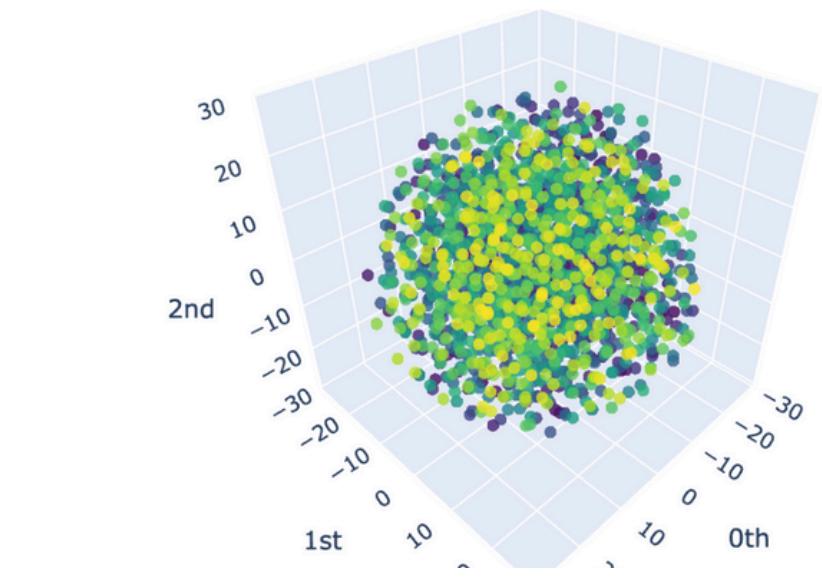
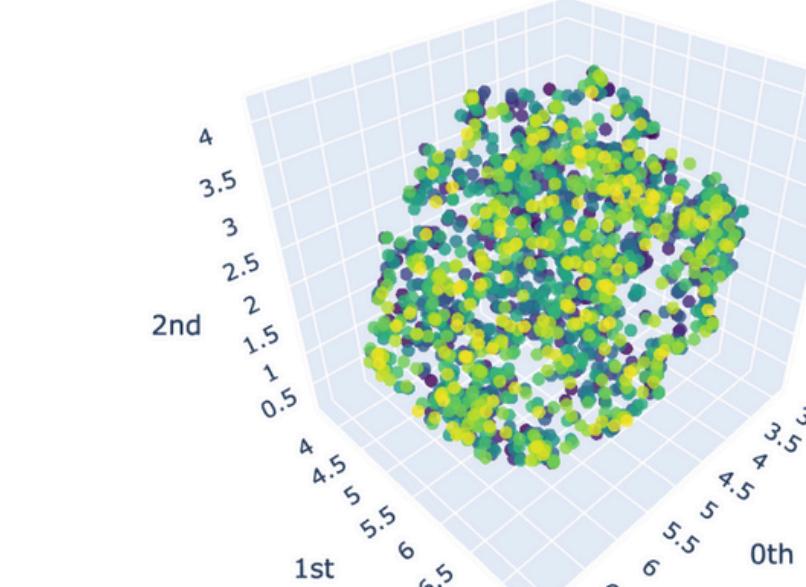
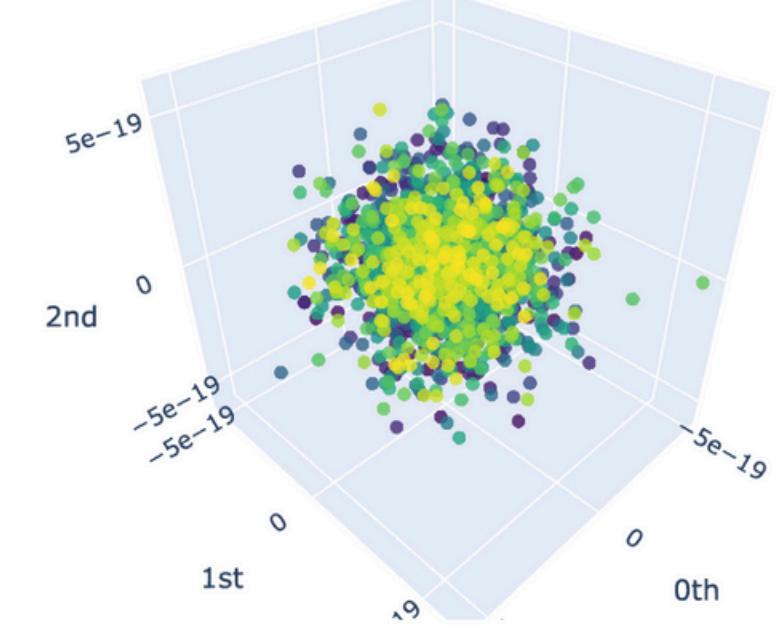
$R = 0.075$



