Descubriendo Periodicidad, Reconocimiento de Patrones y Modelos de Pronóstico de Aprendizaje Automático para El Niño-Oscilación del Sur mediante el Análisis Topológico de Datos

Francisco Castorena Salazar, Juan Pablo Echeagaray González, Emily Rebeca Méndez Cruz, José Eugenio Morales Ortiz, and Mario Javier Soriano Aguilera

Ing. en Ciencias de Datos y Matemáticas, Tec de Monterrey

16 de junio del 2023

Resumen

Este artículo explora la aplicación de técnicas de Análisis de Datos Topológicos (TDA, por sus siglas en inglés) en el contexto de los índices del fenómeno El Niño-Oscilación del Sur (ENSO) para identificar la periodicidad, descubrir anomalías y generar pronósticos. Se logró encontrar índices de periodicidad en todas las mediciones de ENSO a través del encaje de *Takens* así como a través del algoritmo *Mapper*. El modelo de pronóstico de ML desarrollado logró capturar la complejidad subyacente y la no linealidad del comportamiento de ENSO, lo que motiva el uso de TDA para obtener información más precisa.

1. Introducción

El ENSO es un fenómeno climatológico que puede ser considerado en 3 partes principales: El niño (calentamiento de la superficie del oceano), La niña (Enfriamiento) o neutro donde hay una temperatura estable. Debido al error de los modelos sistemáticos, limitaciones de los modelos, y también a la dificultad que resultan las mediciones del ENSO debido a que es demasiado sensible a diversos cambios, fue que se opto por otro tipo de herramientas para poder análisis de los datos para predicciones y búsqueda de periodicidad. Este escrito se organiza de la siguiente manera en las secciones 2 y 3 se describen los fundamentos teóricos y prácticos de las herramientas y técnicas a utilizar, así como una descripción detallada de cómo fueron aplicadas al caso de estudio. En la sección 4 se presentan los resultados obtenidos, los cuales son analizados en retrospectiva en la sección 5 y en la sección 6 se presentan las conclusiones del estudio.

2. Marco Teórico

2.1. Takens Embedding

Este método esta basado en la técnica *Time Delay Embedding*, la cual es una representación univariada de una serie de tiempo a partir de una nube de puntos [1].
Se le conoce como *Time Delay Embedding* a la manera de poder transformar una serie de tiempo a una matriz de tiempo la cual depende de pedazos de los datos [2]. Al transformar una secuencia larga de información a un set pequeño dependientes del tiempo, este se convierte en el centro del análisis en lugar de la predicción de variables en particular en una serie de tiempo.
Dada una serie de tiempo discreto $(X_0, X_1, \dots)$ y una secuencia de muestras de tiempo separadas de manera uniforme $t_0, t_1, \dots$, se extrae un set d-dimensional de vectores de la forma $(X_{t_i}, X_{t_i+\tau}, \dots, X_{t_i+(d-1)\tau})$ para $i = 0, 1, \dots$. Este set es el que se conoce como el *Takens Embedding* de la serie de tiempo y la cual puede ser interpretada como una nube de puntos.
La diferencia entre el tiempo t_{i+1} y t_i se conoce como *stride* o paso, τ es conocido como el tiempo de retraso y d se le atribuye a la dimensión del embedding.

2.2. Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection (UMAP) es una técnica de reducción de dimensiones que puede ser utilizado para visualización similar al t-SNE (t-Distributed Stochastic Neighbour Embedding), pero también es funcional con reducción general de dimensiones no lineales, dicho algoritmo funciona bajo 3 suposiciones [3].

- Los datos estan uniformemente distribuidos en un *Riemannian manifold*.
- La métrica *Riemannian* es localmente constante.
- El colector (manifold) esta localmente conectado.

Con estos supuestos se puede entonces dividir la construcción del algoritmo de UMAP en dos fases principales. La primera consiste en construir representación topológica difusa, después en la segunda fase consiste en la optimización de la reducción de dimensionalidad para acercarnos lo más posible a la representación topológica difusa la cual se mide mediante *cross entropy*. Para más detalles se puede consultar la siguiente referencia [4].