$R = 0.65$

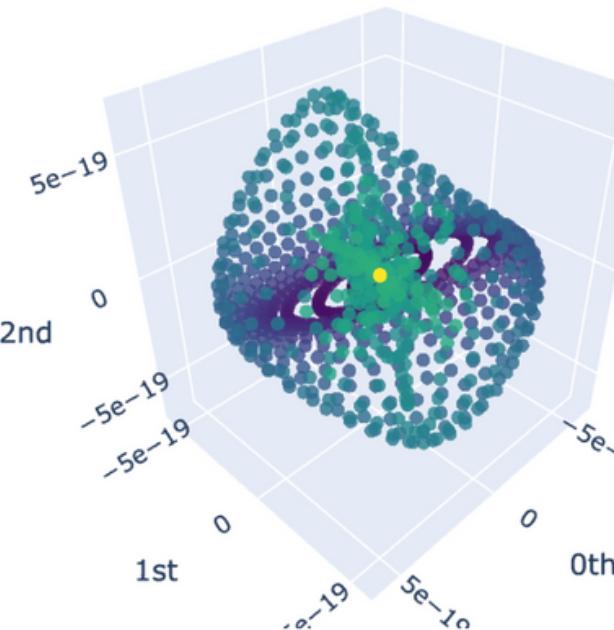
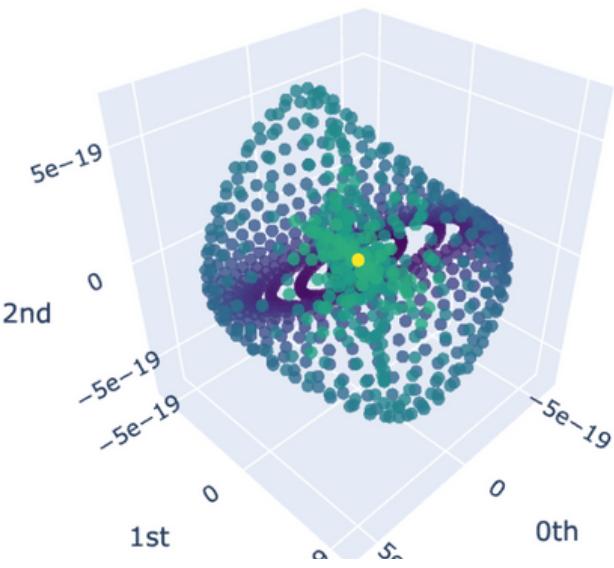
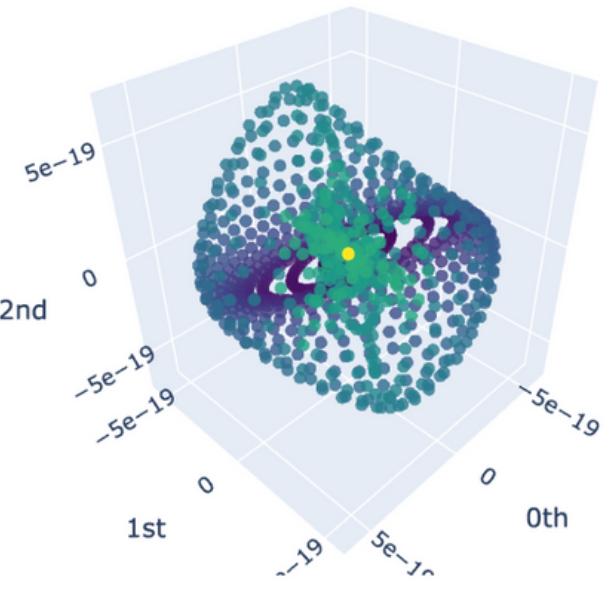