2.3. Spectral Embedding

Spectral Embedding es un método no lineal de reducción de dimensionalidad, forma una matriz de afinidad dada por la función especifica y aplica descomposición espectral a la matriz laplaciana correspondiente, el resultado esta dado por el valor del eigenvector de cada punto de los datos [5].
Spectral Embedding es una aproximación al calculo de un encaje no lineal. La librería de Scikit-Learn utilizada implementa *Laplacian Eigenmaps*, los cuales encuentran la representación de baja dimensión de los

datos usando descomposición espectral de la matriz laplaciana, esta se puede considerar una aproximación discreta. La minimización de costo de la función asegura que los puntos cercanos unos a otros en el manifold sean mapeados cerca en el espacio de baja dimensión, preservando la distancia local [6].

2.4. Complejo Vietoris-Rips

El complejo de Vietoris-Rips es un complejo simplicial definido en un espacio métrico finito, este complejo suele ser usado como una aproximación del complejo de Čech debido a que el cálculo computacional de Vietoris-Rips es más rápido. [7]
Dado un espacio métrico (X, d_x) y $r > 0$ el complejo Vietoris-Rips asociado a r ($VR_r(X)$), tiene a X como su conjunto de elementos y sus complejos simpliciales son todos aquellos subconjuntos finitos no vacíos de X cuyo diámetro es estrictamente menor a r.[8]
La cantidad de complejos simpliciales de Vietoris-Rips puede ser muy grande, donde para una dimensión k con n puntos se pueden tener hasta $O(n^{k+1})$. Por lo cual, en la mayoría de los casos una reducción de dimensionalidad es necesaria.[9]

2.5. XGBoost

XGBoost es una librería que implementa algoritmos de aprendizaje automático bajo la técnica de Gradient Boosting, utilizando Gradient boosting Machines (GBM) para ser capaz de resolver muchos problemas predictivos de forma 'rápida y precisa'.[10]

2.5.1. Gradient Boosting

Este tipo de algoritmo es conocido por su capacidad de encontrar relaciones no lineales entre una variable objetivo y sus variables predictoras, por otra parte puede manejar valores faltantes, atípicos e incluso categóricos sin necesidad de tratamiento especial. Es una técnica de ensamble que combina múltiples modelos llamados débiles para poder crear un modelo fuerte, los débiles son arboles de decisión que se usan para ajustar el modelo y tener mejores predicciones.[11]

3. Metodología

3.1. Limpieza de bases de datos

Se realizará una limpieza de la base de datos proporcionada por ClimateAI así como de la base de datos general de la NOAA obtenida a través de la siguiente liga. El formato de las bases de datos debe ser adecuado a uno tabular que sea sencillo de interpretar y de utilizar para los algoritmos de aprendizaje automático.

3.2. Análisis de Periodicidad

Dada una serie de tiempo obtenida de la limpieza de datos realizada, se realizará un corte de la serie de tiempo en ventanas arbitrarias; dichas ventanas serán procesadas con el algoritmo de *Takens Embedding* para obtener una representación de la serie de tiempo en un espacio de dimensión superior. La selección de la dimensión de encaje y el retraso de tiempo serán determinados por el algoritmo de *False Nearest Neighbors* y *Mutual Information* respectivamente [12].
La serie de tiempo transformada será sujeta a una reducción de dimensionalidad (UMAP o Spectral Embedding) a un espacio de 3 dimensiones; se estimará un diagrama de persistencia de Vietoris-Rips y la curva de números de Betti para la serie de tiempo procesada. Los resultados de estas operaciones serán visualizados para ser interpretados de forma manual y determinar si existe evidencia periodicidad en la serie de tiempo.

3.3. Reconocimiento de Patrones

El reconocimiento de patrones será acotado al análisis de los índices de anomalías *ENSO* en la región 3.4; no se realizará ningún corte a esta serie de tiempo así como tampoco un tratamiento de valores atípicos, puesto que se trabajará bajo el supuesto de que el algoritmo *Mapper* será lo suficientemente robusto para ignorar dichos valores.
Para enriquecer el agrupamiento a realizar, se generarán 2 series de tiempo nuevas que contengan en sus observaciones el año y mes de ocurrencia de la medición respectivamente. Estas 3 características serán proyectadas al índice de anomalías Nino3.4; donde se generará una cubierta arbitraria en base la selección de cubos y traslape permitido.
Dicha cubierta será agrupada con KMedias dada la distribución de campana que sigue esta variable; finalmente se generará una visualización del grafo compuesto para ser analizado de forma manual.