$R = 0.65 - 0.1K$



# SEÑAL

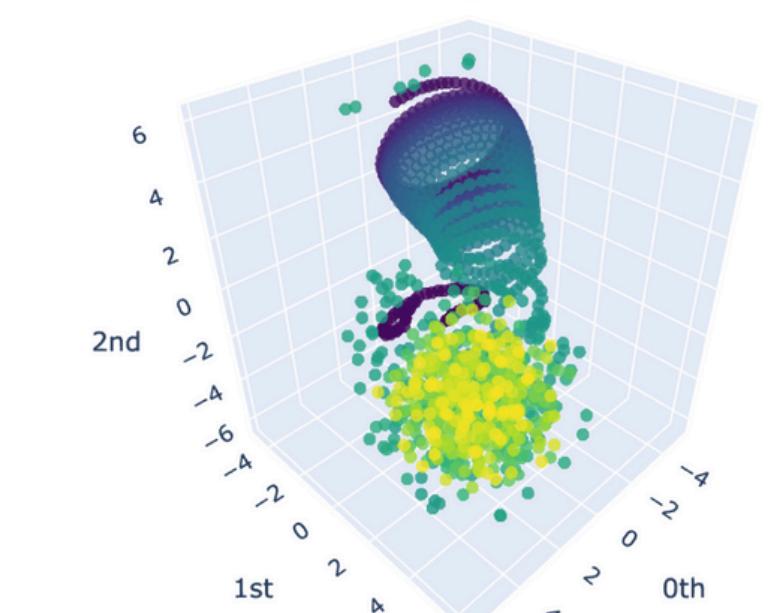
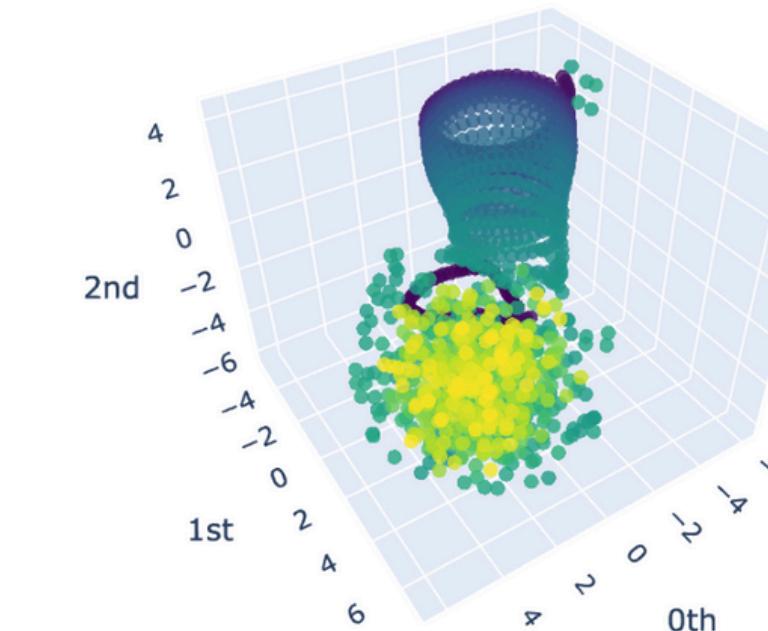
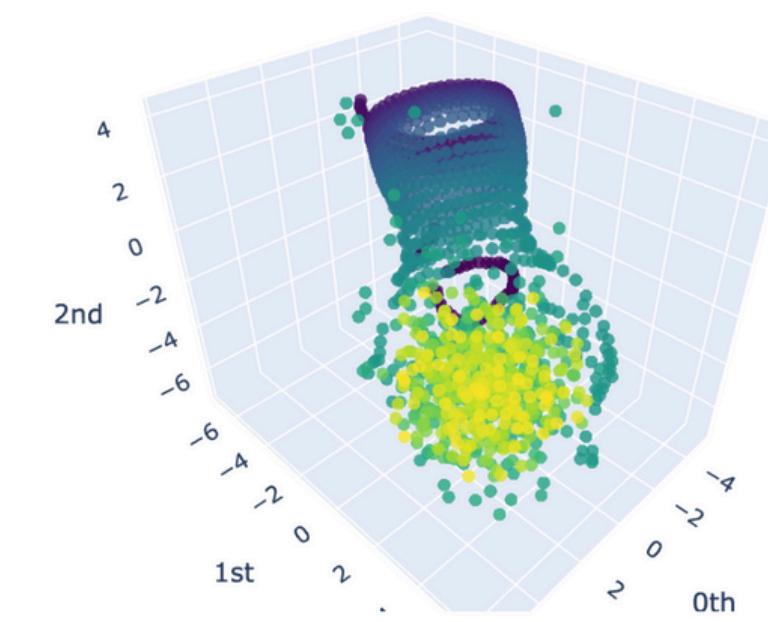