3.4. Modelado de Pronósticos

Se implementará un modelo de ML para pronosticar los índices de anomalías Nino3.4 en un horizonte de 40 meses. El modelo será informado por las propiedades topológicas de la serie de tiempo, en la figura 1 se muestra el pipeline de modelado a implementar. Se generará una división de los datos en entrenamiento y prueba, dejando los últimos 40 meses para la prueba del modelo.
La optimización de hiper-parámetros todavía no es soportada por la librería *giotto*, por lo que se será pospuesta como trabajo futuro.

4. Experimentación y resultados

4.1. Análisis de Periodicidad

En la figura 2 se despliega la serie de tiempo de índices de anomalías Nino 3.4 en una ventana de tiempo de 15 años comenzando en 1980.

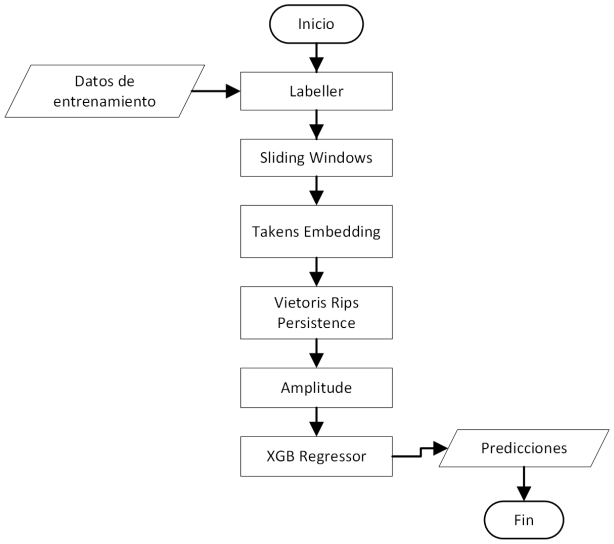

Figura 1: Pipeline de modelado

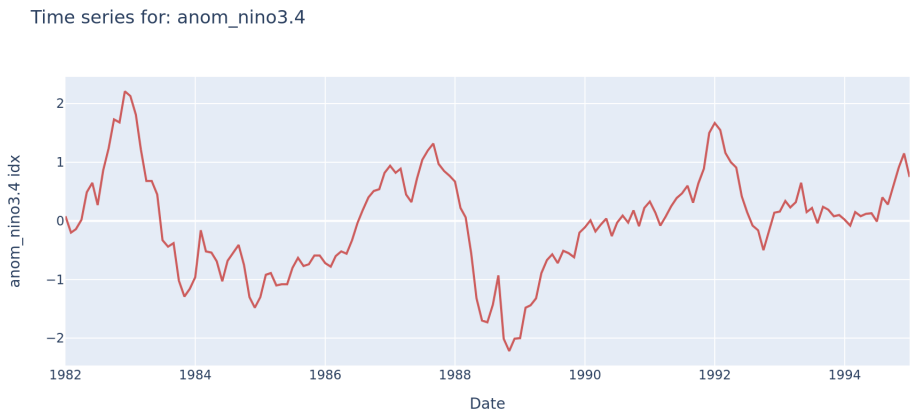

Figura 2: Serie de tiempo de índice de anomalías Nino 3.4, 1980 a 1995

La serie de tiempo fue procesada con el algoritmo de *Takens Embedding*, la serie de tiempo fue proyectada a un espacio de 5 dimensiones con un retraso temporal de 2 observaciones. Dicha proyección fue a su vez reducida a un espacio de 3 dimensiones con el algoritmo de *UMAP* con 4 componentes (figura 3); se seleccionó este número de componentes bajo el supuesto de que la serie de tiempo contiene eventos del Niño, la Niña, eventos neutrales y una componente de ruido. Los diagramas de persistencia fueron calculados con el complejo de Vietoris-Rips (figura4), la complejidad computacional del algoritmo no representó un problema dada la reducción de dimensionalidad originada por el encaje.