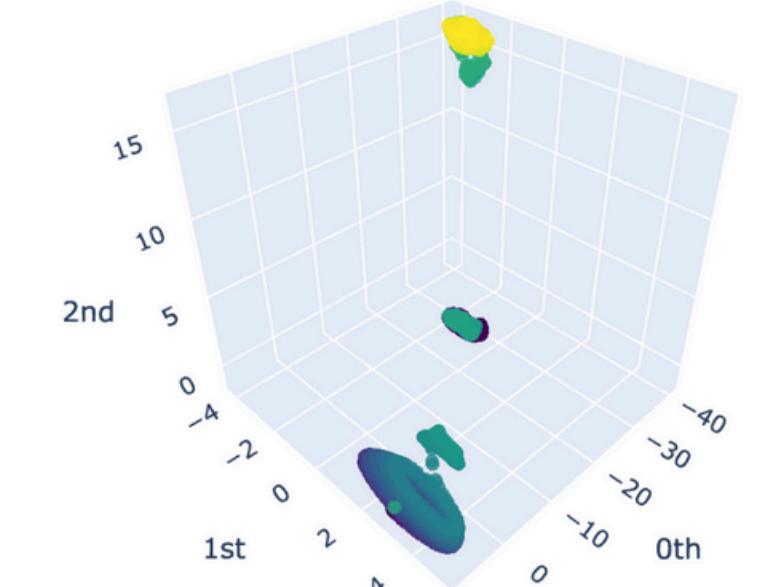
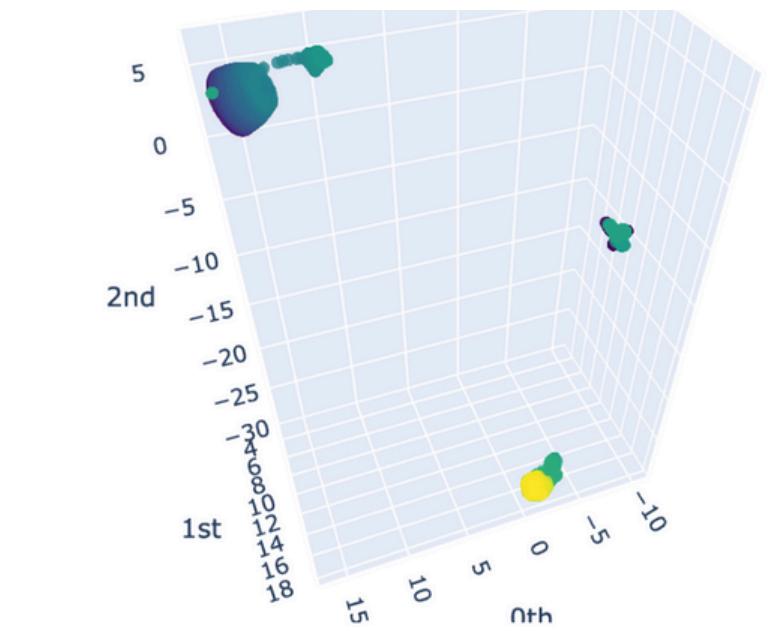
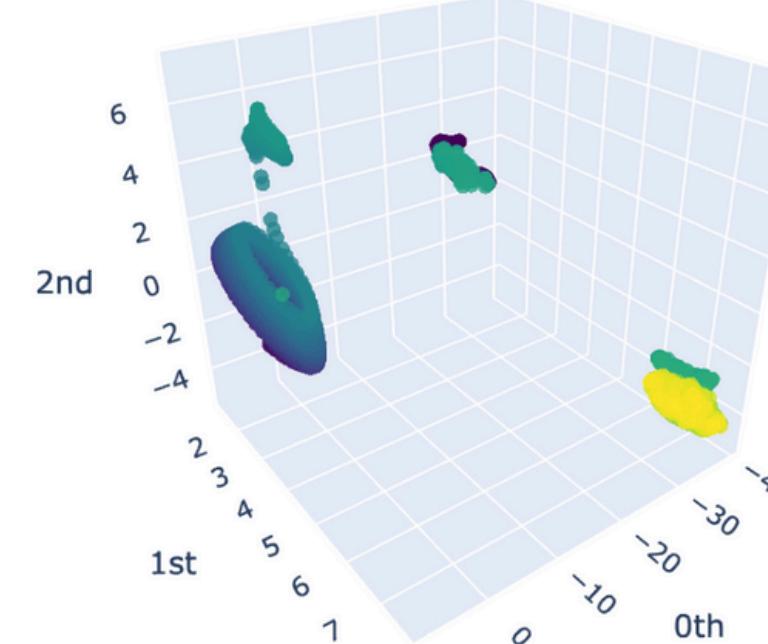
R = 0.075