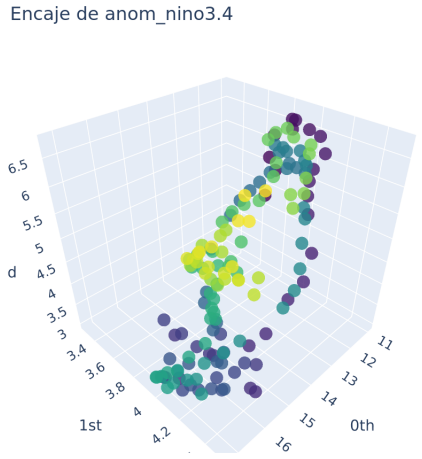

Figura 3: Proyección a 3 dimensiones con UMAP del encaje de Takens de Anom Nino3.4

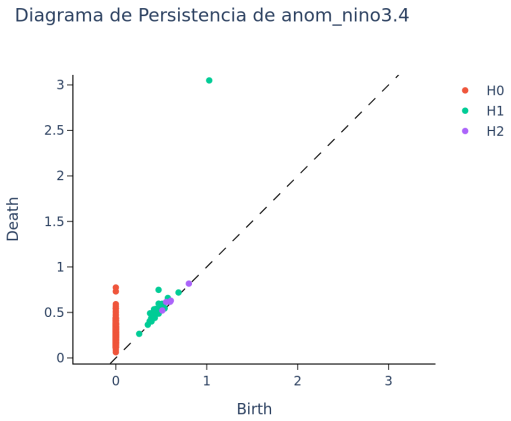

Figura 4: Diagrama de Persistencia de Anom Nino3.4

En el diagrama de persistencia anterior se puede apreciar como se tiene un punto de H_1 que persiste por más tiempo que los demás puntos del tipo H_1 siendo este el hueco más grande que se tiene para la componente conexa generada, presentando evidencia suficiente como para afirmar que existe periodicidad en la serie de tiempo, la cual podemos comprobar de forma heurística al observar la forma de dona de la figura 3. De forma general se encontraron índices de periodicidad en todas las series de tiempo proporcionadas; en la tabla 1 se presenta un sumario de los parámetros utilizados, y en la sección **??** se presentan los diagramas de persistencia de cada serie de tiempo.

Variable	**Ventana de tiempo**	**Proyección**	**Dimensión de encaje**	**Retraso temporal**
Nino 1.2	1980-2023	Spectral Embedding	8	3
Anom Nino1.2	2000-2020	UMAP	6	2
Nino3	1992-2005	UMAP	6	1
Anom Nino3	1980-1995	UMAP	7	3
Nino3.4	1982-2006	UMAP	7	4
Anom Nino3.4	1980-1995	UMAP	5	2
Nino4	1980-1997	UMAP	7	3
Anom Nino4	1980-2023	UMAP	8	10

Cuadro 1: Resumen de encajes de Takens y proyecciones

4.2. Reconocimiento de patrones

En la figura 5 se despliega el grafo creado por el algoritmo *Mapper* al tratar de agrupar las observaciones de anomalías Nino 3.4. Se destaca que el grafo generado tiene forma de cuerda, donde no quedó ningún nodo desconectado de la red, y con cada nodo teniendo una distribución de valores distinta, representada por el índice de anomalías.

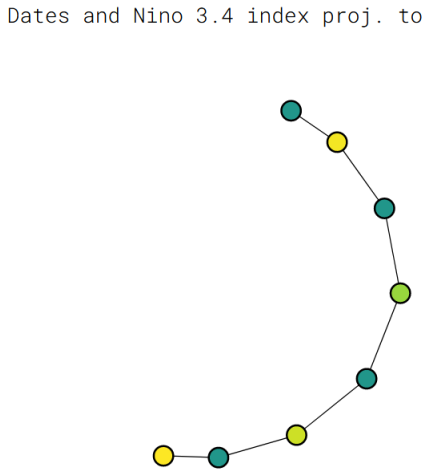

Figura 5: Grafo de línea generado por Mapper en Anom Nino3.4

La secuencia de nodos *calientes* a *fríos* que sucede en todo el grafo lleva a pensar que *Mapper* tal vez esté capturando la naturaleza periódica del fenómeno. En la figura 6 se despliegan las series de tiempo de 4 nodos (caliente a frío).

Notamos que al agrupar los nodos en tuplas (caliente, frío) los meses de la serie de tiempo siempre son respetados, garantizando una continuidad entre todas las observaciones; aunado a esto, se logra llegar a una separación de observaciones mayoritariamente frías de las calientes; el equipo conjetura que estas agrupaciones naturales en conjunto con la continuidad de las observaciones son evidencia considerable de que el algoritmo *Mapper* captura de forma adecuada la naturaleza periódica del fenómeno.