PCA

UMAP

t-SNE



# Metodología TDA

1	Takens Embedding	Embedding de Takens a partir de series de tiempo.
2	embedding_dimension	Dimensión del espacio para incrustar los datos.
3	embedding_time_delay	Retardo temporal entre observaciones.
4	stride	Paso entre puntos consecutivos.
5	PCA	Reducción de dimensionalidad con tres componentes.
6	Persistence Diagram	Persistencia de Vietoris-Rips para homología en dimensiones 0 y 1.
7	Scaler	Normaliza o escala los datos.
8	Entropy	Calcula la entropía de la persistencia para medir complejidad.
9	Pipeline	Pipeline secuencial con todos los pasos anteriores.

# TDA con PCA + Amplitud

## Propósito General:

Amplitud mide la "amplitud" de los diagramas de persistencia, proporcionando un resumen cuantitativo de las características topológicas representadas en estos diagramas.

## Métrica de 'Bottleneck':

Esta métrica evalúa la distancia más grande entre dos diagramas de persistencia considerando el emparejamiento óptimo de sus puntos.

Mide la diferencia más significativa en las características topológicas que capturan.



```
from gtda.diagrams import PersistenceEntropy, Scaler, Amplitude
entropy = PersistenceEntropy(normalize=True, nan_fill_value=-10)
# Añadir un cálculo de la amplitud para evaluar la importancia de
amplitude = Amplitude(metric='bottleneck')

# Define los transformadores que actúan en paralelo
parallel_processing = FeatureUnion([
    ("amplitude", amplitude),
    ("entropy", entropy)
])

# Ahora construimos el pipeline incluyendo esta transformación en
steps = [
    ("embedder", embedder),
    ("pca", batch_pca),
    ("persistence", persistence),
    ("scaling", scaling),
    ("combined_features", parallel_processing)
]

topological_transfomer = Pipeline(steps)
```