4.3. Modelo de pronóstico

Una vez hecha la separación en datos de entrenamiento y prueba, se entrenó un regresor XGB sobre las características topológicas de la serie de tiempo Anom Nino 3.4; en la figura 7 se despliega una muestra de predicciones que va desde 1880 a 1940. Destacamos que la serie de tiempo no presenta un sobre-ajuste a los datos, pero sí logra capturar la forma misma de la serie; indicando que las características extraídas del TDA fueron lo suficientemente informativas como para que el modelo de pronóstico pudiera aprender la forma de la serie de tiempo.

En la figura 8 se presentan las 40 predicciones qe realizó el modelo en los datos de prueba. En este caso aislado notamos que el modelo logró capturar la tendencia de la serie de tiempo con éxito, más no así los valores exactos de la misma. El *MAPE* de este modelo fue de 1.58 %.

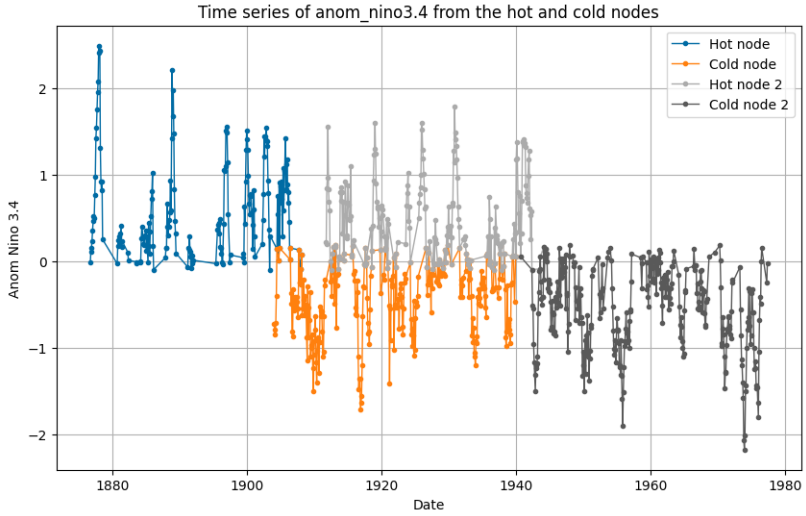

Figura 6: Muestra de clústers definidos por Mapper en Anom Nino3.4

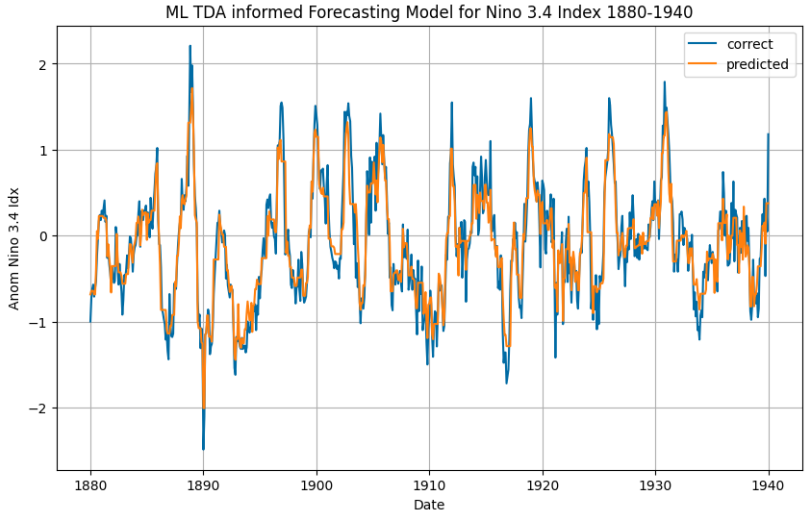

Figura 7: Pronóstico de Anom Nino3.4

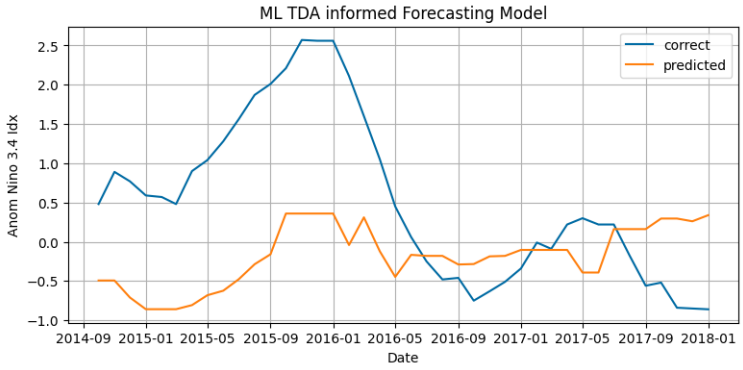

Figura 8: Pronóstico en datos de prueba

5. Discusión de resultados

5.1. Generalización de análisis de periodicidad

Actualmente el análisis implementado para la búsqueda de periodicidad para todas las variables a excepción de Anom Nino3.4 quedaron confinadas a observaciones a partir de 1980; la inclusión de datos históricos que cubran desde 1880 (como lo es con Anom Nino3.4) enriquecería el estudio realizado, así como también permitiría una mejor generalización de los resultados obtenidos.

5.2. Funciones de proyección para Mapper

El agrupamiento con Mapper realizado también se vio mermado por el grado de información disponible para las mediciones del índice de anomalías de Nino 3.4; se considera que al disponer de un mayor número de observaciones para las demás variables, se podría hacer uso de funciones de proyección y agrupamiento más complejas en el proceso de generación de cubos y clusters, incrementando aún más el grado de análisis de las relaciones de interdependencia entre las variables.

5.3. Optimización de hiper-parámetros de modelo de pronóstico

El pipeline de características topológicas de la serie de tiempo que hemos utilizado todavía no soporta la optimización de hiper-parámetros en un conjunto de datos de validación; sin embargo consideramos que este es un límite suave que está en vías de ser solventado en la próxima iteración de la librería *Giotto-TDA*.