# Señal - Ruido Chico

R = 0.075

	<b>Regresión Logística</b>	<b>Árboles de decisión</b>	<b>Bosques Aleatorios</b>	<b>KNN</b>	<b>Naive Bayes</b>	<b>Promedio</b>
TDA con PCA	Accuracy: 0.52 ROC AUC: 0.505	Accuracy: 0.56 ROC AUC: 0.558	Accuracy: 0.523 ROC AUC: 0.525	Accuracy: 0.55 ROC AUC: 0.544	Accuracy: 0.507 ROC AUC: 0.494	Accuray: 0.532 ROC AUC: 0.5252
Solamente PCA	Accuracy: 0.52 ROC AUC: 0.5	Accuracy: 0.52 ROC AUC: 0.5	Accuracy: 0.52 ROC AUC: 0.5	Accuracy: 0.507 ROC AUC: 0.497	Accuracy: 0.517 ROC AUC: 0.48	Accuracy: 0.5168 ROC AUC: 0.4954
TDA con UMAP	Accuracy: 0.517 ROC AUC: 0.474	Accuracy: 0.47 ROC AUC: 0.47	Accuracy: 0.48 ROC AUC: 0.476	Accuracy: 0.467 ROC AUC: 0.458	Accuracy: 0.523 ROC AUC: 0.521	Accuracy: 0.4914 ROC AUC: 0.4798
TDA con t-SNE	Accuracy: 0.613 ROC AUC: 0.453	Accuracy: 0.573 ROC AUC: 0.574	Accuracy: 0.48 ROC AUC: 0.46	Accuracy: 0.5 ROC AUC: 0.494	Accuracy: 0.373 ROC AUC: 0.504	Accuracy: 0.5078 ROC AUC: 0.497
TDA con PCA + Amplitud	Accuracy: 0.513 ROC AUC: 0.523	Accuracy: 0.547 ROC AUC: 0.546	Accuracy: 0.533 ROC AUC: 0.545	Accuracy: 0.523 ROC AUC: 0.527	Accuracy: 0.527 ROC AUC: 0.497	Accuracy: 0.5286 ROC AUC: 0.5276

# Señal - Ruido Intervalo

R = 0.65 - 0.1K

	Regresión Logística	Árboles de desición	Bosques Aleatorios	KNN	Naive Bayes	Promedio
TDA con PCA	Accuracy: 0.56 ROC AUC: 0.571	Accuracy: 0.576 ROC AUC: 0.582	Accuracy: 0.54 ROC AUC: 0.552	Accuracy: 0.524 ROC AUC: 0.559	Accuracy: 0.58 ROC AUC: 0.57	Accuray: 0.556 ROC AUC: 0.5668
Solamente PCA	Accuracy: 0.544 ROC AUC: 0.5	Accuracy: 0.544 ROC AUC: 0.5	Accuracy: 0.456 ROC AUC: 0.5	Accuracy: 0.544 ROC AUC: 0.568	Accuracy: 0.604 ROC AUC: 0.557	Accuracy: 0.5432 ROC AUC: 0.525
TDA con UMAP	Accuracy: 0.544 ROC AUC: 0.576	Accuracy: 0.516 ROC AUC: 0.513	Accuracy: 0.508 ROC AUC: 0.5	Accuracy: 0.52 ROC AUC: 0.521	Accuracy: 0.608 ROC AUC: 0.592	Accuracy: 0.5392 ROC AUC: 0.5094
TDA con t-SNE	Accuracy: 0.593 ROC AUC: 0.567	Accuracy: 0.54 ROC AUC: 0.54	Accuracy: 0.513 ROC AUC: 0.551	Accuracy: 0.567 ROC AUC: 0.612	Accuracy: 0.513 ROC AUC: 0.526	Accuracy: 0.5452 ROC AUC: 0.5592
TDA con PCA + Amplitud	Accuracy: 0.546 ROC AUC: 0.62	Accuracy: 0.508 ROC AUC: 0.507	Accuracy: 0.628 ROC AUC: 0.652	Accuracy: 0.604 ROC AUC: 0.628	Accuracy: 0.576 ROC AUC: 0.612	Accuracy: 0.5724 ROC AUC: 0.6038