6. Conclusiones

El análisis de periodicidad y pronóstico de los índices de El Niño sigue siendo uno de los principales desafíos en los cuáles el Análisis Topológico de Datos puede ofrecer una perspectiva de estudio completamente distinta

En este trabajo se ha logrado demostrar y cuantificar la presencia de periodicidad en las series de tiempo de El Niño a través de técnicas de TDA; dicha periodicidad fue también descubierta al hacer uso de métodos de clusterización con cualidades topológicas, las mismas funciones que permitieron el descubrimiento de periodicidad facilitaron la creación de un modelo de pronóstico que logró capturar la forma de la serie de tiempo.

Referencias

[1] Giotto-tda, "Singletakensembedding." [Online]. Available: https://giotto-ai.github.io/gtda-docs/0.3.0/modules/generated/time_series/embedding/gtda.time_series.SingleTakensEmbedding.html

[2] T. Von Oertzen and S. M. Boker, "Time delay embedding increases estimation precision of models of intraindividual variability," *Psychometrika*, vol. 75, pp. 158–175, 2010.

[3] "Umap: Uniform manifold approximation and projection for dimension reduction." [Online]. Available: https://umap.scikit-tda.org/

[4] "How umap works." [Online]. Available: https://umap.scikit-tda.org/how_umap_works.html

[5] scikitlearn, "Manifold spectralembedding." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.SpectralEmbedding.html

[6] ——, "Spectral embedding." [Online]. Available: https://scikit-learn.org/stable/modules/manifold.html#spectral-embedding

[7] G. library, "Rips complex." [Online]. Available: https://gudhi.inria.fr/doc/3.0.0/group__rips__complex.html

[8] O. B. O. Sunhyuk Lim, Facundo Memoli, "Vietoris-rips persistent homology, injective metric spaces, and the filling radius," *arXiv*, vol. 5, no. 1, pp. 3–4, 2022.

[9] Y. ZHANG, "Persistent homology and sparse vietoris-rips filtration," p. 4, 2017.

[10] xgboost developers, "Xgboost documentation." [Online]. Available: https://xgboost.readthedocs.io/en/stable/#

[11] T. Masui, "All you need to know about gradient boosting algorithm - part 1. regression." [Online]. Available: https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502

[12] J. Perea, "Lecture 16 - Topological Methods for the Analysis of Data," 3 2020. [Online]. Available: https://www.youtube.com/watch?v=DZwK2gT-d8g

A. Código

Consultar código en la siguiente liga.