# Señal - Ruido Grande

R = 0.65

	Regresión Logística	Árboles de decisión	Bosques Aleatorios	KNN	Naive Bayes	Promedio
TDA con PCA	Accuracy: 0.674 ROC AUC: 0.741	Accuracy: 0.628 ROC AUC: 0.629	Accuracy: 0.67 ROC AUC: 0.719	Accuracy: 0.648 ROC AUC: 0.699	Accuracy: 0.684 ROC AUC: 0.743	Accuray: 0.6608 ROC AUC: 0.7062
Solamente PCA	Accuracy: 0.476 ROC AUC: 0.5	Accuracy: 0.476 ROC AUC: 0.5	Accuracy: 0.476 ROC AUC: 0.5	Accuracy: 0.506 ROC AUC: 0.538	Accuracy: 0.568 ROC AUC: 0.551	Accuracy: 0.5004 ROC AUC: 0.4954
TDA con UMAP	Accuracy: 0.572 ROC AUC: 0.631	Accuracy: 0.54 ROC AUC: 0.541	Accuracy: 0.558 ROC AUC: 0.583	Accuracy: 0.562 ROC AUC: 0.566	Accuracy: 0.584 ROC AUC: 0.576	Accuracy: 0.5632 ROC AUC: 0.5794
TDA con t-SNE	Accuracy: 0.553 ROC AUC: 0.568	Accuracy: 0.487 ROC AUC: 0.486	Accuracy: 0.5 ROC AUC: 0.51	Accuracy: 0.513 ROC AUC: 0.536	Accuracy: 0.507 ROC AUC: 0.535	Accuracy: 0.512 ROC AUC: 0.527
<b>TDA con PCA + Amplitud</b>	Accuracy: 0.782 ROC AUC: 0.853	Accuracy: 0.708 ROC AUC: 0.708	Accuracy: 0.762 ROC AUC: 0.854	Accuracy: 0.746 ROC AUC: 0.816	Accuracy: 0.776 ROC AUC: 0.827	<b>Accuracy: 0.7548 ROC AUC: 0.8116</b>

# Mejores parámetros para el TDA

	emb_dim	emb_time	stride_variable	pca_components	tam_entrenamiento
1	20	25	5	5	0.75
2	30	20	7	4	0.8
3	50	15	10	3	0.85
4	100	10	3	2	0.9
5	150	5	13	1	0.85

Average Accuracy on valid: 0.8343999999999999

Average ROC AUC on valid: 0.8879081403553787

Best model by Accuracy: Random Forest with a score of 0.864

Best model by ROC AUC: Random Forest with a score of 0.925

# Mejores hiperparámetros por modelo

Al obtener los mejores parámetros para el TDA, se probaron hiperparámetros para cada modelo.

Se probaron 5 configuraciones distintas para Regresión Logística y para Bosques Aleatorios, mientras que para los otros tres modelos se probaron 3 configuraciones distintas cada uno.

```
rf_model_1 = RandomForestClassifier(n_estimators=50, max_features='sqrt', max_depth=5)
```

## RANDOM FOREST

Accuracy on valid: 0.88  
ROC AUC on valid: 0.93



Accuracy on valid: 0.872  
ROC AUC on valid: 0.926

Accuracy on valid: 0.864  
ROC AUC on valid: 0.925

Accuracy on valid: 0.868  
ROC AUC on valid: 0.924

Accuracy on valid: 0.872  
ROC AUC on valid: 0.926

## REGRESIÓN LOGÍSTICA

Accuracy on valid: 0.724  
ROC AUC on valid: 0.839

Accuracy on valid: 0.852  
ROC AUC on valid: 0.922

Accuracy on valid: 0.87  
ROC AUC on valid: 0.927

Accuracy on valid: 0.802  
ROC AUC on valid: 0.919

Accuracy on valid: 0.844  
ROC AUC on valid: 0.921

## KNN

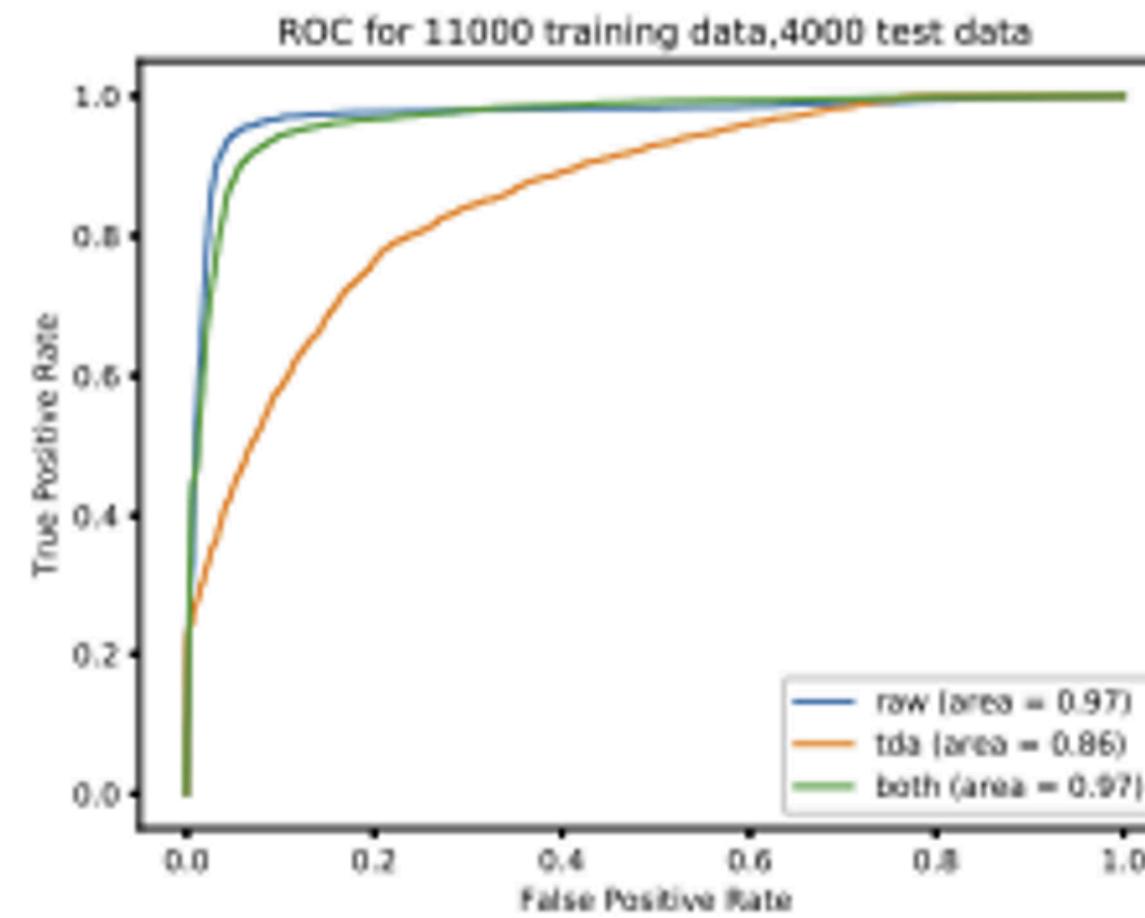
Accuracy on valid: 0.846  
ROC AUC on valid: 0.895

Accuracy on valid: 0.856  
ROC AUC on valid: 0.911

Accuracy on valid: 0.868  
ROC AUC on valid: 0.917



# Comparativa de resultados con el Paper



Random Forest  
Accuracy on valid: 0.88  
ROC AUC on valid: 0.93

# CONCLUSIONES

En general, la mejor configuración para detectar las ondas gravitacionales fue la de **TDA con PCA + Amplitud**, sobre todo cuando la señal - ruido es alta.

También se observó que cuanto mayor sea la señal - ruido de las ondas, mejor se detectan dichas ondas mediante los modelos probados.

En nuestro análisis, los parámetros que nos ayudaron a mejorar el TDA fueron bajar el 'stride' a 3 y el número de componentes del PCA a 2.

Nuestros mejores resultados se dieron entre el modelo de Bosques Aleatorios y de Regresión Logística, siendo el de **Bosques Aleatorios** el mejor en cuanto exactitud y curva ROC AUC.